Portrait feature detection and its applications to clothes matching and caricature synthesis

肖像写真からの特徴検出とその服装検索及び似顔絵生成への応用

山梨大学大学院

医学工学総合教育部

博士課程学位論文

２０１８年９月

李宏林

Abstract

Portrait photos are widely used in social daily life not only in formal occasions but also in cascual occasions, such as applying for job, making certificates, synthesizing caricatures and so on. Recently, with the development of Internet and mobile phone, generating various styles of attractive caricatures according to portrait photos or selffiles is more and more popular in social media and commercial activities, which can bring attractive impression and protect privacy at the same time. Another interesting application is recommending appropriate clothes according to input portraits, that is to say, to collocate portrait faces with upper clothes for personality recommendation. In addition, since collar is one of the most important components in upper clothes due to its horizontal perspective position closest to the observers’eyes, retrieving clothes with desired collars in online shops is of practical meaning.

Portrait feature detection for upper clothes (especially for collars), hair, and face, which are the core elements in portrait photos, is very important for conducting the above applications. In Chapter 2, Chapter 3 and Chapter 4, we introduced the hand-crafted features specially designed for collar detection, capturing the characteristics of facial components and hair styles, and deep features related to the collocation of upper clothes and portrait

faces, respectively. Using these features, three applications are developed or designed which are similar collar retrieval, attractive caricature synthesis, and appropriate upper clothes recommendation, in Chapter 2, Chapter 3, and Chapter 4, respectively. Details of the three applications are shown as following.

Collar is a crucial part of the clothes due to its horizontal perspective position closest to the observers' eyes and serving as the frame for one's face in portrait photos. Searching for clothes with desired collars is very important for recommending clothes collocation. Content-Based Image Retrieval (CBIR) methods are developed to help people find what they desire based on preferred images instead of linguistic information. Chapter 2 focuses on retrieving clothes with similar collars according to input clothes images based on three types of combined features called Spread-Sift, Spread-Saliency-Sift and Spread-Saliency. In addition, a prototype of clothes image retrieval system based on Relevance Feedback approach and the Optimum Forest (OPF) algorithm is also developed to improve the query results iteratively. A series of experiments are conducted to test the qualities of the three types of combined features and validate the effectiveness and efficiency of the RF-OPF prototype from multiple aspects.

Portrait caricatures have been serving as icons, avatars and so on in real life and internet. Most caricature synthesis systems can be classified into three categories based on the approaches they used, which are

photo-transformed, example-based and style-transfer-based frameworks. Recently, caricature synthesis techniques based on deep learning have also been developing rapidly. Chapter 3 proposes a new caricature synthesis system under the example-based framework based on the feature deviation matching method, a cross-modal distance metric, which employs the deviation from average features rather than the values of features themselves to search for similar components from caricature examples. The proposed system does not require a paired photo-caricature database, and the designed geometric features can effectively capture the visual characteristics of input portrait photos. In addition, the proposed system can control the exaggeration of individual facial components, and provide several similarity-based candidates to satisfy users' different preferences. Extensive experiments are conducted to evaluate the results: 1. Similarity of the three types of caricatures (expressive, photo-realistic and drawing) with input portrait photos. 2. Comparison with the paired-example method. 3. Effectiveness of the designed hair and facial component features.

Upper clothes part is an important element in portrait photo. How to coordinate upper clothes with given female portraits is valuable for daily life. Traditional clothing coordination focuses on clothing component collocation or situation-clothing collocation without considering human personalities, such as human face and hairstyles. By regrading features of the female hair-face part and the upper clothes part as coming from different domains,

we adapted deep learning and cross-modal retrieval techniques for upper clothes recommendation. The Deep Canonical Correlation Analysis (DCCA) method is one of the traditional cross-modal retrieval methods. It uses deep learning frameworks to project extracted features from different modalities into a common feature space. By maximizing the correlations between the paired projected features from different modalities, the DCCA framework tries to make them as similar as possible, which can thus be mutually retrieved directly in the common feature space. The Unsupervised Domain-Adversarial (UDA) method is inspired by Generative Adversative Networks (GAN) and used for cross-modal retrieval. By embedding a label predictor into the feature projector to preserve the orginal semantic distribution and conducting adversarial learning between the feature projector and the modality classifier to reduce the gap across different modalities, the UDA method is effective to generate similar common feature representations from different modalities maintaining their original distributions. In Chapter 4, we proposed two preliminary frameworks based on the DCCA and UDA methods, respectively, for upper clothes recommendation, which focused on investigating the collocation relations between upper clothes and female faces and hairstyles. Firstly, we divide the portrait photos into two parts, hair-face and upper clothes parts, and extract features from them with deep learning frameworks. Secondly, feature projectors made up of fully connected layers or linear projectors are used to

embed the extracted upper clothes features and hair-face features into a shared common feature space. Thirdly, the DCCA and UDA methods are used in the first and second preliminary frameworks, respectively, to make the projected feature representations from upper clothes and hair-face as similar as possible.At last, by conducting direct distance comparison in the common feature space, for each input portrait face, appropriate upper clothes will be recommended which is with the smallest distance to it.

Keywords: Collar retrieval, Relevance Feedback, Optimum-Path Forest, Caricature synthesis, Cross-modal distance metric, Feature deviation matching, Upper clothes recommendation, Deep learning, Cross-modal retrieval.

Directory

Chapter 1 Introduction

Portrait photos are widely used in social daily life not only in formal occasions but also in cascual occasions, such as applying for job, making certificates, synthesizing caricatures and so on. Recently, with the development of Internet and mobile phone, generating various styles of attractive caricatures according to portrait photos or selffiles is more and more popular in social media and commercial activities, which can bring attractive impression and protect privacy at the same time. Another interesting application is recommending appropriate clothes according to input portraits, that is to say, to collocate portrait faces with clothes for personality recommendation. In addition, since collar is one of the most important components in upper clothes due to its horizontal perspective position closest to the observers'eyes, retrieving clothes with desired collars in online shops is of practical meaning.

Based on the above considerations and inspired by the fact that most of standard portrait photos and selfiles are made up of hair, face and the upper clothes, We conduct the below projects for similar collar retrieval, attractive caricature generation and appropriate upper clothes recommendation,as shown in Fig.1.

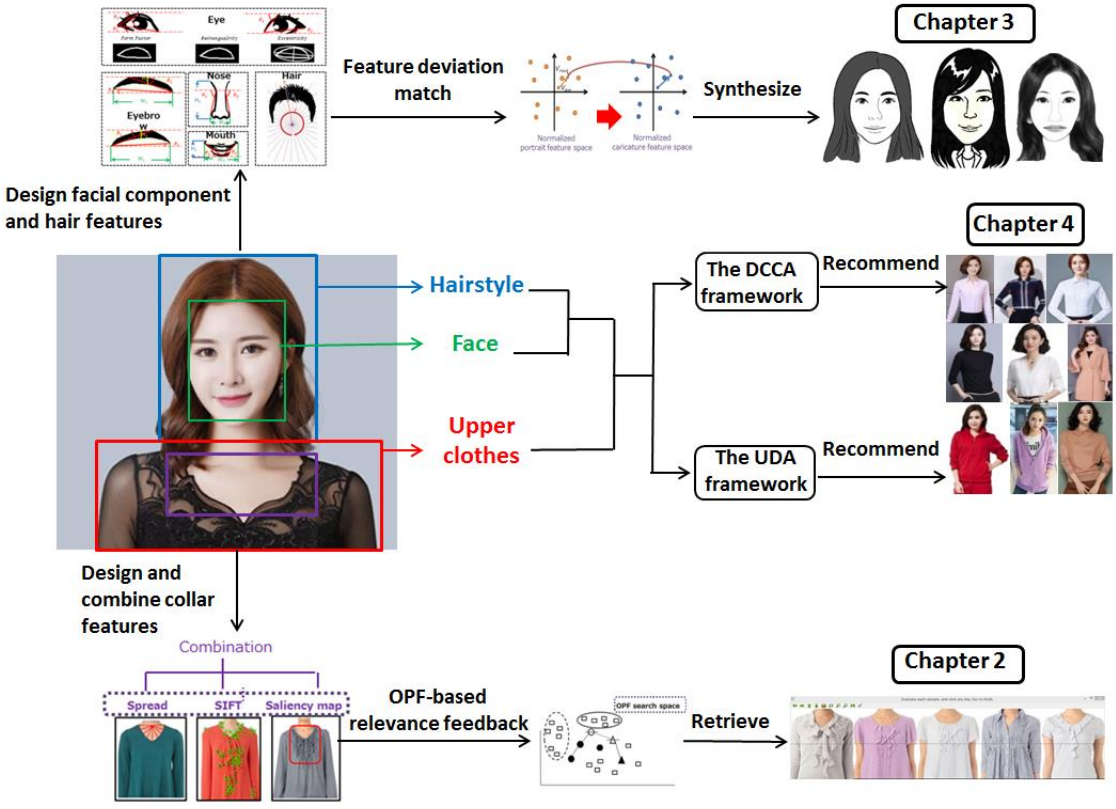


Fig.1 Proposed research projects in this thesis

1.1 Retrieval of Clothing Images based on Relevance Feedback with Focus on Collar Designs

As e-commerce continues to gain momentum in the market, more and more consumers are searching

online shops for clothing items. However, sifting through the massive amounts of available products to find an item that suits one’s tastes and preferences can be an arduous, time-consuming task. In addition, it is difficult to detect human body size online automatically, which is very important for clothes collocation. Collar is a crucial part of the clothes due to its horizontal perspective position closest to the observers’ eyes and serving as the frame for one’s face in portrait photos. In Chaper 2, we focus on searching for clothes with desired collars. Content-Based Image Retrieval (CBIR) method [1] involves expressing image content in feature vectors and then comparing the similarities between various images. Although CBIR methods eliminate the need to build queries out of keywords or other linguistic information and allow users to search for visual information with visual information input, their efficiencies largely depend on the quality of the designed features. Some researchers [2] have attempted to use color and texture, but few researchers have delved into the possibilities of developing feature vectors to capture the designs of clothing in detail.

The main contribution of Chapter 2 is two-fold. One is a novel method for extracting the detailed designed features of collars from 2D clothes images. The other is a prototype of clothes image retrieval system based on Relevance Feedback approach and Optimum Forest algorithm for allowing users to find clothes images with preferred designed collars. HSV color model, Tamura texture, Gabor texture, SIFT feature, SURF feature, MSER (Maximally Stable Extremal Regions) and so on, which are colors, texture, shape and local features, are usually used to detect prominent features for upper clothes.In our study, we consider collar spread, turnover and the front design as three important design factors of a collar and develop new methods for automatically extracting those features from clothes images. For the turnover and front design, different feature vectors based on Sift feature and Saliency Map are designed. We further apply Optimum-Path Forest algorithm to perform image search procedures. By incorporating the Relevance Feedback approach into the OPF process [3], our method enables users to search for images based on his or her subjective preferences on collar design. A series of experiments are conducted to validate the above proposed ideas and methods.

1.2 Caricature Synthesis with Feature Deviation Matching under Example-Based framework

Portrait caricatures have been serving as icons, avatars and so on in real life and internet. Since the early 1980s, many computer-based methods have been developed for synthesizing caricatures [4]. These can be roughly classified into photo-transformed [5-8, 14-16**,** 27] and example-based [9-13, 17-20] techniques. Recently, style-transfer-based [21-26] and deep-learning [27] based approaches are developing quickly. In Chapter 3, we propose a new example-based caricature generation technique that can synthesize stylized caricatures with a small number of unpaired examples of portrait photos and caricatures. It is very important to design appropriate features to describe facial and hair components before applying similar component retrieval and synthesis.For hair feature detection, some researchers deploy shape matching and classification using height functions, and some further align their outer contours with Horn’s quaternion-based method to achieve robust hair matching. The PCA (Principal Component Analysis) method is used to extract prominent features from the whole face. Various methods, such as HOG (Histogram of Oriented Gradients), Fourier spectrum and landmark-based distance, were employed for detecting local component features for eye, eyebrow, mouth, nose and face contour. Recently, deep learning techniques are deployed to extract features from various objects based on huge number of samples. In Chapter 3, we proposed some effective hand-crafted geometric features for describing hair and facial components.

The proposed caricature synthesis system incorporates component-specific learning based on feature vectors that intuitively match the features that people employ to perceive or communicate the characteristics of faces, which can also provide users with control over the individual facial components. By using a new cross-modal distance metric called feature deviation matching, we can compare the component features of different modalities in different feature spaces directly. Thus, we search caricature components similar to input portrait facial components and synthesize the output caricatures.

The main contributions of Chapter 3 are shown as below:

1. The newly proposed cross-modal distance metric called feature deviation matching technique makes it possible to generate various styles of caricatures under the conventional example-based framework without requiring paired photo–caricature training sets.
2. By focusing only on the perceptually prominent features, the designed feature vectors are robust and effective for capturing the visual facial features of input portrait photos.
3. The proposed system enables users to control the exaggeration of individual facial components. Various combinations of individual facial and hairstyle components, based on different exaggeration coefficients and similarity rankings, can provide users with different candidates to satisfy their particular preferences; this has not been achieved in most existing style-transfer-based and deep-learning-based approaches.

Extensive experiments are conducted to validate effictiveness of the proposed ideas and methods from the following aspects: 1. Similarity of the three types of caricatures (expressive, photo-realistic and drawing) with input portrait photos. 2. Comparison with the paired-example method. 3. Effectiveness of the designed hair and facial component features.

1.3 Preliminary frameworks for upper clothes recommendation based on deep learning and cross-modal retrieval techniques

In Chapter 4, we proposed two preliminary frameworks for upper clothes recommendation according to input portraits based on deep learning and cross-modal retrieval techiniques. Different from Chapter 2 and Chapter 3, we employ deep learning frameworks to extract features from portrait and upper clothes photos automatically instead of hand-crafted features.

These two prelimary frameworks are designed for upper clothes recommendation according to given female portrait faces based on the Deep Canonical Correlation Analysis (DCCA) [28] and Unsupervised Domain-adversarial (UDA) [96] methods. There are two core ideas in DCCA and UDA methods. Firstly, both of them use deep learning frameworks to extract cross-modal features and project them into a common feature space. Secondly, the CCA and UDA modules are employed, respectively, to make the projected features from different modalities as similar as possible for direct comparison. Thus, the mutual retrieval in the common feature space can be used for recommending appropriate upper clothes according to input female portrait faces.

These two proposed frameworks are in progress and only preliminary frameworks by now.

1.4 Structure of the thesis

Portrait photos taken by cameras and mobile phones are widely used in real life and internet for certificates, communication, entertainments and so on. This thesis focuses on researches of portrait feature extraction and its applications for desired collar retrieval, extractive caricature generation and appropriate clothes with desired collars recommendation.

In Chapater 1, we make a short introduction for the research background on the above issues at first. Then the corresponding reseach backgrounds and the congtributions of the proposed three projects are discussed. At last, the structure of the thesis is displayed.

In Chapter 2, we make a detailed introduction for the first project: Retrieval of Clothe Images based on Relevance Feedback with Focus on Collar Designs. Three types of hand-crafted features (Sift feature, Saliency feature and Saliency-Sift feature) are applied to capture prominent features of collars for clothes retrieval. In addition, we employ a relevant feedback system based on optimum forest (OPF) algorithm to improve the query results iteratively.

In Chapter 3, an example-based caricature synthesis framework based on feature deviation cross-modal distance metric is employed to synthesize three types of caricatures (expressive, photo-realistic and drawing) according to input portrait photos. Exaggeration control and similarity-ranking based functions make it possible to satisfy different users' preferences.

In Chapter 4, two preliminary frameworks based on deep learning and cross-modal retrieval techniques are proposed to recommend upper clothes according to input female portrait faces.

In Chapter 5, we make a conclusion and summary for the whole thesis and discuss the future research work.

The structure of this thesis is shown as Fig.2.

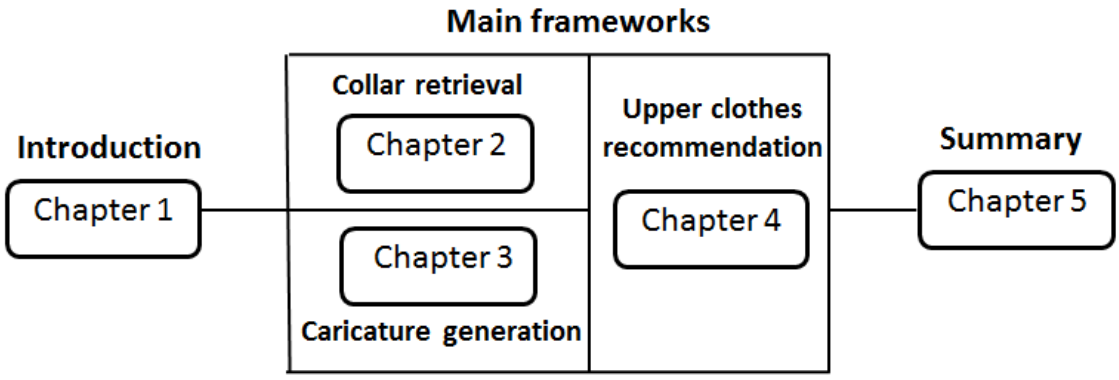


Fig.2 Structure of this thesis

Chapter 2 Retrieval of Clothing Images based on Relevance Feedback with Focus on Collar Designs

Chapter 2 mainly describes details of the proposed clothes retrieval framework. The proposed framework combines the OPF algorithm to improve the query results iteratively online.

2.1 Introduction

As e-commerce develops quickly, more and more consumers are searching online shops for clothes items. Many sites support keyword-based searches, but items in online shops often lack specific design-related tags and include technical names that few shoppers are familiar with. One image search approach that researchers have proposed as visual query-based alternatives to keyword-driven searches is the Content-Based Image Retrieval (CBIR) method, which involves expressing image content in feature vectors and then comparing the similarities between various images. Although CBIR methods eliminate the need to build queries out of keywords or other linguistic information and allow users to search for visual information with visual information input, their efficiencies largely depend on the quality of the feature vectors, and it remains to be challenged to extract the image features capturing the design of clothes well. Some researchers have attempted to use color and texture, but few researchers have delved into the possibilities of developing feature vectors to capture the designs of clothing in detail.

The main contribution of this paper is two-fold. One is a novel method for extracting the detailed design features of collars from 2D clothes images. The other is a prototype of clothing image retrieval system based on Relevance Feedback approach and Optimum Forest algorithm for allowing users to find clothes images with preferred designed collars. We focus on collars because it is a well-known fact for garment designers that collar is a crucial part of a garment as it serves as the frame for one’s face [29-31]. Since face is the most important visual attribute for characterizing a person, the design of a collar can largely affect the look of a garment and the overall impression of a person. In addition, the artistic formation of a collar is usually the most eye-catching part when people look for clothes due to its horizontal perspective position closest to the observers' eyes [32]. Although computer assisted 3D garment design technologies [33, 34] and virtual fitting systems [35, 36] become available recently, E-commerce websites for clothes shopping mainly show 2D clothing images for users to choose. Our proposed method allows users to retrieve the clothes with preferred designed collars from such online shopping sites first, which will narrow down the candidates and speed up the procedure of virtual fitting based on image rendering technique [36].

In this Chapter, we consider collar spread, turnover and the front design as three important design factors of a collar and develop new methods for automatically extracting those features from clothes images. For the turnover and front design, three dependent and combined feature vectors based on Sift and Saliency Map methods are designed. We further apply Optimum-Path Forest algorithm to improve image query results iteratively. By incorporating the Relevance Feedback approach into the OPF process, our proposed framework enables users to search for clothes images based on his or her subjective preferences on collar design. To validate the above proposed ideas and methods, a series of experiments are conducted. The

experiment results presented in [37] could not provide sufficient validation to the proposed methods. First, only four subjects were involved in the experiments and thus the results might not be reliable enough. Secondly, neither the effects of the different collar types on the evaluation scores nor the bias caused by the difference in the number of images of different collar types was considered. Thirdly, the improvements of the RF-OPF prototype on the three feature extraction methods were not discussed. In this work, we redesign the experiments and succeed in achieving reliable and detailed results to validate the effectiveness and efficiency of the proposed methods for all referred collar types.

2.2 Related researches

The ongoing spread of e-commerce, among other factors, has prompted numerous researchers to explore the possibilities of applying search methods to clothes. Liu et al. [38], for example, proposed a method that makes it possible to use snapshots to search for clothes available in online marketplaces. The method developed by Liu et al. involves taking a picture of a person's body, separating the body into its constituent parts (feet and legs, for example), and determining the features of each part to enable users to search for images on a fashion website. Bossard et al. [39] proposed a method for classifying apparel in photographs. Using SVM and Random Forest allows their method to establish clothes categories like long skirts and coats and classify clothes according to sleeve length, material and other attributes, but these classification schemes were the ultimate purpose of the research; Bossard et al. did not include the idea of searching for clothing based on specific design elements. A project by Hsu et al. [2] used images with uniform backgrounds as queries for retrieving a limited scope of clothing items that one might find in an online shop. With a piece of clothing serving as the input for the method, their approach involved comparing items based on the features−color, texture, Sift features, and outline−that the pixels in the clothing regions of the given images form. Sift features were also used in our study, which aimed to extract the designs and other characteristics of collars. None of these existing methods are capable of searching for detailed information that collar designs represent. One recent study proposed the idea of using sketches to search for clothing items with the desired design [40]. However, the method presents problems for people who lack sufficient sketching skills.

In the CBIR field, meanwhile, Relevance Feedback (RF) method has been drawing substantial attention for its use of dialogic feedback between the users and the system for learner-driven learning and searching purposes. Searching based on Relevance Feedback method makes it possible to update classifiers by showing results to users. Researchers have already tested this approach in searching for images with ambiguous thematic content, such as ocean scenes, cats, and sunsets. One study has proposed a method that produces high-quality results via minimal amounts of feedback by incorporating different types of classifiers and reusing past classification results [41]. By employing a prototype with the RF approach for learning and the OPF algorithm for searching procedures, our method enables users to search for collars that align with their personal preferences, which may have positive effects on choosing the whole clothes.

2.3 Collar design

Generally, image searches operate on the similarity of visual attributes of multiple images. Finding a collar that suits one's tastes, however, would require the matching of design elements in greater detail. For our study, we begin by interviewing instructors at fashion colleges about the design elements of collars. Then we take their suggestions into consideration to design feature vectors that can enable fine-tuned

search functionality. Clothes experts have suggested that collars generally come in the following ten types (Fig. 3) [42].

Fig.3 Main collar types

Three important elements should be taken into consideration in describing the collar designs. The first is the collar spreads-the ways they open. The second is Turnover. Fig. 4 shows two examples of the collar design with turnover. The existence and shape of the turnover is one of the most distinctive features affecting the preferences of many consumers looking for buying clothing. The third is front designs, which refers to the types of ribbons and frills as shown in Fig. 5.

Fig. 4 Turnover collars

Fig. 5 Collar with Front designs

Collar design preferences vary considerably according to buyers' tastes and needs. A person looking for work clothing, for example, would probably prefer a simple, clean look to a busy, loud design. On the other hand, a buyer trying to find something flashier to wear for a fancy occasion might opt for an ensemble that features frills or a gather. Personal tastes affect people's clothes choices in a wide variety of other ways, for instance, the likelihood of a person with a reserved, quiet personality buying a frilly design is rather slim. In hopes of enabling a search approach that reflects user preferences, we design feature vectors to describe detailed collar designs and build a search system using the Relevance Feedback method based on OPF algorithm. For the preliminary feature extraction process, our method involves obtaining feature vectors describing three designed elements-collar spread, turnover, and front design-for each clothes item in the clothes database. Three kinds of feature extraction methods based on Sift, Saliency and Saliency-Sift, respectively, are implemented. In the runtime phase, OPF algorithm is used to classify the images into relevant or irrelevant based on the initial training images. The OPF classifier is refined through iterative relevance feedback from the users. The proposed clothes retrival framework is shown as Fig.6.

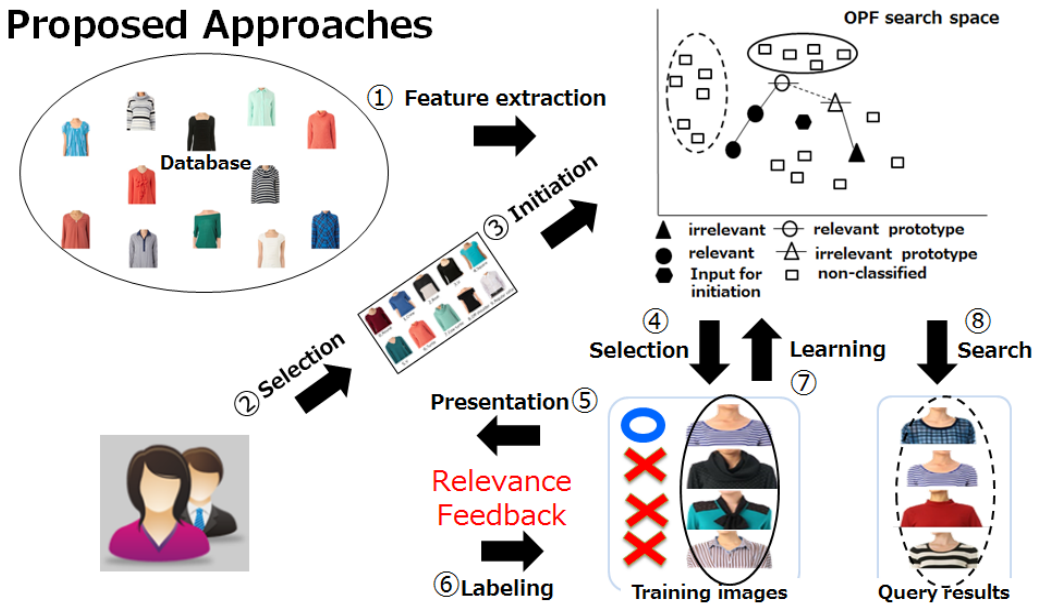


Fig. 6 The proposed clothes retrival framework

2.4 Designand extract feature vector

In order to adopt the RF approach and use the OPF algorithm to search the items of preferred collar design, it is necessary to build a feature vector space that effectively reflects the design features of collars. To describe the three important elements, which are spread, turnover and front designs, the feature vectors are designed and extracted as below.

2.4.1 Collar spread

For our study, we limit our scope to images of tops that showed the wearer's upper body only. We also assume that each image has a monochromatic background showing a human subject from the front and the color of which is at least somewhat different from that of the subject's skin. Incorporating more advanced image processing technologies would make it possible to ease these restrictions, but doing so would deviate from the main focus of our study. Fig.7 illustrates the process of extracting feature vectors that represent the collar spread (opening). We use the following steps to extract features.

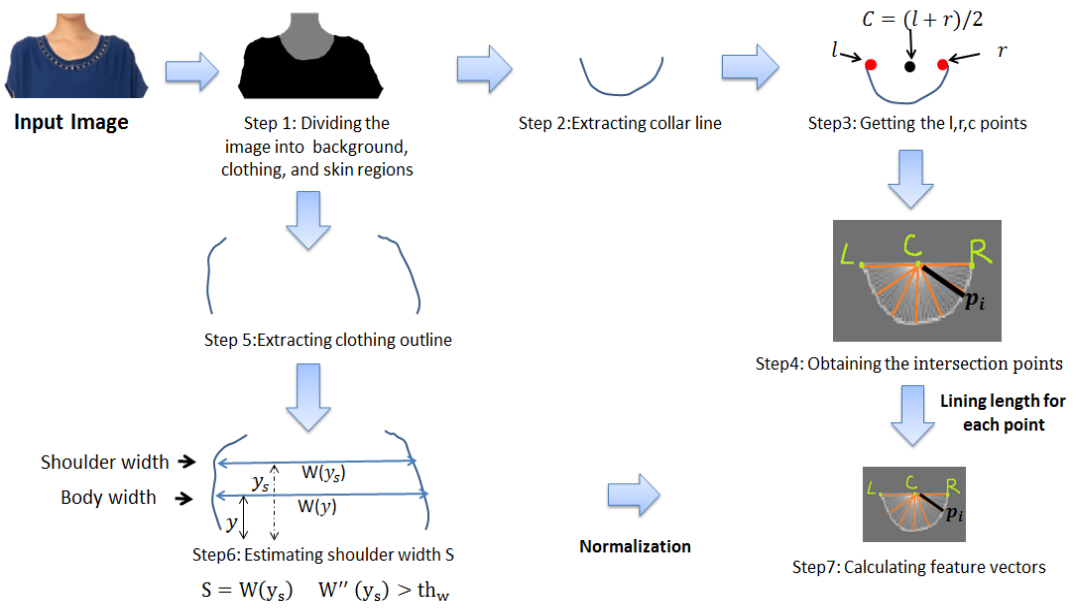


Fig. 7 Compute the feature vectors of the collar spread

Step 1: Dividing the image into background, clothes and skin regions.

