Noname manuscript No.
(will be inserted by the editor)

Caricature Synthesis with Feature Deviation Matching under Example-Based framework

Honglin Li · Masahiro Toyoura · Xiaoyang Mao

Abstract Example-based caricature synthesis techniques have been attracting large attentions for being able to generate attractive caricatures of various styles. This paper proposes a new example-based caricature synthesis system using a feature deviation matching method as a cross-modal distance metric. It employs the deviation values from average features across different feature spaces rather than the values of features themselves to search for similar components from caricature examples directly. Compared with traditional example-based systems, the proposed system can generate various styles of caricatures without requiring paired photo-caricature example databases. The newly designed features can effectively capture visual characteristics of the hairstyles and facial components in input portrait images. In addition, this system can control the exaggeration of individual facial components, and provide several similarity-based candidates to satisfy users' different preferences. Experiments are conducted to prove the above ideas.

Keywords Caricature synthesis · Example-based · Cross-modal distance metric · Feature deviation matching

1 Introduction

Since the early 1980s, many computer-based methods have been developed for synthesizing caricatures [1]. These can be roughly classified into photo-transformed and example-based techniques. Photo-transformed approaches [2-5, 12-14, 28] 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 require a large number of photo-caricature pairs as example data [6-10, 15, 16, 18, 20]. 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 [28] 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

H. Li · M. Toyoura · X. Mao
University of Yamanashi,
4-3-11, Kofu, Yamanashi, 400-8511, Japan
E-mail: {g15dhl01|mtoyoura|mao}@yamanashi.ac.jp

do not provide users with any control over the details of stylization, such as the degree of exaggeration of individual facial components.

In this paper, 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 [6-10, 15, 16, 18, 20] 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.

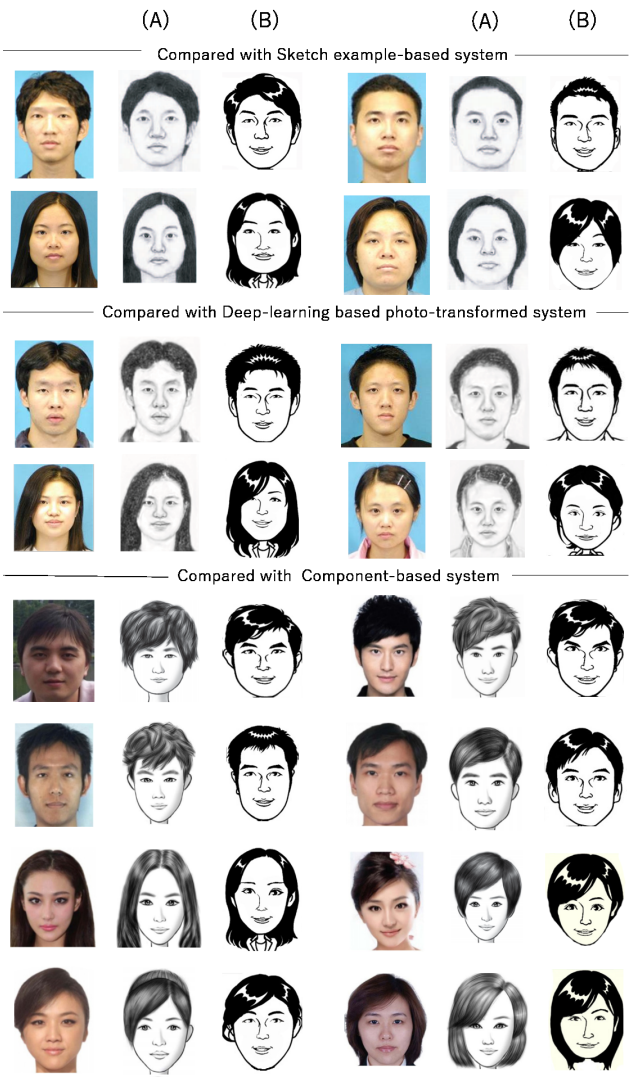


Fig. 1 Comparison of the expressive style synthesized by our proposed system with those generated by three state-of-art systems (a sketch example-based system [13], a deep-learning based photo-transformed system [28] and a component-based system [20]), A and B columns show the results of the other systems and our system respectively.

