Inter-modality Image Synthesis and Recognition

ZHANG, Wei

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Information Engineering

The Chinese University of Hong Kong
July 2012

Abstract

Inter-modality image synthesis and recognition has been a hot topic in computer vision. In real-world applications, there are diverse image modalities, such as sketch images for law enforcement and near infrared images for illumination invariant face recognition. Therefore, it is often useful to transform images from a modality to another or match images from different modalities, due to the difficulty of acquiring image data in some modality. These techniques provide large flexibility for computer vision applications.

In this thesis we study three problems: face sketch synthesis, example-based image stylization, and face sketch recognition.

For face sketch synthesis, we expand the frontier to synthesis from uncontrolled face photos. Previous methods only work under well controlled conditions. We propose a robust algorithm for synthesizing a face sketch from a face photo with lighting and pose variations. It synthesizes local sketch patches using a multiscale Markov Random Field (MRF) model. The robustness to lighting and pose variations is achieved with three components: shape priors specific to facial components to reduce artifacts and distortions, patch descriptors and robust metrics for selecting sketch patch candidates, and intensity compatibility and gradient compatibility to match neighboring sketch patches effectively. Experiments on the CUHK face sketch database and celebrity photos collected from the web show that our algorithm significantly improves the performance of the state-of-the-art.

For example-based image stylization, we provide an effective approach of transferring artistic effects from a template image to photos. Most existing methods do not consider the content and style separately. We propose a style transfer algorithm via frequency band decomposition. An image is decomposed into the low-frequency (LF), mid-frequency (MF), and high-frequency(HF) components, which describe the content, main style, and information along the boundaries. Then the style is transferred from the template to the photo in the MF and HF components, which is formulated as MRF optimization. Finally a reconstruction step combines the LF component of the photo and the obtained style information to generate the artistic result. Compared to the other algorithms, our method not only synthesizes the style, but also preserves the image content well. We demonstrate that our approach performs excellently in image stylization and personalized artwork in experiments.

For face sketch recognition, we propose a new direction based on learning face descriptors from data. Recent research has focused on transforming photos and sketches into the same modality for matching or developing advanced classification algorithms to reduce the modality gap between features extracted from photos and sketches. We propose a novel approach by reducing the modality gap at the feature extraction stage. A face descriptor based on coupled information-theoretic encoding is used to capture discriminative local face structures and to effectively match photos and sketches. Guided by maximizing the mutual information between photos and sketches in the quantized feature spaces, the coupled encoding is achieved by the proposed coupled information-theoretic projection forest. Experiments on the largest face sketch database show that our approach significantly outperforms the state-of-the-art methods.

摘要

跨模態圖像的合成和識別已成為計算機視覺領域的熱點。實際應用中存在各種各樣的圖像模態，比如刑偵中使用的素描畫和光照不變人臉識別中使用的近紅外圖像。由於某些模態的圖像很難獲得，模態間的轉換和匹配是一項十分有用的技術，為計算機視覺的應用提供了很大的便利。

本論文研究了三個應用：人像素描畫的合成，基於樣本的圖像風格化和人像素描畫識別。

我們將人像素描畫的合成的前沿研究擴展到非可控條件下的合成。以前的工作都只能在嚴格可控的條件下從照片合成素描畫。我們提出了一種魯棒的算法，可以從有光照和姿態變化的人臉照片合成素描畫。該算法用多尺度馬爾可夫隨機場來合成局部素描圖像塊。對光照和姿態的魯棒性通過三個部分來實現：基於面部器官的形狀先驗可以抑制缺陷和扭曲的合成效果，圖像塊的特征描述子和魯棒的距離測度用來選擇素描圖像塊，以及像素灰度和梯度的一致性來有效地匹配鄰近的素描圖像塊。在CUHK人像素描數據庫和網上的名人照片上的實驗結果表明我們的算法顯著提高了現有算法的效果。

針對基於樣本的圖像風格化，我們提供了一種將模板圖像的藝術風格傳遞到照片上的有效方法。大多數已有方法沒有考慮圖像內容和風格的分離。我們提出了一種通過頻段分解的風格傳遞算法。一幅圖像被分解成低頻、中頻和高頻分量，分別

描述內容、主要風格和邊緣信息。接著中頻和高頻分量中的風格從模板傳遞到照片，這一過程用馬爾可夫隨機場來建模。最後我們結合照片中的低頻分量和獲得的風格信息重建出藝術圖像。和其它算法相比，我們的方法不僅合成了風格，而且很好的保持了原有的圖像內容。我們通過圖像風格化和個性化藝術合成的實驗來驗證了算法的有效性。

我們為人像素描畫的識別提出了一個從數據中學習人臉描述子的新方向。最近的研究都集中在轉換照片和素描畫到相同的模態，或者設計復雜的分類算法來減少從照片和素描畫提取的特征的模態間差異。我們提出了一種新穎的方法：在提取特征的階段減小模態間差異。我們用一種基於耦合信息論編碼的人臉描述子來獲取有判別性的局部人臉結構和有效的匹配照片和素描畫。通過最大化在量化特征空間的照片和素描畫的互信息，我們設計了耦合信息論投影森林來實現耦合編碼。在世界上最大的人像素描畫數據庫上的結果表明我們的方法和已有最好的方法相比有顯著提高。

Acknowledgement

This thesis would not have been possible without the support of many individuals, to whom I would like to express my gratitude. First and foremost, I would like to thank my supervisor, Prof. Xiaoou Tang, for the opportunity of working with him, and for his continuous guidance, motivations, and encouragement in the past five years. His invaluable and generous help played a significant role in various stages of my way towards a research career. I would also like to thank my mentors, Prof. Xiaogang Wang, Prof. Jianzhuang Liu and Prof. Shifeng Chen. They supported my thesis research and provided a lot of help. Also, special thanks to my thesis committee members, who provided a lot of valuable comments.

In addition, my thanks are given to the professors in IE Department, and also my previous and present colleagues in Multimedia Laboratory. They helped me in many aspects during these years. I will never forget the joyful days with them. I am also grateful to all of my friends, wherever they are, for their encouragement and concerns. Finally, I would like to thank everyone, who supports me, and makes my life enjoyable as well as my work.

To my parents

Contents

List of Figures

List of Tables

Chapter 1

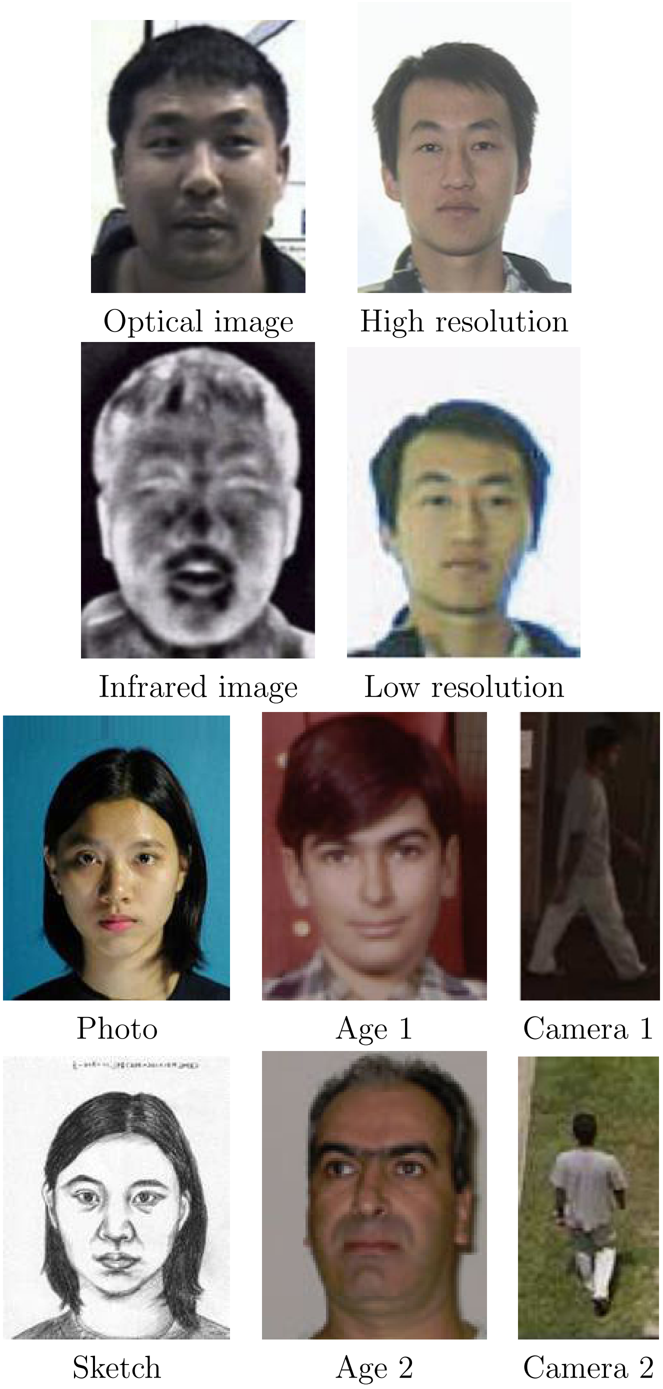
Introduction

1.1 Multi-Modality Computer Vision

The area of computer vision has experienced a rapid progress during the past decades. Understanding images is a central topic in computer vision. In practical applications, images are usually collected in quite different environments or even captured with different equipments. E.g., in a surveillance system, infrared cameras are employed to reduce the effect of illumination, so that the system can work from morning until night, under both strong and weak illuminations, in an adverse outdoor environment. The system queries a database of optical face photos taken with digital cameras, to recognize people captured by infrared cameras. Here optical photos and infrared photos are in different modalities, because they are from different sources and thus have different visual appearances and pixel values even for exactly the same content. Different modalities include optical and infrared images, high-resolution and low-resolution images, face photos and face sketches, faces of a person at different ages, and images under different camera views (Fig. 1.1).

Conventional computer vision techniques did not consider inter-modality differences. In many applications, the differences are very large and they affect the performance of computer vision algorithms heavily. So in recent years, there have been more and more interests in studying images of different modalities. Two major categories of problems are inter-modality image synthesis, i.e., transforming images from a modality to another modality, and inter-modality image recognition, i.e., matching images from different modalities. Inter-modality image synthesis and recognition are of great importance in computer vision, because it is often difficult to acquire image data in some modalities in real-world applications. These techniques provide large flexibility for developing real computer vision systems.

In this thesis, we visit three popular problems: face sketch synthesis from photos, unsupervised image stylization, and face



Optical image

High resolution

Infrared image

Low resolution

Photo

Age 1

Camera 1

Sketch

Age 2

Camera 2

Figure 1.1: Images of different modalities.

photo-sketch recognition. The first two are synthesis problems and the last one is a recognition problem. We focus on these applications, while the proposed approaches can be extended to other applications.

1.2 Face Sketches

Faces are one of the most commonly used biometrics and important in the individual's social interaction. In order to achieve computer-based face perception, numerous topics, such as face detection [75, 63, 73], face recognition [86, 65] and face hallucination [67, 42], have been studied extensively within the last several decades. With the advance of these automatic techniques, it becomes possible to build up systems for various applications, such as access control, video surveillance, and human computer interaction, to save human labors and provide better user experiences. However, these techniques are only for face photos and videos captured by cameras.

In many applications, face sketching is a popular technique that has been widely used. Psychology studies [4] show that a sketch captures the most informative part of a face, in a much more concise and potentially robust representation. We list two typical applications as follows.

- **Law enforcement**. The success of using face sketches to identify criminal suspects has often been publicized in the media coverage [23]. The sketches are drawn by artists based on the recollection of eye-witnesses, and then manually matched with a police mug-shot database comprising face photos of known individuals.
- **Entertainment**. Sketch artists are usually regarded as popular professionals for their ability to provide people personalized face drawings. The sketches can be put up on the wall, or used as people's identities in the digital world, such as through the MSN avatar. In the movie industry, sketch artists also play an active role in drawing cartoon faces.

In these applications, the expertise of the sketch artists is the key to the success, and it often costs a great amount of time for the artists to draw a sketch. It becomes unaffordable, if the size of the applications increase, e.g., drawing sketches for a thousand people. Therefore, it is desirable to have an automatic system to assist humans.

1.2.1 Face Sketch Synthesis

Automatic sketch synthesis system can save a great amount of time for artists on drawing face sketches. Due to the large demand in real-world applications, face sketch synthesis has drawn a great deal of attentions in recent years [60, 43, 22, 69, 9]. Face sketch synthesis also has its root in computer graphics, known as image stylization, a hot topic of non-photorealistic rendering. A comprehensive literature review can be found in Chapter 2.1.

Previous research in sketch synthesis can be categorized in the following two aspects.

- **Example-based approaches v.s. image processing based approaches**. Popular sketch synthesis methods are mostly example-based, which generate a sketch from an input face photo simulating the artistic style of a set of training face photo-sketch pairs [60, 43, 69]. Example-based approaches have the large flexibility of synthesizing sketches of different styles by choosing training sets of different styles. In contrast, image processing based sketch generation methods (e.g., [36]) support only limited artistic styles, and for a new style much human knowledge and experiences are required.

- **Styles with rich textures v.s. line drawing styles**. Recent face sketch synthesis research studies sketch styles with both contours and shading textures, which are more expressive than line drawings without texture [36, 19].

Following the above discussions, this thesis focuses on example-based sketch synthesis with rich textures, which is popular but challenging.

1.2.2 Face Sketch Recognition

Face sketch recognition is to match a face sketch drawn by an artist to one of many face photos in the database. In law enforcement, it is desired to automatically search photos from police mug-shot databases using a sketch drawing when the photo of a suspect is not available. Directly applying existing state-of-the-art face recognition algorithms leads to poor performance in this application, because the photos and sketches are in different modalities [60, 35]. Sketches are a concise representation of human faces, often containing shape exaggeration and having different textures from photos. Face sketch recognition is a special and challenging application of inter-modality face recognition. Other examples of inter-modality face recognition include infrared-optical face recognition, cross-resolution face recognition, and age-invariant face recognition.

Recently, great progress of face sketch recognition has been made in two directions.

The first family of approaches [60, 43, 69] focused on the *preprocessing* stage and synthesized a pseudo-photo from the query sketch or pseudo-sketches from the gallery photos to transform inter-modality face recognition into intra-modality face recognition. Face photo/sketch synthesis is actually a harder problem than recognition. Imperfect synthesis results significantly degrade the recognition performance. The synthesis algorithms usually pursue good visual appearance, while the recognition algorithms requires distinctive biometric information. The mismatch of their goals also reduces the performance of the synthesis-based recognition.

The second family of approaches [41, 38, 35] focused on the *classification* stage and tried to design advanced classifiers to reduce the modality gap between features extracted from photos and sketches. If the inter-modality difference between the extracted features is large, the discriminative power of the classifiers will be reduced.

For more details about inter-modality face recognition, please refer to Chapter 2.3.

1.3 Example-based Image Stylization

Image stylization is a generalization of face sketch synthesis, which focuses on enabling a wide variety of expressive styles for images. The source modality in image stylization is photos, and the target modality is stylized images, such as oil paintings. The stylization problem is categorized into the area of non-photorealistic rendering (NPR) in the graphics community [24, 26]. Much research has been devoted to rendering different styles, e.g., oil painting, watercolor painting and pen-and-ink drawing. However, most of these methods support only limited artistic styles, and for a new style much human knowledge and experiences are required [26].

Example-based image stylization provides an easy way of creating stylized images with a number of styles. The stylized image is synthesized from a real image (e.g., a scene photo) with a given style template (e.g., an oil painting). Formally, given two images as input, a source image A and a template image B^+ whose style is to be simulated, the output A^+ is a synthesized image with the main content in A and the similar style to B^+. This process is also called style transfer, i.e., transferring the style of a template image B^+ to a source image (or video) A. Different from face sketch synthesis, the style template is usually a single painting, because it is difficult to collect many paintings with several consistent styles and the size of the template is usually large enough to provide rich style information. A review of existing approaches will be given in Chapter 2.2.

1.4 Contributions and Summary of Approaches

In addition to improving existing models to achieve better performance, we feel that it is important to look at the problems in a bigger picture. The central of our study is how to match patterns from two different modalities, under complicated real-world scenarios. In particular, we want to expand the frontier of inter-modality synthesis and recognition in the following aspects:

- **Exploring real-world face sketch synthesis**. Real-world face photos have various lightings and poses. Mismatching frequently occurs for inter-modality photo-sketch patch matching using state-of-the-art methods such as multiscale Markov random field [69]. We propose a lighting and pose robust face sketch synthesis algorithm, which includes several components, such as preprocessing for photo-to-photo patch matching, descriptor-based photo-to-sketch patch matching and face shape priors, to reduce mismatching and improve the robustness of the state-of-the-art. Our approach has superior performance on both controlled face databases and internet face photos.

- **Studying style transfer with content and style separation**. A critical problem of previous approaches is that they do not separate the style and content in the style transformation process. Only luminance is transferred from B^+ to A [27, 12, 52], which brings two drawbacks. First, the luminance of two input images may not be in the same dynamic range. To address this problem, a linear mapping that matches the means and variances of the two luminance distributions is often adopted. But usually good correspondences cannot be found for some input images. Second, the content of B^+ may appear in the output images. To break these limitations, we introduce frequency band decomposition, to decompose an image into low-frequency, mid-frequency and high-frequency components. Then the style information in the mid-frequency and high-frequency components are transferred from the template image to the source image. Experiments show that the new approach obtains better synthesis results than previous state-of-the-art.

- **Proposing a new direction based on feature representations for face photo-sketch recognition**. Recent research has focused on transforming photos and sketches into the same modality for matching or developing advanced classification algorithms to reduce the modality gap between features extracted from photos and sketches. We propose a new inter-modality face recognition approach by reducing the modality gap at the feature extraction stage. A new face descriptor based on coupled information-theoretic encoding is used to capture discriminative local face structures and to effectively match photos and sketches. Guided by maximizing the mutual information between photos and sketches in the quantized feature spaces, the coupled encoding is achieved by the proposed coupled information-theoretic projection tree, which is extended to the randomized forest to further boost the performance. We demonstrate the effectiveness of this novel method with the largest face sketch database in the world.

1.5 Thesis Road Map

The remaining part of this thesis is organized as follows. Chapter 2 reviews related work on face sketch synthesis and recognition, image stylization and inter-modality face recognition. Details of our algorithm developed for each application and its experimental results are presented in Chapter 3 – 5. In Chapter 6, we conclude the thesis and discuss future work.

□ **End of chapter.**

Chapter 2

Literature Review

Inter-modality image synthesis and recognition is a field which attracts researchers from both computer vision and computer graphics.

2.1 Related Works in Face Sketch Synthesis

Computer-based face sketch synthesis is different from line drawing generation [36][19]. Line drawings without texture are less expressive than sketches with both contours and shading textures. Popular sketch synthesis methods are mostly example-based, which generates a sketch with rich textures from an input face photo based on a set of training face photo-sketch pairs [60, 43, 69]. These approaches can synthesize sketches of different styles by choosing training sets of different styles.

Existing face sketch synthesis techniques can be divided into global approaches and patch-based approaches (Table 2.1).

Global approaches learn a global mapping from face photos to face sketches. Tang and Wang [60] proposed to apply the eigentransform globally to synthesize a sketch from a photo. The global linear model does not work well if the hair region is included, as the hair styles vary significantly among different people. Lin and Tang [40] proposed coupled bidirectional transform utilizing embedded spaces to estimate the transforms. Liu *et al.* [44] developed a Bayesian tensor inference model for image style transformation. Another global approach proposed by Gao *et al.* [22] was based on the embedded hidden Markov model and the selective ensemble strategy. The common limitation of the global approaches is that it is difficult to learn a global mapping to handle face photo-sketch transformation, because high dimensionality of the input and output images makes the underlying mapping complicated and nonlinear. It is known as curse-of-dimensionality in pattern recognition [25].