This step is the preparation work for identifying the collar line (the clothes-skin boundary) of Step 2 and determining the outline of the clothes in the image (the clothes-background boundary) of Step 5.The collar spread features can be expressed according to the depth value for each angle emanating from the subject's neck, as indicated by the result of Step 7 in Fig. 7.

We separate the background and the foreground of the input image. To extract the skin color region, we use skin color S, the dominant color of the area between the neckline and the chin, as our learning data and extract the pixels with the shortest Mahalanobis' generalized distances from skin color S and use their color t as the skin color. And thus we can define the collar line and clothes outline by locating the skin, clothes and background boundaries. The Watershed algorithm was used to place single seeds in the skin, clothes and background areas respectively and divide the image into the three regions based on these seeds.

Step 2: Extracting collar line

With the image divided into three regions, we then extract pixels from where the skin and clothes regions meet to identify the collar line.

Step 3: Getting the l, r, c points

This step involved extracting the coordinates of the left end 1, center c, and right end r of the collar line, which will be used in Step 7 to compute the feature vectors representing the spread of collar. We use chain codes to trace the contour of the entire collar line identified in Step 2, which allow us to determine the coordinates of left end 1 and right end r. Then the coordinates of 1 and r are used to determine the center c of the collar.

Step 4: Obtaining the intersection points

Then we extend radial lines from the center c in the direction of $\alpha_i = i \times \pi/36 (i = 0, \cdots, 36)$ until the lines intersect with the collar line that we extracted in Step 2. The intersections are defined as points $p_i (i = 0, \cdots, 36)$.

Step 5: Extracting clothes outline

With the image divided into three regions in Step 1, we extract pixels from where the clothing and background regions meet to generate the clothing outline.

Step 6: Estimating shoulder width S

Using the clothes outline founded in Step 5, we next calculate shoulder width. Its value serves as the basis for normalizing the feature vector calculations in Step 7. By normalizing the feature vectors, we ensure that the feature quantities obtained are independent of image size.

To calculate shoulder width, we use the differential of body width. Denoting the body width at height y in the clothing outline as $W(y)$, the shoulder width S is given as $W(y_s)$ at the height y_s where the second order derivative of W(y) exceeds a given threshold.

$$S = W(y_s) \quad W''(y_s) > th_w \tag{1}$$

Step 7: Calculating feature vectors

With the intersection points $p_i(i = 0, \cdots, 36)$ obtained in Step 4 and the center c obtained in Step 3, the lengths $|c - p_i|(i = 0, \cdots, 36)$ are computed and normalized based on shoulder width obtained in Step 6. Finally, the normalized values constitute the 36-d feature vector of the collar.

Fig.8 shows the searching results by using the above spread feature vector only. We can see the collar spread feature is well captured (the first two rows), but the feature of turnover and front design cannot be distinguished (the last two lows).

Fig. 8 Comparison of the result by Spread feature vectors on clothes with V opening ,Square opening, Turn over and Front design collar

2.4.2 Turnover and Front design

To capture the features of turnover and front design, we design three different feature vectors based on Sift, Saliency map and Saliency map plus Sift, which are described in the following as Sift feature, Saliency feature and Saliency-Sift feature.

Sift feature

Sift is the most commonly used size and orientation invariant feature. Using a gradient histogram around arranged points makes it possible to capture local details of image contents. It can be expected to capture ribbon shapes and other design details. To accelerate the feature matching in image retrieval, we combined the Bag-of-visual-word method [44-46] with the Sift feature extraction. First we compute the 128-d local Sift feature. Then K-means method is used to cluster the Sift feature vectors into 500 clusters (500 here is empirically given). The centers of clusters are used as codewords. Finally, each Sift feature vector is mapped to the closest codeword to obtain the histogram of the codewords, which is a 500-d feature vector. We compute the above feature vector for the region from the collar to the chest (Collar-Sift) and the whole image (Whole-Sift).

Saliency feature

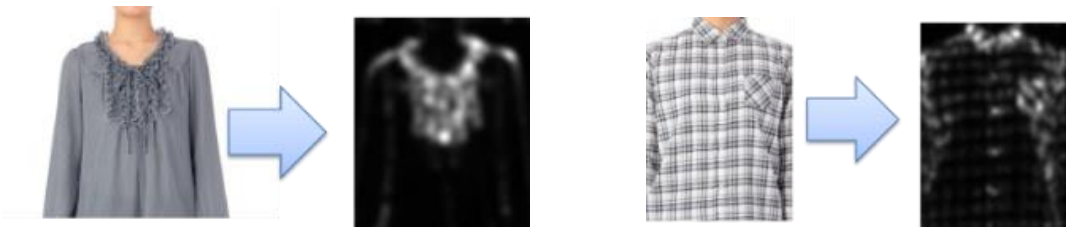
Fig.9 illustrates the comparison of searching results between a piece of plain front design clothes and a striped V collar one using Sift feature. While this approach manages to find relatively good matches (the first row) for the plain clothing image, it fails to produce the same quality results (the second row) for a striped one. The second row results shows that images with similar clothing textures-not collar designs-appear in the top search results. In other words, the Sift features can be too dominated by texture features.

Fig. 9 Comparison of the result by Spread feature vectors on clothes with V opening ,Square opening, Turn over and Front design collar

In order to reduce the effects of the texture features on the collar designs, we propose a new feature extraction method using Saliency Map proposed by Xiaodi etal [46]. Drawing on the mechanisms of human visual attention, the Saliency Map method allows users to determine the area of an image that observers are most likely to focus on. A Saliency Map calculates the degree of attention not based on the presence of patterns, but on the differences between a given location of an image and its surroundings. This method can be expected to be effective for obtaining features from the front designs as well as turnover, which are the decorative embellishments serving for attracting attention. As long as the design is distinct in some way from its surroundings, the Saliency Map method can get good matches even if the clothes

consist of textures.

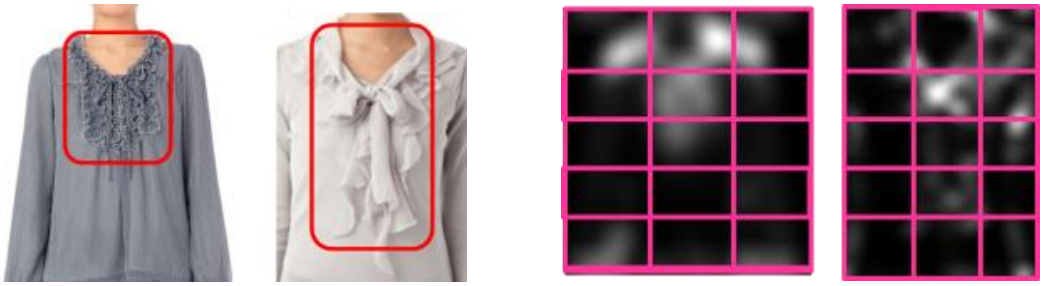
Fig. 10 shows the results from two examples consisting of textures. The frill area of the example shown in Fig.10 (a) exhibits prominent differences from its surroundings on the Saliency Map. Even though the clothes consists of large checked textures, the turnover can still be captured by the Saliency Map（Fig.10 (b) ）.



(a) (b)

Fig. 10 The results of Saliency Map acting on two clothing images. (a) An example with front design. (b) An example with turnover

The appearance of a front design, meanwhile, depends heavily on the size, length and breadth of their attention-drawing elements. Various shapes and sizes of ribbons and frills have an impact on personal preference. Frills that cover a considerable area on a piece of clothes, for example, create quite a different visual impression from minimal, dainty frills on the top of the collar. One can also locate areas with the highest concentrations of elements that diverge from the general look of a given clothes item. The frills in the images of Fig. 11 (a), for example, spread out across the width of the wearer's chest to create a relatively showy impression. The ribbons in the images of Fig.11 (b), meanwhile, give the clothes a more extravagant but vertically oriented appearance. To capture these design-related differences, the proposed method lays a given grid (3×5) over the Saliency Map as shown in Fig. 11 (b). We obtain a 15d feature map, where the pixels values' sum in each grid cell represents a dimension of the feature vector.



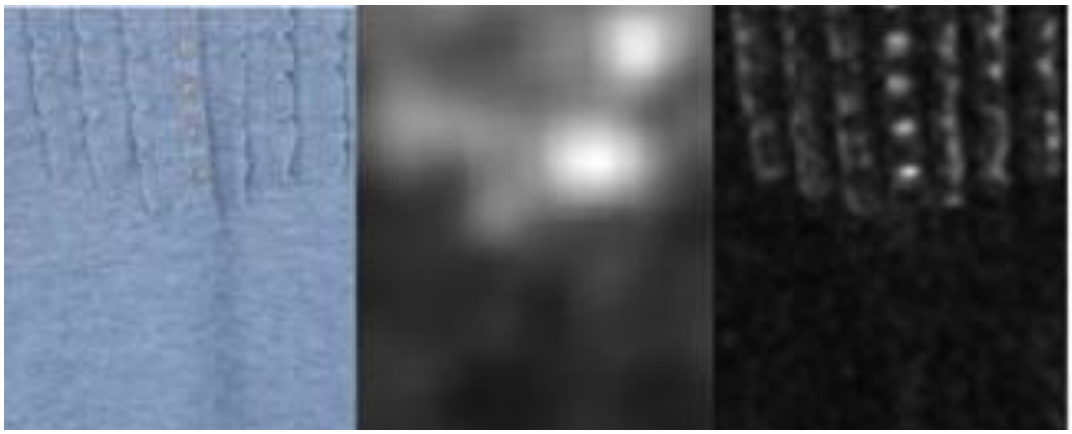
(a) (b)

Fig. 11 Feature vector design for capturing front designs. (a) Examples of ribbons. (b) 3× 5 Grid arranged on Saliency map for capturing the spatial distribution of front designs

Saliency-Sift feature

This approach, however, can capture the overall shape of front design only. If the clothes in the input images feature a small group of frills, the design details of the frills cannot be captured. We thus propose a method that combines the Saliency Map and Sift feature. Fortunately, Saliency Map proposed by Xiaodi et. al allows controlling the level of detail. Using the high frequency band to compute the difference from the average, we obtain results that retain the frill details (Fig. 12(c)). Fig.13 shows the result of using a

Saliency map over a striped shirt with a turnover collar. If we are to apply normal edge detection to this image, with its prominent pattern, it will be difficult to determine where the turnover is. Using a Saliency Map with the proper level of detail, however, allows us to limit the impact of the pattern to a certain degree. Then by computing the Sift features from the saliency map image, we are able to obtain the details of collar design while eliminating the influence of texture features of the clothes item.



(a) **(b)** **(c)**

Fig. 12 Saliency map computed with the method given in [20]. (a) Input image. (b) Saliency Map computed using low frequency band. (c) Saliency Map computed using high frequency band

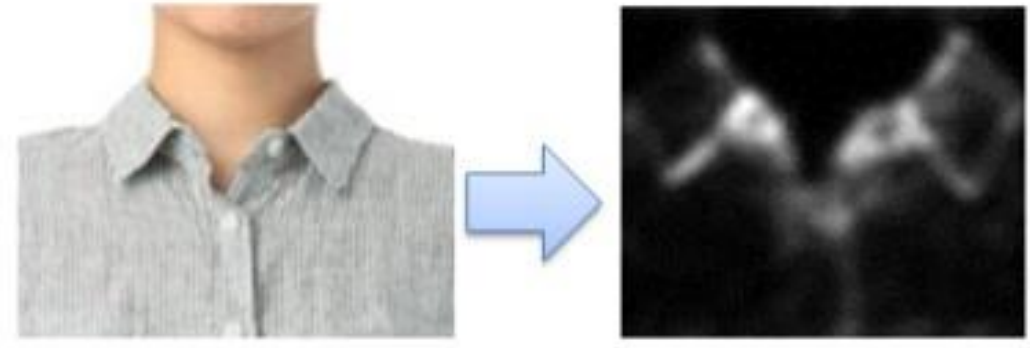


Fig. 13 Extract turnover features in a high-detailed Saliency Map

In Fig. 14, the images of the top row are the search results by using Sift feature only. The clothes items with round neck are also included in the results as those items have the gather pattern similar to the strip texture of the input clothes. Most of those items are eliminated in the results shown in the bottom row by combining Saliency Map and Sift feature. In addition, more items with turnover collar or similar collar spread are included in the search results. However, the query results are still not good enough. So a prototype combining Relevance Feedback approach with OPF classifiers is proposed to improve the qualities of the query results as follows.

Fig. 14 Results by Sift feature (top) and Saliency-Sift (bottom)

In Fig. 14, the images of the top row are the search results by using Sift feature only. The clothes items with round neck are also included in the results as those items have the gather pattern similar to the strip texture of the input clothes. Most of those items are eliminated in the results shown in the bottom row by combining Saliency Map and Sift feature. In addition, more items with turnover collar or similar collar spread are included in the search results. However, the query results are still not good enough. So a prototype combining Relevance Feedback approach with OPF classifiers is proposed to improve the qualities of the query results as follows.

2.5 The RF-OPF prototype

Relevance feedback (RF) method is a critical component in our CBIR prototype, which makes it possible for users to interact with the system and thus reflects their design preference in the query. The classifiers, another critical component of our CBIR system, are used for processing queries. Their efficiency (related with response time) and effectiveness (related with users' satisfaction) are very important for evaluating the quality of this CBIR system. In our prototype, we use the Optimum-Path Forest (OPF) [3, 48, 49] classifier for query and classification. OPF works by modeling the classification as a graph partition in a given feature space. It starts as a complete graph, whose nodes represent the feature vectors of all images in the database. All pairs of nodes are linked by arcs which are weighted by the distances between the feature vectors of the corresponding nodes (referred as costs here and after). As illustrated by Fig.15, given a set of training nodes, a minimum spanning tree (MST) can be generated from the complete graph. Then the adjacent training nodes are marked as prototypes if they belong to different classes, which are relevant and irrelevant in our case. The partition of the graph is carried out by the competitions process among prototypes, which offer optimum paths to the remaining nodes of the graph. The optimum paths from the prototypes to the other samples are computed by the algorithm of the image foresting transform (IFT), which is essentially Dijkstra's algorithm modified for multiple sources and more general path-value functions. At last, all the non-prototypes are connected with a prototype directly or indirectly with the minimum costs. With the prototypes as the roots and the non-prototypes as the intermediate and terminal nodes, the Optimum Trees are built, which constitute the Optimum-Path Forest (OPF). Compared with SVM, ANN-MLP and K-NN, OPF is usually superior to ANN-MLP and K-NN in accuracy and significantly outperforms SVM in computation time [3, 48, 49], which is very important in a prototype based on RF approach that generates results in a dialogic fashion.

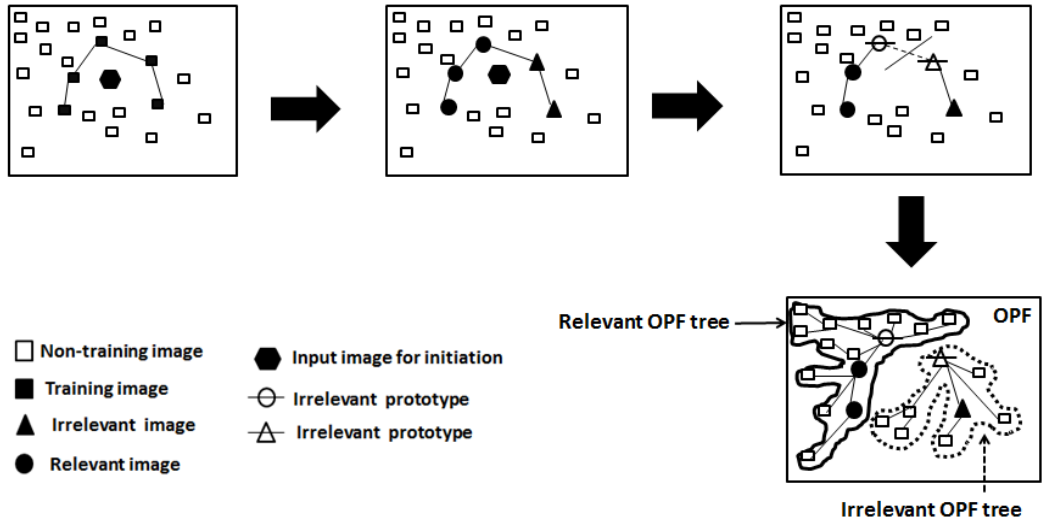


Fig. 15 Generate the OPF Classifier

All the images in the database are represented by the feature vectors extracted with one of the three methods referred in the section 2.4. The first is Spread+Collar-Sift+Whole-Sift represented by 1036-d (36-d+500-d+500-d). The second is Spread+Saliency represented by 51-d (36-d+15-d). The last is Spread+Saliency-Sift represented by 1036-d (36-d+500-d+500-d).

Based on the OPF classifier described above, we build our RF-OPF prototype, which works as following steps:

1. The initial training set containing images of 10 different collar types plus front design type is presented to user. When the users choose one desired collar type from the initial training set, the five images with smallest L2 distance to the chosen image in the feature spaces are returned to the users.

2. The user evaluates the results. If he/she is satisfied with the initial query result, the RF ends up without using the OPF classifier. If not satisfied, he/she should mark the images with × representing irrelevant or O representing relevant.

3. The first five images marked with × or O constitute the original training set for building an OPF classifier mentioned above. The procedure is illustrated as Fig.14 below and the detail of the OPF algorithm was described in [48]. Then we use the OPF classifier to sort the unclassified images of the database into two classes, relevant and irrelevant.

4. The RF-OPF prototype then chooses five images with the maximum degree of relevance as the query results and five images with the minimum degree of relevance as the training samples from the relevant class and returns them to the users. The new marked training samples will be merged into the former training samples to build a new OPF classifier for the next RF phase if the users are not satisfied. This procedure continues until the users are satisfied.

The framework and operating mechanism of our RF-OPF prototype is shown in Fig.6 and Fig.16, respectively.

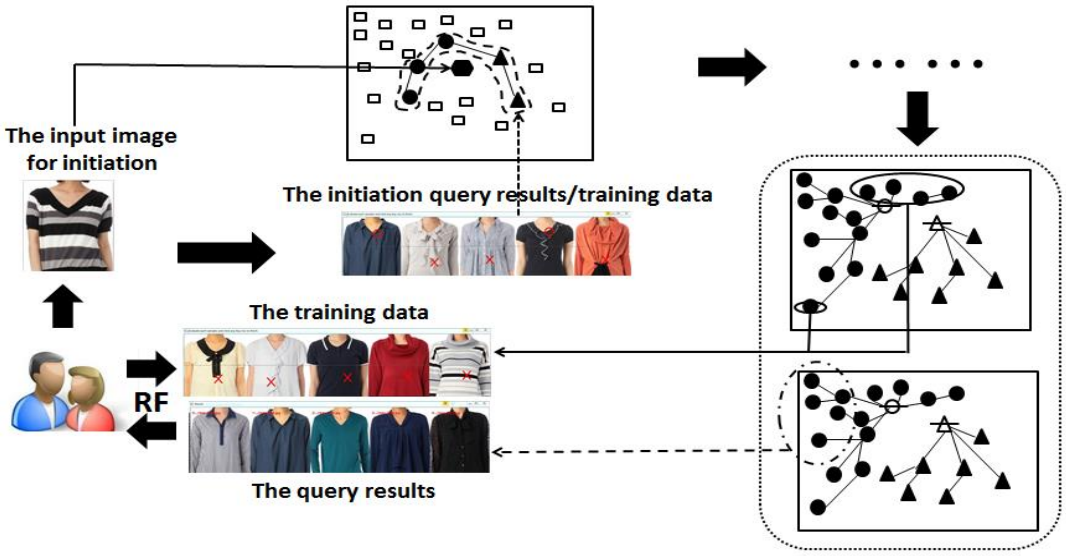


Fig. 16 The operating mechanism of the RF-OPF prototype

When selecting the query results and the training samples, we compare the costs of paths from all non-training images to all relevant and irrelevant prototypes. The five images which belong to relevant class with the largest ratio of costs to the relevant prototypes over costs to the irrelevant prototypes are chosen as the query results. The training samples are the five images which belong to relevant class and having the smallest ratio of costs to the relevant prototypes over costs to the irrelevant prototypes. In our implementation, the cost of the arc connecting two adjacent nodes of the OPF feature space is calculated with the L2-norm. And the cost of a path is the maximum value of the costs of all arcs constituting the path.

Assuming the number of relevant prototypes and irrelevant prototypes to be k and m and denote the k relevant prototypes and m irrelevant prototypes as p_i(i=1, 2 ⋯k) and q_j(j=1, 2 ⋯m), we consider k×m pairs of (p_i, q_j) in computing the ratio of the costs of paths to the relevant and irrelevant prototypes. Let $CR_{U\to p_i}$ and $CI_{U\to q_j}$ represent the costs of the path from a non-training sample U to the relevant prototype p_i and the irrelevant prototype$q_j$respectively. $Relevance_{U\to(p_i,q_j)}$, which represents the ratio of $CR_{U\to p_i}$ over $CI_{U\to q_j}$, is computed as:

$$Relevance_{U\to(p_i,q_j)} = \left\| CR_{U\to p_i} - CI_{U\to q_j} \right\| \quad (2)$$

We use subtraction instead of ratio to avoid encountering the overflow problem when $CI_{U\to q_j}$ is very small.

2.6 Experiment and discussion

2.6.1 Experiment

The images used for experiments are gathered from the Internet. All images have monochromatic background and show upper bodies from the front. Totally there are 274 images including ten different collar types and the clothes with front design as shown in Fig. 3 and 5. The number of images for each of the ten collar types and the clothes with front design is shown in Fig.17. Because it should be easier to find

a certain collar type if the number of it in the database is high, we take into consideration the proportion of image numbers of different collar types in evaluating the efficiency of proposed methods.

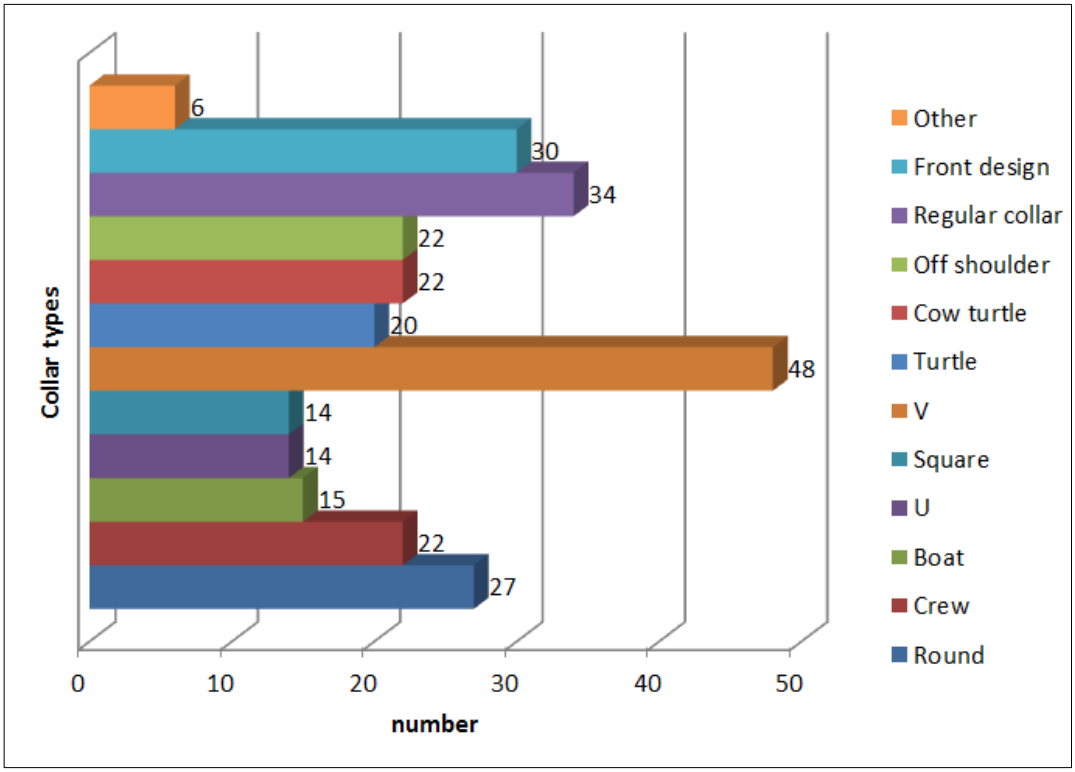


Fig. 17 Number of images for the ten collar types and clothes with front design

For all images, we pre-computed their Spread, Sift, Saliency and Saliency-Sift feature vectors. The Spread feature vector has 36 dimensions representing the distance from the neck center to the collar line measured every $\pi/36$ as shown in Fig. 7. The Sift feature vector is computed for the collar area and the whole upper clothing area, which consists of 500 dimensions separately (Fig. 9). When being used alone, the Saliency feature vector has 15 dimensions (Fig.11). When combined with Sift, it has 500 dimensions.

As described in Section 2.5, our RF procedure starts by letting the users to choose one desired image out of an initial training set containing images of 10 different collar types plus front design type. This step is used to build the initial OPF classifier and has big impact on the results of the following RF steps. In other words, the type of collar the users choose at the initiation step should be an important parameter in designing the experiment for two reasons. One is its relationship with the feature vectors; the effectiveness of the proposed feature vectors should be different for different collar types. The other is its number used in the experiment as it should be easier to find a certain collar type if there are more images of such type in the database. The experiments presented in [37], however, ignored these factors completely. To solve the problems, in our new experiments, each subject is asked to test all the ten collar types plus one front designs for the three different combinations of feature vectors: Spread+Sift, Spread+Saliency and Spread+Saliency-Sift. Therefore, each subject performed 33 tests in total and we compared the effects of the three feature vectors for each collar type respectively. To eliminate the bias caused by the population of collar types in the database, the number of images of each collar type is used to weight the score inversely when compare the average score of the three feature vectors for all images. While previous experiments used only 4 subjects [37], we improved the reliability of experiment results by expanding the number of subjects to 10. The 10 subjects are female college students from school of nursing. For each of the three

feature vectors, they were asked to initiate the RF with each collar type in the initial training set. At each step of RF, the subjects were asked to mark each of the 5 training images as "relevant" (O) and "irrelevant" (×) and evaluate each of the query results as "satisfactory" (O) and "unsatisfactory" (×). Then a score ranging 0-5, which corresponds to the number of satisfactory images, is automatically computed for the query results of each step.

2.6.2 Comparisons of feature vectors (Spread+Sift, Spread+Saliency-Sift, Spread+Saliency)

Fig.18 shows the evaluation scores (averaged by ten subjects' scores) of the initial query results for the three feature vectors. We can observe that Sift and Saliency-Sift perform well at the first six collar types especially for the Square and the V types. Saliency-Sift works well for the Crew type also and Saliency is excellent for the Off shoulder type. We use the weighted (inversely proportional to their number percentage of total) average scores of the collar types to compare the overall qualities of the three methods, which will reduce the bias caused by the different number of images of different collar types. As is shown in Fig.19, the weighted score of the Saliency-Sift method is higher than those of the other two. One-tailed paired t-test reveals that the average score of Saliency-Sift in Fig.19 is significantly higher than that of Sift and Saliency at significance level 5% (p=0.05).