In summary, the contributions of this paper 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. 1 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 [13]. The third and fourth rows: Comparison with deep-learning based photo-transformed system [28]. The fifth to eighth: Comparison with component-based system [20]). 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 [20] method for similarity and are comparable with [13, 28] for as an expressive style.

The remaining part of the paper is organized as follows: Section 2 reviews related works. Section 3 presents the proposed method. Section 4 demonstrates some results and describes the evaluation experiments. Section 5 presents conclusions from the study.

2 Related researches

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. [12] 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. [4] 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 [3]. Recently, Zhang et al. proposed a content-adaptive method for generating portrait sketches based on deep-learning techniques under the photo-transformed framework [28], 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 [13] and image denoising techniques [14], 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 [6]. 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 [18]. 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 [22], 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 [25]. 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 [26]. 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 [27].

Photo-transformed and sketch-based systems are good at generating caricatures very similar to the original portrait photos. Style-transfer-based and example-based systems are good at generating caricatures of various artistic styles at the price of losing some of the original photograph identity. 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.

2.2 Feature deviation applications and cross-modal comparisons

Based on the variation within a population of faces, one could determine an average face [31]. 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 [32]. 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 [33]. 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 [34]. Xu et al. investigated the borderline between likeness and unlikeness through applying the EDFM rule to gradually alternate the face shape at subject study [35]. Cosker et al. learned animation parameters from human video performance and reused them to animate multiple types of facial model [36], 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 [32-36] 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 [37, 38]. Recently, Deep-learning techniques are also widely applied in cross-modal analysis between different modalities [39, 40]. 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. [41-43] 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 Proposed method

3.1 Overview

Fig. 2 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 (4) to (8) show. Feature deviation matching is performed component-by-component across photo and caricature. The ASM (Active shape models) algorithm [11] 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 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.

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 [21] 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 spaces. By comparing the feature deviation values of the input portrait 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. 3, and the evaluation results shown in Section 4 are of comparable quality.

Eye: Two angles, form factor, rectangularity, and eccentricity are used to constitute a 5d feature vector $\{\theta_1, \theta_2,$ 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 (1). Rectangularity is the result of the eye contour area divided by its corresponding minimum bounding box area as formula (2). The eccentricity