To overcome the limitation of global approaches, patch-based approaches were proposed by dividing the image into local patches and learning mappings from photo patches to sketch patches. In this category of methods, Liu *et al.* [43] proposed locally linear embedding (LLE) based reconstruction, and Chang *et al.* [9] proposed sparse representation based reconstruction. Gao *et al.* [21] proposed to jointly training dictionaries for sparse representation based reconstruction. Ji *et al.* [32] investigated and compared different regression models, such as k-nearest-neighbor regression, least squares, ridge regression and lasso [25] in this application. The drawback of these approaches is that the patches are synthesized independently, ignoring their spatial relationships, such that some face structures cannot be well synthesized and the resulting sketches are not smooth enough. In addition, face sketch synthesis through regression, i.e., representing output sketch patches as linear combinations of training sketch patches, causes the blurring effect.

A state-of-the-art approach using a multiscale Markov random field (MRF) model has been proposed recently [69] and achieved good performance under well controlled conditions (i.e. the testing face photo has to be taken in the frontal pose and under a similar lighting condition as the training set). This approach has some attractive features: (1) it can well synthe-

Table 2.1: Existing approaches for face sketch synthesis.

Category		Method
Global approaches		[60], [40], [44], [22]
Patch-based approaches	Synthesizing patches independently	[43], [9], [21], [32]
	Synthesizing patches with neighboring compatibility	[69], [87] and ours

size complicated face structures, such as hair, which are difficult for previous methods [60]; (2) it significantly reduces artifacts, such as the blurring and aliasing effects, which commonly exist in the results of previous methods [60, 43]. Zhou *et al.* [87] proposed Markov weights field, which is similar to MRF but allows linear combinations of patches. It was formulated as a convex quadratic programming problem. Other than selecting local sketch patches from a set of training data, it can synthesize new sketch patches via linear combinations.

2.2 Related Works in Example-based Image Stylization

Non-photorealistic rendering (NPR) is an area of computer graphics that focuses on enabling a wide variety of expressive styles for digital art. In the field of NPR, there has been a rich amount of research on example-based image stylization. The essential problem in this field is to find the mapping from the local pixels/patches of the source image to those of the stylized image. Previous methods assume three different settings: supervised, semi-supervised, and unsupervised (see Table 2.2), which are terms borrowed from the area of machine learning [25].

In the supervised setting, the ground-truth image B corresponding to B^+ is given. Hertzmann et al. proposed the framework of image analogies to estimate the mapping, using the source image B to the stylized image B^+ [27]. Image stylization for highly specialized problems has also been attempted for faces [60, 69, 82], using very large sets of training data.

In the semi-supervised setting, some parts of A^+ are available as training data, and thus ground-truth, i.e., some corresponding parts between A and A^+ are known. Cheng et al. proposed to use a semi-supervised component to exploit this setting. The similarities between the source patches are utilized to propagate information from the sources patches with stylized counterpart to the source patches without stylized counterpart.

In the unsupervised setting, the ground-truth source image A is not given. In real-world problems, people usually have a painting without the corresponding real image. Therefore, the unsupervised setting is more user friendly, but more difficult as well. To deal with the difficulty of lacking ground-truth B, Rosales et al used a finite set of patch transformations [52]. They formulated the problem as inferring the latent variables. Wang et al.'s method requires the user-specified blocks in B^+ as sample textures, and then the textures are applied on segmented image A [64]. Our method also belongs to this category. Comparing to them, our method requires the least information: neither user input [64] nor assuming a set of transformations [52]. In addition, we utilize the full image of B^+ instead of a very small subset of B^+ [64]. The required supervision information of the above-mentioned approaches is summarized in Table 2.2.

As mentioned in Chapter 1.4, our approach separates the style and content using frequency band decomposition, and thus avoids the severe problems of applying the existing approaches in example-based image stylization [27, 12, 52, 64] to real-world applications. There were some existing methods exploring content and style decomposition. Tenenbaum and Freeman [62] separated style and content with bilinear models. Drori et al.'s locally linear model [16] is an example-based synthesis technique

Table 2.2: Comparison of the required supervision information in different example-based image stylization approaches. The input source image and the output image of all the methods are denoted by A and A^+. B is the ground-truth image corresponding to B^+.

Approach	Supervision information	Category
[27]	well aligned B and B^+	Supervised
[12]	available parts of A^+	Semi-supervised
[64]	user-selected parts of B^+	Unsupervised
[52]	B^+ and a set of transformations	Unsupervised
ours	B^+	Unsupervised

that extrapolates novel styles for a given input image. However, these methods used a set of images with the same style to learn content and style decomposition. It is not trivial to apply the solutions to existing example-based image stylization approaches.

Constrained texture synthesis [51, 88] is a topic related to example-based image stylization. However, all existing methods for constrained texture synthesis were supervised. In addition, the mismatching problem between the characteristics of the source image and the template image does not need to be considered in this application.

2.3 Related Works in Face Sketch Recognition

To match a face sketch drawn by an artist to one of many face photos in the database, great progress has been made in two directions.

The first family of approaches [60, 43, 69] focused on the *preprocessing* stage and synthesized a pseudo-photo from the query sketch or pseudo-sketches from the gallery photos to transform inter-modality face recognition into intra-modality face recognition. We introduced the existing face sketch synthesis approaches in Chapter 2.1. The synthesis-based approaches have also been used to solve other problems. In infrared face recognition, Chen *et al.* first transformed an infrared image to a normal optical image and then performed the matching [10]. In face recognition across ages, Suo *et al.* and Park *et al.* first transformed images of different ages to the same age and then performed the recognition [57, 48]. Javed *et al.* proposed to learn the brightness transform functions in order to match objects observed in different camera views [31]. However, face photo/sketch synthesis is actually a harder problem than

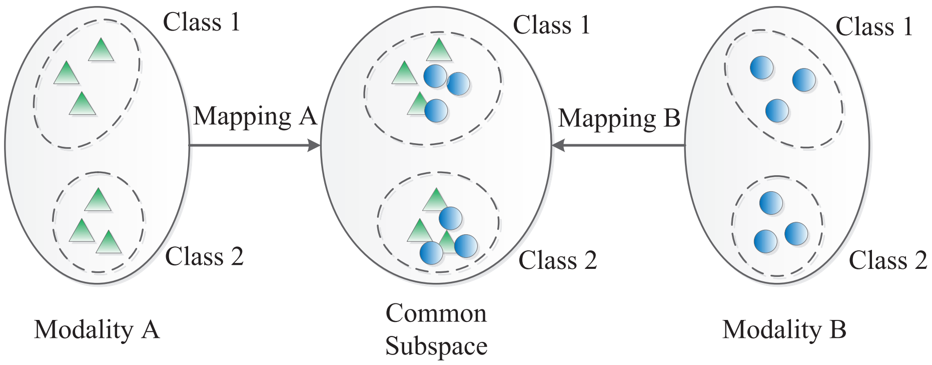
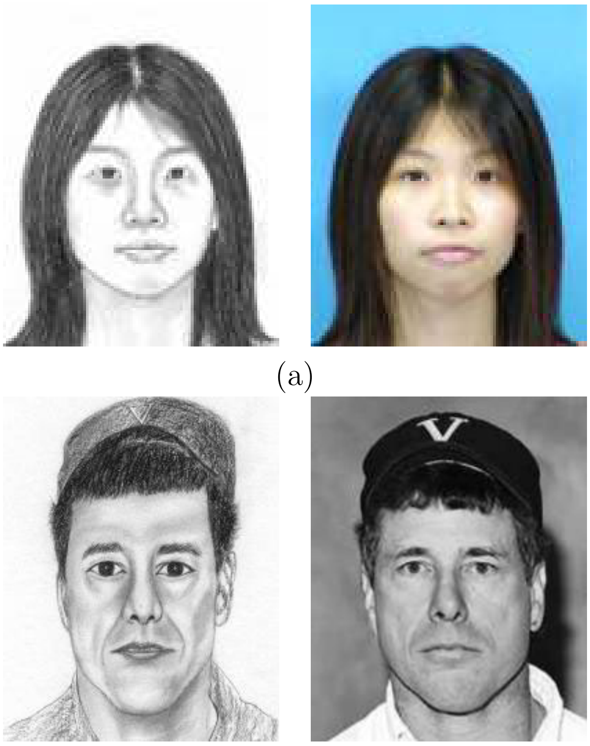


Figure 2.1: Illustration of the common subspace approaches for inter-modality face recognition.

recognition. Imperfect synthesis results significantly degrade the recognition performance. The synthesis algorithms usually pursue good visual appearance, while the recognition algorithms requires distinctive biometric information. The mismatch of their goals also reduces the performance of the synthesis-based recognition.

The second family of approaches [41, 38, 35] focused on the *classification* stage and tried to design advanced classifiers to reduce the modality gap between features extracted from photos and sketches. Several methods have been proposed to mapped feature vectors from two modalities into a common discriminative subspace (as illustrated in Fig. 2.1). The mappings for photo feature vectors and for sketch feature vectors are different, to deal with inter-modality differences. Using these methods, the classification can be performed in the common discriminative subspace. Among them, canonical correlation analysis (CCA) is a classical unsupervised model introduced by Hotelling for correlating linear relationships between two sets of vectors [28]. Lin and Tang [41] incorporated label information utilizing the within-class scatter matrix and between-class scatter matrix. Lei and Li [38] proposed a more computationally efficient approach called coupled spectral regression for learning projections to map data from two modalities into a common subspace. Sharma *et al.* introduced Partial Least Squares [53] for this task. Klare *et al.* [35] proposed local feature-based discriminant analysis (LFDA). They used multiple projections to extract a discriminative representation from partitioned vectors of local binary patterns (LBP) [1] and dense scale-invariant feature transform (SIFT) [45] features. Bhatt *et al.* [5] extracted multiscale extended uniform circular local binary patterns features and used a genetic optimization based approach to find the optimal weights for computing the χ^2 distances between features. for matching sketches with digital face images.

Feature extraction and classification are two major compo-



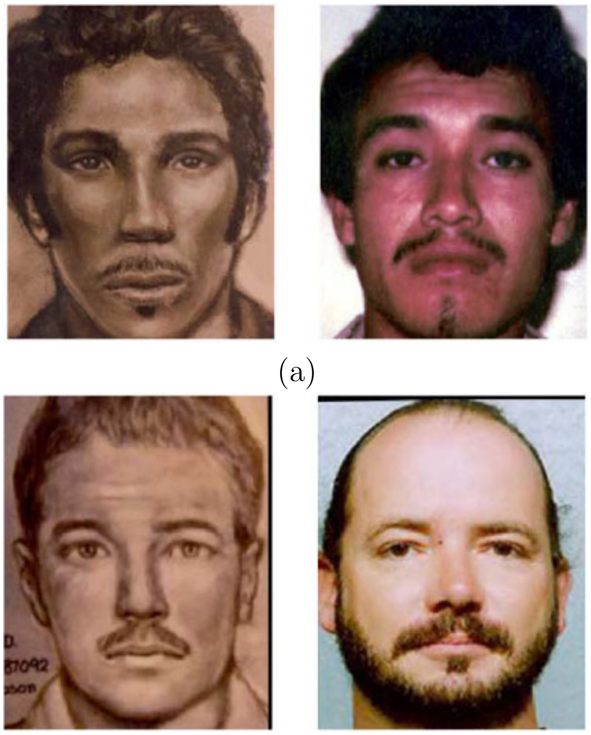
(a)

(b)

Figure 2.2: Example of viewed sketches and corresponding face photos. (a) A sketch and the corresponding photo without shape exaggeration from CUHK face sketch database. (b) A sketch and the corresponding photo with shape exaggeration from CUHK face sketch FERET database.

nents in face recognition. Although the classification-based approaches utilized different features in their papers, these approaches are independent of the features, i.e., they can be applied to any kind of facial features. If the inter-modality difference between the extracted features is large, the discriminative power of the classifiers will be reduced.

In addition to computer-based methods, Zhang *et al.* studied



(a)

(b)

Figure 2.3: Example of forensic sketches and corresponding face photos. (a) A sketch of good quality and the corresponding photo. (b) A sketch of poor quality and the corresponding photo. Both are adopted from [35].

human perceptions on face sketch recognition [85, 84].

Viewed Sketches v.s. Forensic Sketches. Two different types of face sketches has been introduced into face sketch recognition: viewed sketches (see Fig. 2.2) and forensic sketches (see Fig. 2.3). A viewed sketch is drawn while the artist views a photo of the person, while a forensic sketch is drawn by interviewing a witness to gain a description of the suspect. In this research, we adopt viewed sketch databases for testing synthesis and recognition algorithms, because (1) in some applications, such as entertainment, view sketches are used instead of forensic sketches; (2) the quality of forensic sketches are usually uncontrollable (e.g, some sketches may have as good qualities as the one Fig. 2.3 (b) and some may have as poor qualities as the one in Fig. 2.3 (b)), due to the approximate, possibly incomplete, descriptions provided by the eye-witness. So it is difficult to identify which factors affect the performances of algorithms; (3) forensic sketch databases are much smaller than viewed sketch databases.

□ **End of chapter.**

Chapter 3

Lighting and Pose Robust Sketch Synthesis

Automatic face sketch synthesis has drawn a great deal of attention in recent years [60, 61, 43, 22, 69] due to its applications in law enforcement and digital entertainment. For example, in law enforcement, it is useful to develop a system to search photos from police mug-shot databases using a sketch drawing when the photo of a suspect is not available. By transferring face photos to sketches, inter-modality face recognition is made possible [61]. In the movie industry, artists can save a great amount of time on drawing cartoon faces with the assistance of an automatic sketch synthesis system. Such a system also provides an easy tool for people to personalize their identities in the digital

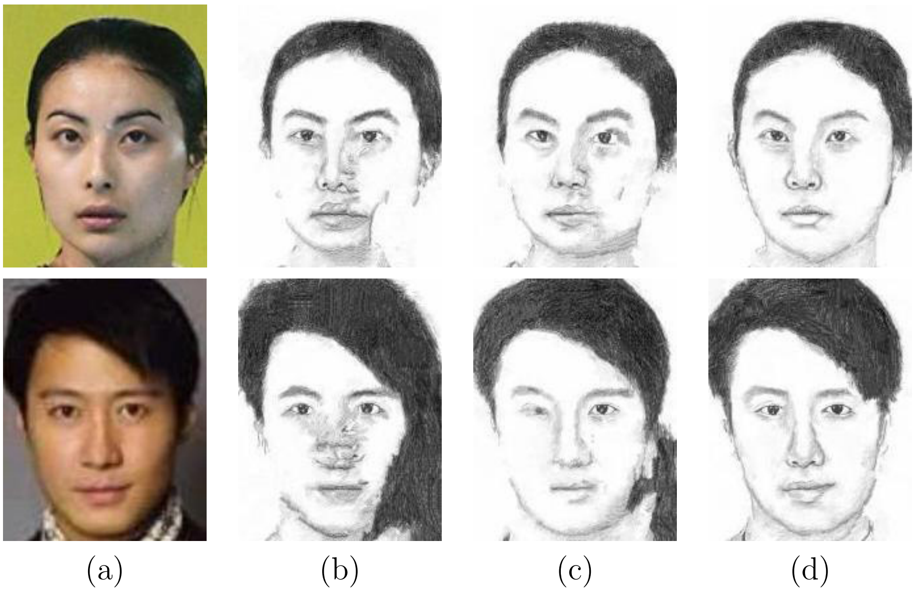


Figure 3.1: Examples of synthesized sketches from web face photos. (a) Test photos; (b) Sketches synthesized by [69]; (c) Sketches synthesized by [69] with luminance remapping [27]; (d) Sketches synthesized by our method. Note that luminance remapping refers to zero-mean unit-variance normalization of the luminance channel of all photos in our implementation. This simple technique was found to be better than non-smooth mappings in image style transformation, such as histogram matching/equalization [27]. The results are best viewed on screen.

world, such as through the MSN avatar.

Although great progress has been made in recent years (see 2.1), previous methods only work under well controlled conditions and often fail when there are variations of lighting and pose. If the testing face photo is taken in a different pose or under a different lighting condition (even if the lighting change is not dramatic) from the training set, the problem could be challenging. Some examples are shown in Fig. 3.1. Due to the variations of lighting and pose, on the synthesized sketches by [69] some face structures are lost, some dark regions are synthesized as hair, and there are a great deal of distortions and artifacts. This is also a serious problem not addressed by other approaches [60][43][22]. It limits their applications to real-world problems.

In face recognition studies, some preprocessing techniques such as histogram equalization, and features such as Local Binary Patterns (LBP) [1], were used to effectively recognize face photos under lighting variations. In the area of nonphotorealistic rendering, luminance remapping was introduced to normalize lighting variations [27]. However, experiments show that simply borrowing these techniques is not effective in face sketch synthesis. See examples in Fig. 3.1.

In this chapter, we address this challenge: *given a limited set of photo-sketch pairs with frontal faces and normal lighting conditions, how to synthesize face sketches for photos with faces in different poses (in the range of* $[-45^o + 45^o]$*) and under different lighting conditions.* We adopt the multiscale MRF model whose effectiveness has been shown in face sketch synthesis [69] and many low-level vision problems [18]. In order to achieve

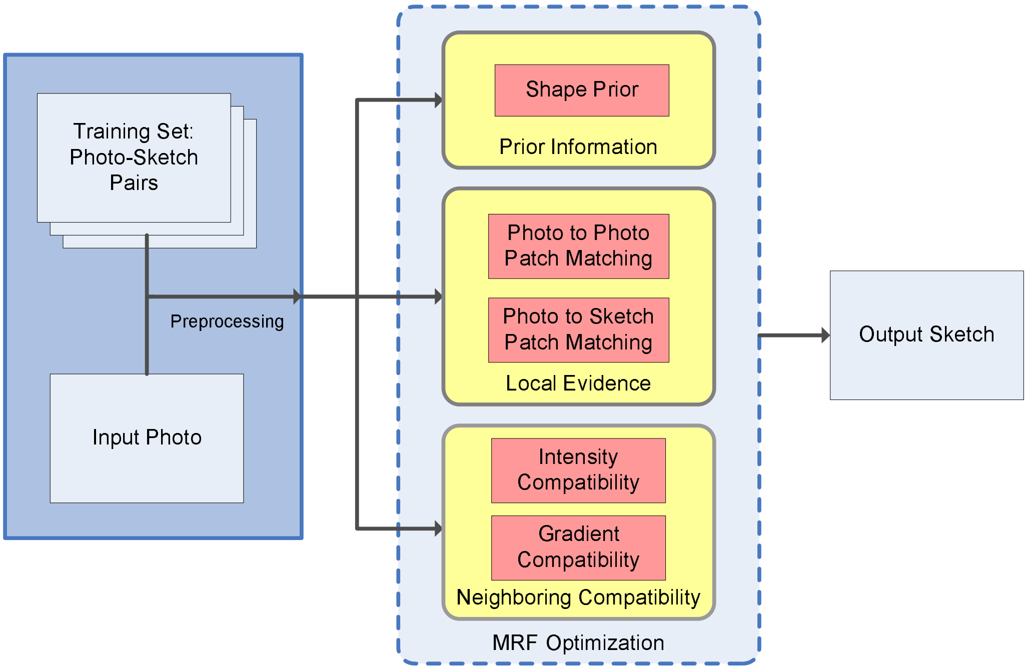


Figure 3.2: Illustration of our framework.

the robustness to variations of lighting and pose, some important improvements are made in the design of the MRF model as summarized in Fig. 3.2. Firstly, a new term of shape priors specific to face components are introduced in our MRF model. It effectively reduces distortions and artifacts and restores lost structures as shown in Fig. 3.1. Secondly, patch descriptors and metrics which are more robust to lighting variations are used to find candidates of sketch patches given a photo patch. In addition to photo-to-photo patch matching, which was commonly used in previous approaches [43][69], our "local evidence" term also includes photo-to-sketch patch matching, which improves the matching accuracy with the existence of lighting and pose variations. Lastly, a smoothing term involving both intensity compatibility and gradient compatibility is used to match neighboring sketch patches on the MRF network more effectively.