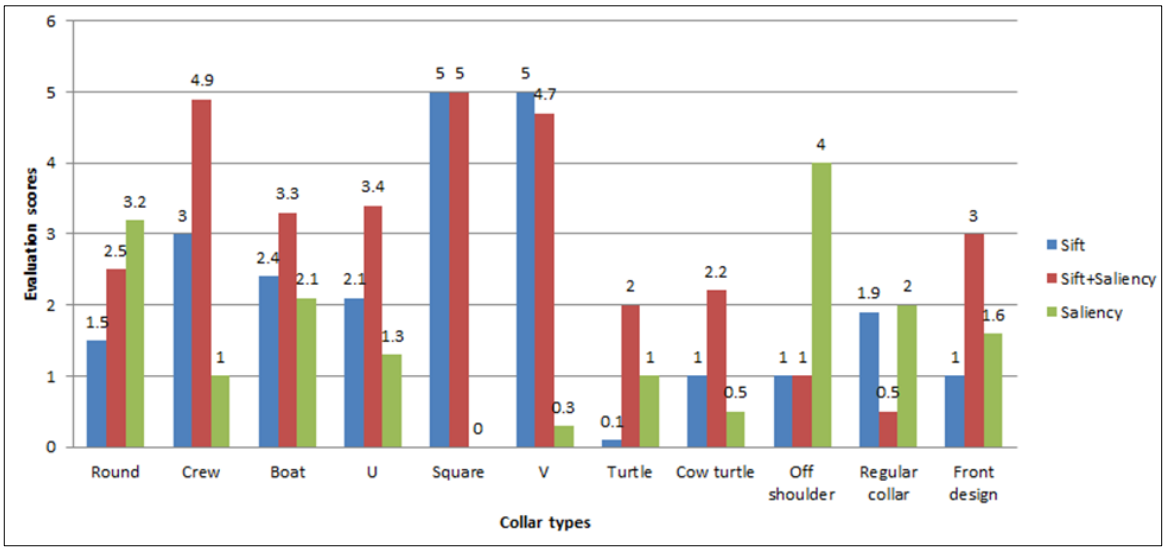


Fig. 18 Comparison of the average scores of the three methods on the collar types

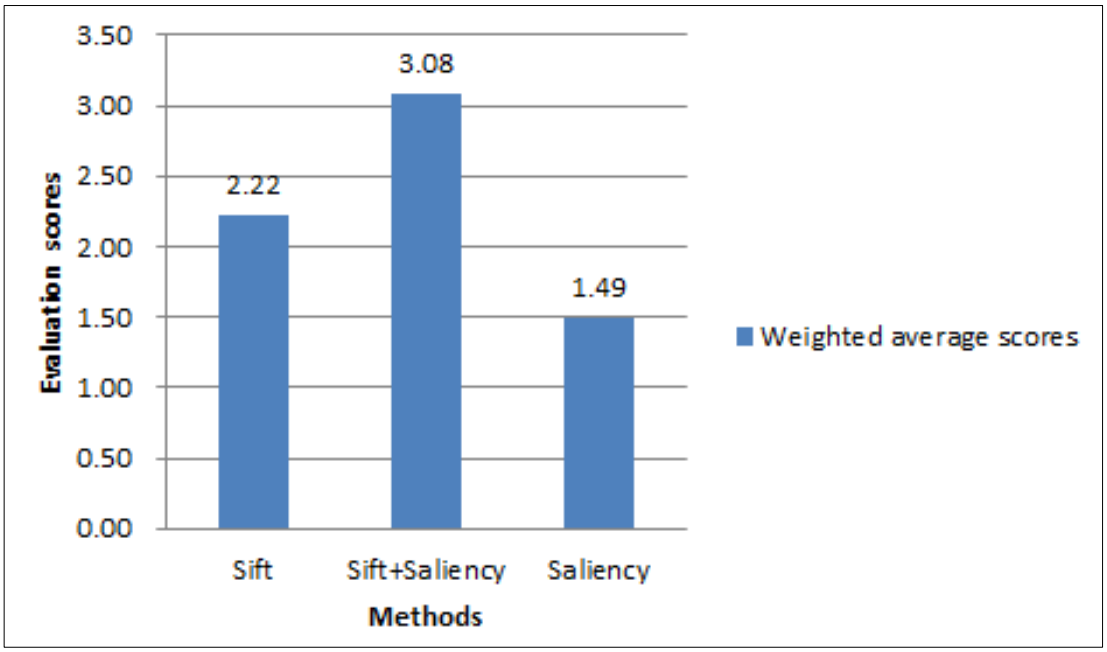


Fig. 19 Comparison of the weighted average scores of the three methods on the collar types

2.6.3 Improvement by RF-OPF

Although the results shown in Fig. 19 demonstrate that Saliency-Sift is more effective than the other two feature vectors, the initial query results of each feature vector is not good enough. It is because that the weighted average score of even the best case (Saliency-Sift) is just 3.08 against the full score of 5. We expect that continuing the RF iterations illustrated in section 2.5 can further improve the qualities of the query results.

So if the subjects get few satisfactory images from the first query results, they are asked to continue the query and evaluation process until they are satisfied with the number of the satisfactory images returned by the RF-OPF prototype. The evaluation scores at each RF step are also averaged by the ten subjects and inversely weighted by the number of collar types as we have done in dealing with the initial query results. Since the subjects always terminated the RF when the query results contain 4 or 5 desired images (the evaluation scores are correspondingly 4 or 5), the times of RF iteration they took also reflects the efficiency of our RF-OPF prototype. To validate whether RF-OPF is useful for improving the query results, we compare the average evaluation scores of all RF steps with the first query scores. Please note that not all collar types are used in computing the average evaluation scores of all RF steps. It is because that the results of the Square and the V collar types with the Sift and the Saliency-Sift methods satisfy the subjects at the initial query phase, and the Crew collar type with the Saliency-Sift method satisfies the subjects at the initial query phase too. On the other hand, Square collar type with the Saliency cannot get any suitable image at the initial query phase, which generates a score of zero that cannot be compared in this experiment. Therefore, as is shown in the left title of Fig.20, there are 10, 8 and 9 collar types with the above three feature vector combinations to be compared. Fig.20 shows that the RF-OPF prototype makes great improvement on the initial query results for all the three feature vectors, which are 107.69%, 43.28% and 95.47% respectively. But for those collar types not involved in the RF phases, the effectiveness of the RF-OPF prototype cannot be validated in this way. We partially address this problem with another experiment described later.

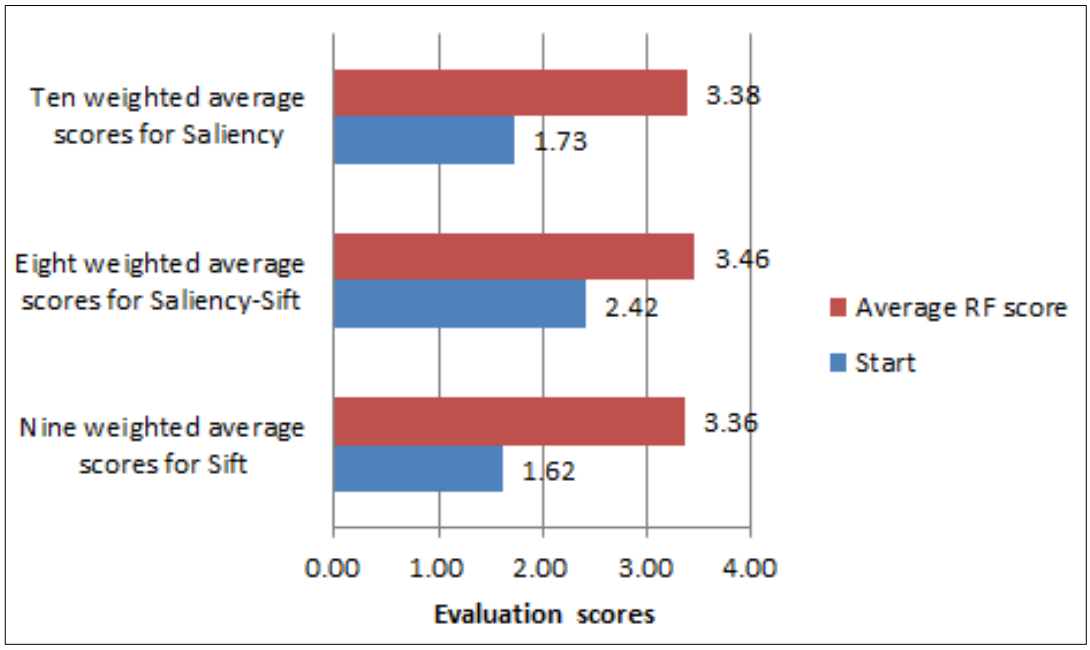


Fig. 20 The improvement of the RF-OPF prototype on the proposed features

The average RF iteration times the subjects take to reach their objects are the evidences to validate the efficiency of the RF-OPF prototype directly and the corresponding feature vectors indirectly. We compare the average RF times of the three methods from the ten subjects on each collar type in Fig.21. From this figure, some useful information about the efficiency of the RF-OPF prototype on the collar types can be observed, such as the Square collar type being still very difficult to be found with the Saliency feature vector even with the help of RF-OPF prototype. Fig.22 shows that the Saliency-Sift feature vector needs the least average RF times (1.30) to make the users satisfied, which demonstrates the efficiency of it. A one-tailed paired t-test for the RF times in Fig.22 shows that the difference between the Sift and Saliency-Sift and that between the Saliency-Sift and Saliency are statistical significance (p=0.05).The difference between the Sift and Saliency is statistical significance at some extent (p=0.1).

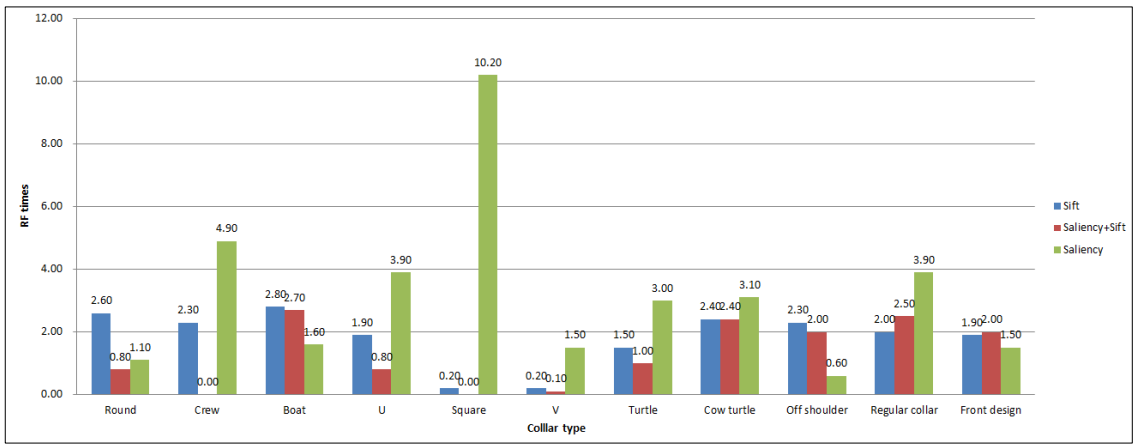


Fig. 21 Comparison of the RF times of the three methods on the collar types

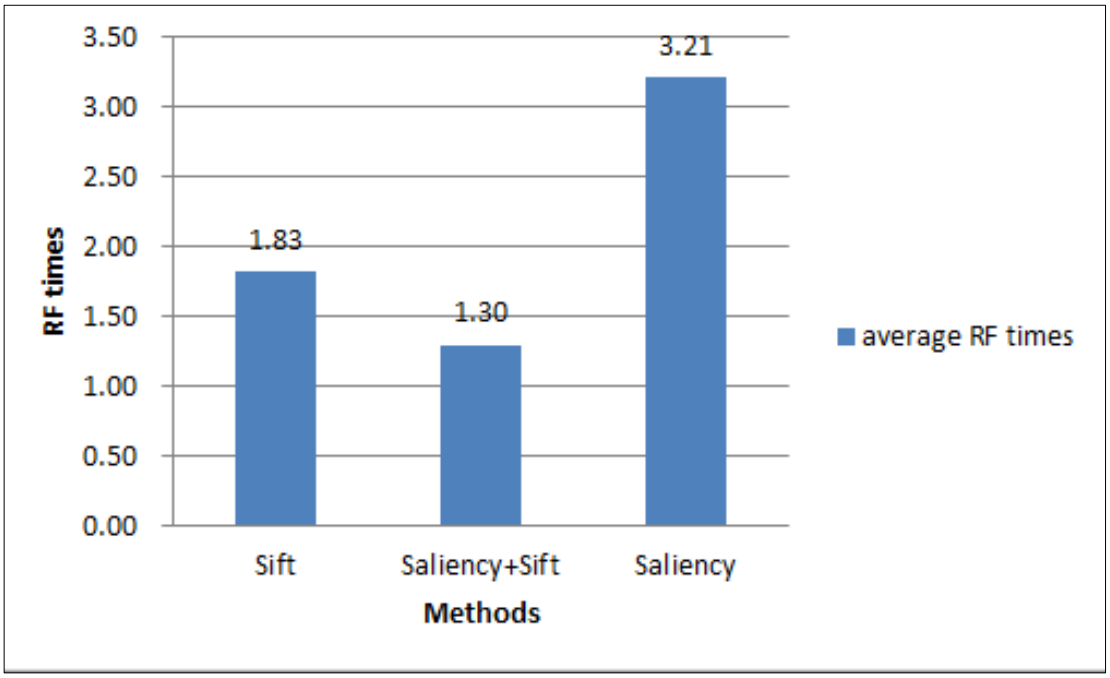


Fig. 22 Comparison of the average RF times of the three methods

Since the experiment results of Fig.20 do not contain the entire collar types, the effectiveness of the RF-OPF prototype cannot be validated completely. To solve this problem, we ask all the subjects to perform 5 times RF steps no matter what the results are at each step. Since the highest average RF times of the three feature vectors shown by Fig.22 is 3.21, 5 times is considered to be enough for this experiment. With the evaluation scores of the initial query and the five RF steps' results, we get Fig.23. It shows that the RF-OPF prototype improves the first query scores of the three feature vectors from 2.15, 2.82 and 1.53 to 4.54, 4.77 and 4.28 respectively. The improvements are significant and the scores trends are incremental in general.

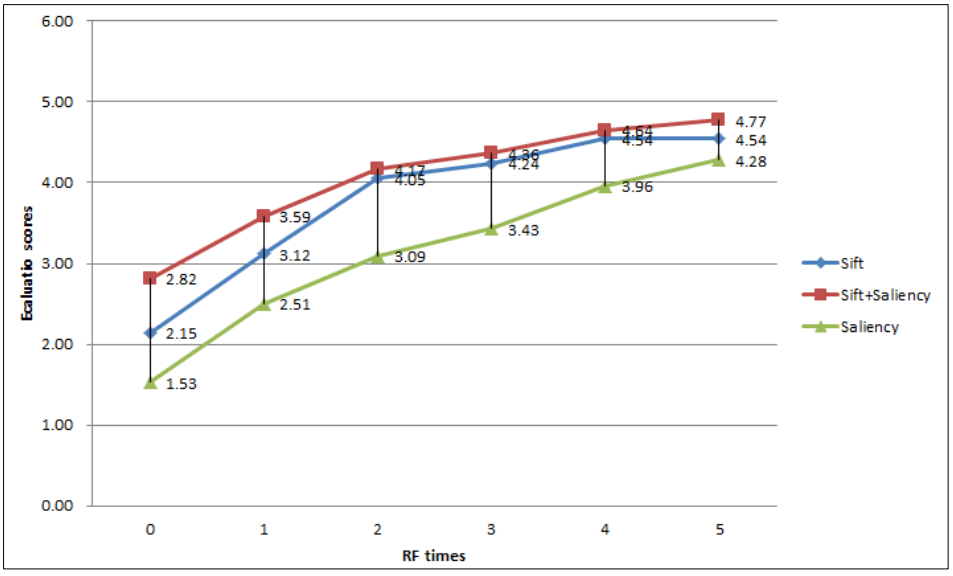
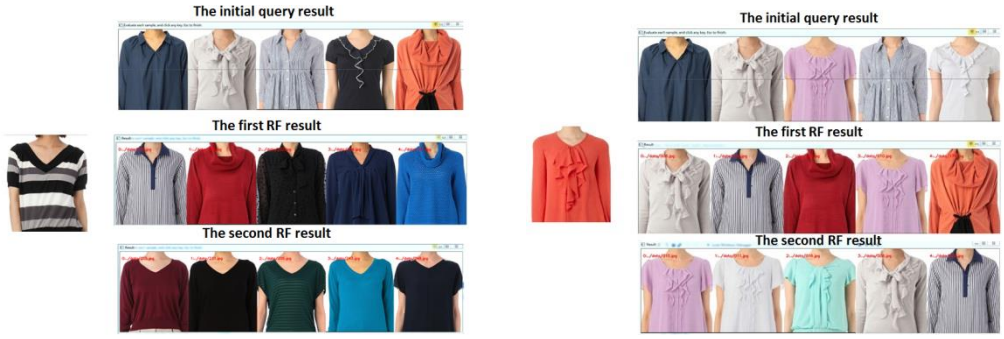


Fig. 23 Scores trends of five RF steps on the three feature vectors

At last we use the Saliency-Sift feature and the RF-OPF prototype to query similar collar images for the

two given images mentioned in Fig. 9. The results are shown in Fig.24, which shows that the striped V collar type image can get five image of similar collar design at the second RF step and the image with front design can get four images of similar design at the second RF step too. Note that, although the initial query results or even the first RF results may be very unsatisfactory, we can always get better results after several RF iterations.



(a) (b)

Fig. 24 (a) The top five results for a striped V collar image with the Saliency-Sift method by the second RF phase. (b) The top five results for a front design image with the Saliency-Sift method by the second RF phase

2.7 Conclusion

Collar design plays an important role when people choose the desired clothing. In this paper, we focus our research on the retrieval of clothes image based on the collar design. Three different feature vectors Sift, Saliency and Saliency-Sift combined with Spread for capturing the detailed features of collar design and a CBIR prototype based on RF method and OPF classifiers are proposed. Through experiments, it is proved that Saliency-Sift is the best among the three feature vectors in terms of both effectiveness and efficiency. Our experiment results also demonstrated that the proposed RF-OPF prototype improves the qualities of the query results significantly.

But there are still some technical problems to be solved. 1. The Saliency-Sift feature vector still does not react well to some collar types, such as the Off shoulder and the Regular collar types. 2. The subjects may be confused by some similar collar types, such as the Round and the U collar type, the Crew and the Boat collar type during the RF phases in the experiment. In addition, the collars combing multiple design features, such as a collar combing turnover with V spread or front design are difficult to be classified by the subjects. 3. The scores of the query results are computed as the number of satisfactory images without considering the degree of satisfactory of each image in the query results. To solve the first problem, we plan to develop and experiment with other new feature vectors for capturing the collar design. The second problem can be solved by improving the spread feature vectors for better describing the roundness, width and angle, so as to better discriminate collar types of similar shape, like the Round, Crew, Boat and U collar types. The 3rd problem can be solved by asking subjects to score each of the retrieved images and analyze the trend of highest and average scores of RF iterations. The current image database is relatively small. We need to gather more images for improving the reliability of the experiment results. Another important issue is that the initial image sets for letting user to select one collar type to initiate the RF have large impact on the results of succeeding steps. We need to explore some new methods which can always

avoid misleading the building of the classifier.

3D Garment design is a rapidly developing field [33, 34]. Presenting 3D clothing images in clothes-shopping websites will help people find the desired clothes more easily. So it is very important for us to extend our research to deal with 3D collar design features in the future. Although collar is very important for choosing clothing, Color, texture and the design of other parts of garments also affect how people choose clothes at different extent. We are going to extend our retrieval system by considering other features. It should be easy to capture color and texture features with existing computer vision technologies. The technology provided in [38] can be employed for capturing the overall shape of garments. Moreover, the proposed RF-OPF system should have high potential to deal with more design factors as it is reported that OPF is superior to conventional learning algorithms such as ANN-MLP, K-NN and SVM in both computation time and accuracy especially in complex situations, i.e., with a large amount of overlapped regions [48,50].

Chapter 3 Caricature Synthesis with Feature Deviation Matching under Example-Based framework

Chapter 3 mainly describes details of the proposed caricature synthesis framework, which is an example-based framework based on a cross-modal distance metric called feature deviation. It can generate various types of caricatures without paired example databases.

3.1 Introduction

Since the early 1980s, many computer-based methods have been developed for synthesizing caricatures [4]. These can be roughly classified into photo-transformed and example-based techniques. Photo-transformed approaches [5-8, 14-16**,** 27] achieve caricature styles by applying certain kinds of image filtering or geometric deformation to input portrait photos. In these systems, filters and deformations are usually tailor-designed for a particular style, and hence cannot be generalized to different styles without changing the underlying algorithms. Example-based systems [9-13, 17-20] require a large number of photo–caricature pairs as example data. In principle, the example-based approach has the advantage that a single framework can be used for generating caricatures of various styles given corresponding styles of example databases. However, obtaining sufficient sets of paired photo–caricature images is usually difficult. Many existing example-based systems asked artists to draw caricatures from a large number of example portrait photos, which is not always possible in real applications. Recently, approaches using deep-learning have attracted substantial attention. While supervised deep-learning approaches [27] may suffer from the problem of requiring even larger number of photo–caricature pairs than traditional example-based methods, unsupervised approaches using the concept of style-transfer-based or image analogies have also been developed [22-27]. With style-transfer-based systems, it is easy for users to apply a specific form of artwork stylization to their own photos for sharing and entertainment purposes. The basic principle of neural-style transferring is to separate a given style from the content of an image by considering different layers of a neural network. Since the stylistic information is mainly represented by low-level textural and color features, these methods are not suitable for achieving geometric stylization, such as deforming (exaggerating) the shapes of individual facial components, which is a technique commonly found in real caricatures. Essentially, deep-learning is an end-to-end approach, and its existing implementations do not provide users with any control over the details of stylization, such as the degree of exaggeration of individual facial components.

In this Chapter, we propose a new example-based caricature generation technique that can synthesize stylized caricatures with a small number of unpaired examples of portrait photos and caricatures. Our technique incorporates component-specific learning based on feature vectors that intuitively match the features that people employ to perceive or communicate the characteristics of faces, which can also provide users with control over the individual facial components. While existing component-specific learning methods [9-13, 17, 18, 51, 52] require paired photo-caricature examples and search for a matching

caricature component via its corresponding photo component, the proposed method searches for the matching caricature components in their feature spaces directly. However, caricatures of expressive styles will not always provide an entirely faithful reflection of the features evident in source portrait photos, and hence a direct comparison of feature vectors between the feature space of caricatures and the feature space of photographs is meaningless. To solve this problem, we propose a new cross-modal distance metric called feature deviation matching. The key idea is that, given fact that a caricature is an expressive representation of a person's prominent features, the feature spaces of both the original photographs and resulting caricatures should show strong correlation between these deviations from their corresponding averaged features. Therefore, the extent of deviation from averaged features across corresponding photo and caricature facial component feature spaces, despite of the modality difference between photo and caricature components, can be used to search for matching facial caricature components directly under the example-based framework. To compute the deviation, the proposed method uses one set of example photos and one set of example caricatures to learn the distributions of their respective feature spaces. The images in these two example sets are not necessarily photo-caricature pairs of the same persons, and the building of such training sets becomes much easier.

In summary, the contributions of this proposed framework are:

1. The newly proposed cross-modal distance metric called feature deviation matching technique makes it possible to generate various styles of caricatures under the conventional example-based framework without requiring paired photo–caricature training sets.

2. By focusing only on the perceptually prominent features, the designed feature vectors are robust and effective for capturing the visual features of input portrait photos.

3. The proposed system enables users to control the exaggeration of individual facial components. Various combinations of individual facial and hairstyle components, based on different exaggeration coefficients and similarity rankings, can provide users with different candidates to satisfy their particular preferences; this has not been achieved in most existing style-transfer-based and deep-learning-based approaches.

Fig.25 shows the comparison of our synthesized caricatures of expressive style with three state-of-art systems (The first and second rows: Comparison with sketch example-based system [15]. The third and fourth rows: Comparison with deep-learning based photo-transformed system [53]. The fifth to eighth: Comparison with component-based system [52].). It can be found that the caricatures generated by the sketch example-based system and the photo-transformed system resemble the input portrait photos the best, but lack artistic feelings. Example or component based systems can provide various styles of caricatures with artistic feelings. The caricatures by our proposed system are competitive with [52] method for similarity and are comparable with [15, 53] for as an expressive style.

The remaining part of the paper is organized as follows: Section 3.2 reviews related works. Section 3.3 presents the proposed method. Section 3.4 demonstrates some results and describes the evaluation experiments. Section 3.5 presents conclusions from the study.

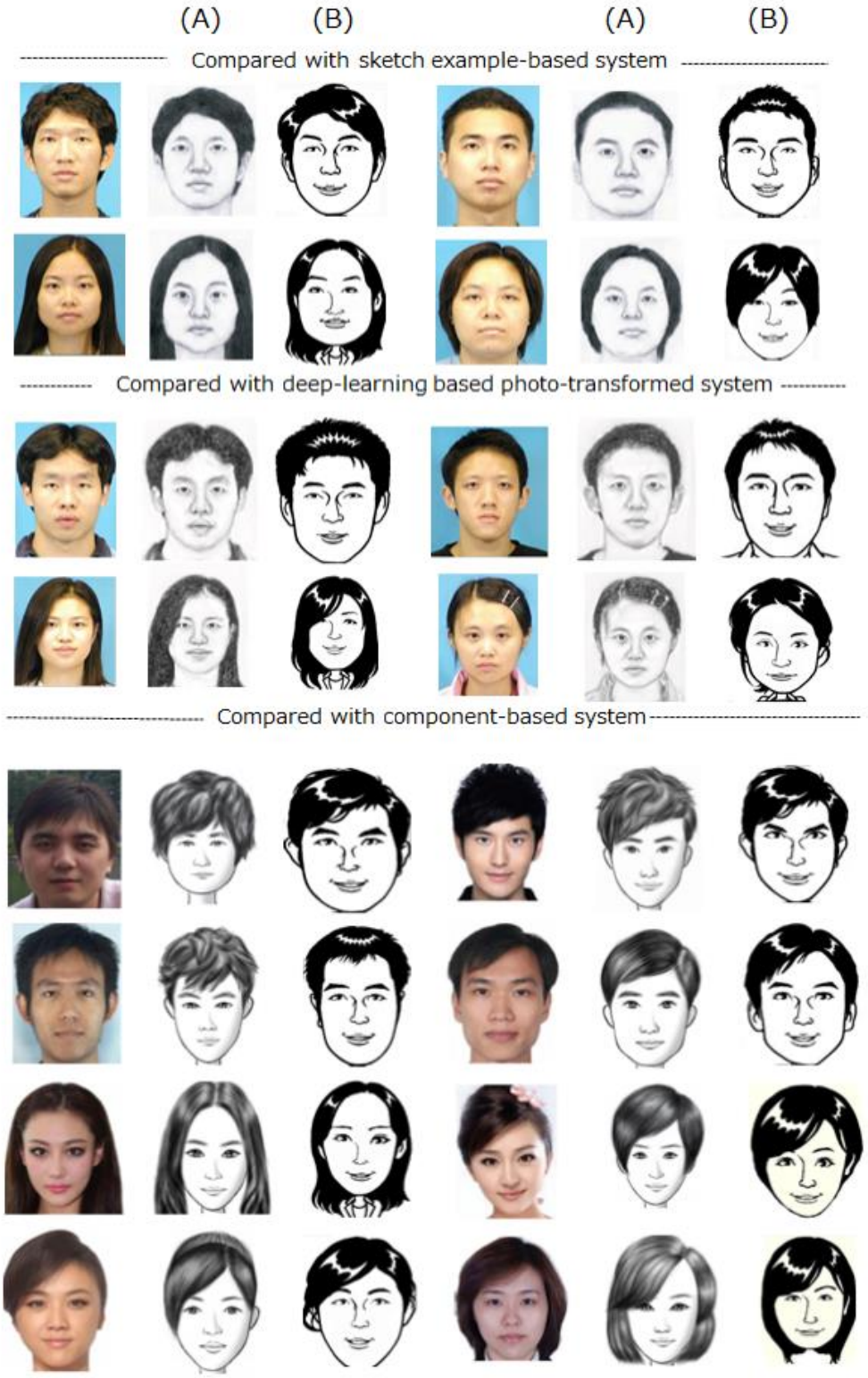


Fig.25 Comparison of the expressive style synthesized by our proposed system with those generated by three state-of-art systems (a sketch example-based system [15], a deep-learning based photo-transformed system [28] and a component-based system [52]), A and B columns show the results of the other systems and our system respectively.