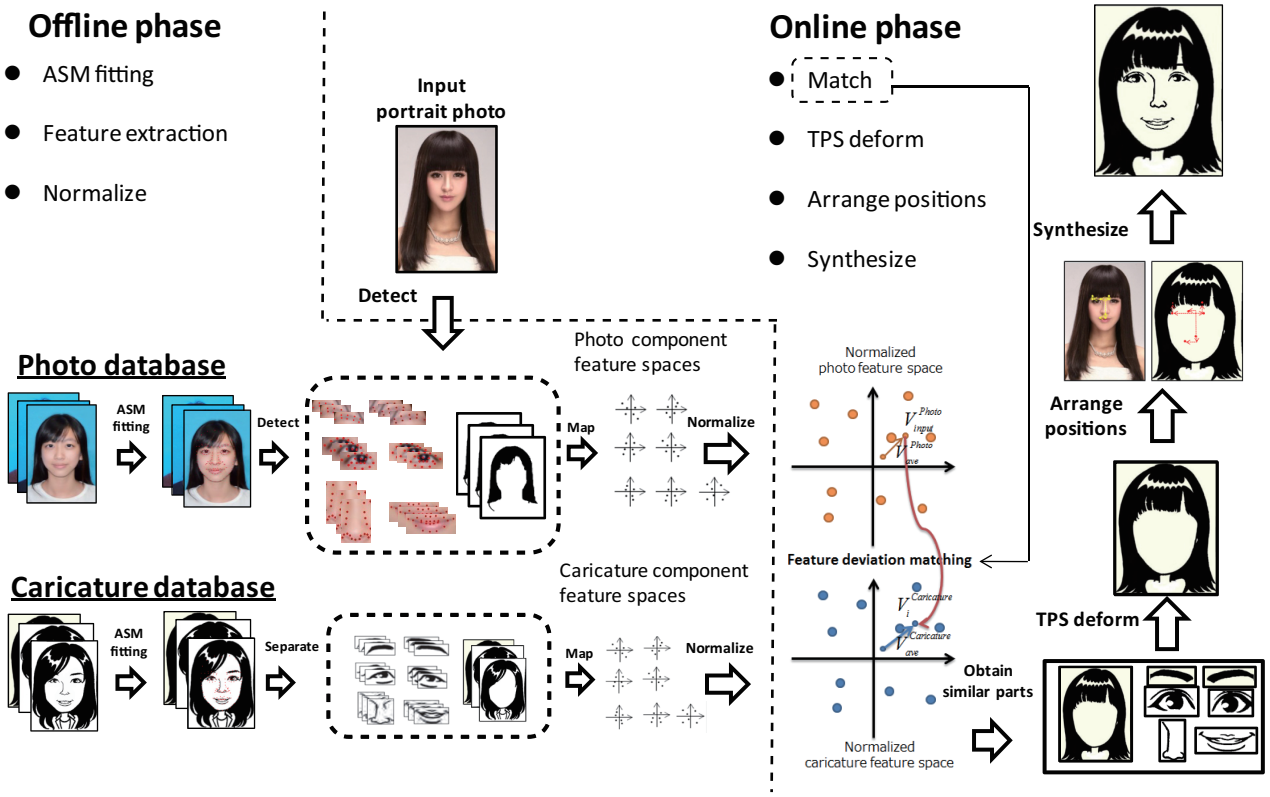


Fig. 2 Framework of the proposed system

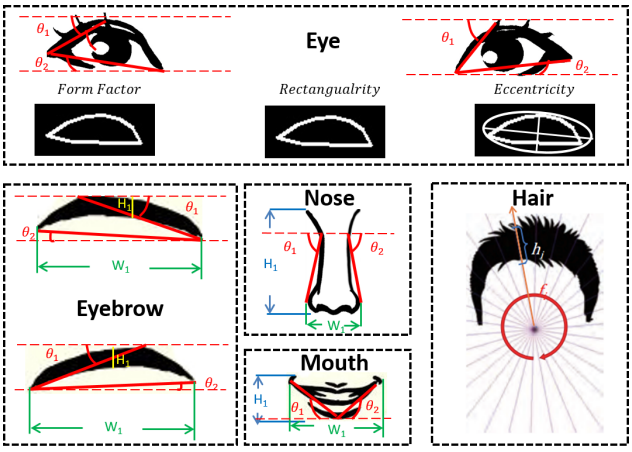


Fig. 3 Designed feature vectors of facial and hairstyle components

value of the eye is calculated by dividing the long axis by the short axis as formula (3).

$$Form\ Factor = \frac{4\pi \times Area}{Perimeter^2} \tag{1}$$

$$Rectangualrity = \frac{Area_{object}}{Area_{bounding-box}} \tag{2}$$

$$Eccentricity = \frac{AxisLength_{long}}{AxisLength_{short}} \tag{3}$$

Eyebrows, *Nose*, and *Mouth*: Eyebrows, nose and mouth are represented by 3d feature vectors $\{\theta_1,\ \theta_2,\ \frac{W_1}{H_1}\}$, 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 [17] to detect the hair region, and adopt a feature vector similar to that used in [10] for representing the hairstyle component. Here, the hairstyle is represented by a 120d feature vector $\{\frac{h_i}{f_i}\}(i = 1, 2, \cdots, 120)$, 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 [30].
Step 3: Calculate the 120d feature vectors.

In step 1, as shown in Fig. 4(a), starting with the ASM points on the upper face contour, we deploy the

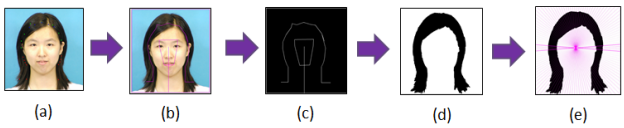


Fig. 4 Procedures for separating hair area from the input portrait photo and obtaining hair feature vectors

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. 4(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. 4(c)), the watershed algorithm is then used to segment the input portrait photo into two parts, the hair and background area (Fig. 4(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. 4(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 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 1 and 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 ($j = 1, 2, \cdots, n; n$ 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 ($k = 1, 2, \cdots, m; m$ 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 $\hat{V}_i^{pho}$, $\check{V}_i^{pho}$, $\bar{V}_i^{pho}$, $\hat{V}_i^{car}$, $\check{V}_i^{car}$, 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 $\hat{\alpha}_i^{pho}$, $\check{\alpha}_i^{pho}$, $\bar{\alpha}_i^{pho}$, $\hat{\alpha}_i^{car}$, $\check{\alpha}_i^{car}$, 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 (4) to (6) so that all values are within the range 0 to 1.