The effectiveness of our approach is evaluated on the CUHK face sketch database which includes face photos with different lightings and poses. We also test on face photos of Chinese celebrities downloaded from the web. The experimental results show that our approach significantly improves the performance of face sketch synthesis compared with the state-of-the-art method [69] when the testing photo includes lighting or pose variations.

3.1 The Algorithm

In this section, we present our algorithm for face sketch synthesis. For ease of understanding, we use the single-scale MRF model in the presentation, instead of the two-scale MRF model in our implementation[1].

[1] We do find that the two-scale MRF model performs better. The details of multiscale MRF can be found in [69]. However, it is not the focus of this chapter.

3.1.1 Overview of the Method

A graphical illustration of the MRF model is shown in Fig. 3.3. A test photo is divided into N overlapping patches with equal spacing. Then a MRF network is built. Each test photo patch x_i^p is a node on the network. Our goal is to estimate the status $y_i = (y_i^p, y_i^s)$, which is a pair of photo patch and sketch patch found in the training set, for each x_i^p. Photos and sketches in the training set are geometrically aligned. y_i^p is a photo patch and y_i^s is its corresponding sketch patch. If patches i and j are neighbors on the test photo, nodes y_i and y_j are connected by an edge, which enforces a compatibility constraint. The sketch of the test photo is synthesized by stitching the estimated sketch patches $\{y_i^s\}$. Based on the MRF model, our energy function is defined in the following form,

$$E(\{y_i\}_{i=1}^N) = \sum_{i=1}^N E_L(x_i^p, y_i) + \sum_{i=1}^N E_{Pi}(y_i) + \sum_{(i,j)\in\Xi} E_C(y_i^s, y_j^s), \tag{3.1}$$

where Ξ is the set of pairs of neighboring patches, $E_L(x_i^p, y_i)$ is the local evidence function (Subsection 3.1.2), $E_{Pi}(y_i)$ is the shape prior function (Subsection 3.1.3), and $E_C(y_i^s, y_j^s)$ is the neighboring compatibility function (Subsection 3.1.4). The shape prior function is specific to face components, which means that different location indicated by i has different E_{Pi}. The above MRF optimization problem can be solved by belief propagation [18] [76].

A MRF model was also used in [69], however, with several major differences with ours. It has no shape prior function which is effective in sketch synthesis. Its local evidence function only computes the sum of the squared differences (SSD) between x_i^p and y_i^p and is sensitive to lighting variations. Our local evidence function uses new patch descriptors which are more robust to lighting variations. Our method includes not only photo-to-photo patch matching (between x_i^p and y_i^p) but also photo-to-sketch patch matching (between x_i^p and y_i^s) to improve the robustness. The neighboring compatibility function in [69] is to minimize SSD between neighboring estimated sketch patches (y_i^s and y_j^s) in their overlapping region, while ours also minimizes the difference of gradient distributions. Details will be explained in the following subsections.

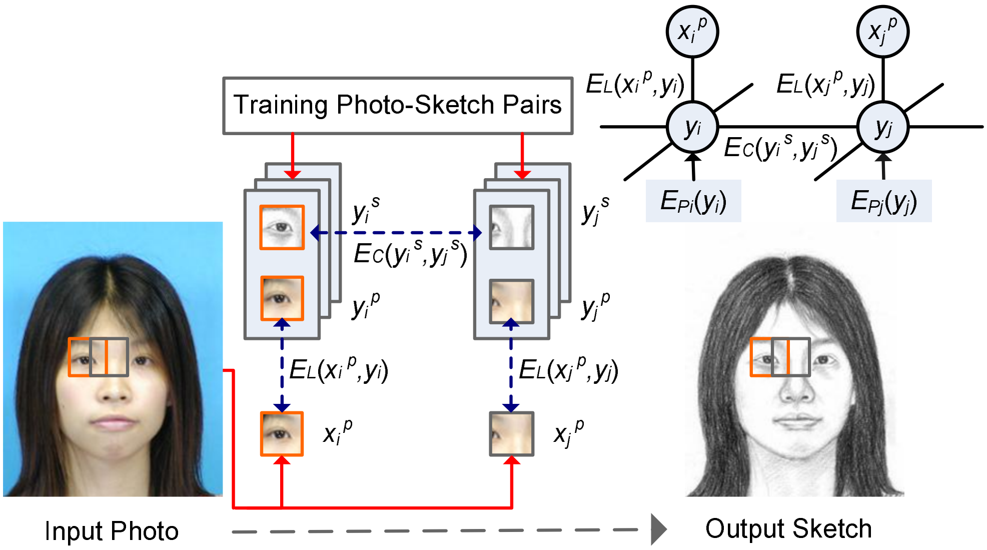


Figure 3.3: Illustration of the MRF model for face sketch synthesis.

3.1.2 Local Evidence

The goal of the local evidence function is to find a sketch patch y_i^s in the training set best matching the photo patch x_i^p in test. However, since photos and sketches are in different modalities, it is unreliable to directly match them. So the training photo patch y_i^p corresponding to a training sketch patch y_i^s is involved. It is assumed that if y_i^p is similar to x_i^p, it is likely for y_i^s to be a good estimation of the sketch patch to be synthesized. We propose to match a testing photo patch with training photo patches and also with training sketch patches simultaneously, i.e. we define the local evidence function as the weighted sum of

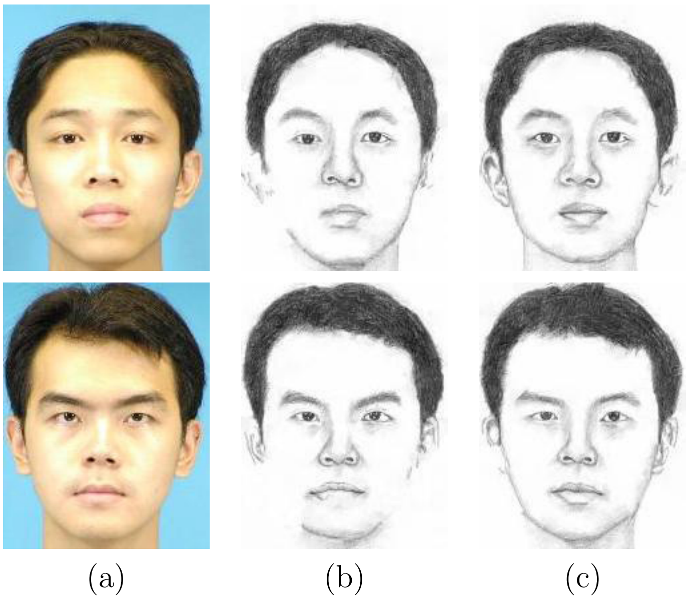
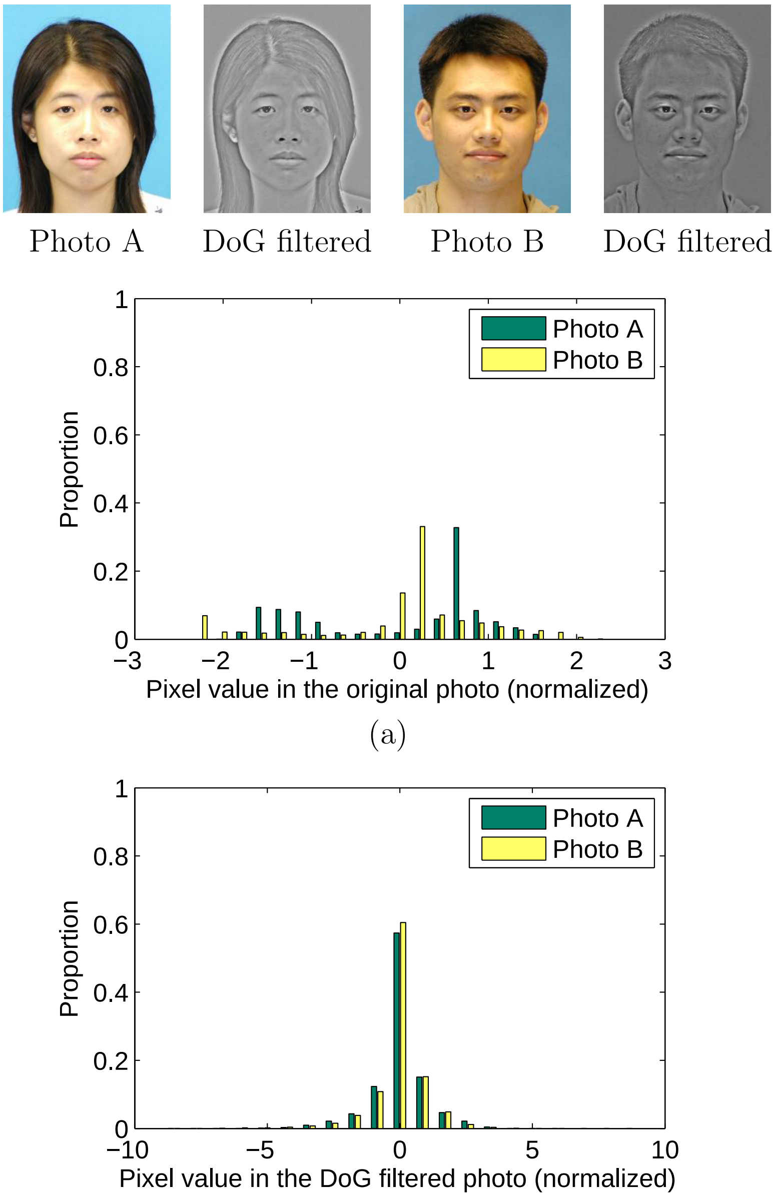


Figure 3.4: Compare the results with/without DoG filtering under a normal lighting condition. (a) Test photos which are under the same lighting as the training set. (b)Synthesized sketch by the method in [69] without DoG filtering. (c) Synthesized sketches by our method with DoG filtering. To evaluate the effectiveness of DoG filtering, other parts, such as shape priors and photo-to-sketch patch matching, in our framework are not used in these examples.

squared intra-modality distance d_{L1}^2 and squared inter-modality distance d_{L2}^2,

$$E_L(x_i^p, y_i) = d_{L1}^2(x_i^p, y_i^p) + \lambda_{L2} d_{L2}^2(x_i^p, y_i^s), \tag{3.2}$$

where λ_{L2} is the weight to balance different terms in the energy function E and it is chosen as 2 in our experiments.



(a)

(b)

Figure 3.5: Examples of DoG filtering with $(\sigma_0, \sigma_1) = (0, 4)$. Photo A is from the training set taken under the normal lighting condition, and Photo B is from the testing set taken under a different lighting condition. The pixel values of DoG filtered photos are scaled to $[0, 1]$ for visualization. (a) Histograms of pixel values of the two photos after luminance remapping. They do not match well. (b) Histograms of pixel values of the two photos after DoG filtering and normalization. They match well.

Photo-to-Photo Patch Matching

A straightforward choice of E_L is the Euclidean distance between x_i^p and y_i^p as used in [69]. However, it does not perform well when the lighting condition varies. Noticing that most of the sketch contours correspond to edges in the photo, we use a difference-of-Gaussians (DoG) filter to process each photo, i.e. convolving each photo with the difference of two Gaussian kernels with standard deviations σ_0 and σ_1, and normalize all pixel values to zero-mean and unit-variance. In our experiments, we find that $(\sigma_0, \sigma_1) = (0, 4)$ or $(1, 4)$ performs the best. DoG filtering has two advantages. First, it can detect and enhance the edges, and thus the synthesized sketch has better facial details. As shown in Fig. 3.4, even for normal lighting, the DoG filtering can improve facial details. Second, subtracting low-frequency component reduces the effect of lighting variations, e.g. shading effects. The example in Fig. 3.6 shows that DoG filtering improves synthesized facial details, especially on the nose and the eyebrows, when there are lighting variations. Luminance remapping [27], which normalizes the distribution of pixel values in an image to zero-mean and unit-variance, is commonly used for lighting normalization. However, its improvement is limited in this application. An example is shown in Fig. 3.5. After luminance remapping, the distributions of pixel values in two photos taken under different lighting conditions still do not match. On the contrary, their distributions after DoG filtering match well. In some cases, photo-to-photo patch matching is not enough and the mismatching problem, such as the hair and profile regions shown in Fig. 3.6 (c), still exists. Thus, photo-to-sketch patch matching is introduced.

Photo-to-Sketch Patch Matching

The intra-modality distance between photo patches does not always work for selecting a good sketch patch. Similar photo patches under the Euclidean distance may correspond to very different sketch patches. Interestingly, people have the ability to directly match photos with sketches. Inspired by this, we propose to use inter-modality distance between testing photo patches and training sketch patches to enhance the selection ability. As the visual appearances of photo and sketch patches are different, it is difficult to directly match them. However, there exists some similarity of gradient orientations between a photo and its sketch. We choose to use the dense SIFT descrip-

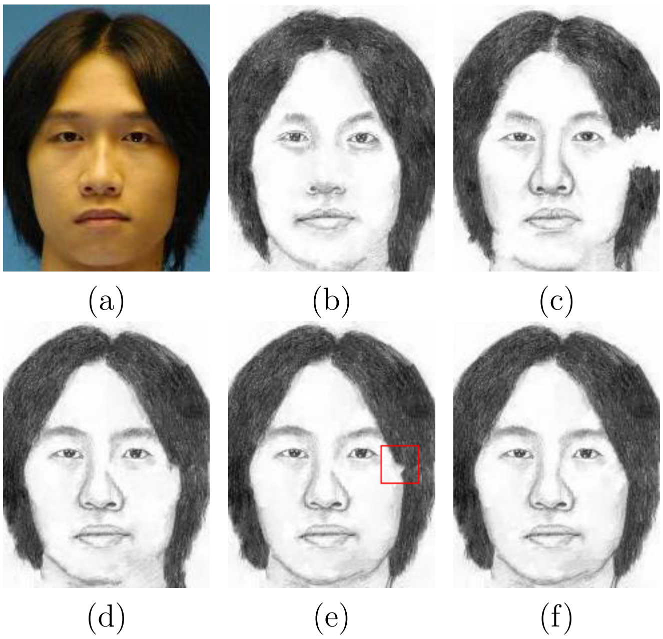


Figure 3.6: Sequential illustration of the roles of each part in our framework. (a) Test photo under a different lighting condition than the training set; (b) Sketch by the method in [69] with luminance remapping as preprocessing [27]; (c) Sketch by our method with P2P+IC; (d) Sketch by our method with P2P+P2S+IC; (e) Sketch by our method with P2P+P2S+prior+IC; (f) Sketch by our method with P2P+P2S+prior+IC+GC. P2P, P2S, prior, IC and GC represent photo-to-photo patch matching, photo-to-sketch patch matching, shape priors, intensity compatibility and gradient compatibility, respectively. The results are best viewed on screen.

tor [45] from the family of histogram-of-orientations descriptors. Our strategy is to assign each patch a dense SIFT descriptor, and use the Euclidean distance between SIFT descriptors of photo patches and sketch patches as the inter-modality distance. To capture structures in large scales, we extract the descriptors in larger regions than patches. For each patch, we extract a region of size 36 × 36 centered at the center of the patch (the size of patch is 10×10), and divide it into 4×4 spatial bins of the same size. 8 orientation bins are evenly spaced over 0°-360°. The vote of a pixel to the histogram is weighted by its gradient magnitude and a Gaussian window with parameter $\sigma = 6$ centered at the center of the patch. So the descriptor is 128 dimensional. The descriptor is normalized by its $L2{-}norm$, clipped by a threshold 0.2 and renormalized as reported in [45]. The synthesis result with photo-to-sketch patch matching is shown in Fig. 3.6 (d). It restores the hair and partial profile lost in Fig. 3.6 (c).

3.1.3 Shape Prior

Face images are a special class of images with well regularized structures. Thus shape priors on different face components can be used to effectively improve the synthesis performance. The loss of some face structures, especially the face profile, is a common problem for the patch-based sketch synthesis methods without referring to global structures. When this happens, the contours of some face components are replaced by blank regions. This problem becomes much more serious when there are variations of lighting and pose. See examples in Fig. 3.1. However, it can be effectively alleviated by using the prior information on different face components to guide the selection of sketch patches. In our approach, a state-of-the-art face alignment algorithm [39] is first utilized to detect some predefined landmarks on both the training sketches and the testing photo. The chosen landmarks locate in regions where loss of structures often happens, especially on the face profile. Shape priors are imposed to these regions but not in other regions. If a landmark f falls into patch i on the test photo, a prior distribution is computed via kernel density estimation,

$$E_{Pi}(y_i) = \lambda_P \ln \left[\frac{1}{\sqrt{2\pi} N_t} \sum_{k=1}^{N_t} \exp \left(- \frac{(\beta(y_i^s) - \beta_{k,f})^2}{h_f^2} \right) \right]. \quad (3.3)$$

N_t is the number of sketches in the training set. $\beta(y_i^s)$ is some statistic on the sketch patch y_i^s. $\beta_{k,f}$ is the statistic on a sketch patch centered at landmark f in sketch image k. h_f is the bandwidth of landmark f and is set as three times of the standard deviation of $\{\beta_{k,f}\}$. The weight $\lambda_P = 0.01$ is to normalize the metric scale of the shape prior term and the performance of our algorithm is robust to λ_P in a fairly large range.

We test several kinds of patch statistics, such as mean gradient magnitude, variance of pixel values, proportion of edge pixels, and find that mean gradient magnitude performs the best and it is chosen as $\beta(\cdot)$. It can well solve the problem of losing structures, as shown in Fig. 3.6 (e).

3.1.4 Neighboring Compatibility

The goal of the neighboring compatibility function is to make the neighboring estimated sketch patches smooth and thus to reduce the artifacts on the synthesized sketch. In our model it is defined as

$$E_C(y_i, y_j) = \lambda_{IC} d_{IC}^2(y_i^s, y_j^s) + \lambda_{GC} d_{GC}^2(y_i^s, y_j^s), \tag{3.4}$$

where the intensity compatibility term d_{IC}^2 is the SSD in the overlapping region between two neighboring sketch patches y_i^s and y_j^s, and the gradient compatibility term d_{GC}^2 is the squared Euclidean distance between the dense SIFT descriptors of y_i^s and y_j^s. The intensity compatibility term is for the smoothness of the output sketch. However, only using this term tends to lose some face structures since two blank regions in neighbors have high intensity compatibility. Thus, we further add the gradient compatibility constraint, which requires that the neighboring patches have similar gradient orientations. The use of gradient compatibility can further alleviate the structural loss, an example of which is given in Fig.s 3.6 (e) and (f) (the region in the red box). We set the weights $\lambda_{IC} = 1$ and $\lambda_{GC} = 0.1$.

3.1.5 Implementation Details

All the photos and sketches are translated, rotated, and scaled such that the two eye centers of all the face images are at fixed position. We crop the images to 250×200 and the two eye center positions are $(75, 125)$ and $(125, 125)$. All color images are converted to grayscale images for sketch synthesis.