3.2 Related researches

3.2.1 Caricature synthesis techniques

Photo-transformed and example-based systems are the two main approaches used in early caricature synthesis systems. Style-transfer-based systems have recently attracted attention, and produce results that combine original portrait photos with various styles of example caricatures.

Photo-transformed systems usually use certain image processing techniques to transform portrait photos into caricatures. Gooch et al. [14] succeeded in creating caricatures that highlight and exaggerate representative facial features, by first generating black-and-white facial illustrations from photographs and then deforming the facial illustrations. Min et al. [7] proposed an automatic portrait system that leveraged the And/Or graph based on existing sketch templates. Transductive learning-based face sketch-photo synthesis technique was proposed to optimize both the reconstruction fidelity of the input photo (sketch) and the synthesis fidelity of the target output sketch (photo), which can efficiently optimize the corresponding probabilistic model by alternating optimization [6]. Recently, Zhang et al. proposed a content-adaptive method for generating portrait sketches based on deep-learning techniques under the photo-transformed framework [26], which can preserve non-facial factors such as hairpins and spectacles better than previous techniques.

Example-based systems generate caricatures by assembling facial and hairstyle caricature components that resemble the input portrait photo's into the output caricature template. Based on Markov random fields [17] and image denoising techniques [18], methods are proposed to create sketches from photos by selecting the most appropriate neighboring patches to synthesize a target patch. Chen et al. employed a large database of photo–caricature pairs and successfully reflected personal features in a visually natural rendering by employing non-parametric sampling techniques [12]. Yang et al. proposed the technique for synthesizing a caricature by searching a database of teaching pairs, which can produce further exaggerated effects by adjusting component sizes and positions [19]. Zhang et al. improved the synthesized results of example-based systems by using machine learning methods to optimize the combinations and positions of facial components [20].

Style-transfer-based systems impose the visual attributes (such as color and texture) of the example caricatures onto input portrait photos to generate various corresponding styles of caricatures. Shih et al. transferred the local statistics of an example caricature onto an input portrait [54], allowing users to easily reproduce the visual styles of renowned artists. Liao et al. used their "Deep Image Analogy" technique to synthesize target caricatures, which finds semantically-meaningful dense correspondences between the example caricature and the input portrait photo by adapting the notion of "image analogy" with features extracted from a deep convolutional neutral network for matching [24]. Fisher et al. performed non-parametric texture synthesis, which retains more of the local textural details of the artistic exemplar compared with other style-transfer-based systems and does not suffer from image warping artifacts caused by aligning the style exemplar with the target face [25]. Furthermore, the system was able to generate perfect animation by combining consistent caricatures.

Taigman et al. proposed an unsupervised cross-domain image generation method based on a domain transfer network (DTN), which can generate caricatures while preserving the identities of the input face images [26].

Although effective in various applications, the above systems have respective shortcomings. Photo-transformed systems are usually based on style-specific algorithms, and hence cannot be generalized to different styles. Large databases of photo–caricature pairs are required for the example-based systems—a drawback that renders them impractical. The caricature synthesis systems based on supervised deep-learning techniques require even larger databases of photo-caricature pairs than traditional example-based approaches. Using either deep-learning or traditional parametric or non-parametric approaches, style-transfer-based systems can only transfer texture or color, whereas the stylization of geometric features, such as the shapes of facial features, are particularly important for achieving the expressive styles of caricatures.

Based on the newly proposed cross-modal distance metric, feature deviation matching, our proposed system can synthesize various styles of caricatures with unpaired photo and caricature databases. Compared with photo-transformed and style-transfer-based systems, the proposed system can provide users with candidate caricatures that have various combinations of individual facial components based on different exaggeration coefficients and similarity rankings.

3.2.2 Feature deviation applications and cross-modal comparisons

Based on the variation within a population of faces, one could determine an average face [55]. Average face and facial component features are widely used for exaggeration control on facial components and areas in caricature generation and other similar applications. Brenman proposed a widely used rule "Exaggerating the Difference From the Mean face" (EDFM) and designed the corresponding system called "Caricature Generator", which exaggerated a graphic representation of a subject face according to the differences from an average face computed from a face dataset [56]. Koshimizu et al. defined another EDFM rule and applied it to their interactive system (PICASSO), which can generate an output caricature from a source image based on certain deviation value of the source image from the average image [57]. Mo et al. used normalized deviation from the average model to exaggerate the distinctive features, based on the consideration that the DFMs (Difference-From-Mean) for different components are different [58]. Xu et al. investigated the borderline between likeness and unlikeness through applying the EDFM rule to gradually alternate the face shape at subject study [59]. Cosker et al. learned animation parameters from human video performance and reused them to animate multiple types of facial model [60], which successfully used feature deviation for remapping the facial expression parameters between different appearance models and between the appearance models and 3D models. Firstly, it employed PCA to different areas of face image to generate corresponding features. Then calculated their deviation values from the corresponding average features for generating shape-free features (deviation features) to act as the training data since facial expressions can be regarded as dynamic changes of the corresponding facial areas, which can be related with the exaggeration level represented by deviation values from average features. Although [56-60] used feature deviation values to build the distribution of feature spaces or control the exaggeration level of facial components and areas, the feature deviation values came from a same feature space.

Canonical Correlation Analysis (CCA) has been a very popular method for embedding multimodal data in a shared space to analyze the linear relation between different modalities [61, 62]. Recently, Deep-learning techniques are also widely applied in cross-modal analysis between different modalities [63, 64]. Deep Canonical Correlation Analysis (DCCA) was proposed to analyze the non-linear relation between different modalities, such as image with audio, image with text, audio with text, and so on. [65-67] used different sub DNNs composed of fully connected layers to covert features of different modalities into low dimensional features in a shared D-dimensional semantic space and then calculate their similarities by

CCA function. But the above methods need to prepare a great number of data and require high computational cost.

Traditional paired-example based systems use labels to relate the example photo components to the paired example caricature components and then search for the similar caricature component by comparing the input photo components with the example photo component, which can avoid the cross-modal problem. But it is not easy to prepare large paired-example databases for different styles of caricatures, which requires different artists' hard work. To alleviate drawback of the paired-example based approach, we proposed a new cross-modal distance metric based on feature deviation matching, which compares the input photo components with the example caricature components directly. While most of the existing methods use feature deviation in one feature space, our proposed method focuses on using feature deviation for matching across different feature spaces. Although CCA and DCCA methods are effective to solve the multimodal problems, such as between image and audio, image and text and so on, they require high computational cost and large dataset. Our proposed method is much simpler to implement and effective for the cross-modal matching between portrait photos and different types of caricatures.

3.3 Proposed methods

3.3.1 Overview of framework

Fig. 26 provides an overview of the proposed system. In the offline phase, we use one photograph database and one caricature database to learn the distribution of feature spaces—in other words, to compute the extent and average of feature vectors in each feature space respectively, as formulas (6) to (10) show. Feature deviation matching is performed component-by-component across photo and caricature. The ASM (Active shape models) algorithm [68] is used to detect the feature points required for computing the feature vectors of facial components from both photos and caricatures. By focusing only on perceptually prominent features, the designed feature vectors are robust against ASM fitting error and can effectively capture the visual features of the input portrait photos. However, the fitting results of ASM are instable for some caricatures occasionally. Fortunately, it is not necessary to detect the feature points of caricature components during the online phase. Therefore, we manually adjust the instable fitting results of ASM to locate the feature points at the correct positions on the caricatures. It takes about from several seconds to several minutes at most to adjust instable fitting results for each caricature. In the online phase, given an input portrait photo, ASM is applied to detect the feature points required for computing the feature vectors of individual facial components from the input portrait photo.

Then, for each component, values are calculated representing the deviations of actual features from averaged feature vectors in their corresponding feature spaces. By comparing the feature deviation values of the input portrait photo components with those of the caricature components, the most similar caricature components are found, as formula (9) shows. After deforming the facial contours of the output caricature, deciding the component positions, and adjusting their sizes, the searched facial components are composited into the output caricature. Details of the procedures are described in the following subsections.

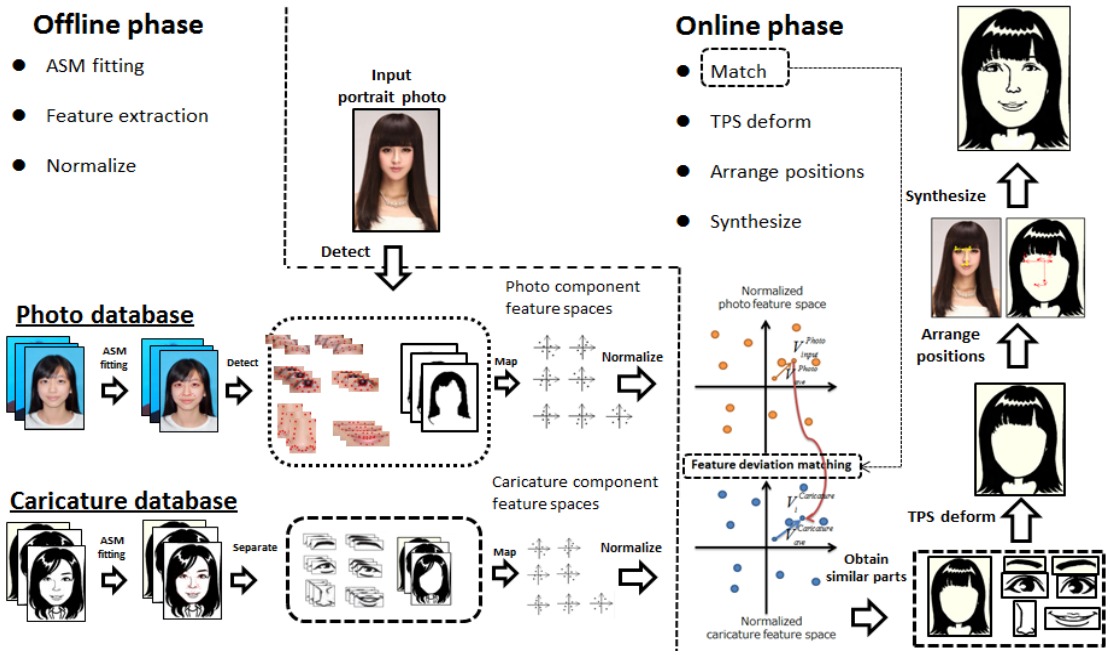


Fig. 26 Framework of the proposed caricature synthesis system

3.3.2 Feature vectors

There are external and internal features exist in human portraits. Hairstyle and face shape can be regarded as the external features, and eye, eyebrow, mouth and nose are considered as internal features. One of the largest advantages of caricatures is that they can emphasize a person's most prominent facial features to make it easy for observers to identify the subject at a glance. For creating such caricatures, feature vectors should be carefully designed to capture the prominent features of faces. By observing many types of caricatures, it can be found that mostly eyebrow and nose components contain little internal details. The positions of eyeballs and the status of eyelids and lips are indeed very important. But in our currently processed caricatures, most of the eyeballs are in the center. It is better to deal with the status of eyelids and lips by preparing two sets of corresponding components (closed and open) and using threshold values like [20] to determine which set to be used. Our previously designed features in [69] tried to capture the internal and external details of facial components simultaneously. But the experiment results showed that they usually canceled out each other. In addition, humans perceive faces based more on characteristic information—small, thin, and drooping, for example—than on precise shape information. Due to such considerations, we have designed feature vectors mainly composed of simple geometric characteristics that reflect the overall information of individual facial components.

Using only a minimum set of characteristic features can also alleviate the negative effect of occasionally inaccurate ASM fitting results to some extent. The newly designed features are illustrated in Fig. 27, and the evaluation results shown in Section 3.4 are of comparable quality.

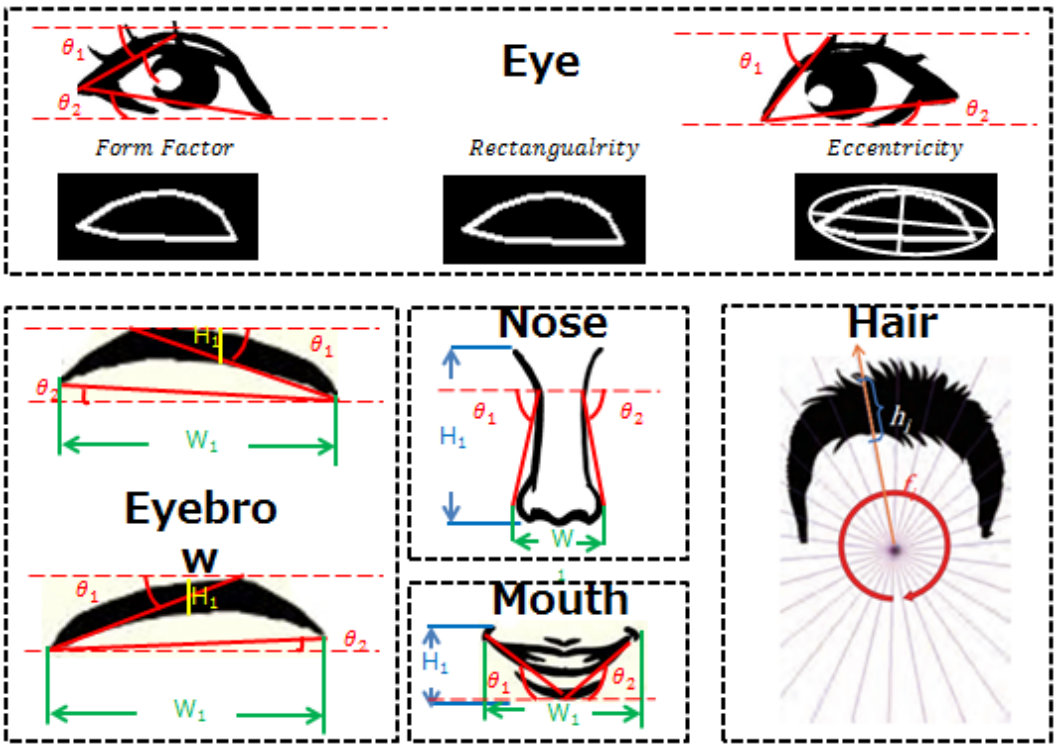


Fig. 27 Designed feature vectors of facial and hairstyle components

Eye: Two angles, form factor, rectangularity, and eccentricity are used to constitute a 5d feature vector { θ_1, θ_2, Form Factor, Rectangularity, Eccentricity} for representing the eyes. After drawing the eye contours by connecting the corresponding detected ASM feature points of eyes, we calculate the areas and perimeters of the eye contours and the areas of their corresponding bounding boxes. The lengths of the long and short axes of their corresponding smallest circumscribed ellipses are also computed. Form factor features can be obtained by using the product of 4π with the eye contour area, divided by the square of its perimeter as formula (3). Rectangularity is the result of the eye contour area divided by its corresponding minimum bounding box area as formula (4). The eccentricity value of the eye is calculated by dividing the long axis by the short axis as formula (5).

$$Form\ Factor = \frac{4\pi \times Area}{Perimeter^2} \tag{3}$$

$$Rectangualrity = \frac{Area_{object}}{Area_{bounding-box}} \tag{4}$$

$$Eccentricity = \frac{AxisLength_{long}}{AxisLength_{short}} \tag{5}$$

Eyebrows, Nose, and Mouth: Eyebrows, nose and mouth are represented by 3d feature vectors {θ_1, θ_2, $\frac{W_1}{H_1}$}, which are made up of two angles and their width/height ratios.

Hairstyle: Hairstyle is the most influential and discriminative component in the facial image. We use the method in [70] to detect the hair region, and adopt a feature vector similar to that used in [13] for

representing the hairstyle component. Here, the hairstyle is represented by a 120d feature vector $\{\frac{h_i}{f_i}\}$ $(i=1,2,...120)$, which is achieved with the following three steps:

Step 1: Deploy hair seed lines.

Step 2: Separate the hair area from the face image via the watershed algorithm [71].

Step 3: Calculate the 120d feature vectors.

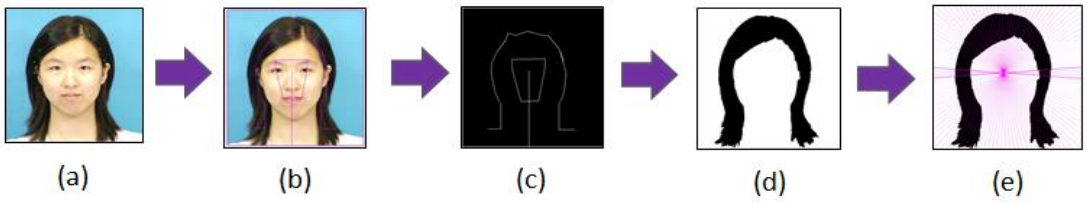


Fig. 28 Procedures for separating hair area from the input portrait photo and obtaining hair feature vectors

In step 1, as shown in Fig. 28 (a), starting with the ASM points on the upper face contour, we deploy the upper hair seed points according to the color difference of adjacent pixels. In addition, we also detect the bottom-left and bottom-right hair areas to determine whether the subject has long hair, especially for female portrait photos; if so, then the lower hair seed points are also deployed. The center points of the eyebrows, eyes, nose, mouth, and the four corner points of the portrait photo framework act as background seed points. By connecting the hair seed points and the background seed points separately, we generate seed lines (Fig. 28 (b)).

In step 2, after drawing the white seed lines on a black background image of the same size as the input portrait photo (Fig. 28 (c)), the watershed algorithm is then used to segment the input portrait photo into two parts, the hair and background area (Fig. 28 (d)).

In step 3, 120 straight lines are drawn, radiating from the center of the face (the bottom point of the nose, detected via ASM, was used in our paper), as shown in Fig. 28 (e). The points at which the straight lines intersect the hair region at each angle are located, and corresponding ratios are calculated (equal to the hair thickness value divided by the distance from the center point to the first intersection point).

Face shape: The proposed system does not regard face shape as a single component, but instead handles this together with the hair, as described in the following section on the hair-contour component.

3.3.3 Deviation-based feature matching

Even if the proposed system succeeds in calculating the same feature vectors for the portrait photo and caricature components, it is unlikely that the feature vectors from different feature spaces will match up. To deal with this cross-modal problem, traditional example-based systems construct paired photo–caricature databases. By using labels to relate the photo components to the paired caricature components, the feature vectors of the input photo components need only be compared with those in the photo component databases. The most similar caricature components are then found according to the related labels. However, acquiring a large number of paired photo–caricature examples is very difficult and hence limits the use of such systems in real applications.

As described in Section 3.1 and 3.2, the tendencies in photo and caricature component feature spaces would follow the same patterns. We propose a new cross-modal distance metric, namely feature deviation matching, for matching items across different feature vector spaces.

Before defining the formulas to compute the deviation values, the variables were denoted as follows: feature vector for a facial or hairstyle component $i \in$ {left eye, right eye, left eyebrow, right eyebrow, nose, mouth, nose, and hairstyle} of the input portrait photo as V_i^{in} ; the $j-th$ ($j=1,2,...,n,n$ represents the number of photos in the photo component database) image of portrait photo component i in the corresponding photo component database as $V_i^{Pho,j}$; and the $k-th$ ($k=1,2,...,m,m$ represents the number of caricatures in the caricature component database) image of caricature component i in the corresponding caricature component database as $V_i^{car,k}$. The maximum, minimum, and average feature vectors of component i in the corresponding photo and caricature component databases are denoted as $\hat{V}_i^{pho}$, $\breve{V}_i^{pho}$, $\bar{V}_i^{pho}$, $\hat{V}_i^{car}$, $\breve{V}_i^{car}$, and $\bar{V}_i^{car}$. The normalized feature vectors of the input photo components, and those in the photo and caricature component databases, are correspondingly denoted as α_i^{in}, $\alpha_i^{pho,j}$, and $\alpha_i^{car,k}$. The maximum, minimum, and average values of the normalized feature vectors are subsequently denoted as $\hat{\alpha}_i^{pho}$, $\breve{\alpha}_i^{pho}$, $\bar{\alpha}_i^{pho}$, $\hat{\alpha}_i^{car}$, $\breve{\alpha}_i^{car}$, and $\bar{\alpha}_i^{car}$. The feature deviation values are defined as β_i^{in}, $\beta_i^{pho,j}$, and $\beta_i^{car,k}$. The above variables are then calculated by the following formulas:

Firstly, each dimension in the feature vectors of facial and hairstyle components is normalized using formulas (6) to (8) so that all values are within the range 0 to 1.

$$\alpha_i^{in} = \frac{(V_i^{in} - \breve{V}_i^{pho})}{(\hat{V}_i^{pho} - \breve{V}_i^{pho})} \tag{6}$$

$$\alpha_i^{pho,j} = \frac{(V_i^{pho,j} - \breve{V}_i^{pho})}{(\hat{V}_i^{pho} - \breve{V}_i^{pho})} \tag{7}$$

$$\alpha_i^{car,k} = \frac{(V_i^{car,k} - \breve{V}_i^{car})}{(\hat{V}_i^{car} - \breve{V}_i^{car})} \tag{8}$$

After normalizing the feature vectors of the input photo components and those in the photo and caricature component databases, we calculate their corresponding feature deviation values with formulas (9) and (10).

$$\beta_i^{in} = \frac{\alpha_i^{in} - \bar{\alpha}_i^{pho}}{\hat{\alpha}_i^{pho} - \breve{\alpha}_i^{pho}} \tag{9}$$

$$\beta_i^{car,k} = \frac{\alpha_i^{car,k} - \bar{\alpha}_i^{car}}{\hat{\alpha}_i^{car} - \breve{\alpha}_i^{car}} \tag{10}$$

The matching caricature component $\tilde{k}_i$ in the caricature component database can be determined via formula (11).

$$\tilde{k}_i = \arg\min_k \left\| \beta_i^{in} - \beta_i^{car,k} \right\| \tag{11}$$

By applying formula (11), the proposed system searches for the most similar facial and hairstyle caricature components in the caricature database.

In addition, more exaggerated or averaged caricatures can also be easily synthesized similarly to [19] by configuring an exaggeration coefficient in formula (11) as follows:

$$\tilde{k}_i = \arg\min_k \left\| \partial\beta_i^{in} - \beta_i^{car,k} \right\| \tag{12}$$

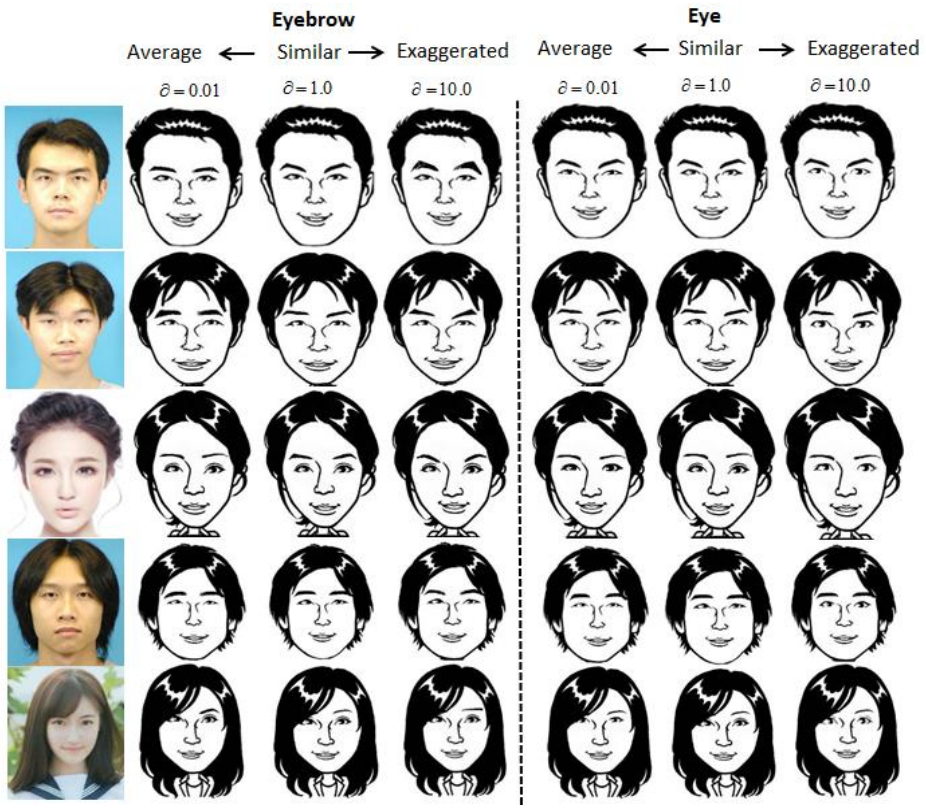


Fig. 29 Synthesized caricatures with different exaggeration coefficients for eyes or eyebrows

The proposed system then searches for exaggerated components (where the deviation from the mean is larger) when $\partial > 1$, similar components (where there is small deviation from the mean) when $\partial = 1$ and averaged components (where the deviation from the mean is smaller) when $0 < \partial < 1$. In the current implementation, the exaggeration coefficient can be applied to the facial components except for the hairstyle component. Fig. 29 shows five examples in which eye and eyebrow features generated using differing coefficients are imposed onto synthesized male and female caricatures. Each example contains

three caricatures with the exaggeration coefficients set to 0.01, 1.0, and 10.0 for eyes or eyebrows separately.

3.3.4 Synthesis of resulting caricature

The proposed system thus synthesizes the output caricatures with the searched caricature components via the following two steps.

3.3.4.1 Deform the output caricature face shapes using the Thin Plate Spline (TPS) algorithm

Since a large part of a facial shape is actually represented as the boundary of the hair region, we treat the hair and face shape as a single hair-contour component when preparing the example databases. The search procedure does not attempt to find specific contours matching the input portrait photos, but uses the feature vector of hairstyle only for feature deviation matching. However, it is known that the overall impression of a face varies considerably according to the face shape [59]. Therefore, in the synthesis phase, we deform the face shape of the searched hair-contour component to match that of the input portrait photo. The deformation procedure is illustrated in Fig. 30.

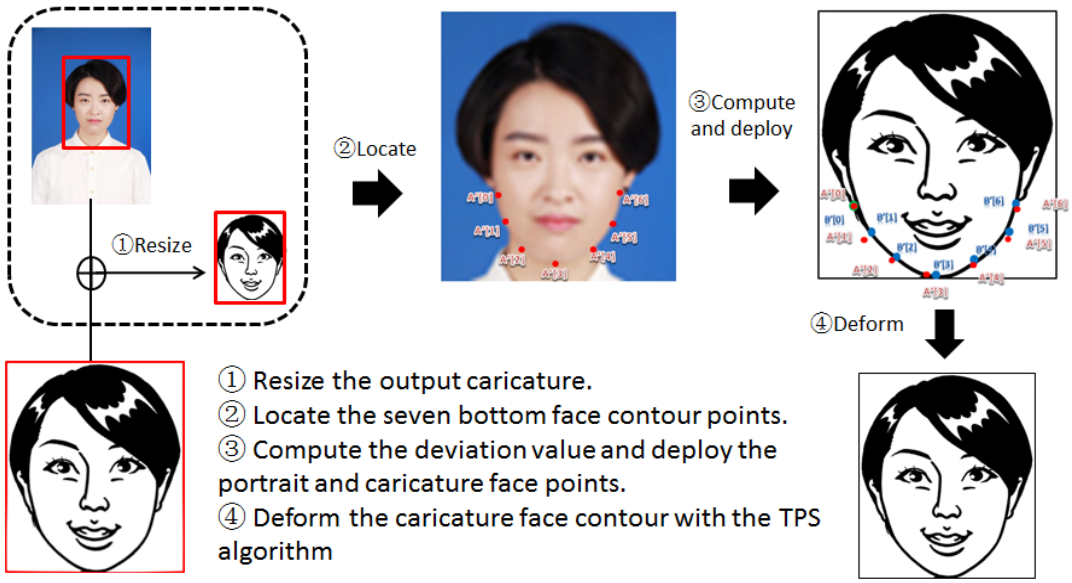


Fig. 30 Deformation of the caricature face shape using the TPS algorithm

At first, we compute the width and height ratios of the input portrait face over the searched caricature *hair-contour* component. According to the computed ratios, the searched caricature *hair-contour* component is resized similarly to the input portrait face. Then, the Thin Plate Spline (TPS) algorithm [72] is applied to deform the resized hair-contour component so that the face contours of the input photo and the output caricature align with each other. TPS deformation uses the seven points on the face-bottom contour, which are detected with ASM fitting.