$$\alpha_i^{in} = \frac{V_i^{in} - \check{V}_i^{pho}}{\hat{V}_i^{pho} - \check{V}_i^{pho}}, \tag{4}$$

$$\alpha_i^{pho,j} = \frac{V_i^{pho,j} - \check{V}_i^{pho}}{\hat{V}_i^{pho} - \check{V}_i^{pho}}, \tag{5}$$

$$\alpha_i^{car,k} = \frac{V_i^{car,k} - \check{V}_i^{car}}{\hat{V}_i^{car} - \check{V}_i^{car}}. \tag{6}$$

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 (7) and (8).

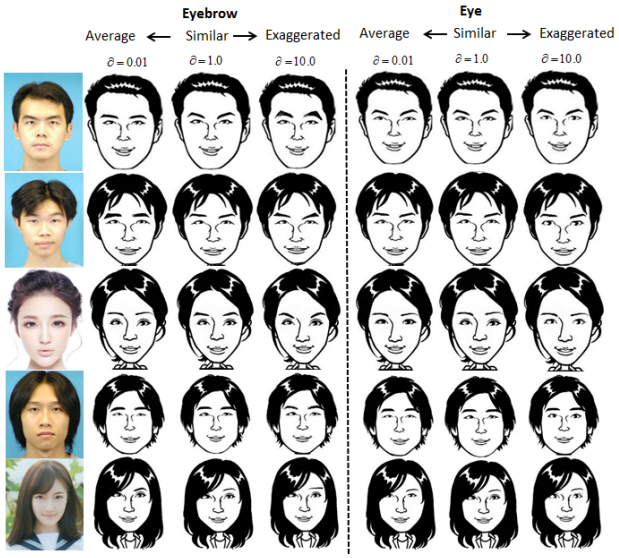


Fig. 5 Synthesized caricatures with different exaggeration coefficients for eyes or eyebrows

$$\beta_i^{in} = \frac{\alpha_i^{in} - \bar{\alpha}_i^{pho}}{\bar{\alpha}_i^{pho} - \breve{\alpha}_i^{pho}}, \tag{7}$$

$$\beta_i^{car,k} = \frac{\alpha_i^{car,k} - \bar{\alpha}_i^{car}}{\bar{\alpha}_i^{car} - \breve{\alpha}_i^{car}}. \tag{8}$$

The matching caricature component $\tilde{k}_i$ in the caricature component database can be determined via formula (9).

$$\tilde{k}_i = \arg\min_k ||\beta_i^{in} - \beta_i^{car,k}||. \tag{9}$$

By applying formula (9), 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 [18] by configuring an exaggeration coefficient in formula (9) as follows:

$$\tilde{k}_i = \arg\min_k ||\partial\beta_i^{in} - \beta_i^{car,k}||. \tag{10}$$

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

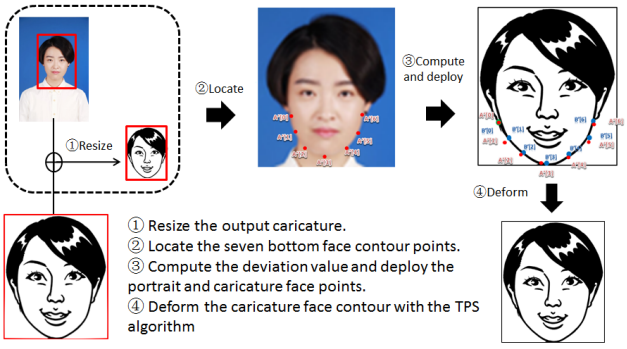


Fig. 6 Deformation of the caricature face shape using the TPS algorithm

can be applied to the facial components except for the hairstyle component. Fig. 5 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.

3.4 Synthesis of resulting caricature

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

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 [35]. 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. 6.

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 [19] 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.