- **Preprocessing on Test Photos**. Empirically, when lighting is near frontal, our algorithm can work well without the preprocessing step. However, for side light, we need to use Contrast Limited Adaptive Histogram Equalization (CLAHE) [50] for preprocessing.[2] We use the setting that the desired histogram shape is Rayleigh distribution (parameter $\alpha = 0.7$).
- **Candidate Selection**. In order to save computational

[2] CLAHE improves the method in [69] little and deteriorates its performance in some cases. So we choose to report their results without the preprocessing.

cost, a step of candidate selection as suggested in [18] is used before optimizing the MRF model. For each test photo patch x_i^p, top K $(K = 20)$ photo-sketch pairs with the smallest energy of $E_L(x_i^p, y_i) + E_{Pi}(y_i)$ are selected from the training set as candidates. In order to take the advantage of face structures, candidates are searched within a 25×25 local region around patch i instead of in the entire images. The final estimation y_i on node i is selected as one of the K candiates through joint optimization of all the nodes on the MRF network.

- **Two-scale MRF**. We use two-scale MRF with the same setting as in [69]. Patch sizes at the two layers are 10×10 and 20×20, respectively. MAP estimate is used in the belief propagation algorithm [18].

- **Stitching Sketch Patches**. To avoid blurring effect, we use a minimum error boundary cut between two overlapping patches on their overlapped pixels as what is usually done for texture synthesis [17].

3.1.6 Acceleration

The bottleneck of accelerating the algorithm is patch candidate selection (see Subsection 3.1.5). In order to find the photo-sketch patch pairs which best match the input photo patch x_i^p, the distances $d_{L1}^2(x_i^p, y_i^p)$ and $d_{L2}^2(x_i^p, y_i^s)$ in the local evidence term Eqn. (3.2) have to be computed for all possible patch pairs y_i in the training database.

- **Integral histogram for photo-to-photo patch distances**. The photo-to-photo patch distance

  $$d_{L1}^2(x_i^p, y_i^p) = \sum_{s \in R} (I_t^p(s) - I_0^p(s + s_0))^2,$$

  where I_t^p and I_0^p are DoG filtered test and training photos. $s \in R$ is the spatial location of pixels in test patch x_i^p and $s_0 \in [-12, 12] \times [-12, 12]$ is the shifting amount of training patches. Fixing s_0, computing the distances between all x_i^p and their corresponding y_i^p can be speeded up using integral computation [63]. We first compute an integral image of the squared difference between I_t^p and shifted I_0^p, and then compute statistics over the rectangle regions over the image.

- **Compression of SIFT descriptors for photo-to-sketch patch distances**. We use linear projections to reduce the dimensionality of SIFT descriptors. Coupled linear projections are trained using regularized Canonical Correlation Analysis [83] for training photo-sketch patch pairs. Then we can use the photo projection vector and sketch projection vector to compress photo patches and sketch patches, respectively. Note that the projection vectors and the compressed SIFT descriptors of the training sketch patches can be computed offline and stored. As shown in Fig. 3.7, we compare the results with and without SIFT descriptor compression. We find that the visual quality of synthesized sketches does not change when reducing 128-dimensional SIFT to 64-dimensional.

It takes about 10 seconds running our optimized C++ implementation, while 90 seconds running a naive C++ implementation to synthesize a sketch on a computer with 3.20 GHz CPU.

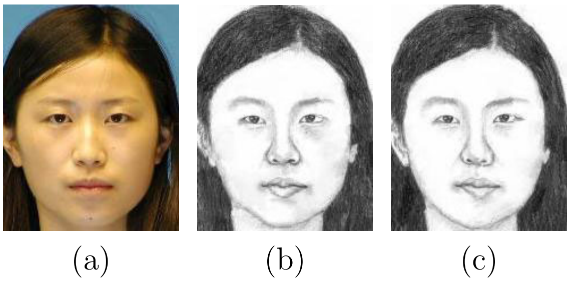


Figure 3.7: Representative results on the baseline set. (a) Test photo; (b) Sketch synthesized without SIFT descriptor compression; (c) Sketch synthesized with 50% SIFT descriptor compression. The results are best viewed on screen.

3.2 Experimental Results

We conduct experiments on the CUHK database [69] commonly used in face sketch synthesis research, and a set of celebrity face photos from the web. In all the experiments, 88 persons from the CUHK database are selected for training, and each person has a face photo in a frontal pose under a normal lighting condition, and a sketch drawn by an artist while viewing this photo. In the first experiment, 100 other persons are selected for testing. We have three data sets: the baseline set, the lighting variation set, and the pose variation set. The baseline set includes 100 face photos taken in a frontal pose under the same lighting condition as the training set. The lighting variation data set includes three photos with faces in a frontal pose with three dif-

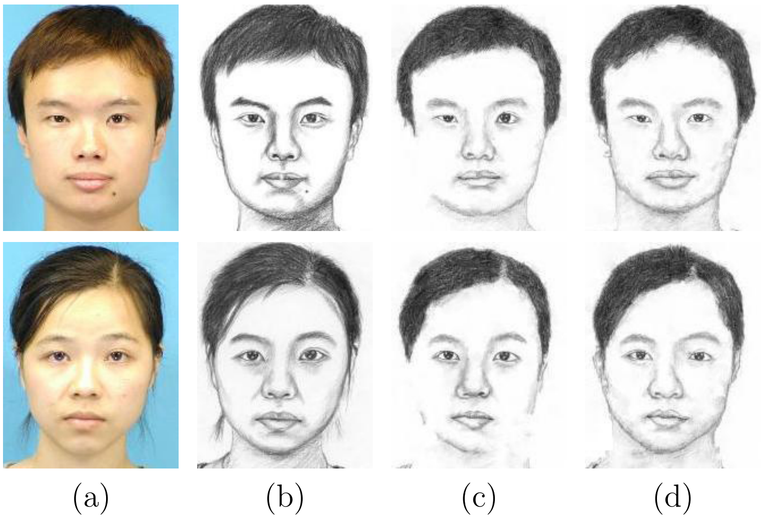


Figure 3.8: Representative results on the baseline set. (a) Test photo; (b) Sketch drawn by the artist while viewing the normal lighting photo; (c) Sketch by the method in [69]; (d) Sketch by our method. The results are best viewed on screen.

ferent lightings (dark frontal/dark left/dark right) for each person. And the pose variation set includes two photos with faces in left and right poses (with 45 degrees) under a normal lighting condition for each person. In the second experiment, some face photos of Chinese celebrities with uncontrolled lighting conditions and poses are downloaded from the web.[3] All photos are with a neutral expression. Parameters are fixed throughout the experiments. Due to the thesis length, only a limited number of examples are shown in this paper.

[3] The CUHK database cannot be used as a training set for photos of people from other ethnic groups, partially due to the human perception.

Table 3.1: Rank-1 and Rank-10 recognition rates using whitened PCA [74]. The whitened PCA model is trained on the 100 sketches drawn by the artist while viewing the baseline set. It performs better than standard PCA without whitening on all the tasks. The reduced number of dimension is 99, and it is the best for all the tasks.

Rank-1 recognition rates					
Testing set	[69]	[69] with LBP	[69] with HE	[69] with LR	Ours
Baseline	96%	-	-	-	99%
Front Light	58%	58%	70%	75%	84%
Side Lights	23.5%	25.5%	38%	41.5%	71%

Rank-10 recognition rates					
Testing set	[69]	[69] with LBP	[69] with HE	[69] with LR	Ours
Baseline	100%	-	-	-	100%
Front Light	87%	87%	95%	96%	96%
Side Lights	56%	75.5%	80.5%	78.5%	87.5%

3.2.1 Lighting and Pose Variations

We first investigate the effect of lighting and pose variations separately on the CUHK database. A preliminary test is on the baseline set. Our algorithm performs as well as the method in [69]. On some photos, our algorithm can produce even better face sketches as shown in Fig. 3.8. To give a quantitative evaluation of the performance, we test the rank-1 and rank-10 recognition rates when a query sketch synthesized from a test photo is used to match the sketches drawn by the artist. The

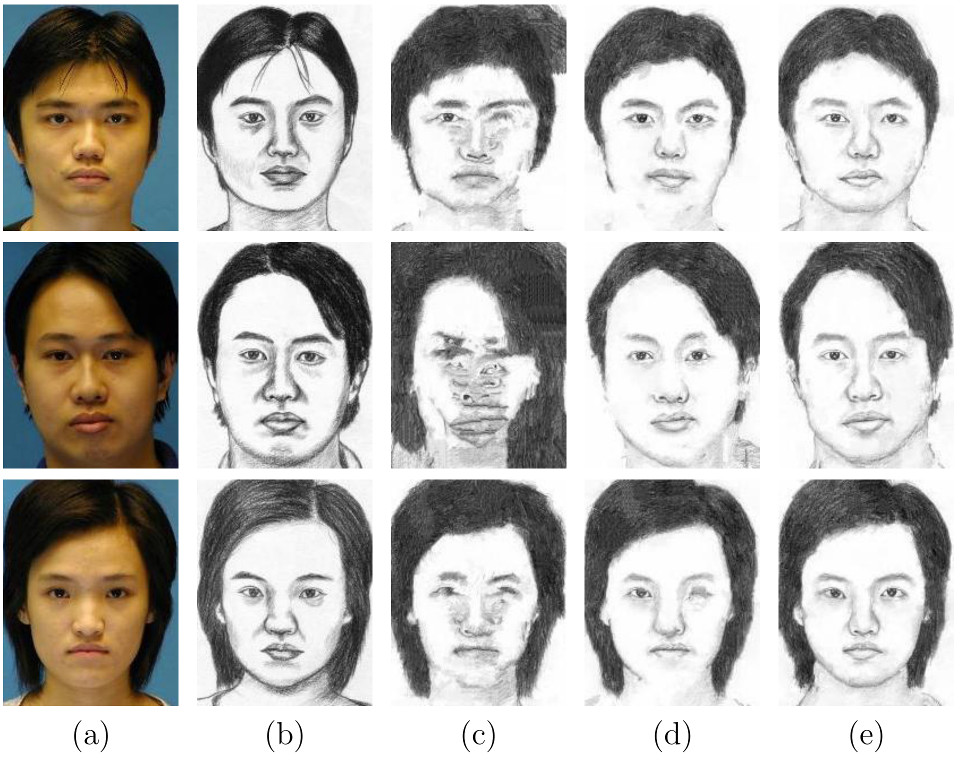


Figure 3.9: Representative results on photos under the dark frontal lighting. (a) Test photo; (b) Sketch drawn by the artist while viewing a normal lighting photo; (c) Sketch by the method in [69]; (d) Sketch by the method in [69] with luminance remapping [27]; (e) Sketch by our method. The results are best viewed on screen.

results are shown in Table 3.1.[4] Our algorithm slightly beats the previous method by 3%.

Lighting

Although the previous method performs well on the normal lighting set, their performance degrades dramatically when the

[4] Recognition rates cannot completely reflect the viual quality of synthesized sketches. It is used as an indirect measurement to evaluate the performance of sketch synthesis since no other proper quantitative evaluation methods are available.

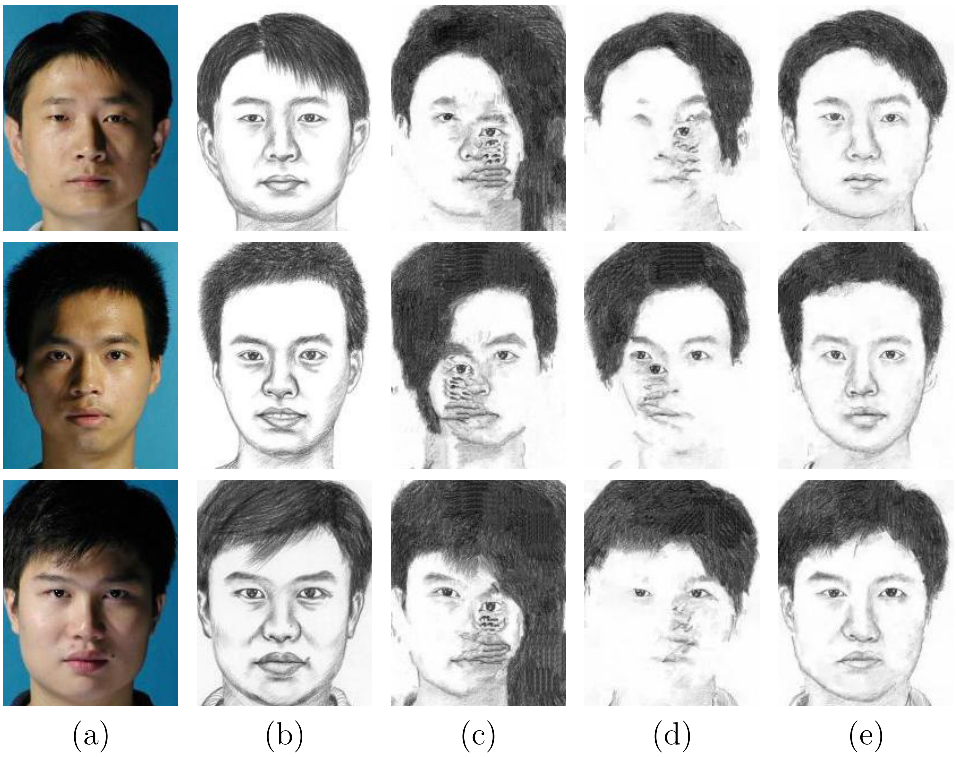


Figure 3.10: Representative results of photos under dark side lightings. The notations (a)–(e) are the same as Fig. 3.9. The results are best viewed on screen.

lighting changes. Our method performs consistently well under different lighting conditions. To make a fair comparison, we also report the results of [69] with several popular illumination normalization methods, including histogram equalization (HE) and luminance remapping (LR) [27], and with LBP [1], an illumination invariant feature.

On the recognition rate, our method beats all the others, as shown in Table 3.1. The method in [69] performs very poorly without any preprocessing. LR and HE improve the method in [69], but LBP improves little. LR performs better than HE and LBP. As hair and background are included in face photos, previous illumination normalization methods, such as HE, do not perform well. By converting a patch to its LBP feature, information to distinguish different components, which is important for sketch synthesis, may be lost and thus mismatching often occurs. In addition, we find that dark side lighting conditions are more difficult than dark frontal lighting, and under dark side lightings, our method beats all the others by a large amount on the rank-1 recognition rate.

On the visual quality, LR improves the method in [69], but as shown in Fig.s 3.9 and 3.10, the facial details and profile are still

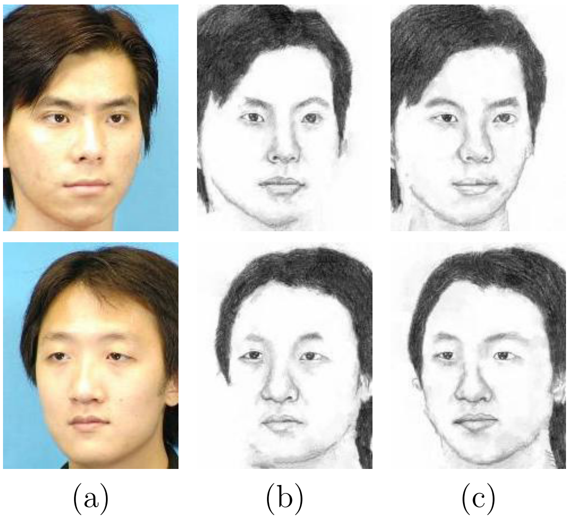


Figure 3.11: Representative results of photos with pose variations. (a) Photo; (b) Sketch by the method in [69]; (c) Sketch by our method. The results are best viewed on screen.

much worse than those given by our method. Under dark frontal lighting, their results usually have incorrect blank regions and noisy details. Under dark side lightings, the preprocessing helps only a little as it processes the photos globally. See the failed results shown in Fig. 3.10.

Pose

To test the robustness of our method to pose variations, we use the pose set with the similar lighting condition as the training set. As shown in Fig. 3.11, our method performs better than the method in [69].[5] With pose variations, the major problem of the results by [69] is to lose some structures especially on the profile. This problem can be efficiently alleviated by the shape priors, photo-to-sketch patch matching and gradient compatibility designed in our model.

3.2.2 Celebrity Faces from the Web

The robustness of our method is further tested on a challenging set of face photos of Chinese celebrities with uncontrolled lighting and pose variations from the web. They even have a variety of backgrounds. As shown in Fig. 3.12 and Fig. 3.13, the method in [69] usually produces noisy facial details and distortions, due to the uncontrolled lightings and backgrounds, and the large variations of pose and face shape. However, our method performs reasonably well.

3.3 Conclusion

We proposed a robust algorithm to synthesize face sketches from photos with different lighting and poses. We introduced shape

[5] As we do not have the sketches drawn by the artist for different poses, the recognition rates are not tested.

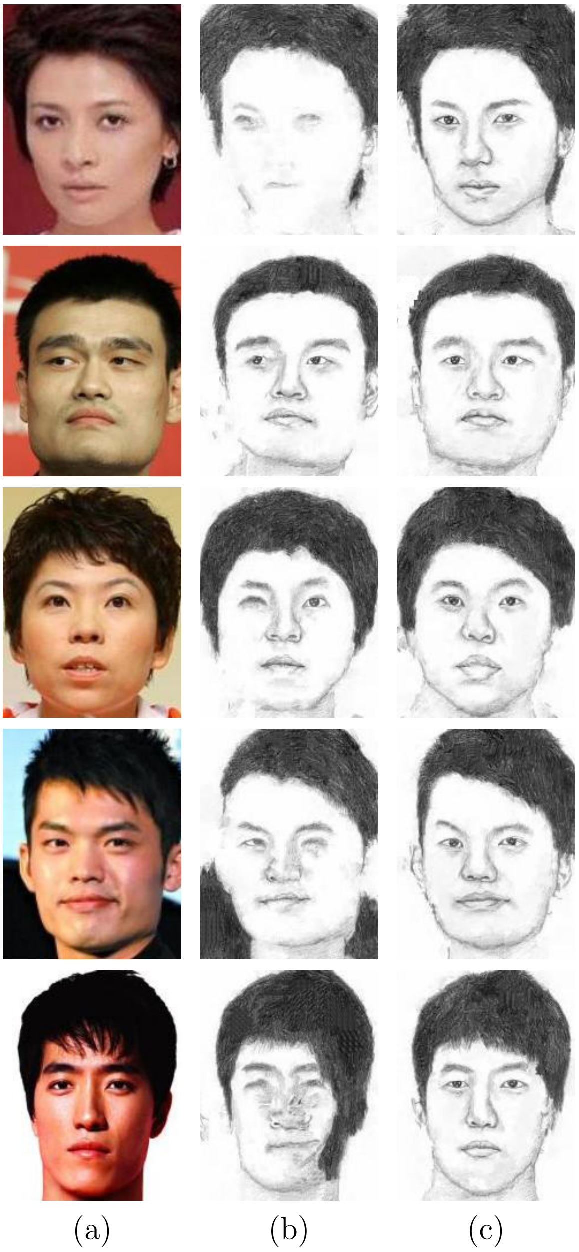


Figure 3.12: Results of Chinese celebrity photos. (a) Photo; (b) Sketch by the method in [69] with luminance remapping [27]; (c) Sketch by our method. The results are best viewed on screen.