3.3.4.2 Arrange the components into the output caricature

With the deformed hair-contour caricature component as the template, we arrange the searched facial components onto it appropriately to create the output caricature. Before doing so, the proposed system delineates circumscribed rectangles on the template for the individual facial components. The circumscribed rectangles around the corresponding facial components are defined as the bounding boxes of

all the feature points of the components. To maintain the relative sizes and shapes of searched components in the output caricature, formulas (13) and (14) are applied to calculate the widths (W) and heights (H) of the circumscribed rectangles.

$$W_i^{out} = \frac{W_i^{in}}{W_{face}^{in}} \times W_{face}^{out} \tag{13}$$

$$H_i^{out} = \frac{H_i^{sea}}{W_i^{sea}} \times W_i^{out} \tag{14}$$

In addition, the proposed system uses the following predefined rules to make the output caricature face layout resemble that of the input portrait photo.

1. Define the starting center point of the face.

The proposed system uses the bottom point of the nose as the center point, as this point is always most correctly fitted by ASM.

2. Locate the positions of the upper-left corners of the facial component rectangles.

To maintain consistency of layouts between the input portrait photo and output caricature face, the relative distances among their facial components are calculated by formulas (15) and (16) (x and y represent horizontal and vertical distances between individual facial components).

$$d_i^{out,x} = \frac{d_i^{in,x}}{W_{face}^{in}} \times W_{face}^{out} \tag{15}$$

$$d_i^{out,y} = \frac{d_i^{in,y}}{H_{face}^{in}} \times H_{face}^{out} \tag{16}$$

As shown in Fig. 31, starting from the bottom point of nose (regarded as the center point), the positions of the eyes and mouth can be located by their distances from it. Subsequently the eyebrow positions are deduced from the positions of the eyes.

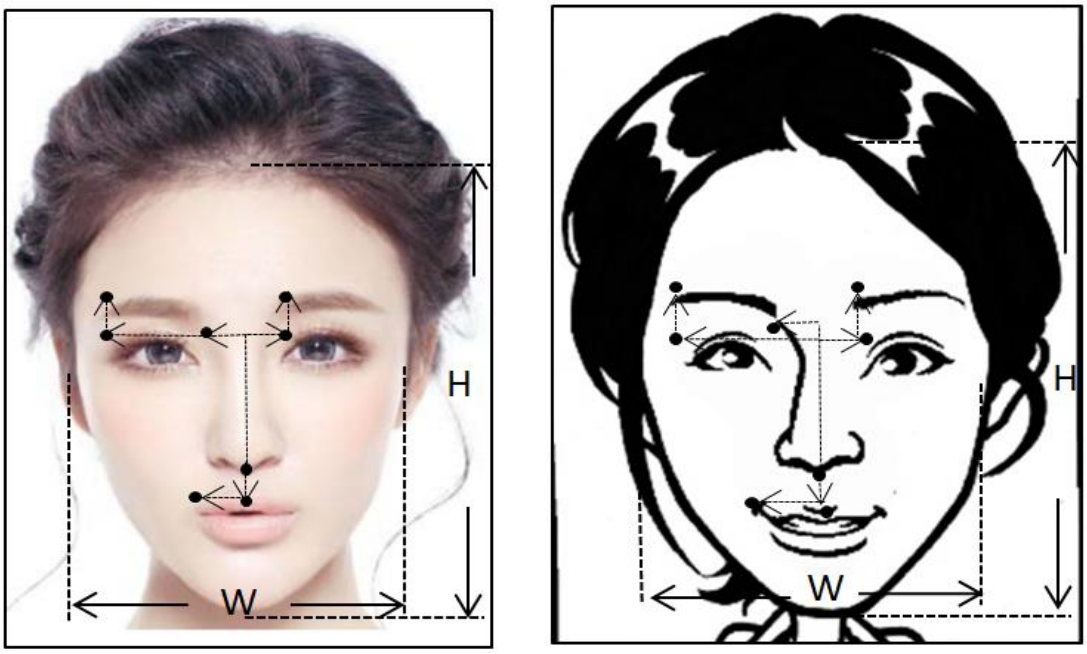


Fig. 31 Procedure for locating positions of facial components

After obtaining the final sizes and positions of the synthesized caricature components, the proposed system deploys them to the output caricature template. It is possible to improve the final synthesized results by combing the method from [20], with which the combinations of facial components and adjustments of positions trained by machine learning are considered. The method used in [20] can also determine eyelid types, and whether subjects are wearing spectacles. In addition, by using a Gabor filter to detect the texture value of the cheeks in the portrait photos, and setting a threshold value, it is possible to determine whether the input portrait shows an elderly subject. The distance between the upper and lower lips indicates whether a subject's mouth is open. Combining the above measures with the proposed system may improve the synthesized results in future studies.

3.4 Results and evaluation

3.4.1 Results

To validate the effectiveness of the proposed cross-modal distance metric, namely feature deviation matching, and caricature synthesis algorithm, we conducted experiments with four example sets of three different caricature styles: a male example set of expressive style, three female example sets of expressive styles, photo-realistic style and drawing style. For comparison with the method using paired photo–caricature examples, the male example set is constructed from paired photo–caricature examples, although the paired relationship is not used to search for similar caricature components in our system. Fig. 32, Fig. 33, Fig. 34 and Fig. 35 show some results based on the four different datasets respectively.

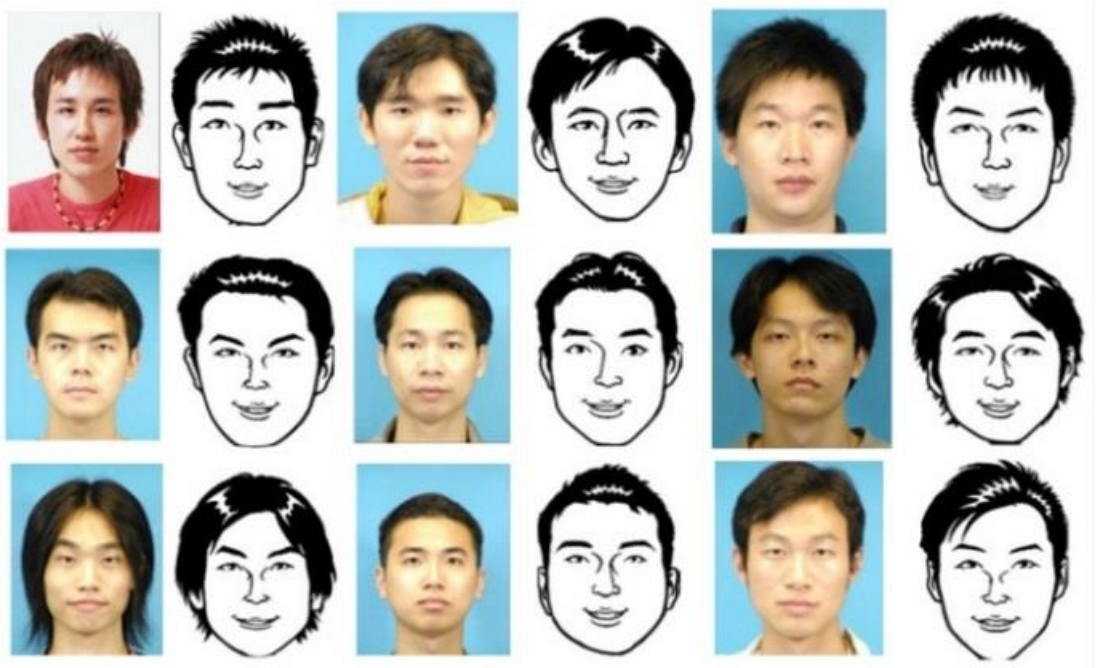


Fig. 32 Examples of the male caricatures of the expressive style

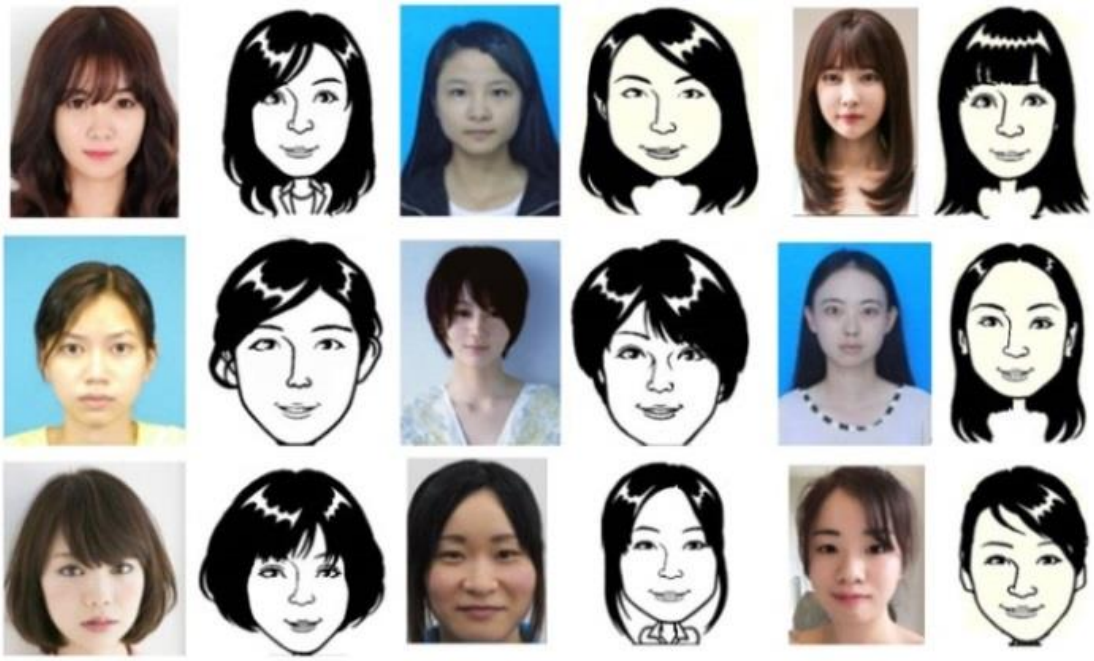


Fig. 33 Some examples of female caricatures of the expressive style

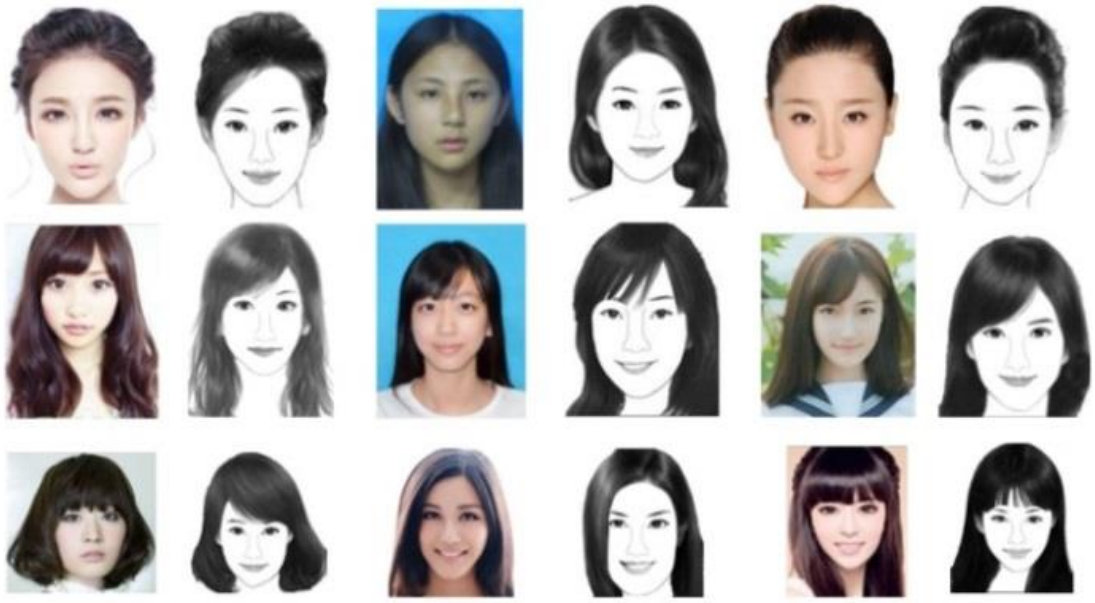


Fig. 34 Some examples of female caricatures of the photo-realistic style

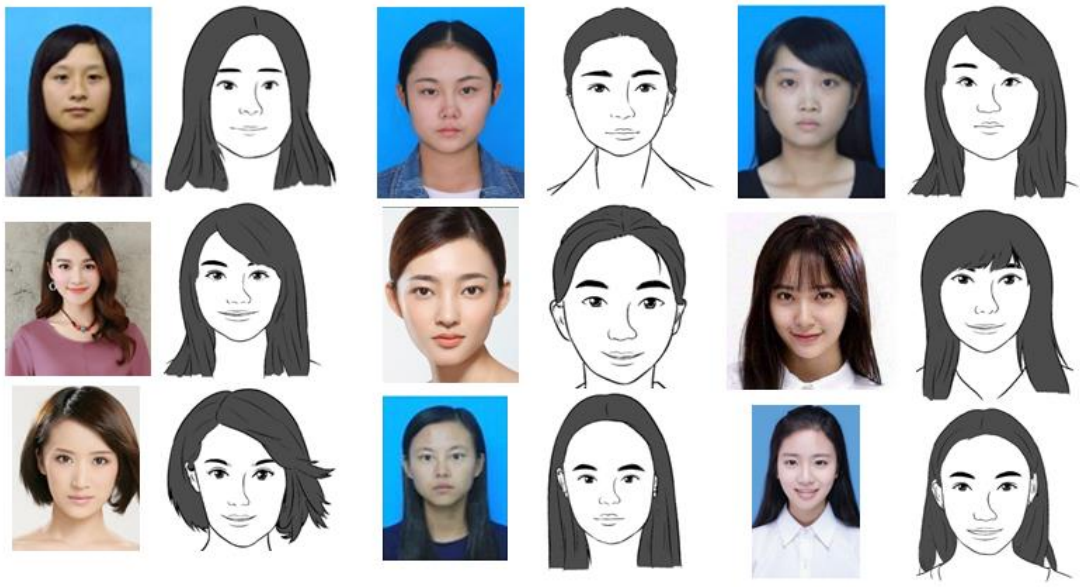


Fig. 35 Some examples of female caricatures of the drawing style

For practical applications, similarity is not the only factor considered. That is to say, the most similar caricature may not be the most desirable, since users' perceptions of the face can be highly subjective. In other words, in some cases, the second or third most similar caricatures may be preferred rather than the most similar caricatures provided by the system. Our proposed system can provide users with different candidate caricatures by controlling the exaggeration coefficients for corresponding facial components, as Fig.29 shows. Furthermore, various combinations with similarity-based facial and hairstyle components can satisfy users' differing preferences in another way. Fig.36, Fig. 37, Fig. 38 and Fig. 39 show three candidate caricatures combined with the searched facial and hairstyle components for each input portrait photo. Caricatures in column (A), (B), and (C) are synthesized with the most similar, second-most, and third-most similar components, respectively. Note that the overall impression of a caricature is determined not only by the similarities of the components themselves but also by their combinations. Training them

through machine learning [20] or other methods may generate caricatures with more attractive combinations of components. By providing several similarity-based candidates, it is also possible to combine this system with the relevance feedback system based on the OPF (Optimum-Path Forest) algorithm to improve the synthesized results interactively and iteratively for online training [73].

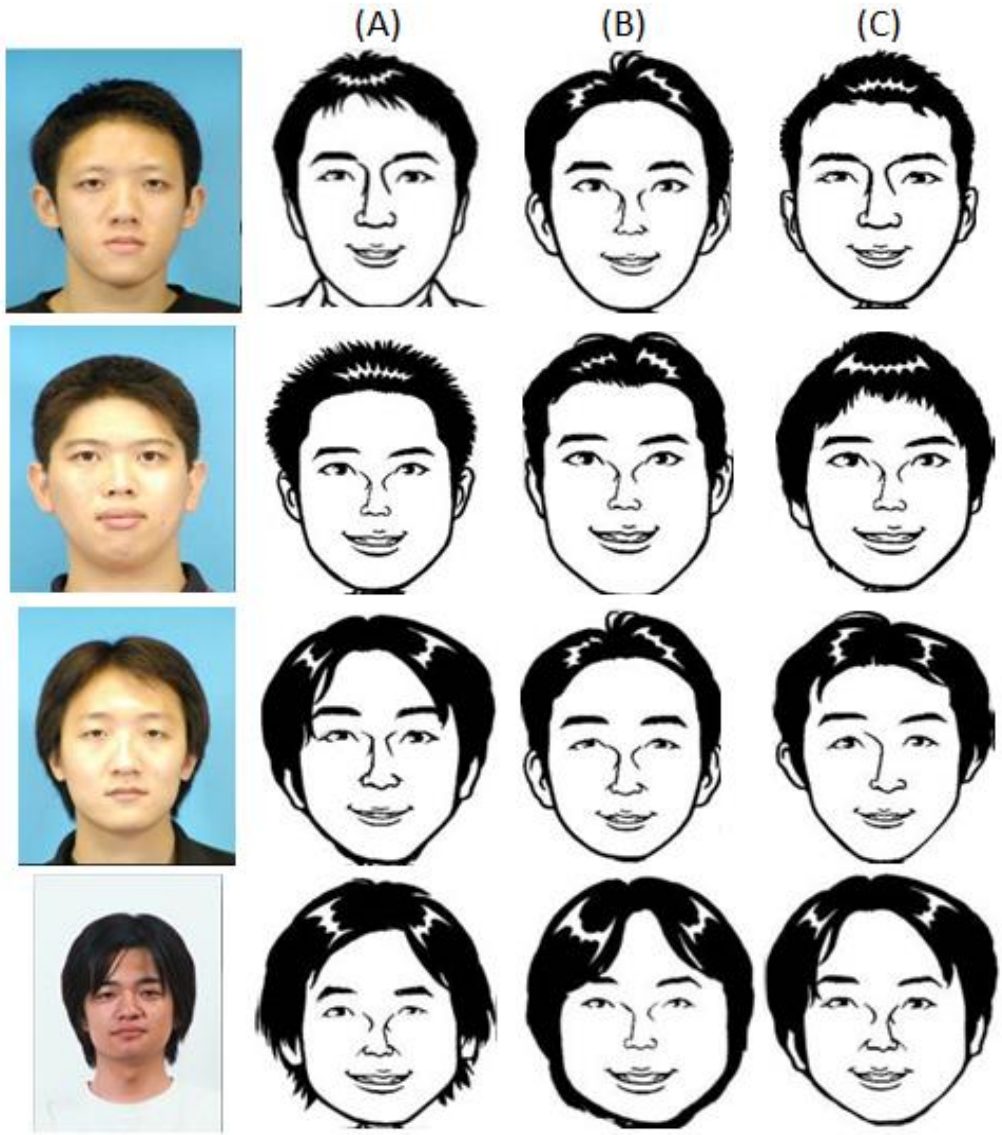


Fig. 36 Some examples of male caricatures of the expressive style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

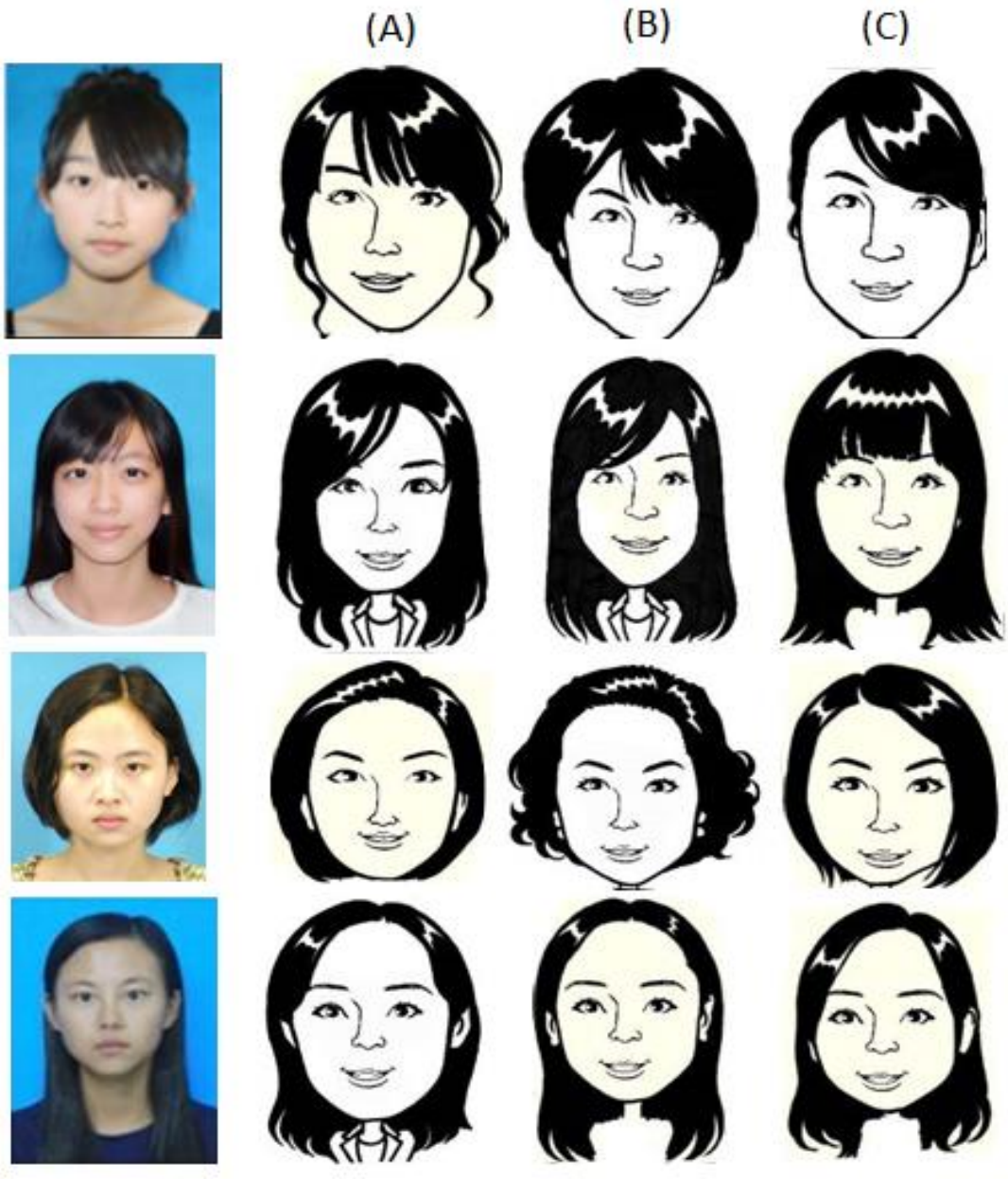


Fig. 37 Some examples of female caricatures of the expressive style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

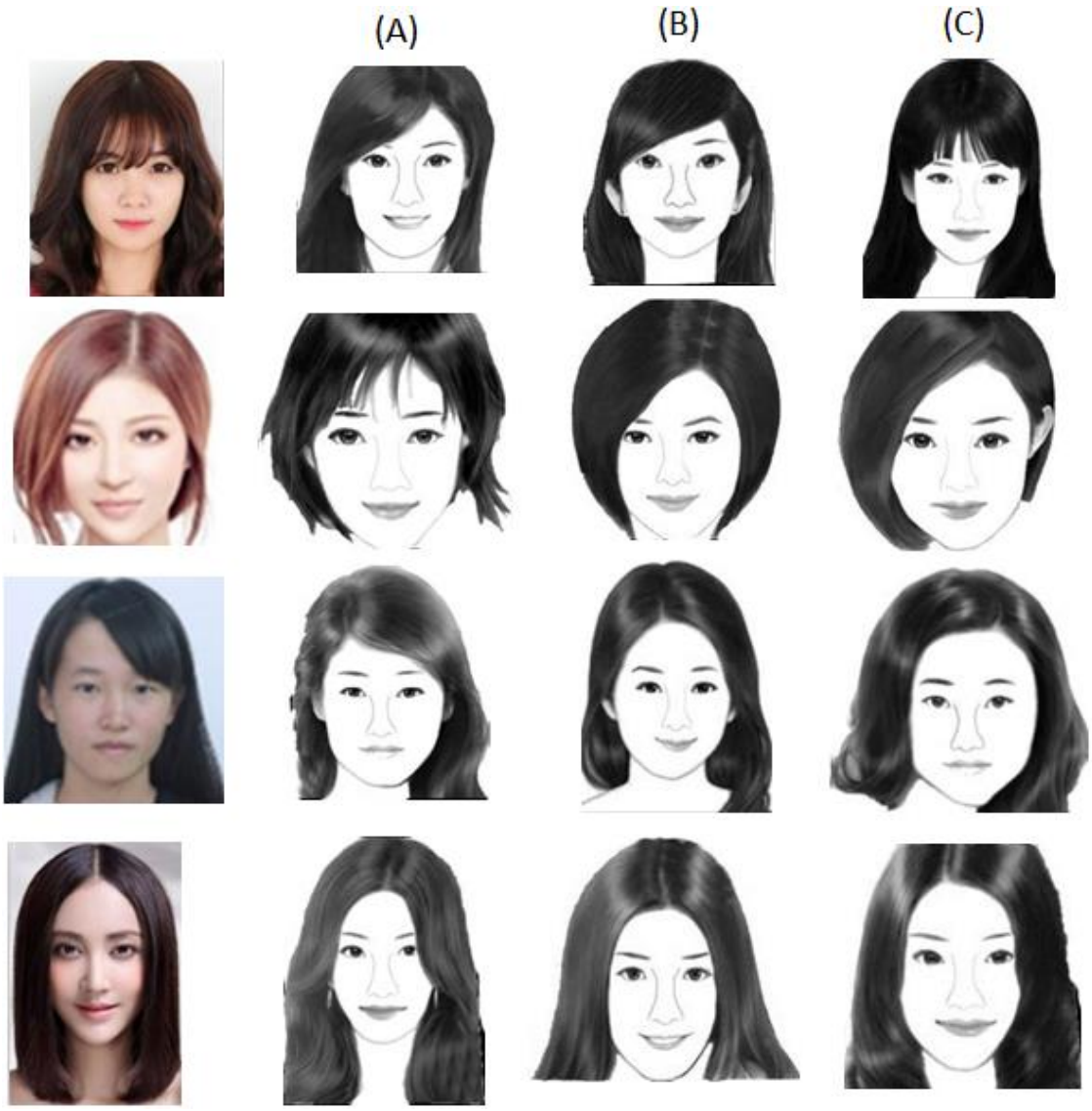


Fig. 38 Some examples of female caricatures of photo-realistic style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

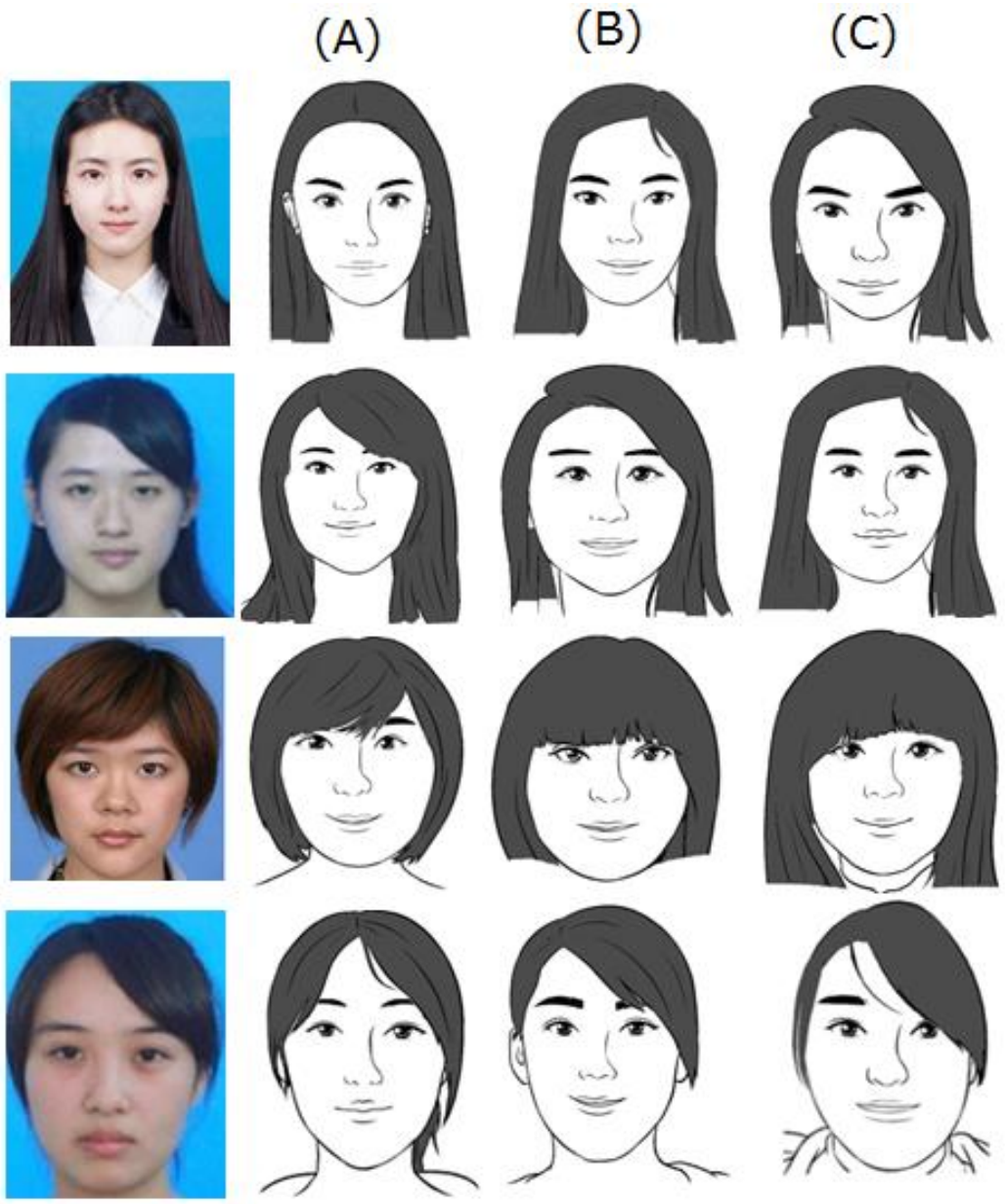


Fig. 39 Some examples of female caricatures of the drawing style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

3.4.2 Evaluation

3.4.2.1 Experiment I: Similarity

Experiment I is used to evaluate whether the proposed system can synthesize similar caricatures according to the input portrait photos without a paired photo–caricature database and be generalized to generate various styles of caricatures.