3.4.2 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 (11) and (12) are applied to calculate the widths (W) and heights (H) of the circumscribed rectangles.

$$W_i^{out} = \frac{W_i^{in}}{W_{face}^{in}} \times W_{face}^{out}, \tag{11}$$

$$H_i^{out} = \frac{H_i^{sea}}{W_i^{sea}} \times W_i^{out}. \tag{12}$$

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 (13) and (14) (x and y represent horizontal and vertical distances between individual facial components).

$$d_i^{out,x} = \frac{d_i^{in,x}}{W_{face}^{in}} \times W_{face}^{out}, \tag{13}$$

$$d_i^{out,y} = \frac{H_i^{in,y}}{H_{face}^{in}} \times H_{face}^{out}. \tag{14}$$

As shown in Fig. 7, 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. 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.

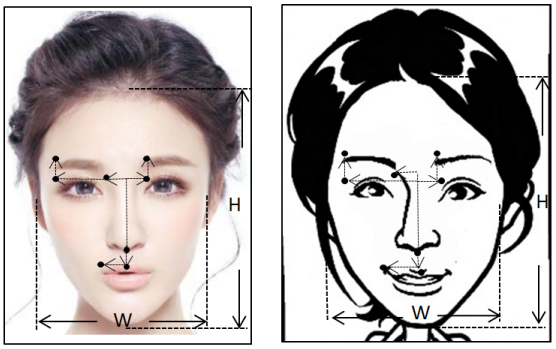


Fig. 7 Procedure for locating positions of facial components

4 Results and evaluation

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.