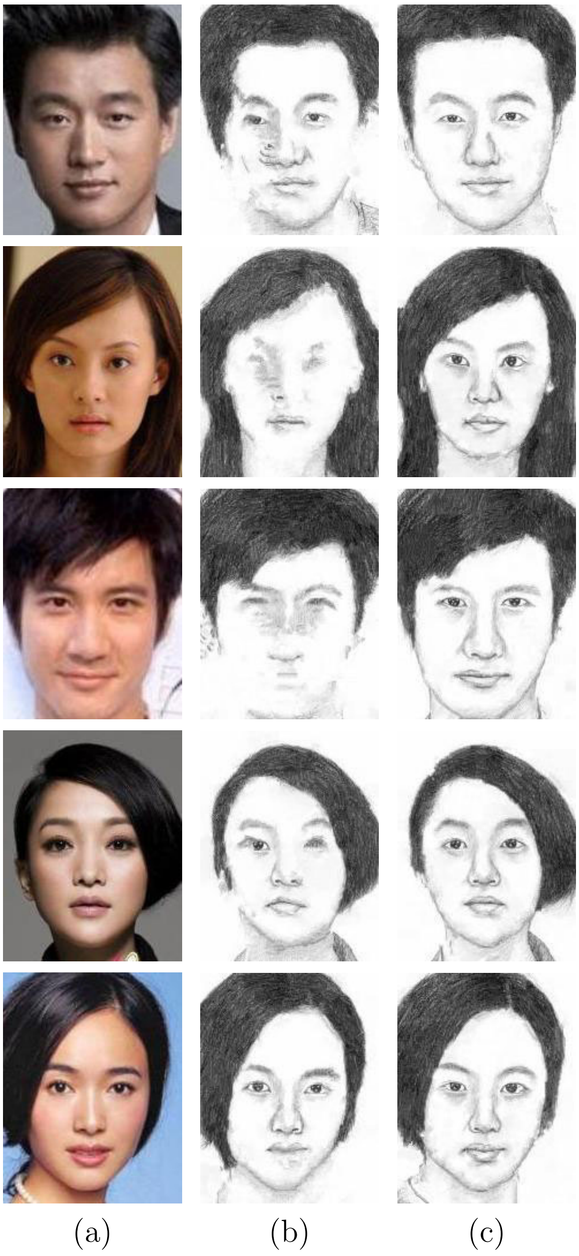


Figure 3.13: More results of Chinese celebrity photos. (a) Photo; (b) Sketch by the method in [69] with luminance remapping [27]; (c) Sketch by our method. The results are best viewed on screen.

priors, robust patch matching, and new compatibility terms to improve the robustness of our method. Our method is formulated using the multiscale MRF. It significantly outperforms the state-of-the-art approach. In the future work, we would like to further investigate face sketch synthesis with expression variations.

□ **End of chapter.**

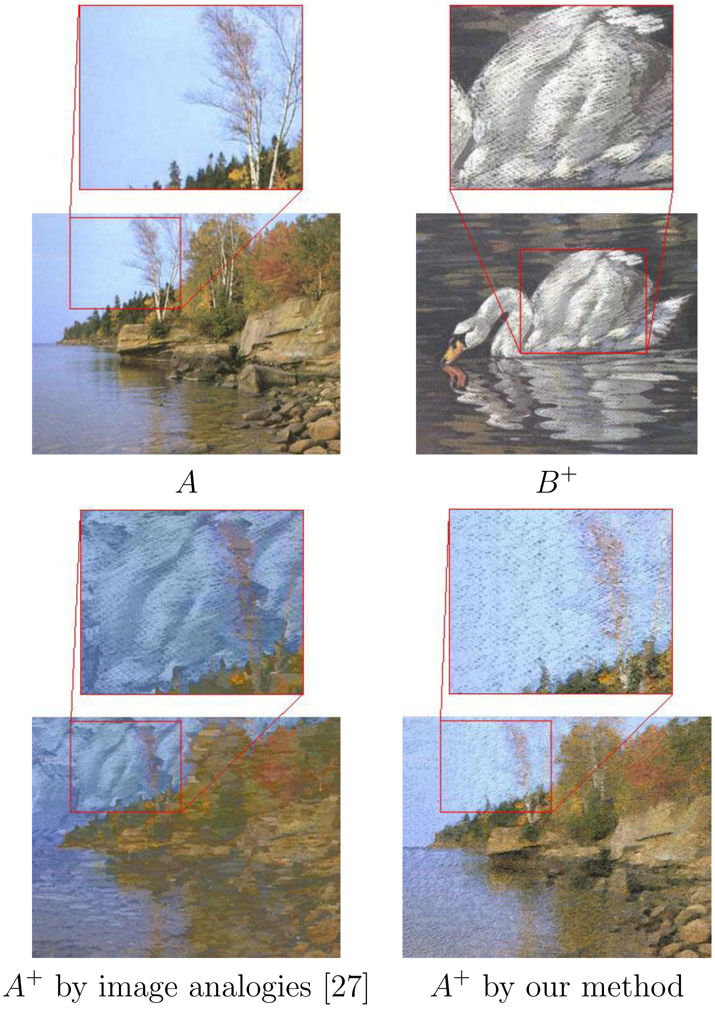
Chapter 4

Style Transfer via Band Decomposition

4.1 Introduction

Image stylization, which focuses on enabling a wide variety of expressive styles for images, has been an emerging technique during the past decade [24, 26]. Much research has been devoted to rendering different styles, e.g., oil painting, watercolor painting and pen-and-ink drawing. However, most of these methods support only limited artistic styles, and for a new style much human knowledge and experiences are required [26].

Example-based image stylization provides an easy way of creating stylized images with a number of styles. The stylized im-



A B^+

A^+ by image analogies [27] A^+ by our method

Figure 4.1: Comparison of luminance-based style transfer and our method. The body of the swan appears in A^+ by image analogies. The source image and style template are downloaded from the project page of image analogies.

age is synthesized from a real image (e.g., a scene photo) with a given style template (e.g., an oil painting). Formally, given two images as input, a source image A and a template image B^+ whose style is to be simulated, the output A^+ is a synthesized image with the main content in A and the similar style to B^+.

This process is also called style transfer, i.e., transferring the style of a template image B^+ to a source image (or video) A.

One critical problem of the existing approaches is that they do not separate the style and content in the style transformation process. In [27], [12], [52], only luminance is transferred from B^+ to A, which brings two drawbacks. First, the luminance of two input images may not be in the same dynamic range. To address this problem, a linear mapping that matches the means and variances of the two luminance distributions is often adopted. But usually good correspondences cannot be found for some input images. We will show some such examples in Section 4.4 for examples. Second, the content of B^+ may appear in the output images. Fig. 4.1 shows such an example.

In this chapter, we propose a novel algorithm to convert a real image to a stylized image with similar artistic style with an arbitrarily given template image. To break the limitations of the previous approaches, the basic idea is to decompose pixels into different components. Both the source image and template image are decomposed into the LF, MF, and HF components, which describe the content, main style, and information along the boundaries. We introduce the pixel grouping technique from image analysis, to simplify the frequency band decomposition greatly. Then the style is transferred from the template image to the source image in the MF and HF components. Style transfer is formulated as a global optimization problem by using Markov random fields (MRFs) [18], and a coarse-to-fine belief propagation algorithm is used to solve the optimization problem. To combine the LF component and the obtained style information, the final artistic result can be achieved via a reconstruction step.

One advantage of our method is its ease of use. For a given arbitrary style template image, it can automatically transform the style of an input image to the template style. It does not require the registered pair of source and stylized images [27, 12], or any user input [64]. The second advantage is that compared with other methods, our algorithm preserves the content of the input image and synthesizes the style, since we solve the problem in different frequency components. The third advantage is that it can be easily extended. We present an application of our image style transfer method: personalized artwork.

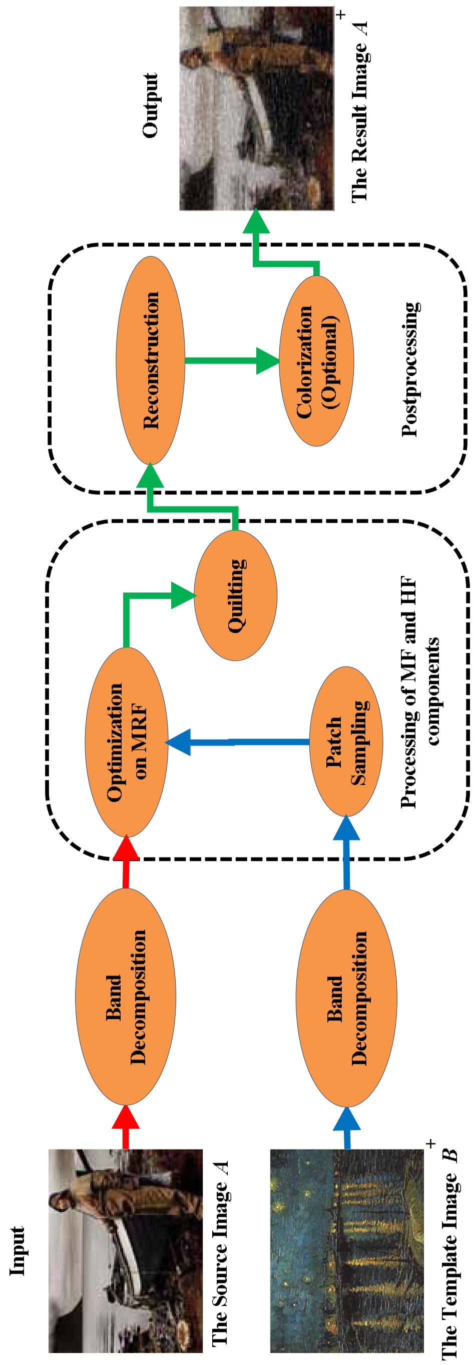


Figure 4.2: The framework of image style transfer.

4.2 Algorithm Overview

Our main idea originates from the classical theory of image analysis. In the view of image analysis, the low-frequency part contains the main content of an image, and the high-frequency part reflects the detail and boundary information of an image. Our style transfer approach separates an image into three frequency components, the LF, MF, and HF components. In our method, the LF part represents the main content of an image, the MF part represents the style (e.g., different artistic styles to different paintings), and the HF part represents the details along the boundaries.

Such a band decomposition solves the style transfer problem for A and B^+ with different luminance distribution (content) elegantly. First, the luminance distributions of A and B^+ are determined by the LF component. Then the MF and HF components of A^+ simulate the characteristics of the corresponding components of B^+. In our algorithm, the simulation step is formulated as a global optimization problem using MRFs [18]. Finally the three components of A^+ are merged to reconstruct the final image A^+. The three-stage process is shown in Fig. 4.2.

4.3 Image Style Transfer

Before the processing, we convert two input images A and B^+ from the RGB color space to the YIQ color space. In our algorithm, all operations are conducted in the Y channel (i.e., the luminance channel), and the I and Q channels are stored for the final color restoration. The reason of selecting only the Y channel for processing is that the artistic style is much more visually sensitive to changes in the Y channel than in the other two.

4.3.1 Band Decomposition

We consider the style transfer problem in different frequency components. We notice that different frequency components in an image contain different information. The LF component is affected by the illuminance and radiation of the image formation process and is almost irrelevant with the style. The MF component contains textures and determines the style of the image. In our algorithm, the HF component represents information along the boundaries. In addition, the LF component in an image is normally much larger in amplitude than the HF component. So, we argue that it is necessary to perform a frequency band decomposition before transferring the style from the template image to the input source image. To our best knowledge, we are the first to consider this problem for example-based image stylization.

In our approach, an image $I(x, y)$ is a combination of LF, MF, and HF components. We denoted it as

$$I(x, y) = I_L(x, y) \oplus I_M(x, y) \oplus I_H(x, y), \tag{4.1}$$

where I_L, I_M, and I_H are the LF, MF, and HF components, respectively.

We design a two-step strategy in our algorithm, instead of using the traditional frequency band decomposition techniques in signal and image processing. The first step is to separate the MF component from the image. It can be achieved via image segmentation, which is simply a partition of an image into contiguous regions of pixels that have similar appearances such as color or texture. Let the segmented image of A be A_S, in which each segment is represented by its mean luminance. A_S contains $A_L(x, y)$ and $A_H(x, y)$, the LF and HF components of

A, which is

$$A_S(x, y) = A_L(x, y) \oplus A_H(x, y). \tag{4.2}$$

We define

$$A_M(x, y) = A(x, y) - A_S(x, y), \tag{4.3}$$

where A_M is the MF component of A.

Mean shift is a *nonparametric* data clustering technique, which does not need to specify the number of clusters, and has been successfully applied to image segmentation [13]. Therefore, it is adopted in our algorithm. In the mean shift image segmentation, each pixel is assigned a feature point in a five-dimensional space, consisting of two spatial coordinates and three color components. The feature points are grouped by the clustering algorithm.

In the second step, we use the gradients of the segmented image A_S to estimate the HF component of A, i.e.,

$$A_H(x, y) = \left[\frac{\partial A_S(x, y)}{\partial x}, \frac{\partial A_S(x, y)}{\partial y} \right]. \tag{4.4}$$

We also perform the same decomposition on B^+ to obtain B_S^+, B_M^+ and B_H^+.

4.3.2 MF and HF Component Processing

In this stage, information is propagated from the MF and HF components of B^+ to those of A, respectively. Patches are sampled from the components of B^+, and organized to be the components of a new image A^+, which is close to A in some measure, i.e., we obtain A^+_M and A^+_H from B^+ and A. In the following description, the reference images are referred to the components of B^+, which the candidate patches come from, and the target images are the corresponding components of A, which need to be covered by patches. Note that the following computation is conducted independently in the MF and HF components.

Global Optimization on Markov Random Fields

We formulate the patch mapping problem as a labeling problem modeled by discrete MRFs [18]. First, the reference image is sampled as a dictionary $\mathcal{P}$ of $w \times h$ patches. Then the target image is divided into overlapping patches with the same size. Construct an undirected graph $G = (V, E)$, where the node set $V = \{v_1, v_2, ..., v_N\}$ contains all the patches in the target image, and E is the set of edges connecting each node to its four neighbors. For each node v_i we assign a patch x_i from the dictionary $\mathcal{P}$. Then the problem is to find the best configuration $X = \{x_1, x_2, ..., x_N\}$ to minimize an energy function defined later, where $x_i \in \mathcal{P}$ $(1 \leq i \leq N)$.

The placement of patches should match the target image and have local consistency. So, our energy function is

$$E(X) = \sum_{v_i \in V} E_1(x_i) + \lambda \sum_{(v_i, v_j) \in E} E_2(x_i, x_j), \tag{4.5}$$

where $E_1(x_i)$ is the penalty cost of assigning the patch x_i to the node v_i, $E_2(x_i, x_j)$ is the consistency cost of a neighboring node pair (v_i, v_j) having labels (x_i, x_j), and λ is a balance factor. We set $\lambda = 5$ in all experiments.

The definition of $E_1(x_i)$ is the sum of the squared differences (SSD) of pixel features between x_i and the region that x_i covers in the target image. $E_2(x_i, x_j)$ is the SSD of pixel features in the overlapping region between x_i and x_j. In summary, E_1 is used to control the reliability of the synthesized image and E_2 helps to produce a seamless synthesized image.

Component Image Quilting

After the placement of the patches, the component image quilting from patches is performed via the minimum cut algorithm.

Image quilting, which aims to produce a seamless image from overlapping patches, has been extensively studied in texture synthesis [17], [37]. The idea of image quilting is to find a seam with the least inconsistencies between neighboring patches with overlapping regions, which is formulated as a minimum cut problem. In our implementation, the patches are placed successively. For the placement of each patch, we construct a graph whose nodes are the pixels that are in the overlapping region of the existing patches and the new patch. A source node and a sink node are added to represent the existing patches and the new patch, respectively. In the graph, the boundary pixels are connected to the source or sink node with an infinite weight and each node is connected to its four neighbors. The weight of the edge connecting a neighboring pixel pair (i, j) is defined as

$$W(i,j) = \|\mathbf{F}(i) - \mathbf{F}_{new}(i)\|^2 + \|\mathbf{F}(j) - \mathbf{F}_{new}(j)\|^2, \tag{4.6}$$

where $\mathbf{F}(\cdot)$ and $\mathbf{F}_{new}(\cdot)$ denote the existing and new features of a pixel, respectively, and $\|\cdot\|$ is the L2-norm. After a cut is obtained, the existing features are updated according to the cut.

Implementation Details

The following implementation details of the above steps are highly related to the quality and speed of synthesis, both of which are critical for real-world applications.

First, to enhance the synthesis quality, we observe that the MF component of the template image usually contains strong style characteristic, which may bring noise in the MRF model. So in the MRF model, the MF component of the template image is processed as (Fig. 4.3)

$$\widetilde{B}_M^+ = \mathbf{Median}(B_M^+), \tag{4.7}$$

and the dictionary $\mathcal{P}$ is taken from $\widetilde{B}_M^+$, where **Median** is the median filter. In the quilting step, we use the corresponding patches from the original MF component B_M^+ to keep the style characteristic.

Second, we design several mechanisms to speed up the MRF optimization. In our application, both the size of dictionary $\mathcal{P}$ and the number of nodes in MRFs are large, due to the large size of images used for stylization. Although the popular belief propagation (BP) is adopted to efficiently solve the energy min-

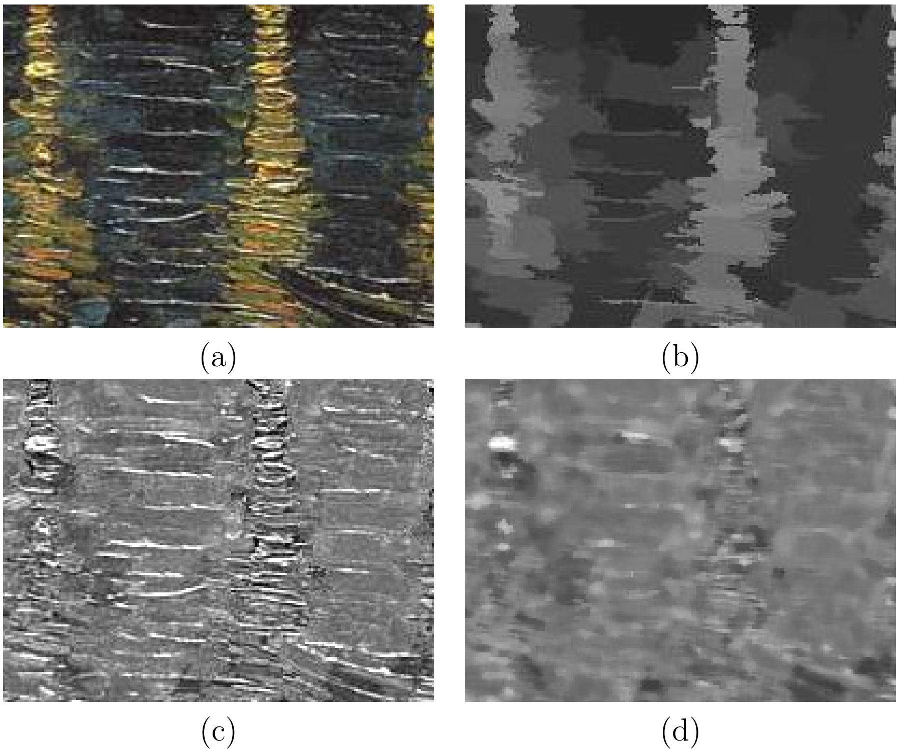


Figure 4.3: Components of a template image. (a) The template image. (b) The segmented image. (c) The MF component. (d) The filtered MF component. The segmented image has strong boundaries, which belong to the HF component. The MF component contains the style characteristic, while a median filter can remove most of the style.