Since the feature deviation is a relative value and its real meaning depends on the distributions of the corresponding feature spaces, it is very important that the feature spaces of the photo and caricature databases should have similar distributions for executing the feature deviation matching. We achieve this

by collecting photos and caricature images with various types of hairstyles and facial components, avoiding distribution bias in their feature spaces.

80 female portrait photos, 60 female caricatures of expressive style, 120 female caricatures of photo-realistic style and 100 female caricatures of drawing style were collected respectively. Ten participants (6 female, 4 male, all aged in their twenties) were asked to evaluate the resulting caricatures generated for 40 randomly chosen female photos which are not included in the example photo database. The test photos and the resulting caricatures of the three styles were presented to each participant simultaneously side by side as Fig. 40 shows, which make participants can compare the caricatures with not only the input portrait photo but also among the caricatures themselves. Then participants were asked to evaluate whether the synthesized caricature resembles the input portrait photo, using a five-point scale. To make participants evaluate the results more reasonably, calibrating the average scores is necessary. They were told score 3 means acceptable in similarity, 1 and 5 represent least similar and most similar, that is to say, using the score 3 as a base line value for comparison to avoid personal difference in scoring metric. The average scores were 3.54, 3.61 and 3.83 for the expressive and two realistic styles.

The experiment results prove that the proposed system can be generalized to synthesize different styles of caricatures similar to the input portrait photos without paired-example databases.

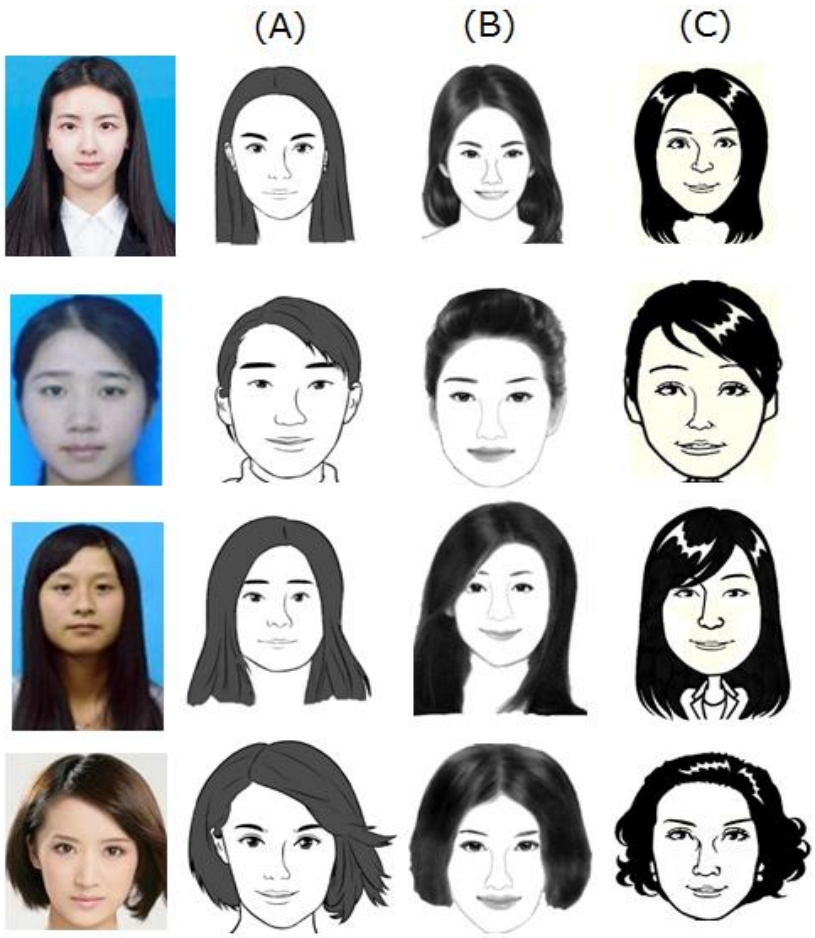


Fig. 40 Some examples displayed to participants in Experiment I (A, B and C represent caricatures of drawing, photo-realistic and expressive styles respectively)

3.4.2.2 Experiment II: Comparison with the paired-example method

Experiment II compares the results of the conventional paired-example matching method with those of the proposed feature deviation matching method.

The comparison should be based on the same paired photo-caricature database. We therefore performed this experiment on the expressive style of male caricatures, which used the paired photo-caricature database containing 83 male portrait photos and 83 paired caricatures. As mentioned at the beginning of this section, the paired relationship is not used to search for the similar caricature components in our system. Caricatures generated for 40randomly chosen male photos were compared between our proposed method and the paired-example method used in [19].

An additional 15 participants (10 female, 5 male, all aged in their twenties) took part in the comparison experiment. Each tested photo was displayed accompanied with two caricatures synthesized by the proposed technique and that used in [19]. As Fig. 41 shows, caricatures generated by the proposed method (first row) were compared with those from the paired-example matching method (second row). The two caricatures were placed randomly when conducting the experiments. Participants were asked which caricature was more similar to the tested photo. In addition, we also asked them to provide similarity scores simultaneously as in Experiment I. All the 15 participants evaluated 40 photos and hence there were 600 trials in total, out of which, in 341 trials (56.83%) the results generated with the proposed method were evaluated to be better than those generated by the paired-example matching method. The similarity scores are 3.82 and 3.66 for our proposed system and the paired-example system respectively. These two experiment results confirm that the proposed system can synthesize caricatures that are competitive with those synthesized with traditional paired-example system.

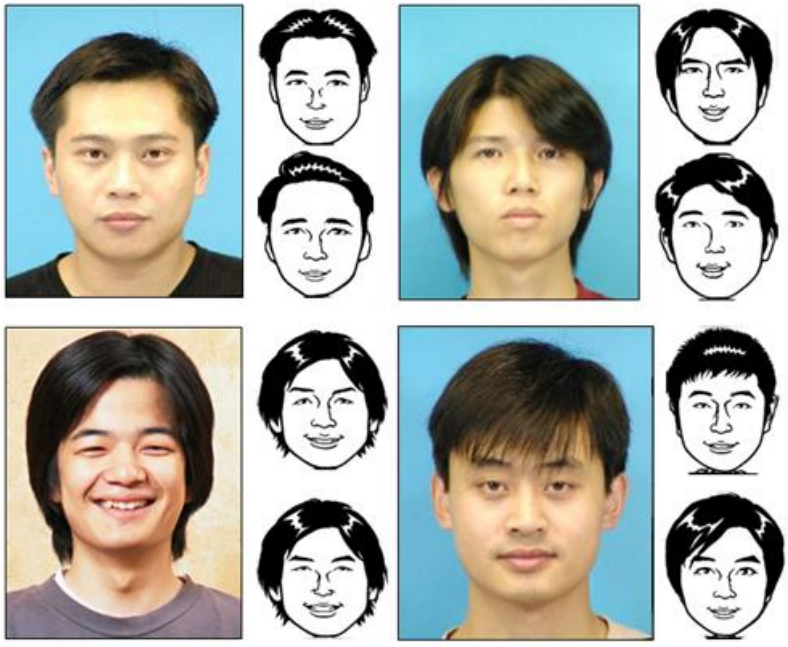


Fig. 41 Some examples in Experiment II on comparison of synthesized caricatures based on the proposed feature deviation matching and the paired-example matching methods

3.4.2.3 Experiment III: Capturing prominent features

Experiment III examines whether caricatures generated by the proposed system can capture distinctive facial features of the input portrait photos. This experiment evaluates the expressive style with male caricature and the photo-realisitic and drawing styles with female caricatures generated by the proposed systems. Four caricatures together with input portrait photo were displayed for 5 seconds (considering the participants should compare the similarities not only between the input portrait photo and the caricatures but also among the caricatures themselves) to another 15 participants (10 female and 5 male, different from participants in Experiment II). One of the four caricatures was synthesized by the proposed system, and the other three were selected from the corresponding caricature databases. Each participant evaluated 40 input portrait photos, giving a total of 600 trials. For each trial, a participant was asked to select the caricature most similar to the input photo from the four displayed. The recognition count increased by one when a participant selected the caricature synthesized by the proposed system.

As mentioned in section 3.3.2, hairstyle and face shape are two types of external features for human portraits. Hairstyle can be regarded as the most influential and discriminative component for human vision [74, 75]. Face shape also influences the impression of human face greatly [59]. The evaluation results may be largely affected by hairstyles and face shape if the remaining three caricatures are random chosen like [19, 20, 69]. For example, if the input female photo shows a short hairstyle, the displayed caricatures with long hair may be excluded immediately by most participants no matter how similar the other facial components are. In order to avoid such factors and evaluate whether the proposed system can capture the internal facialfeatures from the input photos, we therefore conducted adjustments on the experiments of [19, 20, 69]. Firstly, we randomly selected three candidate caricatures from caricature categories with hairstyles similar to the synthesized caricature. Secondly, we manually adjusted the three candidate caricatures' face shapes to be as same as possible with the synthesized caricature. As shown in Fig.42, the upper-left caricature was generated by the proposed system, and the other three were randomly selected from corresponding caricature categories with similar hairstyles and then with their face shapes adjusted. The four caricatures were randomly placed when conducting the experiments. As the result, the overall recognition rates for the expressive style of male caricature, the expressive style of female caricature, the photo-realistic style of female caricature and drawing style of female caricatures were 44.2%, 42.5%, 42.8% and 52.3% respectively. Binominal tests reveal that all the recognition rates are significantly higher than the assumed recognition rate (25%, select one randomly from four) at a significance level of 99% ($p=0.01$), which reasonably reflect the internal feature detection ability of the proposed system.

From the above experiments, it can be concluded that the proposed system based on the feature deviation matching method can synthesize facial caricatures resembling input photos without utilizing a paired photo–caricature database. Using the designed geometric feature vectors, the proposed system can reliably capture prominent facial features. Even if a paired photo-caricature database is available, the proposed system can be used as an alternative, since its synthesized results are comparable with those of the paired-example system. It is very important to ensure that the photo and caricature databases have similar distributions. In the current implementation, we achieve this by using various different components as possible. A strict, automatic method would be more desirable for ensuring better results.

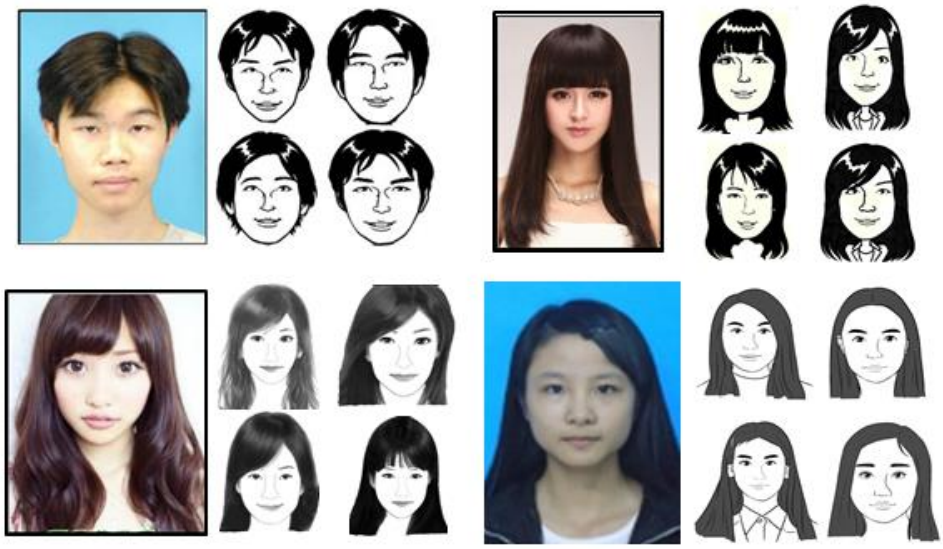


Fig. 42 Some examples in Experiment III on prominent internal feature capturing

3.5 Conclusion

We propose a cross-modal distance metric called feature deviation matching for example-based caricature synthesis. With this new matching method, the proposed system can synthesize caricatures capturing the prominent features of faces without a paired photo–caricature database. Different styles of caricatures can also be generated by employing caricature databases containing corresponding stylistic examples to the proposed system. Moreover, the designed geometric features can effectively capture the prominent features of input portrait photos. Compared with style-transfer-based approaches, our proposed system can better satisfy different users by controlling the details of stylization, such as the degree of exaggeration of individual facial components, and by providing users with a choice between several similarity-based synthesized caricatures. In addition, in future work, the proposed system could be combined with the OPF algorithm to improve the synthesized results interactively and iteratively.

Although the proposed system is robust, flexible, and easily generalized, the results could be further improved by using more a sophisticated algorithm, such as that in [20], in arranging the facial components. As described in the previous section, we maintain similar distributions throughout the photo and caricature component databases by manually confirming that both databases contain the most diverse examples possible. We can expect better results by applying a stricter, automated method.

Future work will consider more detailed features, such as single or double eyelids, closed or open mouths, with or without spectacles, and subjects' age. In addition, the machine learning methods for facial component combinations and position adjustment used in [20] should contribute to the final synthesized results. Combining the proposed system with the relevance feedback approach can further reflect a user's preferences. Since the proposed feature deviation matching method can ignore divergences between the photo and caricature component feature spaces, it is possible to make use of this advantage for the inverse application, i.e., to synthesize photos according to input caricatures, for example by assisting law enforcement agencies to match photographs of known criminals based on hand-drawn caricatures produced by specialist artists.

Chapter 4 Upper Clothes Recommendation according to Portrait Faces based on Deep Learning and Cross-modal Retrieval

Chapter 2 and Chapter 3 focus on collar, face and hair feature extraction and their applications on clothes retrieval and caricature generation. In this Chapter, we propose two prelimary frameworks for upper clothes collocation with female faces based on deep learning and cross-modal retrieval techniques. The proposed frameworks are still in progress.

4.1 Introduction

Retrieving similar or desired clothes images according to users' input clothes images under different circumstances is popular for online shopping as Fig.43 shows, where semantic attributes, hand-crafted and DNN visual features extracted from images are used for predicition and classification tasks [76-78]. Clothes recognition techniques are also necessary to detect clothes area for dealing with images with various backgrounds [79-81]. Collocation recommendation or evaluation systems are advanced applications developed from retrieval modules. Various collocation recommendation systems focus on fashion collocation, which contain single-component collocation [82] and multiple-component collocation [83]. Recently, as Fig.44 shows, fashion collocation recommendation systems based on deep learning frameworks develop quickly [84, 85]. Some researchers used deep learning frameworks to extract features from clothes images to generate item visual factors and combined them with item latent factors for predicting the whole impressions [86]. Situation-clothes collocation [87] is another important application, which recommends appropriate styles of clothes for different situations and collocate them automatically, as Fig.45 shows.

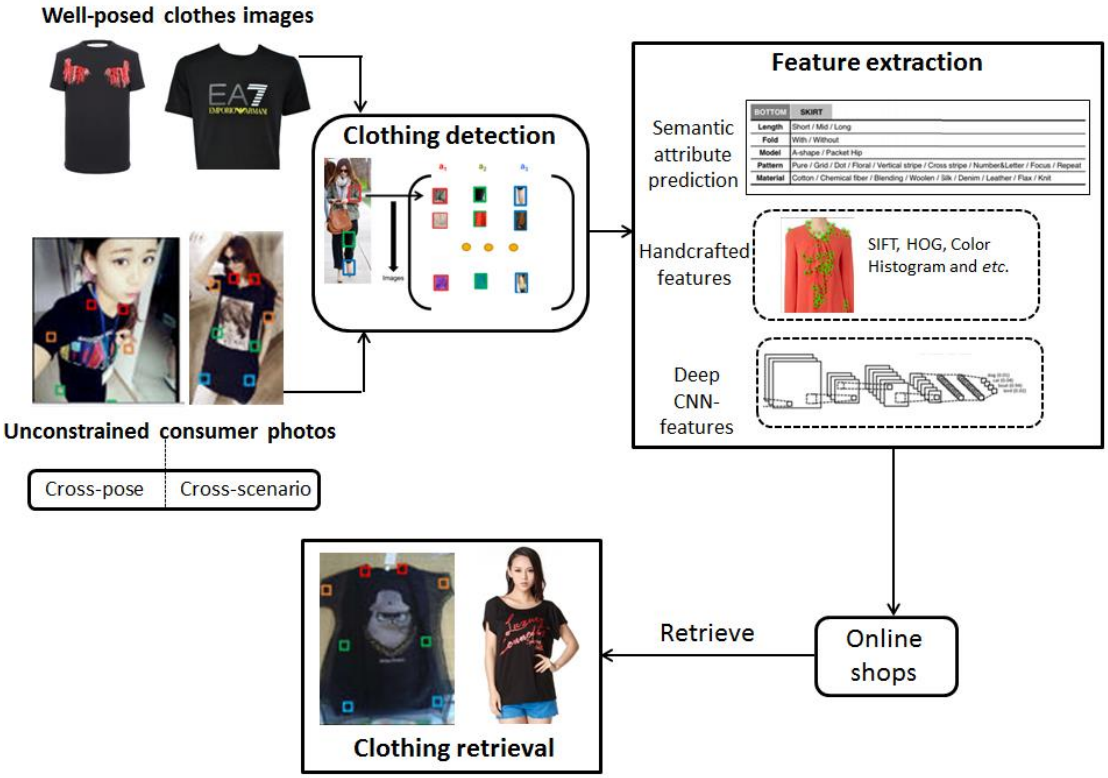


Fig. 43 Retrieve similar or desired clothes after clothing detection for online shopping with well-posed clothes images and unconstrained consumer photos based on semantic attributes, hand-crafted and deep features [76-81]

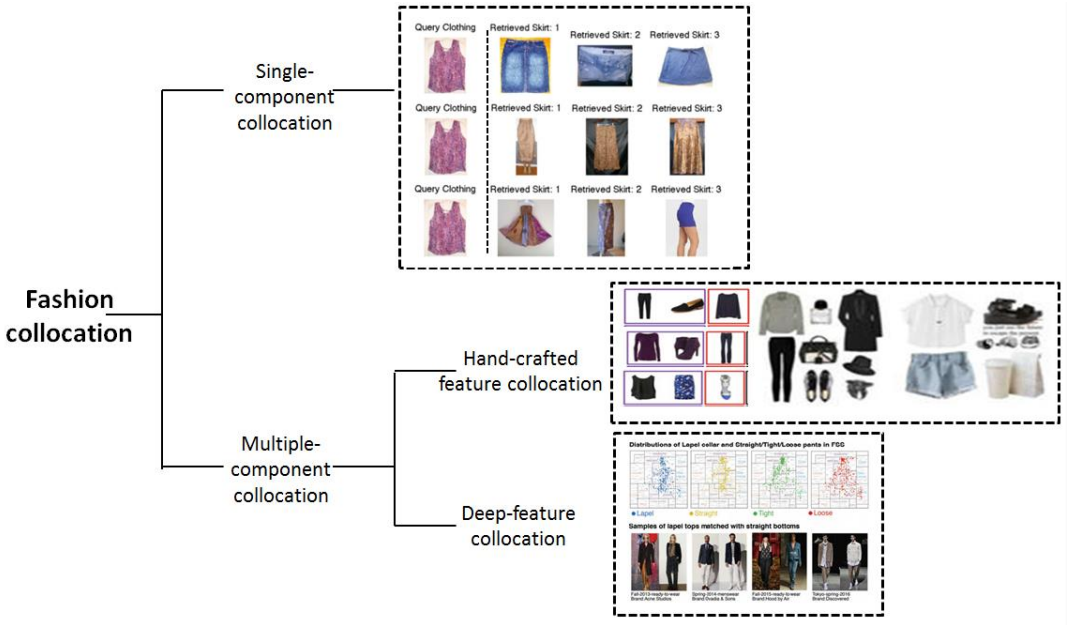


Fig. 44 Clothes collocation according to different components for different fashions [82-85]

Fig. 45 Clothes collocation according to different circumstances [87]

But the above recommendation systems are designed for clothes component or clothes-situation collocations without considering personality. That is to say, they do not consider the factors of personal faces, hairstyles, body sizes and so on. As [88] proposed, hairstyles should be collocated with face contours. A same female with different hairstyles and different clothes will also give us different impressions, as Fig.46 shows. It is very important to recommend or evaluate clothes collocation according to personal faces, hairstyles and body sizes.

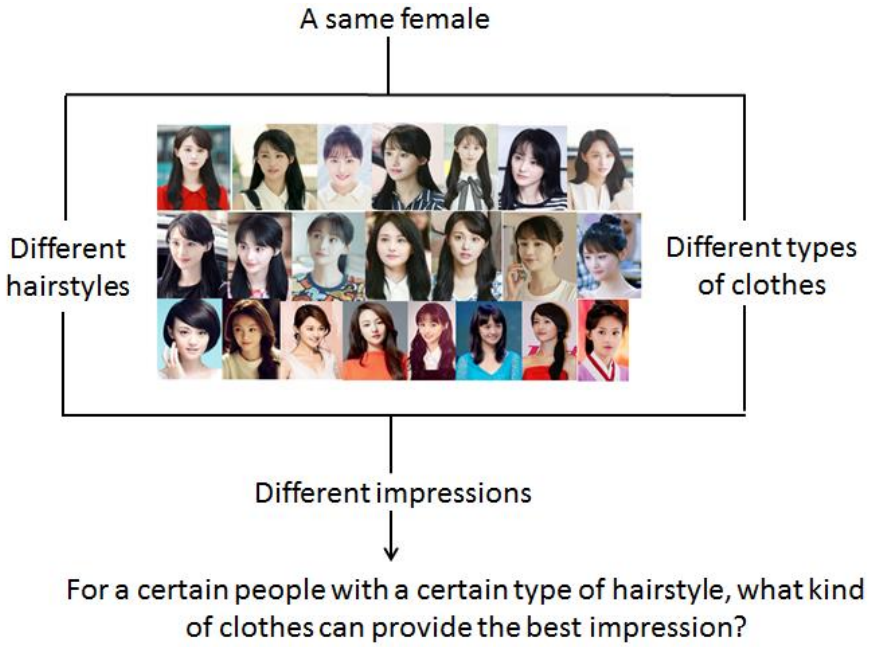


Fig. 46 A same female with different upper clothes and hairstyles provide different impressions

Faces, hairstyles and body sizes are very important for clothes collocation. As is well-known, for clothes shopping, people always put the selected clothes under their faces without wearing them in front of mirrors to see whether they fit their hairstyles and faces or not firstly. After that, they try on the selected clothes one by one to find whether the selected clothes fit their body sizes or not, which are time-consuming things. Clothes fitting systems combined with 3D Kinect devices appear gradually in some real shops to help users try on clothes virtually and consider the above three factors simultaneously as Fig.47 shows. In some online clothes collocation websites, users choose hairstyles, clothes and then manually input their body sizes to try on clothes in real time as Fig.48 shows. Some online websites also contain the functions for detecting users' body sizes in real time, but users need computers connected with client Kinect devices.

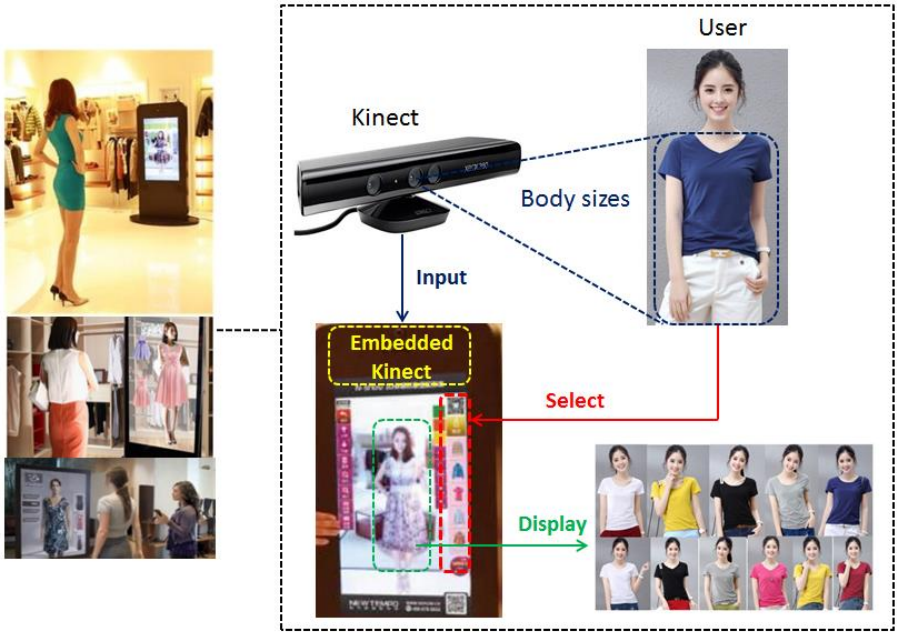
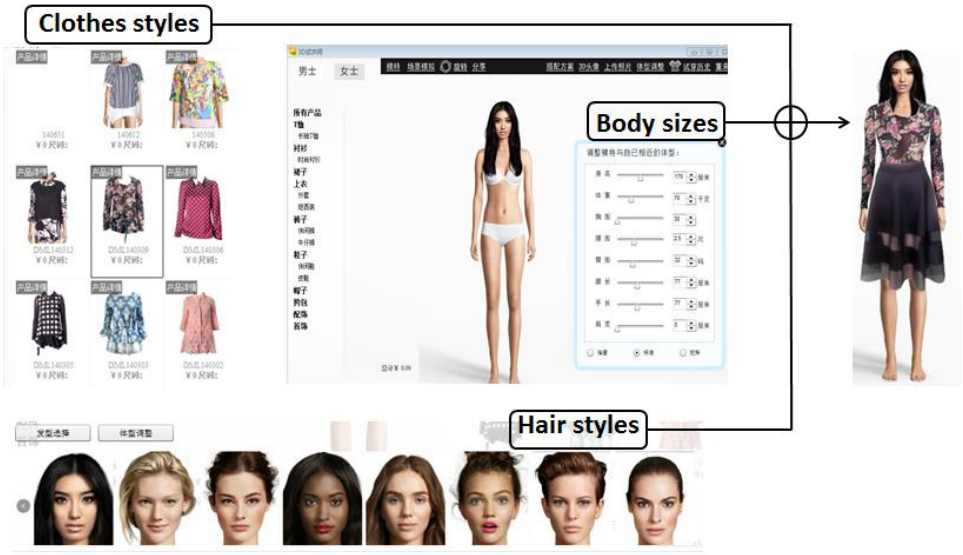


Fig. 47 Clothes fitting systems combined with 3D Kinect devices in real shops



http://3d.oleoad.com/3dshiyi.asp, 2018.7

Fig. 48 Manual clothes collocation in online websites based on clothes styles, body sizes and hair styles (http://3d.oleoad.com/3dshiyi.asp, 2018.7)

As Fig. 49 shows, for clothes collocation in household cases, people first consider what kinds of situations they will attend and then choose the corresponding types of clothes and decorations in their wardrobes which have fitted their body sizes when bought. Then they will change their hairstyles to fit the selected clothes or select other clothes to fit their hairstyles. With body sizes neglected and face unchangeable, clothes collocation are mainly determined by situations and flexible hairstyles.