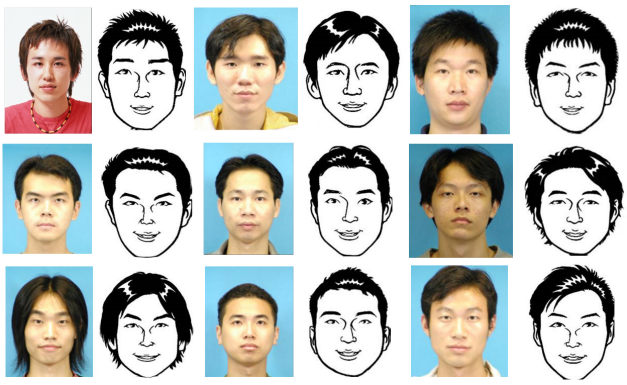


Fig. 8 Examples of the male caricatures of the expressive style

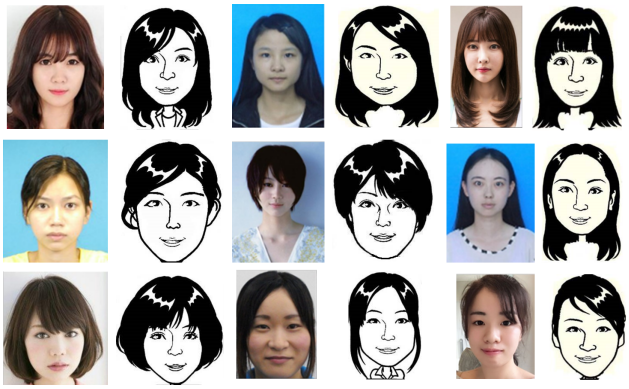


Fig. 9 Some examples of female caricatures of the expressive style

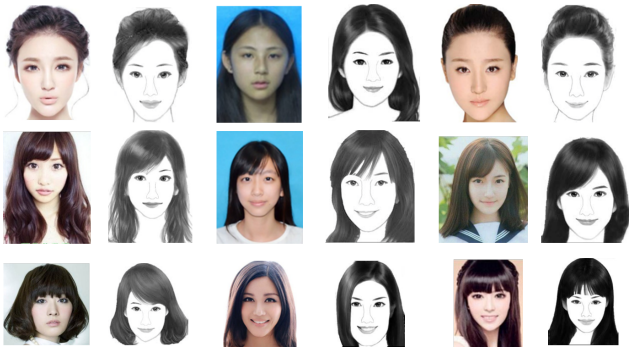


Fig. 10 Some examples of female caricatures of the photo-realistic style

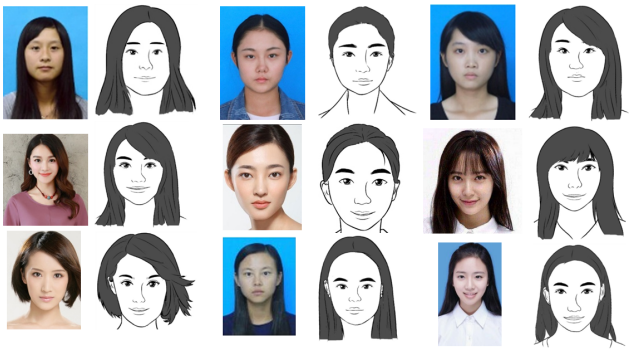


Fig. 11 Some examples of female caricatures of the drawing style

8, Fig. 9, Fig. 10 and Fig. 11 show some results based on the four different datasets respectively.

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. 5 shows. Furthermore, various combinations with similarity-based facial and hairstyle components can satisfy users' differing preferences in another way. Fig. 12, Fig. 13, Fig. 14 and Fig. 15 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 [29].

4.2 Evaluation

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

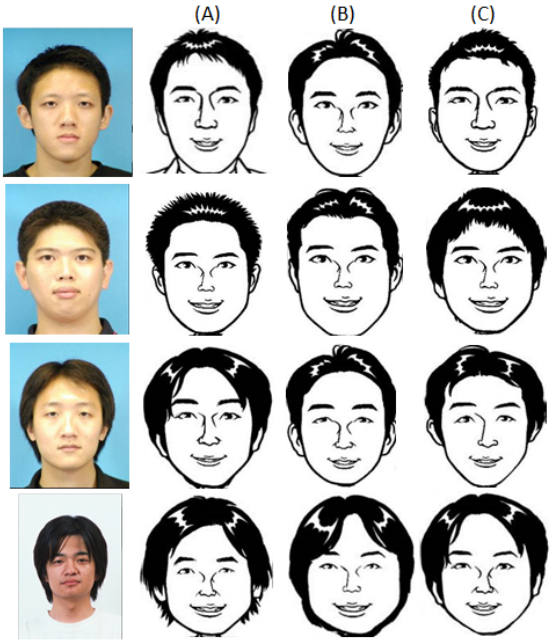


Fig. 12 Some examples of male caricatures of the expressive style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

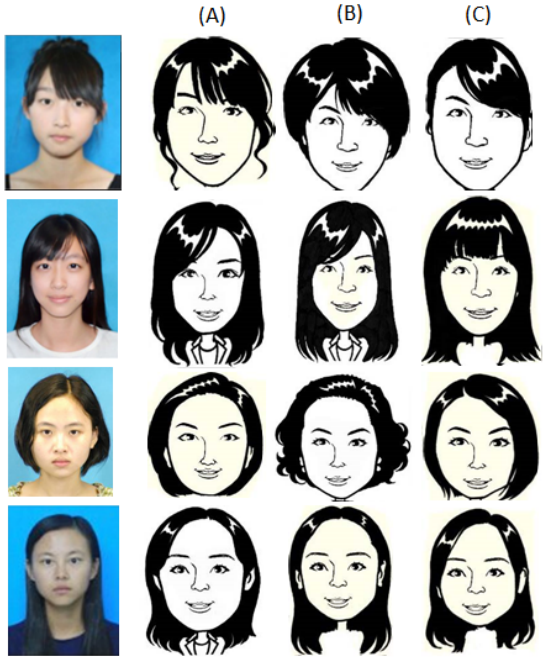


Fig. 13 Some examples of female caricatures of the expressive style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

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

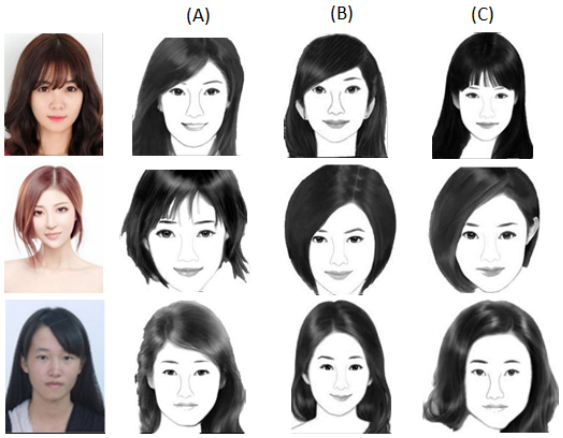


Fig. 14 Some examples of female caricatures of photo-rea-ist-c style-syn-hesized with the most (left), second-most (middle), and third-most (right) similar searched components