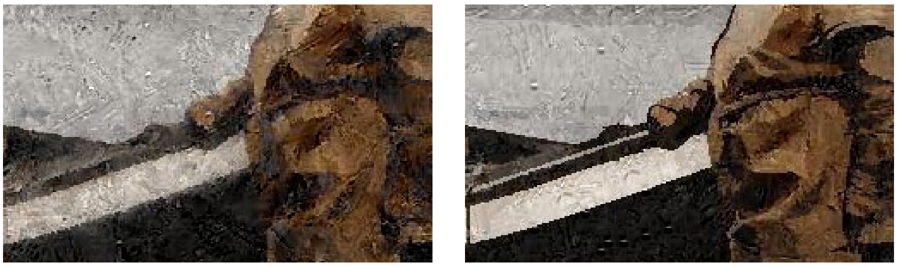


Figure 4.4: Comparison of the output images with MF and HF processing (the left column) and with only MF processing (the right column). The template image is "Freud" in Fig. 4.5.

imization problem, it is still necessary to find some way to speed up the optimization. In this work, we accelerate the algorithm via three aspects as follows. The first two are for reducing the size of dictionary $\mathcal{P}$, as the computational complexity of BP is the square of the number of the patches in the dictionary, and the last one is for reducing the number of nodes in the MRF.

1) In order to reduce the size of the dictionary, we utilize only 50% of the most representative sampled patches. In our implementation of constructing the MF component dictionary, the quality of a patch in representing the style is measured by

$$d_p = \sum_{(x,y)\in p} |\widetilde{B}_M^+(x,y) - B_M^+(x,y)|, \tag{4.8}$$

where $(x,y) \in p$ means that (x,y) is a point in patch p. The larger d_p is, the more style information the patch p contains. We choose the patches with the top 50% d values in the dictionary. It works well for all the artistic styles we test.

2) We use a two-step coarse-to-fine BP algorithm [30] to reduce the computational cost. First, the patches in the dictionary are divided into K clusters with the k-means algorithm. Then, the first BP is applied to find labels in the set of centers of clusters $\mathcal{P}^c = \{c_1, c_2, ..., c_K\}$, where c_i is the center of the i-th cluster. Finally, we perform the second BP for each cluster to select the best patches in the cluster. More details about the two-step BP can be found from [30].

3) In the optimization process for the HF component,

$$if\ \forall(x,y) \in p, A_H(x,y) = 0, then\ A_H^+(x,y) = 0. \qquad (4.9)$$

It means that, for the patches far away from the boundaries in the image, we keep their result all zero in the optimization process, since no boundary information should be transferred to regions far away from the boundaries. Thus the number of unlabeled nodes is greatly reduced in the optimization step for the HF component, which greatly reduces the running time.

Besides, for the MF and HF components, the patch sizes are different. The HF component uses a small size to keep more details, while the MF component uses a relatively large size to make the algorithm faster.

The processing of the HF component described in Section 4.3.2 is to create good appearance along the boundaries. If the HF component of A is copied to the output image A^+ without any processing, the boundaries of the output image A^+ are of high contrast (see Fig. 4.4 for example). Most of the artistic styles do not have such high contrast.

For some styles, the amount of HF components is important. A case in point, would be impressionist art with lots of little strokes. Please notice that keeping the MF of the painting/image does not make the points and little strokes disappear if the image is not in a very low resolution. The HF component is contributed by the boundaries of the points and little strokes only, but not by the whole points and little strokes. Therefore, our algorithm can still handle images with many points and small strokes. One example is the "Rhone" style in Fig. 4.6.

4.3.3 Reconstruction

The reconstruction step is to reconstruct final A^+ from the previous obtained results. There are three steps in our reconstruction scheme. First, A_S^+ is obtained from A_S and A_H^+, where A_S^+ corresponds to A_S (the segmentation result of A). Then, from (4.3) and achieved A_M^+, $A^+ = A_S^+ + A_M^+$ is obtained. As we state previously, the synthesized A^+ is in the luminance channel. Finally, by combining A^+ and the components of the input image

A in the I and Q channels, the final colorful stylized result is obtained.

Among these three steps, steps 2 and 3 can be performed straightforwardly. We discuss step 1 in more details here. Denote that $A_H^+ = [G_x, G_y]$. A_S^+ can be achieved by solving such a least-square (LS) problem that minimizes

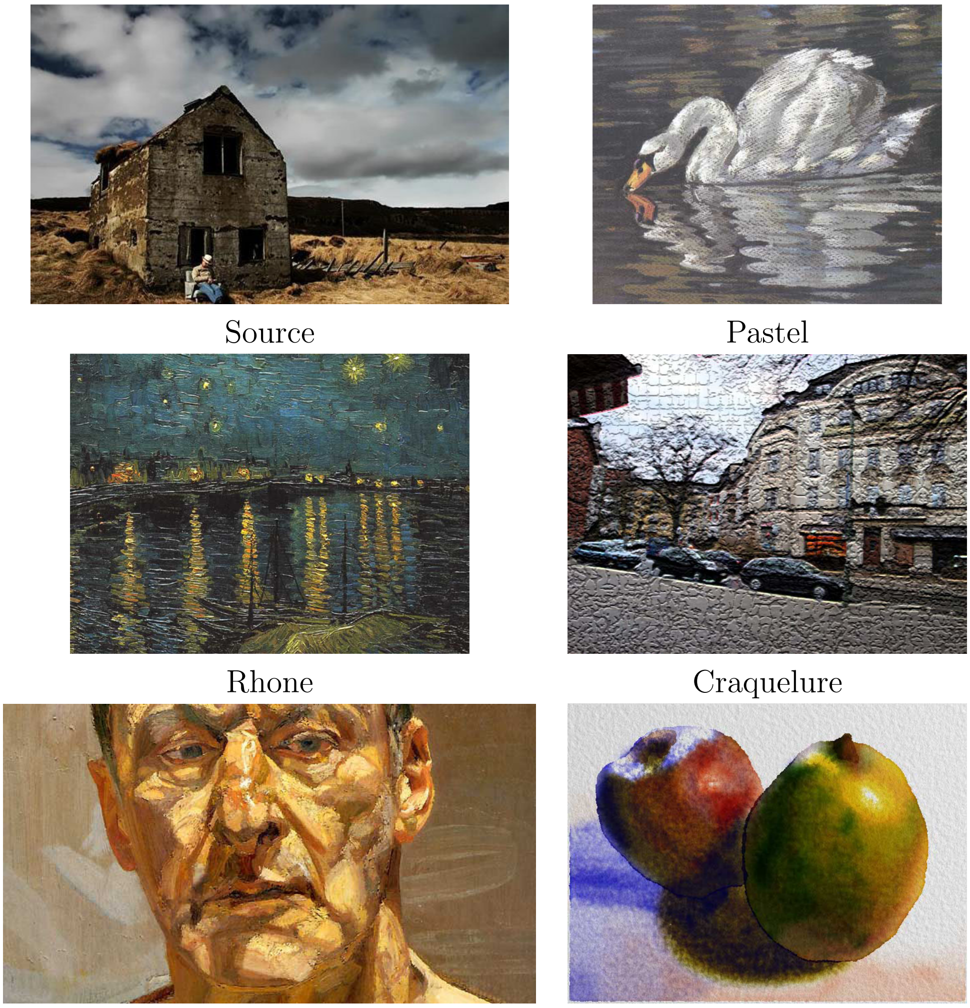
$$\begin{aligned} J(A_S^+) = &\sum_{(x,y)} \Big[\gamma \left(A_S^+(x,y) - A_S(x,y)\right)^2 \\ &+ \left((A_S^+)_x - G_x\right)^2 + \left((A_S^+)_y - G_y\right)^2\Big], \end{aligned} \tag{4.10}$$

where γ is a constant, and $\left(A_S^+\right)_x$, $\left(A_S^+\right)_y$ are partial derivatives of A_S^+ in the x and y direction, respectively. In all experiments, we set $\gamma = 0.01$.

The optimal A_S^+ of the above LS is the solution of the matrix equation

$$\gamma A_S^+ + A_S^+ D_x^T D_x + D_y^T D_y A_S^+ = \gamma A_S + G_x D_x + D_y^T G_y, \tag{4.11}$$

where D_x and D_y are the 1D differential operators, such that $A_S^+ D_x^T = \left(A_S^+\right)_x$ and $D_y A_S^+ = \left(A_S^+\right)_y$. This equation is of the form of the Lyapunov matrix equation $AX + XA = B$, where X is unknown and A and B are given.



Source

Pastel

Rhone

Craquelure

Freud

Watercolor

Figure 4.5: Template images.

4.4 Experiments

4.4.1 Comparison to State-of-the-Art

In this section, we test our method on a variety of source images and template styles, and compare it with the image analogies approach [27]. There are two reasons to compare our algorithm with image analogies. First, image analogies is known to be the best example-based image stylization algorithm so far. Considering the space limitation, we only compare our algorithm with the best performance. Second, since we focus on the automatic image stylization problem for a given arbitrary style template image, we do not compare our algorithm with the ones such as [12], [64], [16] that need manual input or other different input (please refer to Table 2.2).

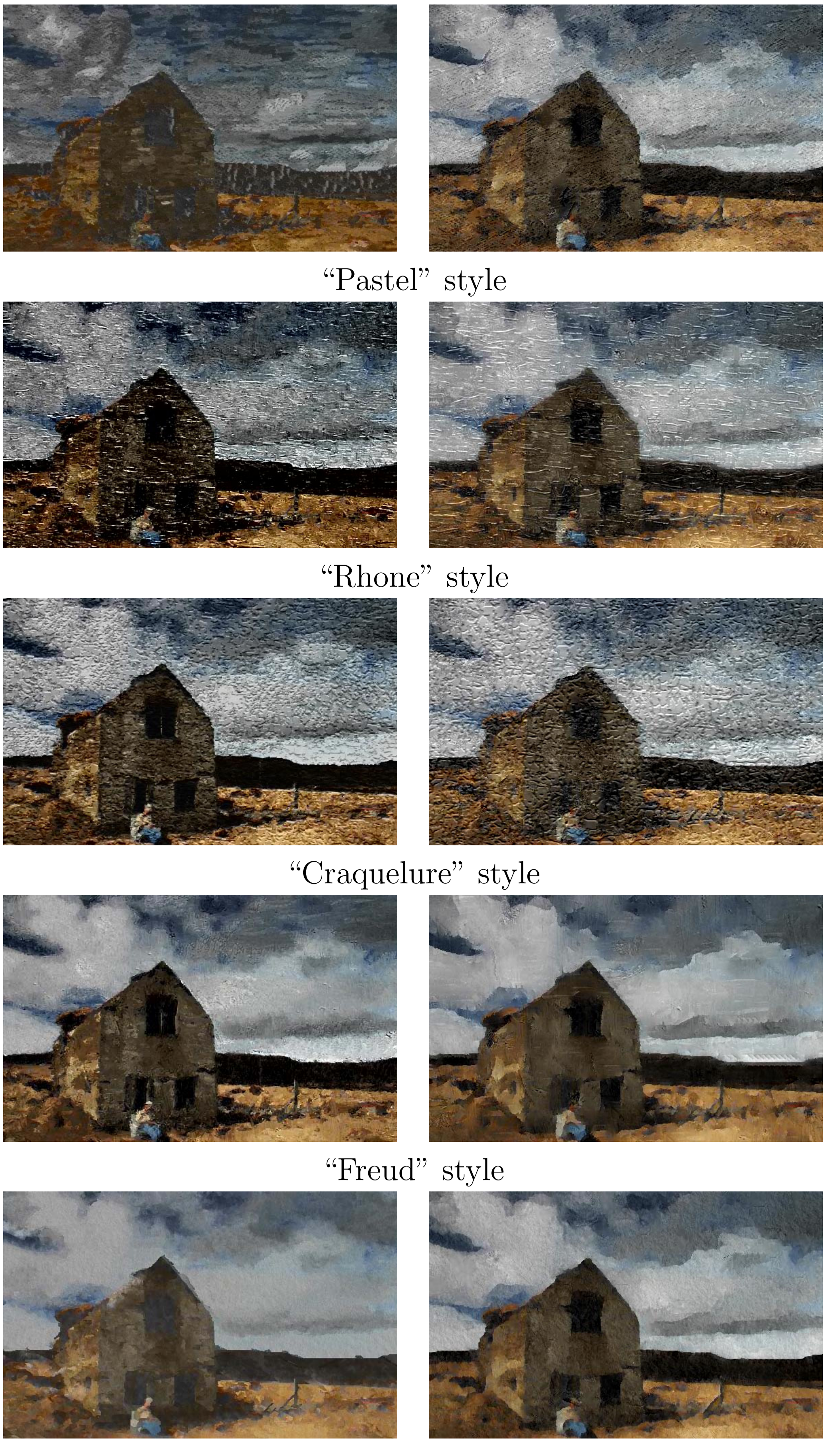
A source image and several style template images are given in Fig. 4.5. In our experiments, there are five template images, called Pastel, Rhone, Craquelure, Freud, and Watercolor, respectively. The Freud style contains many strokes since it is an oil painting, while the Watercolor style contains brushes and diffusion. The difference can be well found on the screen via zooming by, say, 300%. Note that the image analogies approach requires the ground-truth image B, while our method does not.

To our best knowledge, there is no quantitative metric for evaluating the results of image stylization. We evaluate the results visually. The comparison results of our algorithm and image analogies are given in Fig. 4.6.[1] From the results we

[1] We use the executable program of image analogies provided by the authors, available at `http://www.mrl.nyu.edu/projects/image-analogies/`.

can see that our results have better appearances than those of image analogies. The content of the source image is changed much more by using image analogies than our algorithm. For example, for the “Pastel” and “Rhone” styles (Fig. 4.6), image analogies changes the content greatly, while our algorithm preserves the content better. This is because in our algorithm, the LF component representing the image content is extracted and kept unchanged, while in image analogies, no similar scheme is used to preserve the content.

On the other hand, our algorithm synthesizes the style better than image analogies. Because no linear mapping can be found to align the luminance distributions of the input source and style template image perfectly, image analogies cannot synthesize some styles well. The distribution of style patterns relies strongly on the luminance in the result of image analogies. The result on a region with homogenous luminance has the trend of being one with a homogenous pattern. For example, for the style “Craquelure” (see Fig. 4.6), there are weak patterns in the dark regions in the result of image analogies. Another obvious example is for the style of “Pastel”. Some other results of our algorithm are provided in Fig. 4.7 and Fig. 4.8.



“Pastel” style

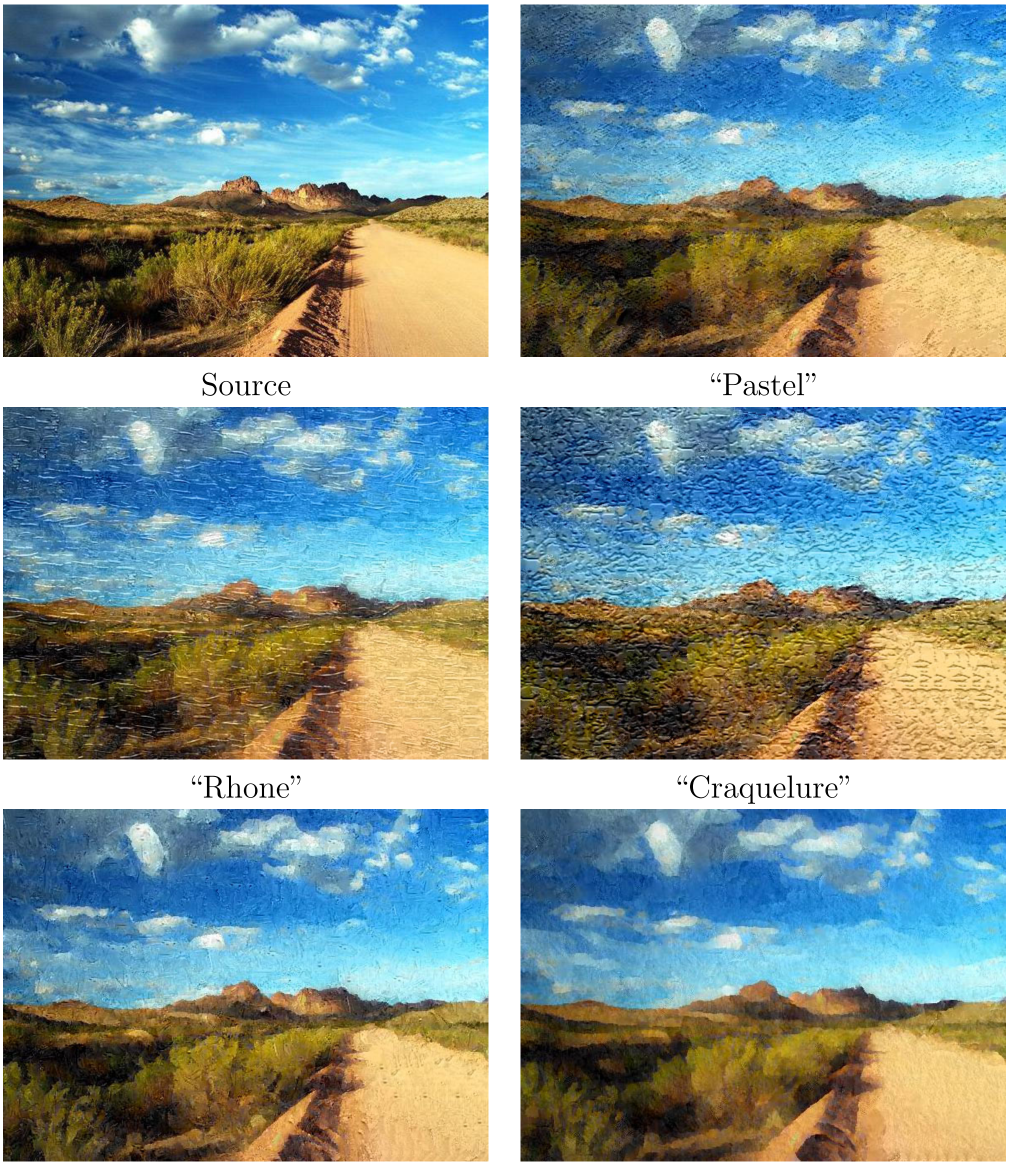
“Rhone” style

“Craquelure” style

“Freud” style

“Watercolor” style

Figure 4.6: Comparison of the results of image analogies (left) and our method (right). The results are best viewed with zooming-in on screen.