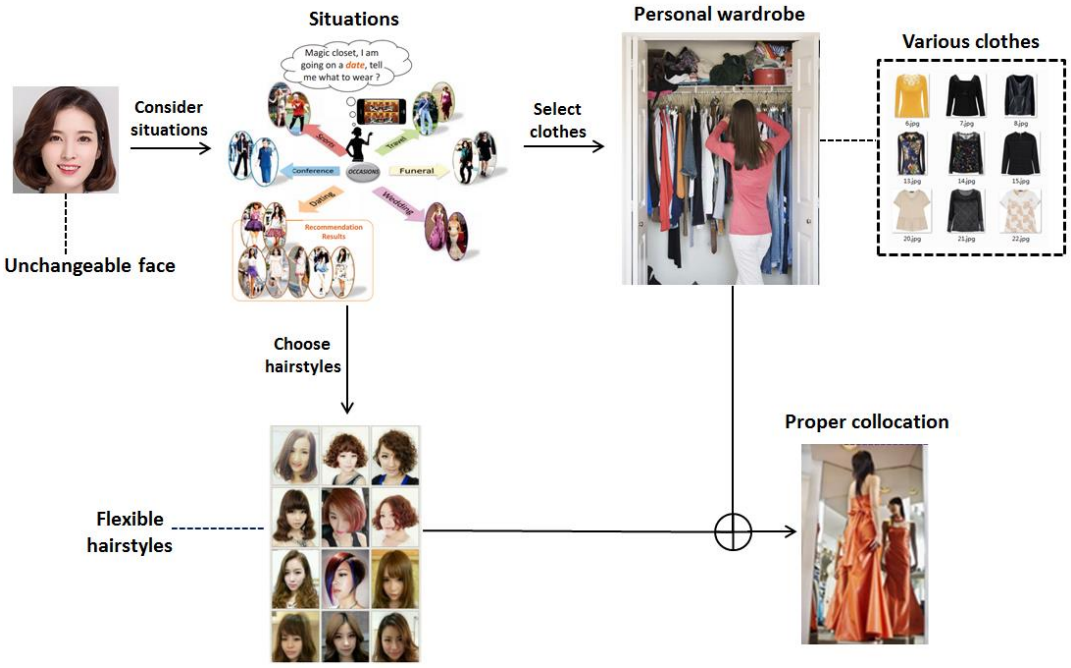


Fig. 49 Users collocate clothes with flexible hairstyles and unchangeable faces according to different situations

Fig. 50 shows phases of the traditional realistic shopping, online shopping and the proposed preliminary framework. Both realistic shopping and online shopping are made up of two key phases, which are face-hairstyle fitting and body size fitting.For realistic shopping, customers always firstly put clothes under their faces to see whether they suit them or not. If feeling good, they will try on the selected clothes in the dressing rooms for body size fitting subsequently. But for online shopping in most cases, customers make decisions on whether to buy clothes or not by only observing the dressing effects from models. And then choose the clothes from limited sizes and color types, such as small, middle and large. There are two problems exist in traditional online shopping. Firstly, the clothes fitting models may not suitable for users themselves. Secondly, it is very difficult for users or models to try on all the clothes one by one if huge number of clothes exists.

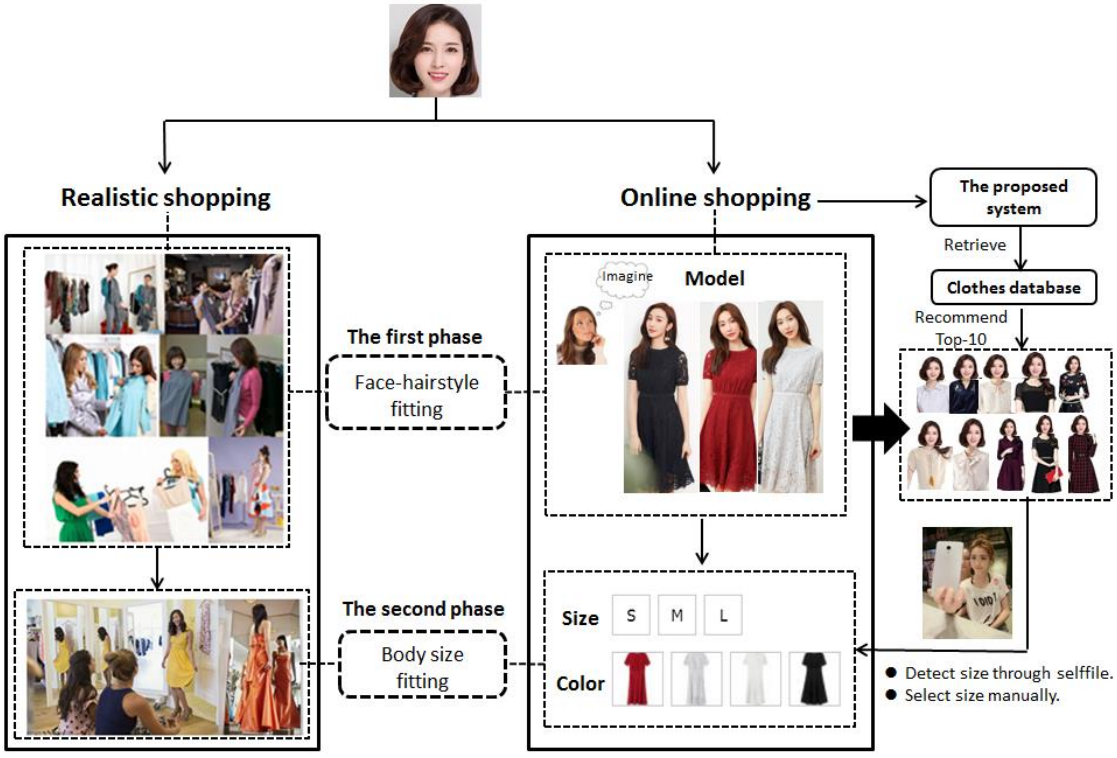


Fig. 50 Phases of the traditional realistic shopping, online shopping and the proposed preliminary framework

Based on the above considerations, we propose two preliminary frameworks on recommending upper clothes according to input portraits for solving the face-hairstyle fitting problem in online shopping, which are based on cross-modal retrieval and deep learning techniques. The body size fitting problem can be dealt with by incorporating some existing systems which can detect personal body sizes through uploaded selfiles or requiring customers to select candidanate sizes manually.

4.2 Cross-modal mutual retrieval, CCA and UDA methods

4.2.1 Cross-modal mutual retrieval methods

Cross-modal mutual retrieval techniques are more and more popular in mutual retrival and recommendation applications. J.Mao et.al used RNN to perform mutual retrieval between images and texts [89], that is to say, to explain images with words or retrieve appropriate images given words. A.Galen et.al conducted experiments on speech analysis for image-audio cross-modal prediction[28], where action features of tongue, lip and jaw extracted from speech fraps are used for mutual retrieval with speech audio features.X.Wu et.al performed image-text cross-modal music retrieval, where music covers are used to retrieve lyrics, and vice versa[90]. H.Hotelling et.al deployed a framework based on DCCA methods to conduct cross-modal music retrieval between audio and lyrics [91]. The refered applications are shown in Fig.51 (a), (b), (c) and (d), respectively.

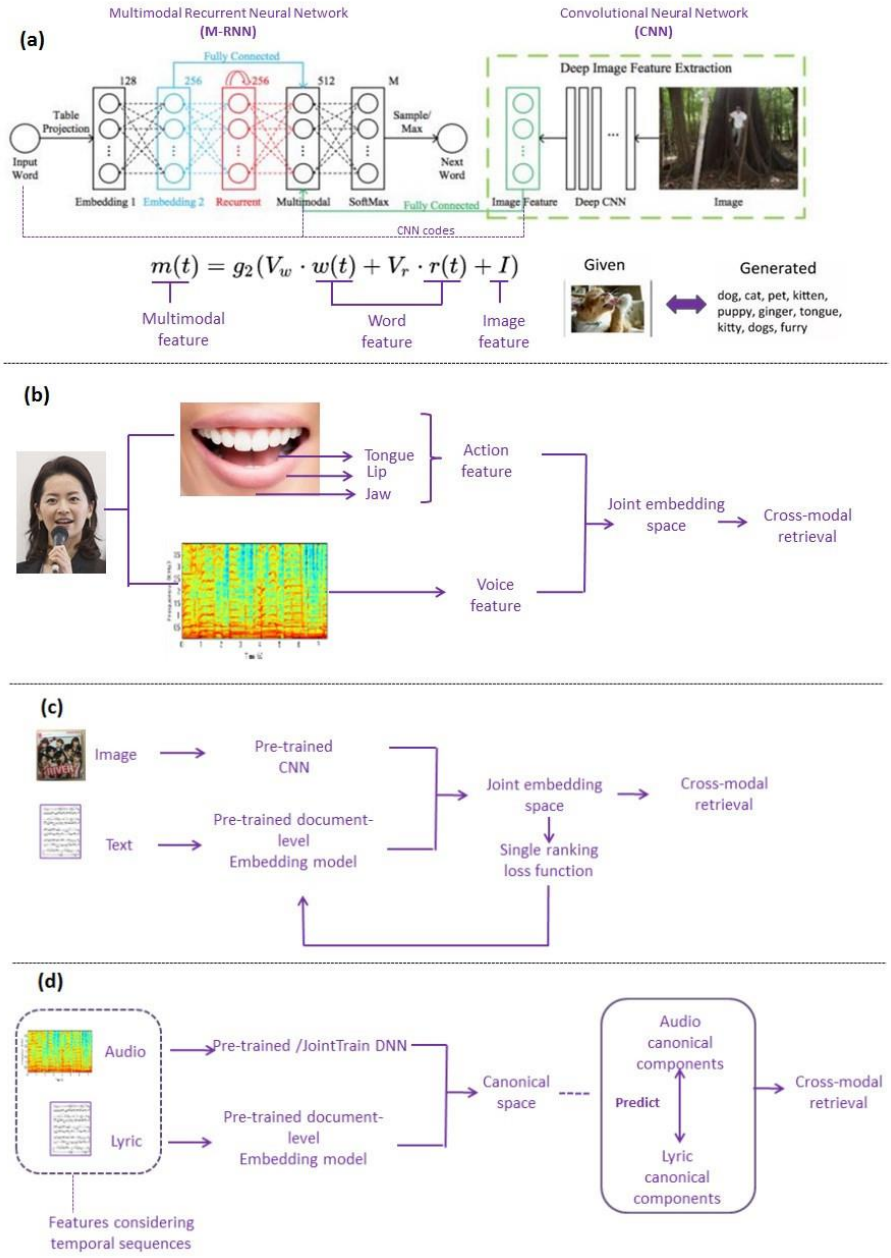


Fig. 51 Examples of cross-modal retrival. (a): Image-text mutual retrival[89];(b):Image-audio mutual retrieval[28]; (c): Image-text music mutual retrival[90] (d):Audio-lyric mutual retrival[91]

4.2.2 The CCA , KCCA and DCCA methods

Canonical correlation analysis (CCA) embeds multimodal data by linear mapping in a shared space for correlation analysis [92]. It tries to find two canonical weights $\mathbf{W_X}$ and $\mathbf{W_Y}$, so that the correlation between the linear projections u= $\mathbf{W_X^T X}$ and v= $\mathbf{W_Y^T Y}$ is maximized.Compared with CCA, kernel canonica correlation analysis (KCCA) uses kernel functions (represented by K in formula (18)), such as radial basis function (RBF) and polynomial, to embed multimodal data into a shared space by nonlinear mapping, which can extract more semantic features [93]. DCCA (deep canonical correlation analysis) tries to calculate non-linear correlations between different modalities by a combination of DNNs and CCA. Compared with KCCA, it has the extra capability of compressing features to a low dimensional space. The formulas of CCA, KCCA and DCCA [28] are shown as (17), (18) and (19), respectively.Note that, for

DCCA, formula (19) contains another three parameters, where H represent filters for convolution layers and ψ represent filters for fully connected layers.

$$(W_x, W_y) = \arg\max_{(W_x, W_y)} corr(W_x^T x, W_y^T y) = \arg\max_{(W_x, W_y)} \frac{W_x^T C_{xy} W_y}{\sqrt{W_x^T C_{xx} W_x W_y^T C_{yy} W_y}} \tag{17}$$

$$(\alpha_1^*, \alpha_2^*) = \arg\max_{(\alpha_1, \alpha_2)} \frac{\alpha_1' K_1 K_2 \alpha_2}{\sqrt{(\alpha_1' K_1^2 \alpha_2)(\alpha_1' K_2^2 \alpha_2)}} = \arg\max_{(\alpha_1' K_1^2 \alpha_1 = \alpha_2' K_2^2 \alpha_2 = 1)} \alpha_1' K_1 K_2 \alpha_2 \tag{18}$$

$$(H_x, H_y, W_x, W_y, \psi_x, \psi_y) = \arg\max_{(H, W_x, W_y, \psi_x, \psi_y)} corr(W_x^T x, W_y^T y) \tag{19}$$

4.2.3 The UDA method

Unsupervised Domain-adaptation (UDA) [96] training is a state-of-the-art technique that combine adversarial learning with cross-modal retrieval. The core idea of it is inspired by Generative Adversative Networks (GAN). GAN is a famous deep learning framework conducting an adversarial game between the generator and discriminator modules, where a joint loss generated by the two modules are maximized and minimized in the discriminator and generator, respectively, for optimizing there parameters iteratively.

In UDA, the feature projector works as the generator to generate common feature representation from extracted cross-modal features. A domain classifier is added as the discriminator for detecting the original modality given an unknown projected feature. But it is too easy for training the generator in this framework, since it is only necessary for it to cheat the discriminator within two types of modalities instead of faking a whole image. To improve the performance of the feature projector, a label predictor is added after it, which is used to detect the semantic labels of the generated common features within their original modalities. That is to say, the generated features should not only try to cheat the Domain classifier but also preserve their original distributions. By embedding the label predictor into the feature projector and optimizing the whole framework with adversarial learning between the improved feature projector and the domain classifier. The proposed framework achieves better improvements.

$$L_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(m_i \times (\lg D(v_i; \theta_D) + \lg D(t_i; \theta_D)) \tag{20}$$

$$L_{emb}(\theta_{imd}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \times (\lg p_i(v_i) + \lg p_i(t_i))) \tag{21}$$

Formula (20) and (21) show the adversarial (adv) loss and embedding (emb) loss generated by the modality classifier and the feature projector, respectively. v_i and t_i ($i = 1, 2, 3 \cdots\cdots n, n$ represents the number of samples in each batch) stand for samples from different modalities, respectively. In Formula

(20), θ_D represents the parameters in the modality classifier. m_i is the ground-truth modality label of each instance,expressed as one-hot vector. $D(v_i;\theta_D)$ and $D(t_i;\theta_D)$ are the generated modality probabilities from the corresponding paired items. In Formula (21), θ_{imd} represents the parameters in the feature projector. y_i is the ground-truth semantic label of each item ,expressed as one-hot vector too. $p_i(v_i)$ and $p_i(t_i)$ are the generated probability distributions from the corresponding paired items. Then, the joint loss combined by the adv loss and the emb loss are maximized in the modality classifier and minimized in the feature projector for optimizing the parameters, respectively.

4.3 The prepared dataset and environments

Based on the above two referred techiniques, we proposed the DCCA and the UDA preliminary frameworks for recommending upper clothes according to the input portrait face photos. Parts of the paired samples are collected from taobao and other online shopping websites, the others are manually selected from deepfashion datasets (http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html)，which are mostly female model photos well coordinated with clothes. Thus face and upper clothes parts (dress are thought of upper clothes parts as a whole) of these photos are manually divided from the neck position with little background as much as possible and regarded as paired samples for cross-modal retrieval.Certainly collecting huge number of ordinary female photos, asking participtants to evaluate them one by one and then selecting the positive samples are theoretically best ideal. But taking this ideal measure should consider portrait privacy. In addition, evaluating huge number of photos would exhaust participtants and resulting in unsuitable evaluation. This collecting method will also cause some problems such as data deviation. That is to say, most of the paired samples come from positive faces. Huge number of various model samples with frontal but different posed faces may relieve this problem. We will also set appropriate threshold values for input portraits in the running phase, which will determine whether the input portraits can be handled by the proposed system or not. Various distance metrics and data visualization methods will be deployed and tested to find appropriate thresholds in the future. Moreover, whether many-to-many coordination for faces and clothes in selected samples will affect the performance of the proposed system should be tested too.

At present, there are totally 30,000 paired samples collected and we are continually expanding the database. Before put into training, these paired samples are transformed into grey images firstly to be focused on the other coordination factors, since as described before in online shopping customers always try to find suitable color types manually after selecting desired clothes.

The training and running environments are listed as below. For hardware, the server is made up of a CPU of I7960, a GPU of GeForce GTX 970 and a 24G memory. For software, the operating system is Utuntu16.04, the graph acceleration softwares are CUDA9.0 plus cuDNN7.1, and they proposed system is run under Python2.7.6, TensorFlow1.8 and Keras2.2.

The first and second frameworks inspired by deep learning and cross-modal retrieval methods adapt the DCCA and UDA frameworks to deal with face and upper clothes coordination applications, respectively.In

the training phases of these two frameworks, CNNs are only used for feature extraction and the training samples do not need to be pre-evaluated by users.

4.4 The preliminary framework for upper clothes recommendation based on the DCCA framework

Fig. 52 shows the offline phase of the proposed DCCA framework. It is an end-to-end system containing a feature extraction module and a CCA module. The feature extraction module is a two-channel DNN networks containing a CNN network, a sub-DNN network and a linear project function, respectively. The CNN network made up of convolution and pooling layers is used to extract semantic features from input images and compress them to a low dimensional space. The sub-DNN network is composed of fully connected layers where local features converge and are conjointed into global features. Another function of the last fully connected layer is to make dimensions of the output upper clothes and portrait face features same. The linear project function is used to project the cross-modal features into a shared feature space, which transfers the output deep features into canonical components. The CCA module is used to make the generated cross-modal canonical components as similar as possible.

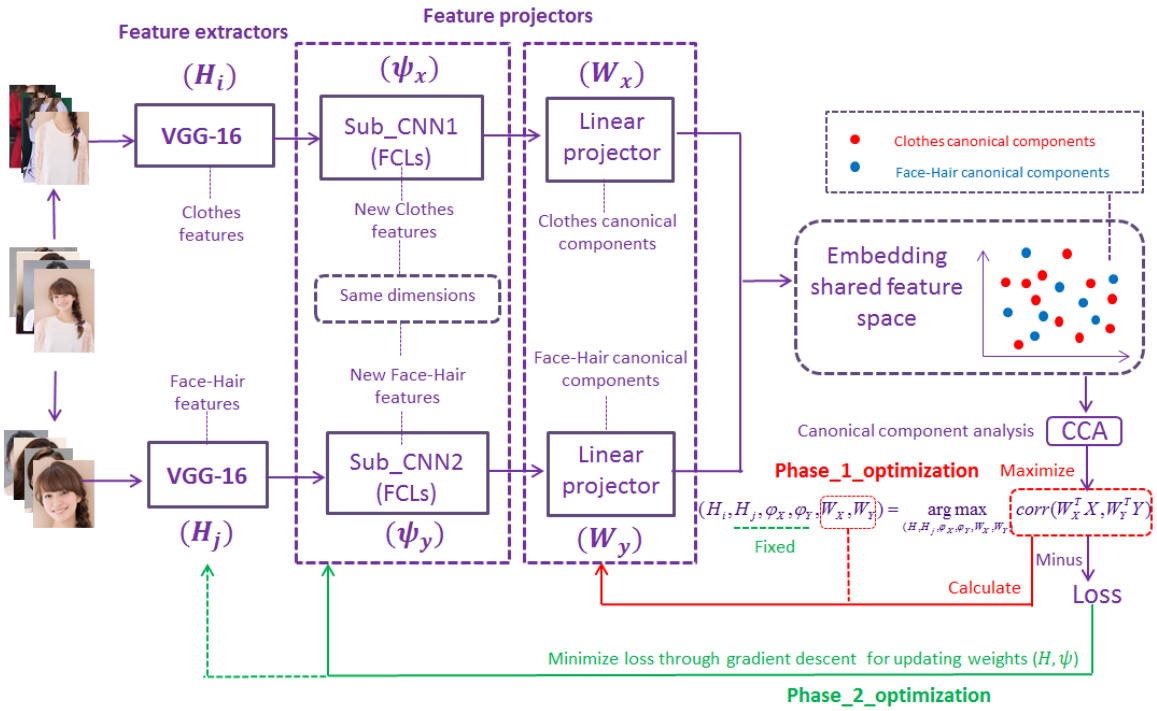


Fig. 52 The preliminary framework for upper clothes or collar recommendation in offline phase based on the DCCA framework

Detailed procedures of the offline phase are shown as below.

Firstly, we deployed two VGG-16 [98] networks to extract features from the cut face and upper clothes images in the prepared paired database, where each image is represented by a 4096-d feature. In the future, we will replace the two VGG16 networks with the DARN network (for clothes) [94] and the Deep ID network (for portrait) [95] to test whether the proposed system can achieve better performance.

Secondly, three fully connected networks are used as nonlinear projectors to transform the extracted 4096-d features into 1024-d features three times, which will try to make the representations more semantic. And a linear function is used to transform the 1024-d features into 10-d features for dimentionality reduction.

Finally, a CCA module is deployed to try to remove difference from the generated cross-modal common features, where an average correlation value between the 10-d clothe and portrait features is generated and maximized as much as possible. The batch size is set to 100, and the epoch turns are set to 75 at present. In practice, the negative value of the correlation is regarded as the customized loss of the proposed system and minimized by traditional gradient descent methods to update weights of the nonprojection network and even finetune the pre-trained VGG-16 networks in the feature extraction module iteratively.

In online phase, we input a portrait face photo and all the candidate upper clothes photos into the pre-trained VGG-16 networks. If necessary, the candidate upper clothes photos can be classified into several categories according to clothes types and users can choose the types they desired for online coordination. Through the trained DCCA framework, the cross-modal canonical components of them are generated and embedded into a shared feature space. Then, using the L2 similarity metric, the distances among the portrait face component and the upper clothes components are calculated. The upper clothes image with the smallest distance is recommended as the most suitable one. Currently, we provide the Top-3 results as candidates for users as Fig.53 shows. In addition, we can also provide the Top-N results to the OPF system introduced in Chapter 2 for online optimization. The detailed procedurea are shown as Fig.54.

Fig. 53 Some examples of the results from the preliminary DCCA framework

In the future, we will deploy the t-SNE method [99] to visualize the feature distributions of the cross-modal canonical components for the first and last epoches in the training phase to observe the effectiveness of training. This method can also be used to analyze whether distributions of the input portrait faces are covered by those of the training database in the online phase.

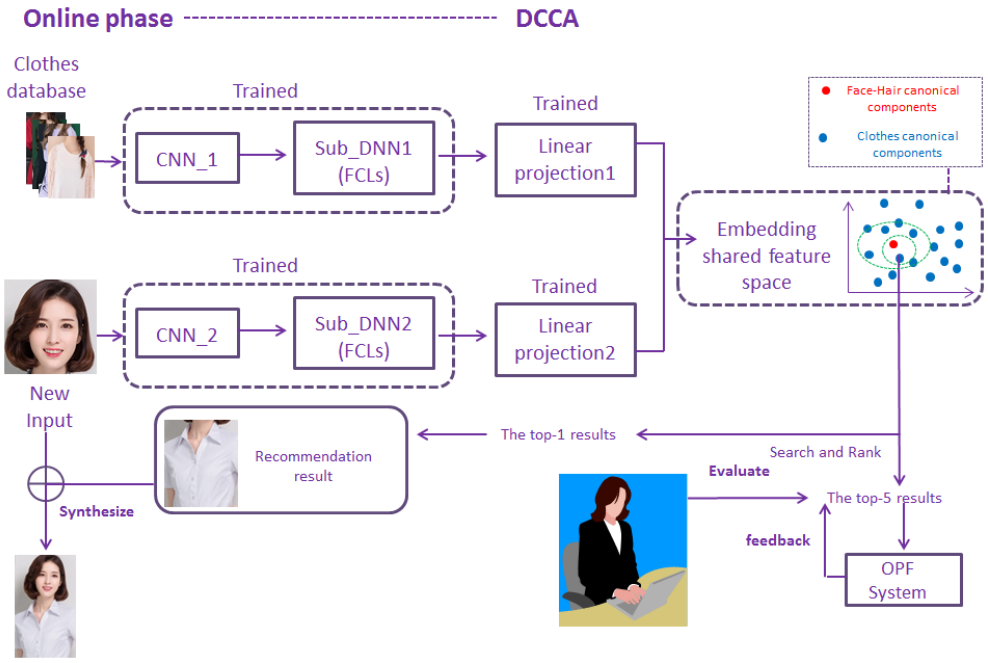


Fig. 54 The preliminary framework for upper clothes or collar recommendation in online phase based on the DCCA framework

We design two types of experiments for evaluation, which are the objective evaluation for retrieval accuracy and the subjective evaluation for coordination results. For the objective evaluation, 1200 paired samples are tested with the MRR1 (Mean reciprocal rank 1), which is an evaluation metric for multiple queries as formula (22) shows, standing for the frequency how often the original paired samples are hit for retrieval. The higher the value is, the more accurate the result is. Since the number of tested samples will greatly affect the MRR1 value, it is necessary to choose an appropriate number of tested samples in the future. For the subjective evaluation, by providing the Top-1 result, we plan to ask several participtants to evaluate 100 tesing input portraits respectively, with a 5-score metric, where 3 means acceptable in coordination, 1 and 5 represent least coordinated and most coordinated, that is to say, using the score 3 as a base line value for comparison to avoid personal difference in scoring metric. In addition, if given the Top-3 results, participtants are asked to mark them with a 3-level metric, where 3 represents best and 1 represents worse, which will be compared with the 3-level metric results provided by the proposed system. We plan to investigate how much the proposed system matches human vision for face-upper clothes coordination by the Top-1, Top-2 and Top-3 matching percents, respectively.

$$MRR1 = \frac{1}{N_q}\sum_{i=1}^{N_q}\frac{1}{rank_i(1)} \tag{22}$$

4.5 The preliminary framework for upper clothes recommendation based on the UDA framework

The first preliminary framework based on DCCA tries to reduce the gap across different modalities by minimizing the correlation loss values, which only considers the similarity between paired samples from different modalities. The UDA framework tries to improve peformances by adapting adversarial learning from GAN, and embedding a label predictor into the feature projector to maintain the original semantic

feature distributions before projected. To use this framework, we plan to classified images of the pre-used paired face and upper clothes databases in the DCCA framework into several correonding categories by the K-means algorithm in the future.