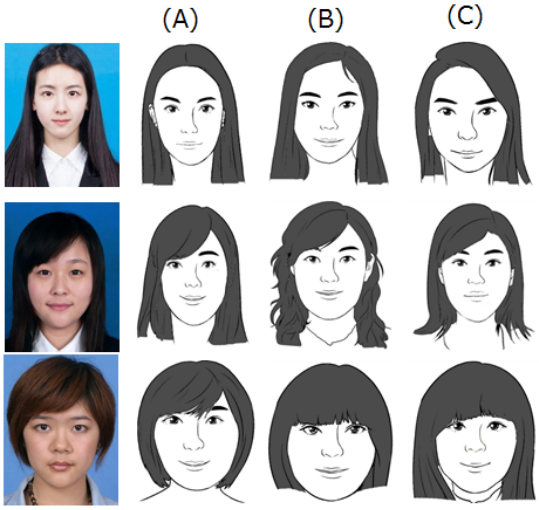


Fig. 15 Some examples of female caricatures of the drawing style synthesized with the most (left), second-most (middle), and third-most (right) similar searched components

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. 16 shows, which

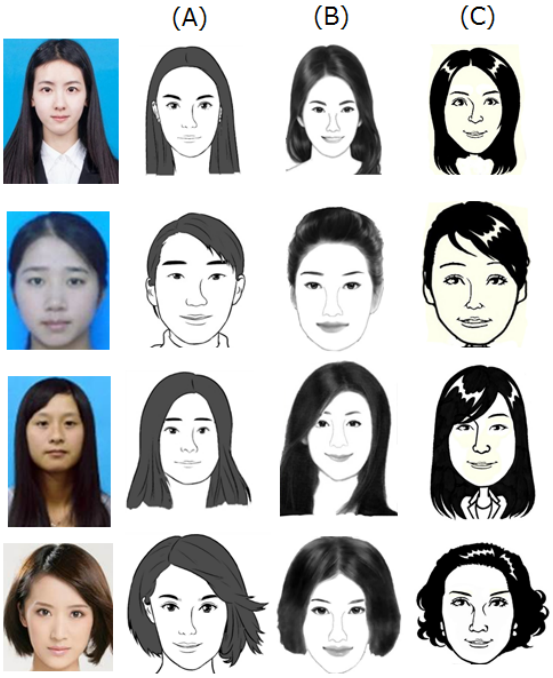


Fig. 16 Some examples displayed to participants in Experiment I (A, B and C represent caricatures of drawing, photo-realistic and expressive styles respectively)

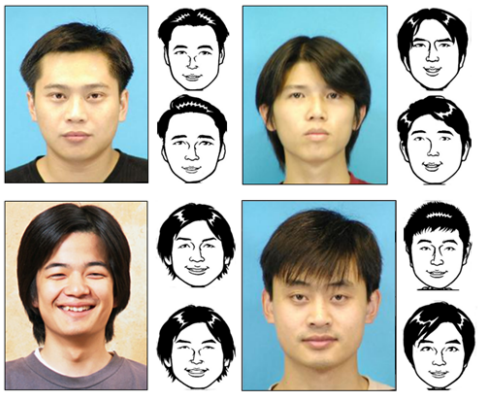


Fig. 17 Some examples in Experiment II on comparison of synthesized caricatures based on the proposed feature deviation matching and the paired-example matching methods

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.

4.3 Experiment II: Comparison with 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 40 randomly chosen male photos were compared between our proposed method and the paired-example method used in [18].

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 [18]. As Fig. 17 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.

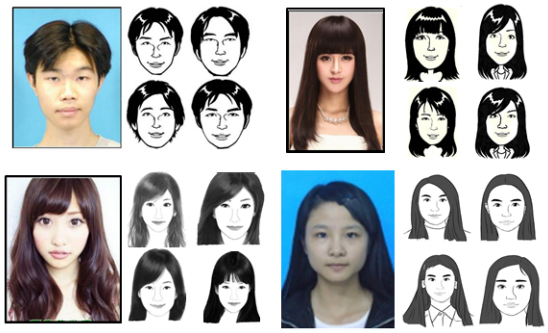


Fig. 18 Some examples in Experiment III on prominent internal feature capturing

4.4 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 3.2, 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 [44, 45]. Face shape also influences the impression of human face greatly [35]. The evaluation results may be largely affected by hairstyles and face shape if the remaining three caricatures are random chosen like [18, 20, 21]. 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 facial features from the input photos, we therefore conducted adjustments on the experiments of [18, 20, 21]. 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. 18, 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.