Source

“Pastel”

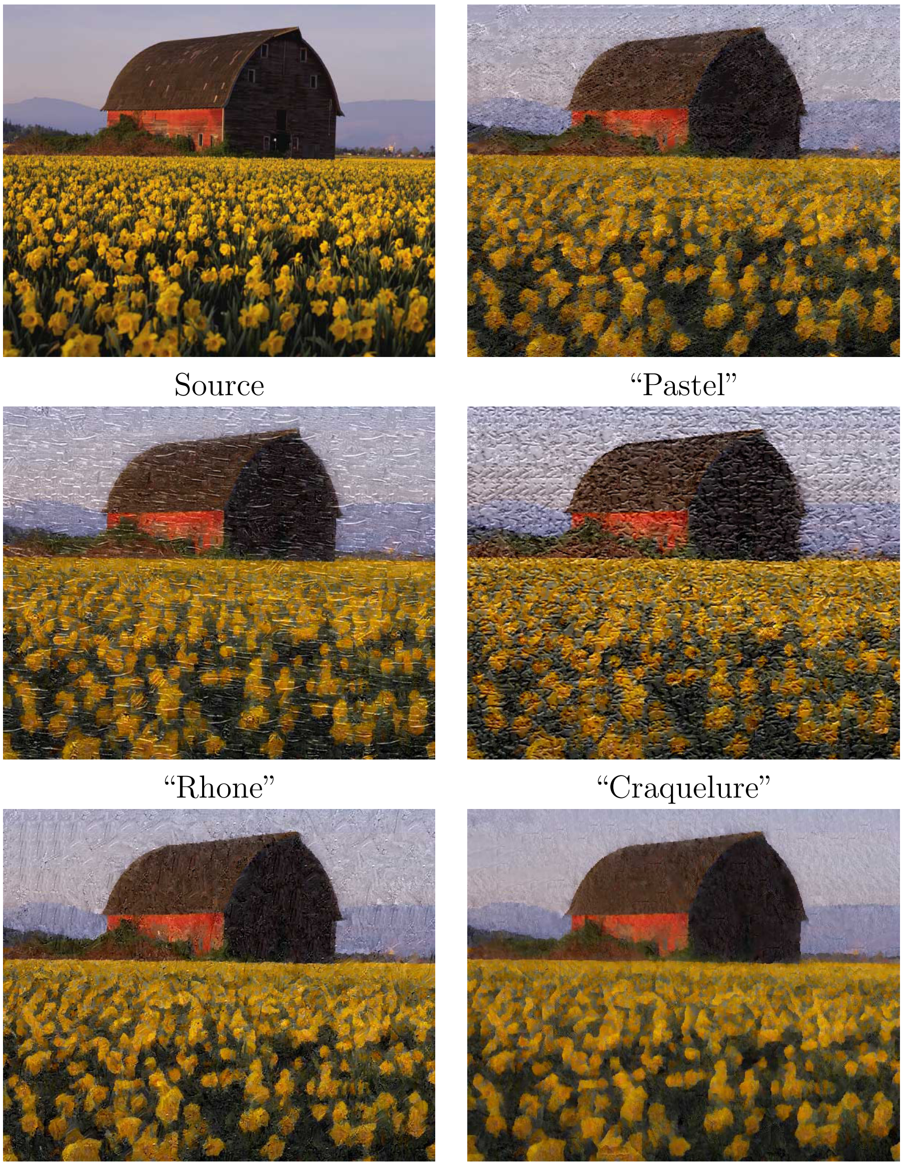
“Rhone”

“Craquelure”

“Freud”

“Watercolor”

Figure 4.7: More results of our method. The results are best viewed with zooming-in on screen.



Source

“Pastel”

“Rhone”

“Craquelure”

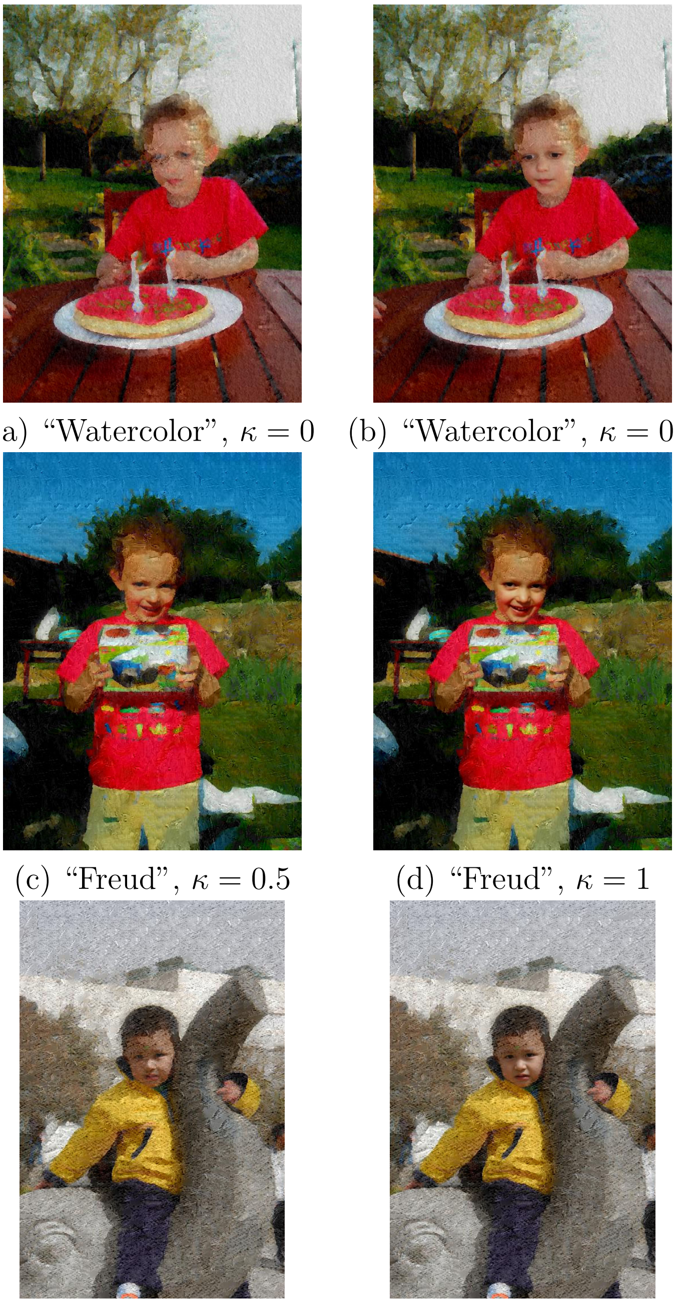
“Freud”

“Watercolor”

Figure 4.8: More results of our method. The results are best viewed with zooming-in on screen.

4.4.2 Extended Application: Personalized Artwork

Personalized artworks appear in our daily life and have drawn more and more attentions with the rapid popularization of digital images. Recently, a novel work called EasyToon has utilized vision and graphics technologies to insert the real face from photos into cartoon images [11]. Inspired by this approach, we propose an alternative way to create a personalized artwork, i.e., to convert a personal photo to an artistic image with the real facial appearance. The face is the most identifiable personal feature and the identity information is usually lost in the stylized face (compare Figs. 4.9(a) and (b)). So we propose the following procedure to create a personalized artwork. First, the face is extracted from the source photograph with the boosting based face detection [73] and active shape model based face alignment [89], and a mask is generated to separate the face and non-facial part. Then, the non-facial part is processed with our image style transfer algorithm. Finally, face blending is applied to synthesize a personalized artwork from the face and stylized non-facial part. As both parts are originally from the same photograph, the only necessary operation in face blending is illumination blending, which is much simpler than EasyToon [11].



(a) “Watercolor”, $\kappa = 0$ (b) “Watercolor”, $\kappa = 0.7$

(c) “Freud”, $\kappa = 0.5$ (d) “Freud”, $\kappa = 1$

(e) “Pastel”, $\kappa = 0.5$ (f) “Pastel”, $\kappa = 1$

Figure 4.9: Results of personalized artworks. The results are best viewed with zooming-in on screen.

Sometimes people may prefer to simultaneously stylize the face and keep the personal facial feature. Therefore, a linear coefficient κ between 0 and 1 is utilized to control the amount of personal information on the stylized face, i.e.,

$$f = \kappa f_{ns} + (1 - \kappa) f_s, \tag{4.12}$$

where f, f_{ns}, and f_s are final, completely stylized, and not stylized face. In Figs. 4.9(b) and (c), where $\kappa = 0.5$, we can recognize the person though the style exists on the face.

4.5 Conclusion

In this chapter, we proposed a frequency band decomposition based approach for transferring the style of an artistic image to real photographs. In our approach, three components, including the low-frequency, mid-frequency, and high-frequency components, are used to describe the content, main style information, and information along the boundaries of an image, respectively. Our algorithm preserves the content of the source image and synthesizes the style by copying style patches from the template. The patch copying process is formulated as a global optimization problem using Markov random fields, and the optimization problem is solved using a coarse-to-fine belief propagation. We further extend our approach to create personalized artwork. Inspired by EasyToon [11], which produces personalized cartoons, we propose a general framework to create personalized artworks from photos.

It is interesting to extend our work to produce stylized videos using an artistic image as the template. Cao *et al.* [7] investigated video stylization and personalization utilizing our algorithm [80].

We find some styles that our current framework does not work well on. One example is highly abstract artworks, e.g., the style of Picasso's surrealistic paintings. To deal with these problems, human interaction may be incorporated. It is attractive to continue our research in this direction.

□ **End of chapter.**

Chapter 5

Coupled Encoding for Sketch Recognition

5.1 Introduction

Face photo-sketch recognition is to match a face sketch drawn by an artist to one of many face photos in the database. In law enforcement, it is desired to automatically search photos from police mug-shot databases using a sketch drawing when the photo of a suspect is not available. This application leads to a number of studies on this topic [59, 60, 61, 69, 22, 35, 9]. Photo-sketch generation and recognition are also useful in digital entertainment industry.

The major challenge of face photo-sketch recognition is to match images in different modalities. Sketches are a concise representation of human faces, often containing shape exaggeration and having different textures than photos. It is infeasible to directly apply face photo recognition algorithms. Recently, great progress has been made in two directions. The first family of approaches [60, 43, 69] focused on the *preprocessing* stage and synthesized a pseudo-photo from the query sketch or pseudo-sketches from the gallery photos to transform inter-modality face recognition into intra-modality face recognition. Face photo/sketch synthesis is actually a harder problem than recognition. Imperfect synthesis results significantly degrade the recognition performance. The second family of approaches [41, 38, 35] focused on the *classification* stage and tried to design advanced classifiers to reduce the modality gap between features extracted from photos and sketches. If the inter-modality difference between the extracted features is large, the discriminative power of the classifiers will be reduced. A literature review can be found in Chapter 2.3.

In this chapter, we propose a new approach of reducing the modality gap at the *feature extraction* stage. A new face descriptor is designed by the coupled information-theoretic encoding,

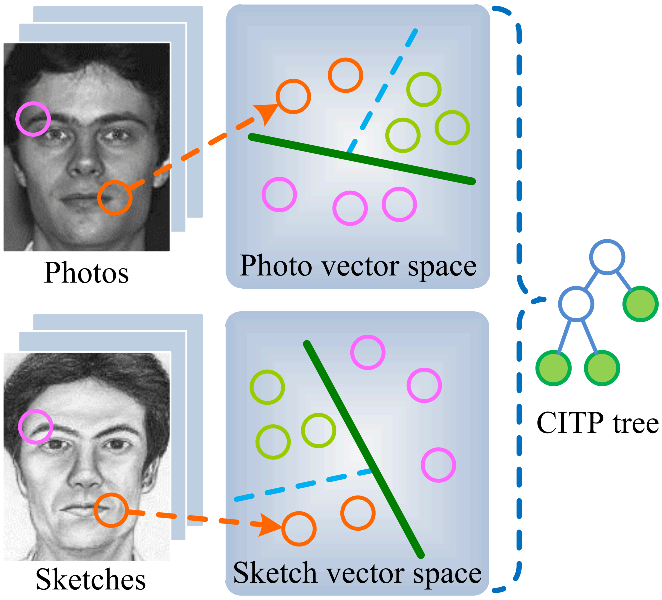


Figure 5.1: A CITP tree with three levels for illustration purpose. The local structures of photos and sketches are sampled and coupled encoded via the CITP tree. Each leaf node of the CITP tree corresponds to a cell in the photo vector space and in the sketch vector space. The sampled vectors in the same cell are assigned the same code, so that different local structures have different codes and the same structures in different modalities have the same code.

which quantizes the local structures of face photos and sketches into discrete codes. In order to effectively match photos and sketches, it requires that the extracted codes are uniformly distributed across different subjects, which leads to high discriminative power, and that the codes of the same subject's photo and sketch are highly correlated, which leads to small inter-modality gap. These requirements can be well captured under the criterion of maximizing the mutual information between photos and sketches in the quantized feature spaces. The coupled encoding is achieved by the proposed randomized coupled information-theoretic projection forest, which is learned with the *maximum mutual information* (MMI) criterion.

Another contribution of this work is to release CUHK Face Sketch FERET Database (CUFSF)[1], a large scale face sketch database. It includes the sketches of $1,194$ people from the FERET database [49]. Wang and Tang [69] published the CUFS database with sketches of 606 people. The sketches in the CUFS database had less shape distortion. The new database is not only larger in size but also more challenging because its sketches have more shape exaggeration and thus are closer to practical applications. Experiments on this large scale dataset show that our approach significantly outperforms the state-of-the-art methods.

5.1.1 Related work

There is an extensive literature on descriptor-based face recognition [1, 58, 71, 90, 70], due to its advantages of computational efficiency and relative robustness to illumination and pose variations. They are relevant to our coupled encoding. However,

[1] Available at `http://mmlab.ie.cuhk.edu.hk/cufsf/`.

those handcrafted features, such as local binary patterns (LBP) [1], dense scale-invariant feature transform (SIFT) [45], learning-based descriptor (LE) [8] and their variants [81, 78, 79, 71, 15], were not designed for inter-modality face recognition. The extracted features from photos and sketches may have large inter-modality variations.

Although information-theoretic concepts were explored in building decision trees and decision forests for vector quantization [2, 47, 54] in the application of object recognition, these algorithms were applied in a single space and did not address the problem of inter-modality matching. With the supervision of object labels, their tree construction processes were much more straightforward than ours.

5.2 Information-Theoretic Projection Tree

Vector quantization was widely used to create discrete image representations, such as textons [46] and visual words [55], for object recognition and face recognition. Image pixels [8, 54], filter-bank responses [46] or invariant descriptors [55, 72] were computed either sparsely or densely on a training set, and clustered to produce a codebook by algorithms such as k-means, mean shift [33], random projection tree [8, 20, 72] and random forest [47, 54]. Then with the codebook any image could be turned into an encoded representation.

However, to the best of our knowledge, it has not been clear how to apply vector quantization to cross-modality object matching yet. In this section, we present a new coupled information-theoretic projection (CITP) tree for coupled quantization across modalities. We further extend the CITP tree to the randomized CITP tree and forest. For clarity of exposition, we present the method in the photo-sketch recognition scenario.

5.2.1 Projection Tree

A projection tree [20] partitions a feature space $\mathbb{R}^D$ into cells. It is built in a recursive manner, splitting the data along one projection direction at a time. The succession of splits leads to a binary tree, whose leaves are individual cells in $\mathbb{R}^D$. With a built projection tree, a code is assigned to each test sample $\mathbf{x}$, according to the cell (i.e. leaf node) it belongs to. The sample is simply propagated down the tree, starting from the root node and branching left or right until a leaf node is reached.

Each node is associated with a learned binary function $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} - \tau)$. The node propagates $\mathbf{x}$ to its left child if $f(\mathbf{x}) = -1$ and to its right child if $f(\mathbf{x}) = 1$.

5.2.2 Mutual Information Maximization

Since quantization needs to be done in both the photo space and the sketch space, we extend a projection tree to a coupled projection tree. In a coupled projection tree, vectors sampled from photos and sketches share the same tree structure, but are input to different binary functions $f_p(\mathbf{x}_p)$ and $f_s(\mathbf{x}_s)$ at each node. A vector $\mathbf{x}_p$ sampled from the neighborhood of a photo pixel is quantized with f_p and a vector $\mathbf{x}_s$ sampled from the neighborhood of a sketch pixel is quantized with f_s. Then the sampled photo vectors and sketch vectors are mapped to the same codebook, but their coding functions represented by the tree are different, denoted by $\mathcal{C}_p$ and $\mathcal{C}_s$, respectively.

To train a coupled projection tree, a set of vector pairs $\mathcal{X} = \{(\mathbf{x}_i^p, \mathbf{x}_i^s), i = 1, ..., N\}$ is prepared, where $\mathbf{x}_i^p, \mathbf{x}_i^s \in \mathbb{R}^D$. In this work, $\mathbf{x}_i^p$ and $\mathbf{x}_i^s$ are the normalized vectors of sampled gradients around the same location[2] in a photo and a sketch of

[2] We sample the gradients (i.e. the first-order derivatives in the horizontal and vertical directions) I_u and I_v for an image I. Please refer to Section 5.3 for details.

the same subject, respectively. Denote that $\mathbf{X}^p = [\mathbf{x}_1^p, ..., \mathbf{x}_N^p]$, $\mathbf{X}^s = [\mathbf{x}_1^s, ..., \mathbf{x}_N^s]$. Since $\mathbf{x}_i^p$ and $\mathbf{x}_i^s$ are sampled from the same subject at the same location, it is expected that they are quantized into the same code by the coupled projection tree. In the meanwhile, in order to increase the discriminative power, it is expected that the codes of $\mathbf{X}^p$ and $\mathbf{X}^s$ are uniformly distributed across different subjects. To achieve these goals, our coupled information-theoretic projection (CITP) trees are learned using the *maximum mutual information* (MMI) criterion (see Fig. 5.2).

Mutual information, which is a symmetric measure to quantify the statistical information shared between two random variables [14], provides a sound indication of the matching quality between coded photo vectors and coded sketch vectors. Formally, the objective function is as follows.[3]

$$I(\mathcal{C}_p(\mathbf{X}^p); \mathcal{C}_s(\mathbf{X}^s)) = H(\mathcal{C}_p(\mathbf{X}^p)) - H(\mathcal{C}_p(\mathbf{X}^p)|\mathcal{C}_s(\mathbf{X}^s)). \quad (5.1)$$

To increase the discriminative power, the quantization should maximize the entropy $H(\mathcal{C}_p(\mathbf{X}^p))$ so that the samples are nearly

[3] The mutual information is originally defined between two random variables $\mathcal{C}_p(\mathbf{x}_i^p)$ and $\mathcal{C}_s(\mathbf{x}_i^s)$. We use the empirical mutual information estimated on the training set throughout this chapter.

uniformly distributed over the codebook. To reduce the inter-modality gap, the quantization should minimize the conditional entropy $H(\mathcal{C}_p(\mathbf{X}^p)|\mathcal{C}_s(\mathbf{X}^s))$.

5.2.3 Tree Construction with MMI

Similar to random projection tree [20], the CITP tree is also built top down recursively. However, it is different in that the CITP tree is not a balanced binary tree, i.e. the leaf nodes are at different levels. So the tree building process consists of searching for both the best tree structure and the optimal parameters at each node.

Tree structure searching. We adopt a greedy algorithm to build the tree structure. At each iteration, we search the node whose splitting can maximize the mutual information between the codes of sampled photo and sketch vectors. The mutual information, given in Eqn. (5.1), can be easily approximated in a nonparametric way. All the sampled photo and sketch vectors in the training set are quantized into codes with the current tree after splitting the candidate node, and the joint distribution of photo and sketch codes is computed to estimate the mutual information. A toy example is shown in Fig. 5.2.