As Fig.55 shows, the second preliminary framework adapts the UDA framework to optimize upper clothes recommendation. At first, we also use the pre-trained clothes CNN and the Deep ID CNN to extract features from the divided clothes and face images, respectively.Secondly, we adapts two CNN, which are fully connected networks, to work as feature projectors for projecting these extracted features into a common feature space.A joint loss combined by the embedding loss came from the feature projectors and the adversarial loss came from the modality classifiers are optimized in the adversarial learning, which is minimized in feature projectors and maximized in the modality classifier.By conducting the above procedures, the clothes and hair-face features are embedded into a common feature space and mutually predicted.

Fig.56 shows the online phase of the second preliminary framework, which is similar to the first preliminary framework.

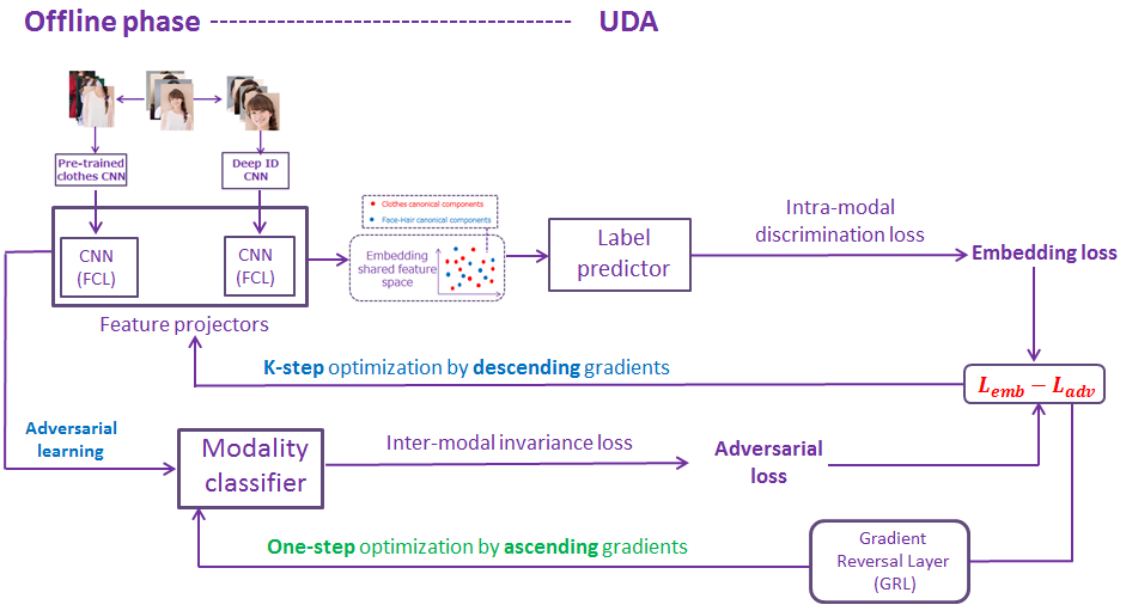


Fig. 55 The preliminary framework for upper clothes or collar recommendation in the training phase based on the UDA framework

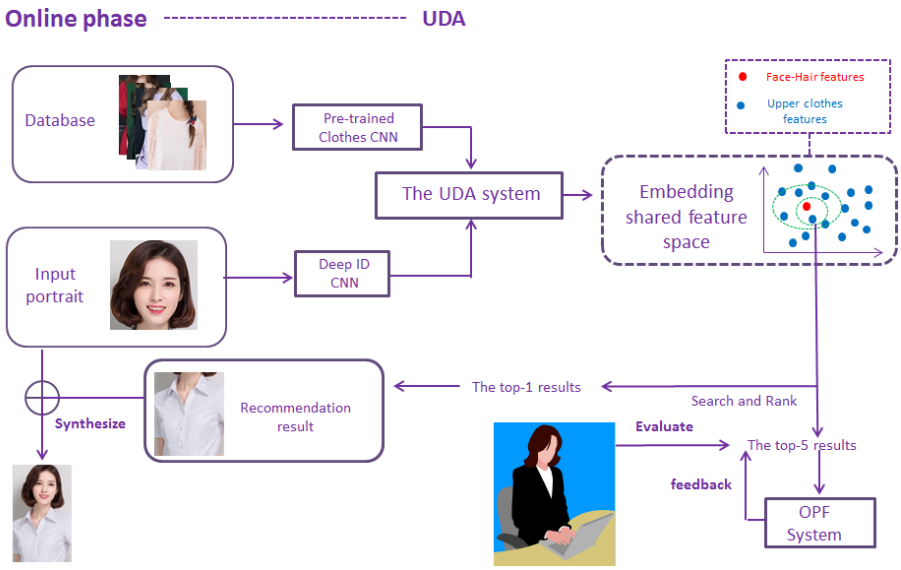


Fig. 56 The preliminary framework for upper clothes or collar recommendation in online phase based on the UDA framework and the OPF system

4.6 Future work

In the future, we will try to do the following work:

1. Evaluate the effects of imposing hand-crafted features designed in project 1 and project 2 and deep features extracted in project 3 on the CCA and UDA frameworks, respectively, as is shown in Fig.57.

2. For the preliminary frameworks, although we can use Deep ID [94] and DARN [95] as the two pre-trained DNN networks to extract portrait face and upper clothes features and fine-tune them with the collecting samples. It is necessary to increase the number of training samples to improve performance for the fine-tunned networks.

3. Compared with the UDA framework, the ACMR framework [97] combined triplet constraints for better performance, which needs positive and negative samples. We will try to add negative samples to the training dataset for applying the ACMR framework on the third project.

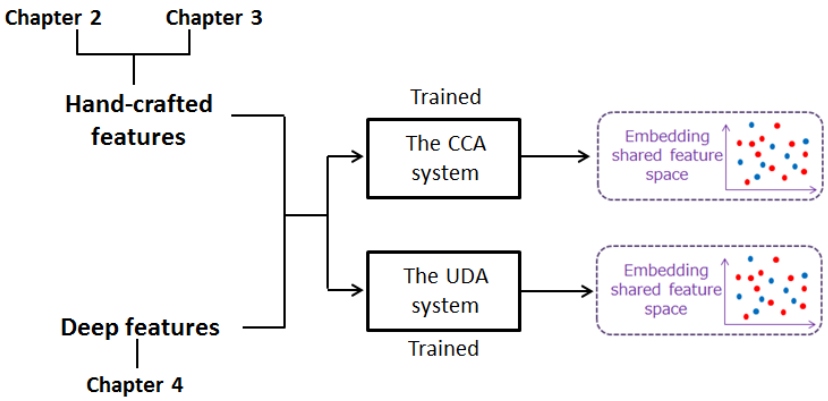


Fig. 57 Comparison for the upper clothes recommendation system based on the CCA and UDA frameworks using hand-crafted and deep features

Chapter 5 Conclusions and discussions

5.1 Conclusions

Portrait photos are widely used in social daily life for various applications. The main contributions of this thesis are as following:

1. Design collar features and use them to retrieve clothes with desired collars. Improve the query results online iteratively by combining the OPF algorithm.
2. Design facial component and hair features and use them to capture prominent features of portrait face photos. Retrieve caricature components similar to the input portrait components based on the proposed feature deviation distance metric directly without paired databases. Synthesize various types of caricatures and provide users several candidates for each input portrait to satisfy their different preferences by exaggeration control and similarity-based ranking methods.
3. Propose preliminary frameworks based on deep learning and cross-modal retrieval techniques for upper clothes or collar recommendation according to input female portrait faces, which focuses on investigating the collocation relations between upper clothes and female faces and hairstyles.The first and second preliminary frameworks are based on the DCCA and UDA frameworks, respectively.

5.2 Discussions

For the first project, there are still some technical problems to be solved.1. The designed features did not perform well for matching some types of collars. 2. Some types of collars are similar and will confuse users when evaluated. 3. The scores of the query results are computed as the number of satisfactory images without considering the degree of satisfactory of each image in the query results. It is necessary to design new features, reclassify types of collars and propose better evaluation metric.

For the second project, we would continue to improve performances of the proposed system by conducting the following measures. 1. We can expect better results by applying a stricter, automated method to maintain similar distributions throughout the photo and caricature component databases. 2. It is possible to improve the synthesized results by employing some techniques to optimize component combination and position arrangement. 3. Considering more details, such as single or double eyelids, closed or open mouths, with or without spectacles, and subjects' age, can effectively improve extractiveness and similarity of the synthesized results. 4. Based on the feature deviation distance metric, some inverse applications, for example, assisting law enforcement agencies to match photographs of known criminals based on hand-drawn caricatures produced by specialist artists, can be realized.

For the third processing project, additional work should be conducted. 1. Expand the training database to

improve performance. 2. Collect negative samples for each paired samples for applying the ACMR framework. 3. Employ the hand-crafted features designed in the first and second projects and the deep features extracted from the third project on the CCA and UDA modules and compare their performances.

List of Publications

Journal Articles

1. Li, H.L.,Toyoura,M.,Shimizu,K.,Yang,W.,Mao,X.Y.:Retrieval of clothing images based on relevance feedback with focus on collar designs. The Visual Computer. **32**(10), pp.1351-1363(2016).
2. Li, H.L., Toyoura, M., Mao, X.Y.: Caricature synthesis with feature deviation matching under example-based framework. The Visual Computer (2018). https://doi.org/10.1007/s00371-018-1495-9

Acknowledgements

I would like to thank my supervisor Professor Mao for her guidance, encouragement, support, inspiration an enthusiasm during my time at University of Yamanashi. With her help, I have learned much knowledge on computer vision research and obtained experience of portrait feature designed and their applications based on traditional and deep learning frameworks.

Associate Professor Toyoura also provided significant support for all researches that relate to this work. He attentively guided me for all technical details of the above research projects, which conducted me to accomplish this work.

Tomomi Shimizu contributed greatly to Chapter 2. Hongchang Sun and Wei Yang did much preliminary work for Chapter 3. Jiayi Xu also provided me much advice for these research projects.

Last but not least, I am deeply grateful to my family. With the mental and financial support from my father and mother, I can try to conquer all the difficulties in my research. Although my father passed away just after I finished the second paper, I will continue to conquer all the challenges in the future in memory of him.

References

Chapter 1

[1] Veltkamp,R.C.,Tanase,M .:Content-based image retrieval systems: A Survey. Technical Report UUCS-2000-34, October (2000).

[2] Hsu,E.,Paz,C.,Shen,S.:Clothing image retrieval for smarter Shopping. EE368, Department of Electrical and Engineering, Stanford University (2011).

[3] Silva,A.T.,Falcao,A.X,.Magalhaes,L.P.: Active learning paradigms for CBIR systems based on optimum-path forest classification. Pattern Recognition 44, 2971-2978 (2011).

[4] Z Sadimon, S.B., Sunar, M.S., Mohamad, D., Haron, H.: Computer generated caricature: a survey. In: Proceedings of International Conference on Cyberworlds, pp. 383-390 (2010).

[5] Chen, H., Zheng, N.N., Liang, L., Li, Y., Xu, Y.Q., Shum, H.Y.: PicToon: a personalized image-based cartoon system. In: Proceedings of the tenth ACM international conference on Multimedia, pp. 171–178 (2002).

[6] Wang, N.N., Tao, D.C., Gao, X.B., Li, X.L., Li, J.: Transductive face sketch-photo synthesis. IEEE transaction on neural network and learning systems. **24** (9), pp. 1364–1376 (2013).

[7] Min, F., Suo, J.L., Zhu, S.C., Sang, N.: An automatic portrait system based on and-or graph representation. In: International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), pp. 184–197 (2007).

[8] Xu, Z.J, Chen, H., Zhu, S.C., Luo, J.B.: A hierarchical compositional model for face representation and sketching. IEEE Transactions on Pattern Analysis and Machine Intelligence. **30** (6), 955–969 (2008).

[9] Chen, H., Liu, Z.Q., Rose, C., Xu, Y.Q., Shum, H.Y., Salesin, D.: Example-based composite sketching of human portraits. In: Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering, pp. 7–9 (2004).

[10] Chen, H., Xu, Y.Q., Shum, H.Y., Zhu, S.C., Zheng, N.N.: Example-based facial sketch generation with non-parametric sampling. In: Proceedings of International Conference on Computer Vision, 2, pp. 433–438 (2001).

[11] Chen, W.J., Yu, H.C., Zhang, J.J: Example based caricature synthesis. Advances in Computer Science & Engineering. **5** (1) (2010).

[12] Liang, L., Chen, H., Xu, Y.Q., Shum, H.Y.: Example-based caricature generation with exaggeration. In: Computer Graphics and Applications, Pacific Conference, pp.386–393 (2002).

[13] Yang, W., Tajima, K., Xu, J.Y, Toyoura, M., Mao, X.X.: Example-based automatic caricature generation. In: Cyberworlds, pp. 237–244 (2014).

[14] Gooch, B., Reinhard,E., Gooch,A.: Human facial illustrations: creation and psychophysical evaluation. ACM Transactions on Graphics (TOG). **23** (1), pp. 27–44 (2004).

[15] Wang, X.G, Tang, X.O.: Face photo-sketch synthesis and recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence. **31**(11), pp. 1955–1967 (2009).

[16] Song,Y.B., Bao, L.C., Yang, Q.X., Yang, M.H.: Real-time exemplar based face sketch synthesis. In: Computer Vision – ECCV 2014, pp. 800–813 (2014).

[17] Yu, L. F., Yeung, S.K., Terzopoulos, D., Chan, T. F.: DressUp! Outfit synthesis through automatic optimization. ACM Transactions on Graphics (TOG). **31** (6), pp. 134:1–134:14 (2012).

[18] Kalogerakis, E., Chaudhuri, S., Koller, D., Koltun, V.: A probabilistic model for component-based shape synthesis. ACM Transactions on Graphics (TOG). **31** (4), pp. 55:1–55:11 (2012).

[19] Yang, W., Toyoura, M., Xu, J.J., Ohnuma, F., Mao. X.Y.: Example-based caricature generation with exaggeration control. The Visual Computer. **32** (3), pp. 383–392 (2016).

[20] Zhang, Y., Dong, W.M., Ma, C.Y., Mei, X., Li, K., Huang, F.Y., Hu, B.G., Deuusen, O.: Data-driven synthesis of cartoon faces using different styles. IEEE Transactions on Image Processing. **26** (1) , 464-478 (2017).

[21] Shih, Y.C., Paris, S., Barnes, C., Freeman, W.T., Durand, Frédo.: Style transfer for headshot portraits. ACM Transactions on Graphics (TOG). **33** (4), 148(2014).

[22] Selim, A., Elgharib, M., Doyle L.: Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (TOG). **35** (4), 129 (2016).

[23] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016.

[24] Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. In: arXiv:1705.01088[cs.CV](2017).

[25] Fišer, J., Jamriška, O., Simons, D., Shechtman, E., Lu, J.W., Asente, P., Lukáč, M., Sýkora, D.: Example-based synthesis of stylized facial animations. ACM Transactions on Graphics (TOG). **36** (4), 155(2017).

[26] Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In ICLR (2017).

[27] Zhang, D.Y., Lin, L., Chen, T.S., Wu, X., Tan, W.W., Izquierdo, E.: Content-adaptive sketch portrait generation by decompositional representation learning. IEEE Transactions on Image Processing. **26** (1) , 328-339(2017).

[28] Galen. A., Arora, R., Bilmes,J., Livescu,K.: Deep Canonical Correlation Analysis. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ser. ICML’13, pp. III-1247-III-1255(2013).

Chapter 2

[29] Hackler,N.: UK COOPERATIVE EXTENSION SERVICE. University of Kentucky- College of Agriculture, CT-LMH.185 (1997).

[30] Onuma,J.:Coordinate Technique-Appealed Edition I. Bunka Publishing, Bunka Fashion College (2001).

[31] Pundir,N.:FASHION TECHOLOGY–Today and Tomorrow. A Mittal Publication, New Delhi (India) (2007).

[32] Wang,L.,Tian,B.:An Analysis of Factors Determining the Shape of Collar. Journal of Panzhihua University Vol.2, 87-89 (2008).

[33] Fang, J.:3D collar design creation. International Journal of Clothing Science and Technology Vol. 15 Iss: 2, pp.88 – 106 (2003).

[34] Liu,Y., Zhang, D., Yuen, M.: A survey on CAD methods in 3D garment design. Computer. Ind. 61, 6, 576-593,(2010).

[35] Zhang,X.,Wong,L.Y.: Virtual fitting: real-time garment simulation for online shopping. In ACM SIGGRAPH Posters (2014).

[36] Hauswiesner,S.,Straka,M.,Reitmayr,G.: Virtual Try-On Through Image-based Rendering. IEEE Transactions on Visualization and Computer Graphics (TVCG), vol.19, No.9, 2 1552-1565(2013).

[37] Shimizu,K.,Yang,W.,Toyoura,M.,Mao,X.:Relevance Feedback Based Retrieval of Cloth Image with Focus on Collar Design.Cyberworlds (2015-10).

[38] Liu,S., Song,Z., Liu,G., Xu,C., Lu,H., Yan,S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. IEEE Conference on Computer Vision and Pattern Recognition, pp.3330-3337 (2012).

[39] Bossard,L.,Dantone,M.,Leistner,C.,Wengert,C.,Quack,T., Gool,L.V.: Apparel classification with style.ACCV12, vol4, pp.321-335 (2012).

[40] Kondo,S., Toyoura, M., Mao, X.:Skirt Image Retrieval based on Sketches. Sketch-Based Interfaces and Modeling (2014).

[41] Tekawa,M.,Hattori,M.:Improvement of Reuse of Classifiers in CBIR using SVM Active Learning. Proc. ICONIP Vol.2, pp.598-605 (2010).

[42] Onuma,J.:Coordinate Technique-Appealed Edition I. Bunka Publishing, Bunka Fashion College (2001).

[43] Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and Learning Object Categories, In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, short course (2007).

[44] Qiu, G.: "Indexing chromatic and achromatic patterns for content-based color image retrieval, Pattern Recognition 35 (8): 1675–1686(2002).

[45] Nister, D., Stewenius,H.:Scalable Recognition with a Vocabulary Tree. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, 2161-2168 (2006).

[46] Hou,X.,Zhang,L.:Saliency Detection: A Spectral Residual Approach. Computer Vision and Pattern Recognition, pp.1-8 (2007).

[47] Hou,X.,Zhang,L.:Saliency Detection: A Spectral Residual Approach. Computer Vision and Pattern Recognition, pp.1-8 (2007).

[48] Papa,J.P.,Falcao,A.X.,Suzuki,C.T.N.:Supervised pattern classification based on optimum-path forest. International Journal of Imaging Systems and Technology 19 (2) 120‑131 (2009).

[49] Silva,A.T., Falcao,A.X., Magalhaes,L.P.: A new CBIR approach based on relevance feedback and optimum path forest classification. Journal of WSCG 18 (1‑3) ,73‑80 (2010).

[50] Papa,J.P., Falcao,A.X.: Optimum-Path Forest: A Novel and Powerful Framework for Supervised Graph-based Pattern Recognition Techniques. Institute of Computing University of Campinas (2010).

Chapter 3

[51] Yang, W., Toyoura, M., Xu, J.J., Ohnuma, F., Mao. X.Y.: Example-based caricature generation with exaggeration control. The Visual Computer. **32** (3), pp. 383–392 (2016).

[52] Zhang, Y., Dong, W.M., Ma, C.Y., Mei, X., Li, K., Huang, F.Y., Hu, B.G., Deuusen, O.: Data-driven synthesis of cartoon faces using different styles. IEEE Transactions on Image Processing. **26** (1) , 464-478 (2017).

[53] Zhang, D.Y., Lin, L., Chen, T.S., Wu, X., Tan, W.W., Izquierdo, E.: Content-adaptive sketch portrait generation by decompositional representation learning. IEEE Transactions on Image Processing. **26** (1) , 328-339(2017).

[54] Shih, Y.C., Paris, S., Barnes, C., Freeman, W.T., Durand, Frédo.: Style transfer for headshot portraits. ACM Transactions on Graphics (TOG). **33** (4), 148(2014).

[55] Harmon,L.D.: The Recognition of Faces. Scientific American. **229** (5), 71-82(1973).

[56] Brennan, S.E.: Caricature generator: the dynamic exaggeration of faces by computer. Leonardo. **18** (3), 170-178(1985).

[57] Koshimizu, H., Tominaga, M., Fujiwara, T, Murakami.K. : On KANSEI facial image processing for computerized facial caricaturing system Picasso. In: IEEE International Conference on Systems , pp. 294-299(1999).

[58] Mo, Z.Y., Lewis, J.P., Neumann, U.: Improved automatic caricature by feature

normalization and exaggeration. In: ACM SIGGRAPH Sketches, pp. 57(2004).

[59] Xu, J.Y., Yang, W., Mao, X.Y., Toyoura, M. Jin, X.G.: A study on perceived similarity between photograph and shape exaggerated caricature. In: International Conference on Cyberworlds, pp. 213-220(2014).

[60] Cosker, D., Roy, S., Rosin, P.L, Marshall, D.: Re-mapping animation parameters between multiple types of facial model. In: International Conference on Computer Vision/Computer Graphics Collaboration Techniques, pp. 365-276(2007).

[61] Hotelling, H.: Relations between two sets of variates. Biometrika. **28** (3/4), 321-377(1936).

[62] Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM international conference on Multimedia, pp. 251-260(2010).

[63] Cao, Y., Long, M.S., Wang, J.M., Liu, S.C.: Collective deep quantization for efficient cross-modal retrieval. In: IEEE International Symposium on Multimedia, pp. 3974-3980(2017).

[64] Zhong, C.L., Yu, Y., Tang, S.H., Satoh, S., Xing, K.: Deep multi-label hashing for large-scale visual search based on semantic graph. In: Spring International Publishing, pp. 169-184(2017).

[65] Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441-3450(2015).

[66] Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on International Conference on Machine Learning, pp. Ⅲ-1247-Ⅲ-1255(2013).

[67] Yu, Y., Tang, S.H., Raposo, F., Chen, L.: Deep cross-modal correlation learning for audio and lyrics in music retrieval. In: arXiv:1711.08976[cs.IR](2017).

[68] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and Application. Computer Vision and Image Understanding. **61** (1), pp. 38–59 (1995).

[69] Li, H.L., Yang, W., Sun, H.C., Toyoura, M., Mao, X.Y.: Example-based caricature synthesis via feature deviation matching. In: CGI'16 Proceedings of the 33rd Computer Graphics International, pp.81-84 (2016).

[70] Yacoob, Y., Davis, L.S.: Detection and analysis of hair. IEEE Transactions on Pattern Analysis & Machine Intelligence. **28** (7), pp.1164–1169 (2006).

[71] Luc,V., Pierre,S.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence. **13** (6) , 583-598(1991).

[72] Bookstein, F.L.: Principal warps: Thin-Plate Splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence. **11** (6), 567–585 (1989).

[73] Li, H.L., Toyoura, M., Shimizu, K., Yang, W., Mao, X.Y.: Retrieval of clothing images based on relevance feedback with focus on collar designs. The Visual Computer. **32** (10), pp. 1351-1363(2016).

[74] Yang, W., Toyoura, M., Mao, X.Y.: Hairstyle suggestion using statistical learning. In: International Conference on Multimedia Modeling, pp.277-287(2012).

[75] Sunhem, W., Pasupa, K.: An approach to face shape classification for hairstyle recommendation. In: Eighth International Conference on Advanced Computational Intelligence, pp.390-394(2016).

Chapter 4

[76] Chen,H.Z., Gallagher,A.,Griod,B.: Describing clothing by semantic attributes. In: Computer Vision – ECCV 2012. ECCV 2012, pp. 609-623(2012).

[77] Wang,X.W., Zhang,T.: Clothes search in consumer photos via color matching and attribute learning. In: MM '11 Proceedings of the 19th ACM international conference on Multimedia, pp.1353-1356(2011).

[78] Shankar,D.,Narumanchi,S.,Ananya,H.A.,Kompali,P.,Chaudhury,K.: Deep learning based large scale visual recommendation and search for E-Commerce. In: arXiv:1703.02344 [cs.CV] (2017).

[79] Liu,Z.W., Luo,P.,Qiu,S.,Wang,X.G.,Tang,X.O.: Deep Fashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: Computer Vision & Pattern Recognition (CVPR), pp.1096-1104(2016).

[80] Girshick,R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: arXiv: 1311.2524 [cs.CV] (2014).

[81] Huang,J., Feris, R., Chen, Q., Yan, Shuicheng.: Cross-Domain image retrieval with a dual attribute-aware ranking network. In: ICCV '15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp: 1062-1070(2015).

[82] Jagadeesh,V.,Piramuthu, R.,Bhardwaj,A.,Di,W.,Sundaresan,N.: Large scale visual recommendations from street fashion images. In: KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1925-1934(2014).

[83] Hu,Y., Yi, X., Davis, L.S.: Collaborative Fashion Recommendation: A full functional tensor factorization approach. In: MM '15 Proceedings of the 23rd ACM international conference on Multimedia, pp.129-138(2015).

[84] Li,Y.C.,Cao,L.L.,Zhu,J.,Luo,J.B.: Mining fashion outfit composition using an end-to-end deep learning approach on set data. IEEE Transactions on Multimedia.**19** (8), pp.1946-1955(2017).

[85] Ma,Y.H., Jia, J., Zhou, S., Fu, J., Liu, Y.J., Tong, Z.J.: Towards better understanding the clothing fashion styles: A multimodal deep learning approach. In: Thirty-First AAAI Conference on Artificial Intelligence, (2017).

[86] He,R., McAuley, J.: Visual bayesian personalized ranking from implict feedback. In: arXiv: 1510.01784(2015).

[87] Liu,S.,Feng,J.S.,Song,Z.,Zhang,T.Z.,Lu,H.Q.,Xu,C.S.,Yan,S.C.: Hi, magic closet, tell me what to wear. In: MM '12 Proceedings of the 20th ACM international conference on Multimedia, pp.619-628(2012).

[88] Yang, W., Toyoura, M., Mao, X.Y.: Hairstyle suggesting using statistical learning. In: MMM: Advances in Multimedia Modelling, pp.277-287(2012).

[89] Mao,J.H.,Xu,W.,Yang,Y.,Wang,J.,Yuille,A.L.:Explaine images with multimodal recurrent neural networks.In: arXiv:1410.1090[2014].

[90] Wu,X.X.,Qiao,Y.,Wang,X.G.,Tang,X.O.:Bridging music and image via cross-modal ranking analysis. IEEE Transactions on Multimedia.**18** (7), pp.1305-1318(2016).

[91] Yu,Y.,Tang,S.H.,Raposo,F.,Chen,L.: Deep cross-modal correlation learning for audio and lyrics in music retrieval. In: arXiv: 1711.08976(2017).

[92] Hotelling, H.: Relations between two sets of variates. Biometrika.**28** (3/4), pp.321-377(1936).

[93] Hardoon, D. R., Szedm ́ak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation. **16**(12), pp.639–2664(2004).

[94] Sun,Y.,Wang,X.J.,Tang,X.O.: Deep learning face representation from predicting 10,000 classes. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp.1891-1898(2014).

[95] Huang, J.S., Feris, R., Chen, Q., Yan, S.C.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: IEEE International Conference on Computer Vision, pp. 1062-1070(2015).

[96] Ganin, Y., Lempitsky, V.: Unsupervised Domain Adaptation by Back-propagation. In: arXiv: 1409.7495(2014).

[97] Wang, B.K., Yang, Y., Xu, X., Hanjalic, A.,Shen,T.H.: Adversarial cross-modal retrieval. In: Proceeding of the 2017 ACM on Multimedia Conference. pp. 154-162(2017).

[98] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: arXiv: 1409.1556(2014)

[99] Maaten, LVD., Hinton, G.: Visualizing data using t_SNE. Journal of Machine Learning.**9**(2605),pp.2579-2605(2008).