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/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.

Acknowledgements

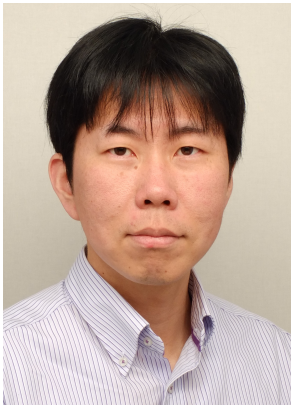
This study was funded by JSPS Grants-in-Aid for Scientific Research (Grant No. 26560006 and 26240015).

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. Chen, W.J., Yu, H.C., Zhang, J.J: Example based caricature synthesis. Advances in Computer Science & Engineering. **5**(1) (2010)
9. 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)
10. Yang, W., Tajima, K., Xu, J.Y, Toyoura, M., Mao, X.X.: Example-based automatic caricature generation. In: Cyberworlds, pp. 237–244 (2014)
11. 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)
12. Gooch, B., Reinhard,E., Gooch,A.: Human facial illustrations: creation and psychophysical evaluation. ACM Transactions on Graphics (TOG). **23**(1), pp. 27–44 (2004)
13. 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)
14. 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)
15. 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)
16. 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)
17. Yacoob, Y., Davis, L.S.: Detection and analysis of hair. IEEE Transactions on Pattern Analysis & Machine Intelligence. **28**(7), pp.1164–1169 (2006)
18. 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)
19. 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)
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. 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)
22. 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)
23. 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)
24. Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016
25. 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)
26. Fišer, J., Jamriška, O., Simons, D., Shechtman, E., Lu, J.W., Asente, P., Luká, M., Sýkora, D.: Example-based synthesis of stylized facial animations. ACM Transactions on Graphics (TOG). **36**(4), 155(2017).
27. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In ICLR (2017)
28. 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)
29. 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)
30. 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)
31. Harmon,L.D.: The Recognition of Faces. Scientific American. **229**(5), 71–82(1973)
32. Brennan, S.E.: Caricature generator: the dynamic exaggeration of faces by computer. Leonardo. **18**(3), 170–178(1985)
33. 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)
34. Mo, Z.Y., Lewis, J.P., Neumann, U.: Improved automatic caricature by feature normalization and exaggeration. In: ACM SIGGRAPH Sketches, pp. 57(2004)
35. 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)
36. 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)
37. Hotelling, H.: Relations between two sets of variates. Biometrika. **28**(3/4), 321–377(1936)
38. 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)
39. 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)
40. 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)
41. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3450(2015)
42. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on International Conference on Machine Learning, pp. -1247–-1255(2013)
43. 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)
44. Yang, W., Toyoura, M., Mao, X.Y.: Hairstyle suggestion using statistical learning. In : International Conference on Multimedia Modeling, pp.277–287(2012)
45. 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)

Honglin Li received the B.Sc. degree in Automation from Central South University and M.Sc. in Computer Science from Huaqiao University in 2003 and 2012 respectively in China. He began to study as a Ph.D candidate in Computer Science in University of Yamanashi from 2015. His research interests include computer vision, pattern recognition and data mining.



Masahiro Toyoura received the B.Sc. degree in Engineering, M.Sc. and Ph.D. degrees in Informatics from Kyoto University in 2003, 2005 and 2008 respectively. He is currently an Associate Professor at Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi, Japan. His research interests are augmented reality, computer and human vision. He is a member of ACM and IEEE Computer Society.

Xiaoyang Mao received her B.Sc. in Computer Science from Fudan University, China, M.Sc. and Ph.D. in Computer Science from University of Tokyo. She is currently a Professor at Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi, Japan. Her research interests include texture synthesis, non-photo-realistic rendering and their application to scientific visualization. She is a member of ACM SIGRRAPH and IEEE Computer Society.