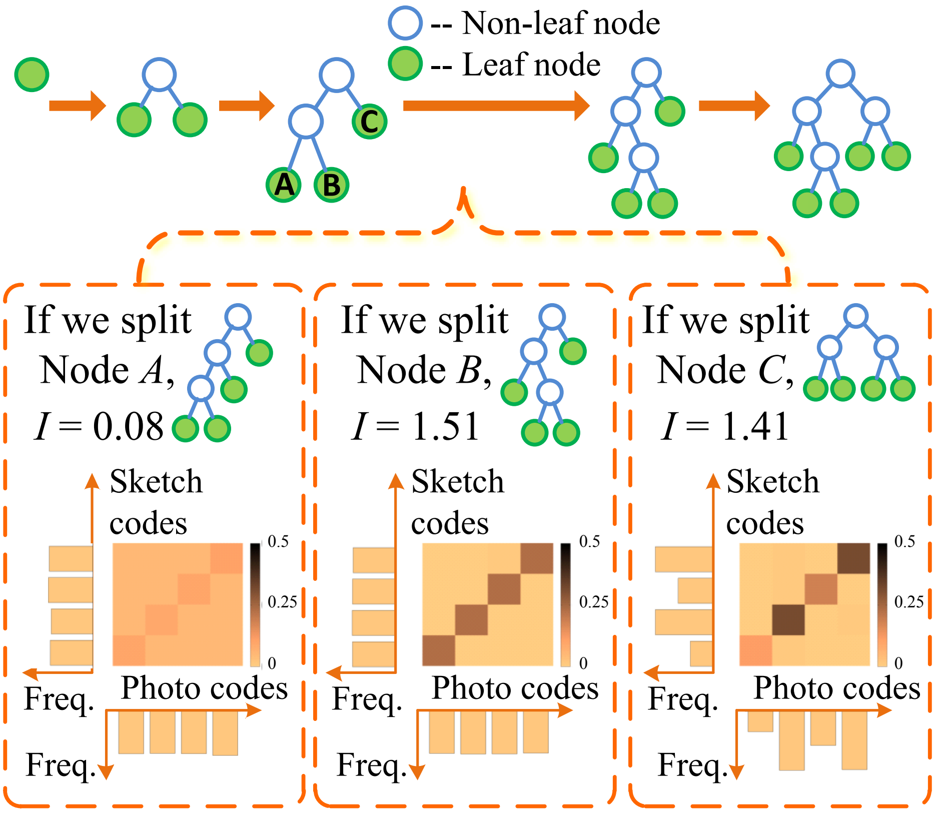

○ -- Non-leaf node
● -- Leaf node
C
A
B
If we split Node A, $I = 0.08$
If we split Node B, $I = 1.51$
If we split Node C, $I = 1.41$
Sketch codes
0.5
0.25
0
Freq.
Photo codes
Freq.

Figure 5.2: An illustration of tree construction with MMI. In each step, all current leaf nodes are tested and the one with the maximum mutual information is selected to split. For a leaf node, we try to split it and obtain a tree to encode photo vectors and sketch vectors. The selected leaf node should satisfy: (1) the codes are uniformly distributed; (2) the codes of photo vectors and corresponding sketch vectors are highly correlated. These requirements can be well captured under the MMI criterion. In this example, if we split node A, requirement (2) will not be satisfied, and if we split node C, requirement (1) will not be satisfied. The corresponding mutual information I of both are relatively small. So node B with the maximum mutual information is selected. The histograms and joint histograms of photo and sketch codes are visualized. In joint histograms, the colors represent the joint probability densities.

Node parameter searching. It is critical to search for optimal parameters of binary functions $f_p(\mathbf{x}_p)$ and $f_s(\mathbf{x}_s)$ to determine how to split the node. Formally, we aim at finding projection vectors $\mathbf{w}_p, \mathbf{w}_s$ and thresholds τ_p, τ_s for node k^4, such that

$$\begin{aligned} y_i^p &= \mathbf{w}_p^T \mathbf{x}_i^p - \tau_p, \quad \widehat{y}_i^p = \operatorname{sign}(y_i^p), \\ y_i^s &= \mathbf{w}_s^T \mathbf{x}_i^s - \tau_s, \quad \widehat{y}_i^s = \operatorname{sign}(y_i^s). \end{aligned} \tag{5.2}$$

Then a binary value $\widehat{y}_i^p$ (or $\widehat{y}_i^s$) is assigned to each vector $\mathbf{x}_i^p$ (or $\mathbf{x}_i^s$), to split the training data into two subsets and propagate them to the two child nodes. The node propagates a training vector pair $(\mathbf{x}_i^p, \mathbf{x}_i^s)$ to its children only if the binary values $\hat{y}_i^p$ and $\hat{y}_i^s$ are the same. Otherwise, the vector pair is treated as an outlier and discarded.

Suppose that the input of a node k is a set of vector pairs $\mathcal{X}_k = \{(\mathbf{x}_{k_i}^p, \mathbf{x}_{k_i}^s), 1 \leq i \leq N_k\}$. Denote that $\mathbf{X}_k^p = [\mathbf{x}_{k_1}^p, ..., \mathbf{x}_{k_{N_k}}^p]$, $\mathbf{X}_k^s = [\mathbf{x}_{k_1}^s, ..., \mathbf{x}_{k_{N_k}}^s]$, $\mathbf{Y}_k^p = [y_{k_1}^p, ..., y_{k_{N_k}}^p]$, $\mathbf{Y}_k^s = [y_{k_1}^s, ..., y_{k_{N_k}}^s]$, $\widehat{\mathbf{Y}}_k^p = [\widehat{y}_{k_1}^p, ..., \widehat{y}_{k_{N_k}}^p]$ and $\widehat{\mathbf{Y}}_k^s = [\widehat{y}_{k_1}^s, ..., \widehat{y}_{k_{N_k}}^s]$. The node is split according to the MMI criterion, i.e. maximizing

$$I(\widehat{\mathbf{Y}}_k^p; \widehat{\mathbf{Y}}_k^s) = H(\widehat{\mathbf{Y}}_k^p) + H(\widehat{\mathbf{Y}}_k^s) - H(\widehat{\mathbf{Y}}_k^p, \widehat{\mathbf{Y}}_k^s). \tag{5.3}$$

Instead of solving the above maximization problem directly, an approximate objective $I(\mathbf{Y}_k^p; \mathbf{Y}_k^s)$ is maximized first. Through

[4] We omit index k of the parameters, for conciseness.

maximizing $I(\mathbf{Y}_k^p; \mathbf{Y}_k^s)$, $\mathbf{w}_p$ and $\mathbf{w}_s$ are estimated without considering τ_p and τ_s. Assume that $y_{k_i}^p$ and $y_{k_i}^s$ are jointly Gaussian distributed. The entropy of a jointly Gaussian random vector $\mathbf{g}$ is $\frac{1}{2}\ln[\det(\mathbf{\Sigma_g})] + \text{const}$ [14], where $\mathbf{\Sigma_g}$ is the covariance matrix of $\mathbf{g}$. Following this, the mutual information can be rewritten in a simple form

$$I(\mathbf{Y}_k^p; \mathbf{Y}_k^s) = \frac{1}{2}\ln\left(\frac{\det(\Sigma_k^p)\det(\Sigma_k^s)}{\det(\mathbf{\Sigma}_k)}\right) + \text{const}, \tag{5.4}$$

where Σ_k^p, Σ_k^s and $\mathbf{\Sigma}_k$ are the covariance of $\mathbf{Y}_k^p$, $\mathbf{Y}_k^s$ and $[(\mathbf{Y}_k^p)^T, (\mathbf{Y}_k^s)^T]^T$, respectively. According to Eqn (5.2),

$$\begin{aligned} &\Sigma_k^p = \mathbf{w}_p^T \mathbf{C}_k^p \mathbf{w}_p, \ \Sigma_k^s = \mathbf{w}_s^T \mathbf{C}_k^s \mathbf{w}_s, \\ &\mathbf{\Sigma}_k = \begin{bmatrix} \mathbf{w}_p^T \mathbf{C}_k^p \mathbf{w}_p & \mathbf{w}_p^T \mathbf{C}_k^{p,s} \mathbf{w}_s \\ \left(\mathbf{w}_p^T \mathbf{C}_k^{p,s} \mathbf{w}_s\right)^T & \mathbf{w}_s^T \mathbf{C}_k^s \mathbf{w}_s \end{bmatrix}, \end{aligned} \tag{5.5}$$

where $\mathbf{C}_k^p$ and $\mathbf{C}_k^s$ are the covariance matrix of $\mathbf{X}_k^p$, $\mathbf{X}_k^s$, respectively, and $\mathbf{C}_k^{p,s}$ is the covariance matrix between $\mathbf{X}_k^p$ and $\mathbf{X}_k^s$.

Substituting Eqn. (5.5) into Eqn. (5.4), we find the equivalence between maximizing (5.4) and the Canonical Correlation

Analysis (CCA) model

$$\max_{\mathbf{w}_p,\mathbf{w}_s} \frac{\mathbf{w}_p^T \mathbf{C}_k^{p,s} \mathbf{w}_s}{\sqrt{\mathbf{w}_p^T \mathbf{C}_k^p \mathbf{w}_p \mathbf{w}_s^T \mathbf{C}_k^s \mathbf{w}_s}}. \tag{5.6}$$

So the optimal $\mathbf{w}_p$ and $\mathbf{w}_s$ are obtained by solving CCA (details are given later). CCA is found with good trade-off between the scalability and performance, when the input set is usually of a large size (about 2.5 million sample pairs in our experiments).

To estimate the thresholds τ_p and τ_s, we use brute-force search to maximize (5.3) in the region $(\tau_p, \tau_s) \in [\hat{\mu}^p - \hat{\sigma}^p, \hat{\mu}^p + \hat{\sigma}^p] \times [\hat{\mu}^s - \hat{\sigma}^s, \hat{\mu}^s + \hat{\sigma}^s]$, where $\hat{\mu}^p = \text{median}_i(y_i^p)$ and $\hat{\sigma}^p = \text{median}_i(|y_i^p - \hat{\mu}^p|)$ are the median and median of absolute deviation of y_i^p, respectively, and $\hat{\mu}^s$ and $\hat{\sigma}^s$ are the median and median of absolute deviation of y_i^s, respectively.

Canonical Correlation Analysis. CCA was introduced by Hotelling for correlating linear relationships between two sets of vectors [28]. It was used in some computer vision applications [77, 34, 56]. However, it has not been explored as a component of a vector quantization algorithm. Blaschko and Lampert [6] proposed an algorithm for spectral clustering with paired data based on kernel CCA. However, this method is not appropriate for quantization, as the kernel trick causes high computational and memory cost due to the very large size of the training set, and the nearest centroid assignment may be unstable (there is no hard constraint to require a pair of vectors in the same cluster).

To solve CCA in (5.6), let

$$\mathbf{S}_m = \begin{bmatrix} \mathbf{0} & \mathbf{C}_k^{p,s} \\ (\mathbf{C}_k^{p,s})^T & \mathbf{0} \end{bmatrix}, \mathbf{S}_n = \begin{bmatrix} \mathbf{C}_k^{p} & \mathbf{C}_k^{p,s} \\ (\mathbf{C}_k^{p,s})^T & \mathbf{C}_k^{s} \end{bmatrix},$$

and then $\mathbf{w} = [\mathbf{w}_p^T, \mathbf{w}_s^T]^T$ can be solved as the eigenvector associated with the largest eigenvalue of the generalized eigenvalue problem $\mathbf{S}_m\mathbf{w} = \lambda(\mathbf{S}_n + \varepsilon\mathbf{I})\mathbf{w}$, where ε is a small positive number for regularization.

The whole algorithm for building a CITP tree is summarized as Algorithm 1.

Algorithm 1 Algorithm of building a CITP Tree

```
Input: a set of vector pairs $\mathcal{X} = \{(\mathbf{x}_i^p, \mathbf{x}_i^s), i = 1, ..., N\}$, where $\mathbf{x}_i^p, \mathbf{x}_i^s \in \mathbb{R}^D$, and the expected number of codes (i.e. leaf nodes) $n_L$.
Create an empty set $\mathcal{S}$, and add the root node to $\mathcal{S}$.
repeat
    for each node $k$ in $\mathcal{S}$ and its associated vector set $\mathcal{X}_k$ do
        Compute the possible node splitting:
        (i) Generate projection vectors $\mathbf{w}_p, \mathbf{w}_s$ and thresholds $\tau_p, \tau_s$ with $\mathcal{X}_k$;
        (ii) For its left child $L$ and right child $R$,
            $\mathcal{X}_L \leftarrow \{(\mathbf{x}_i^p, \mathbf{x}_i^s) | \mathbf{w}_p^T \mathbf{x}_i^p \leq \tau_p, \mathbf{w}_s^T \mathbf{x}_i^s \leq \tau_s\}$,
            $\mathcal{X}_R \leftarrow \{(\mathbf{x}_i^p, \mathbf{x}_i^s) | \mathbf{w}_p^T \mathbf{x}_i^p > \tau_p, \mathbf{w}_s^T \mathbf{x}_i^s > \tau_s\}$,
            $(\mathcal{X}_L \subset \mathcal{X}_k, \mathcal{X}_R \subset \mathcal{X}_k)$;
    end for
    Select the best node splitting with the maximum mutual information in Eqn. (5.1);
    Split the node, remove the node from $\mathcal{S}$ and add its child nodes to $\mathcal{S}$;
until the number of leaf nodes is $n_L$.
Output: the CITP tree with projection vectors and thresholds at each node.
```

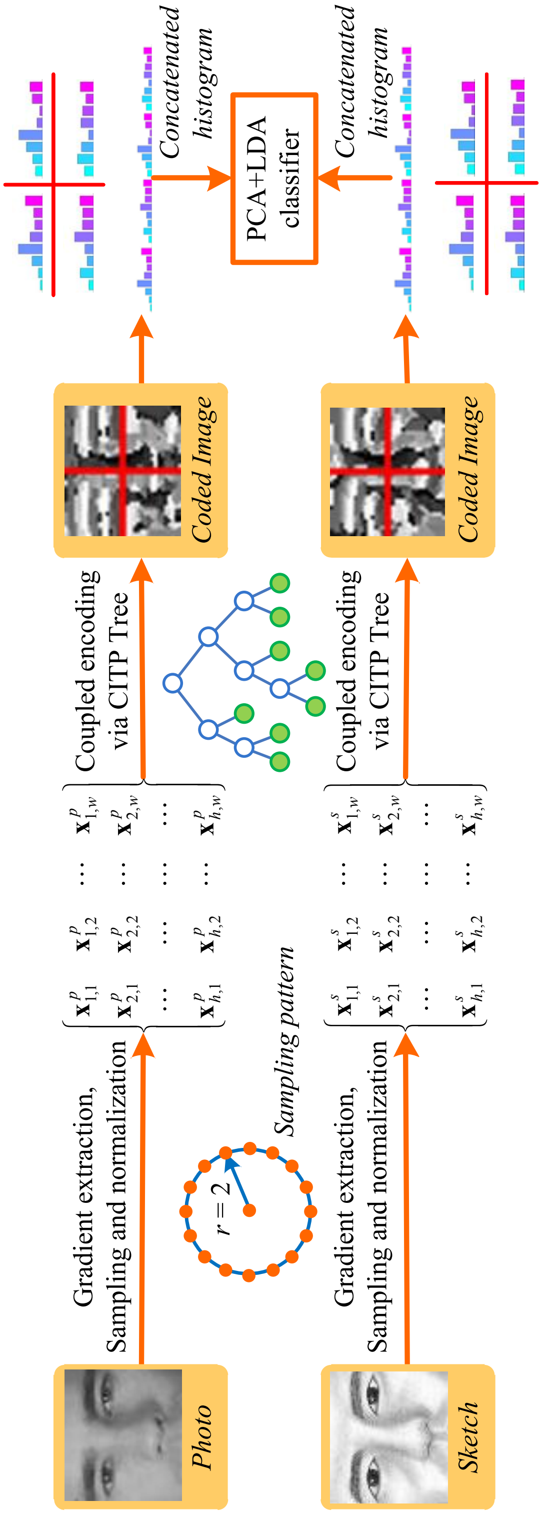


Photo
Sketch
Gradient extraction,
Sampling and normalization
r = 2
Sampling pattern
$\mathbf{x}^p_{1,1}$ $\mathbf{x}^p_{1,2}$ … $\mathbf{x}^p_{1,w}$
$\mathbf{x}^p_{2,1}$ $\mathbf{x}^p_{2,2}$ … $\mathbf{x}^p_{2,w}$
$\mathbf{x}^p_{h,1}$ $\mathbf{x}^p_{h,2}$ … $\mathbf{x}^p_{h,w}$
$\mathbf{x}^s_{1,1}$ $\mathbf{x}^s_{1,2}$ … $\mathbf{x}^s_{1,w}$
$\mathbf{x}^s_{2,1}$ $\mathbf{x}^s_{2,2}$ … $\mathbf{x}^s_{2,w}$
$\mathbf{x}^s_{h,1}$ $\mathbf{x}^s_{h,2}$ … $\mathbf{x}^s_{h,w}$
Coupled encoding
via CITP Tree
Coded Image
Concatenated
histogram
PCA+LDA
classifier

Figure 5.3: The pipeline of extracting CITE descriptors.

5.2.4 Randomized CITP Forest

Randomization is an effective way to create an ensemble of trees to boost the performance of tree structured algorithms [47, 54, 72]. The randomized counterpart of the CITP tree includes two modifications on node splitting as follows.

Randomization in sub-vector choice. At each node, we randomly sample α percent (empirically $\alpha = 80$) of the element indices of the sampled vectors, i.e. use a sub-vector of each sampled vector, to learn the projections. To improve the strength of generated trees, the random choice is repeated for 10 times empirically at each node, and the one with the maximum mutual information in Eqn. (5.3) is selected. The randomization at each node results in randomized trees with different tree structures and utilizing different information from the training data. Therefore, the randomized trees are more complementary.

Randomization in parameter selection. The eigenvectors associated with the first d largest eigenvalues in the CCA model are first selected. Then a set of n vectors are generated by randomly linearly combining the d selected eigenvectors.[5] Ac-

[5] The eigenvectors are orthogonalized with Gram-Schmidt orthogonalization and normalized with L_2-norm.

cording to the MMI criterion in Eqn. (5.3), the best one is selected from the set of n random vectors and used as the projection vectors $\mathbf{w}_p$ and $\mathbf{w}_s$. In our experiments, we choose $d = 3$ and $n = 20$.

The creation of a random ensemble of diverse trees can significantly improve the performance over a single tree, which is verified by our experiments.

5.3 Coupled Encoding Based Descriptor

In this section, we introduce our coupled information-theoretic encoding (CITE) based descriptor. With a CITP tree, a photo or a sketch can be converted into an image of discrete codes. The CITE descriptor is a collection of region-based histograms of the "code" image. The pipeline of photo-sketch recognition using a single CITP tree is shown in Fig. 5.3. The details are given as follows.

Preprocessing. The same geometric rectification and photometric rectification are applied to all the photos and sketches. With affine transform, the images are cropped to 80 × 64, and the two eye centers and the mouth center of all the face images are at fixed positions. Then both the photo and sketch images are processed with a Difference-of-Gaussians (DoG) filter [29] to remove both high-frequency and low-frequency illumination variations. Empirical investigations show that $(\sigma_1, \sigma_2) = (1, 2)$ is the best in our experiments.

Sampling and normalization. At each pixel, its neighboring pixels are sampled in a certain pattern to form a vector. A sampling pattern is a combination of one or several rings and the pixel itself. On a ring with radius r, $8r$ pixels are sampled evenly. Fig. 5.3 shows the sampling pattern of $r = 2$. We denote a CITE descriptor by a sampling pattern with rings of radius $r_1, ..., r_s$ as $\text{CITE}_{r_1,...,r_s}$.

We find that sampling the gradients I_u and I_v results in a better descriptor than sampling the intensities [8]. The gradient domain explicitly reflects relationships between neighboring pixels. Therefore, it has more discriminating power to discover key facial features than the intensity domain. In addition, the similarity between photos and sketches are easier to compare in the gradient domain than intensity domain [82].

After the sampling, each sampled vector is normalized such that its L_2-norm is unit.

Coupled Information-Theoretic Encoding. In the encoding step, the sampled vectors are turned into discrete codes using the proposed CITP tree (Section 5.2). Then each pixel has a code and the input image is converted into a "code" image. The vectors sampled from photos and sketches for training CITP tree are paired according to the facial landmarks detected by a state-of-the-art alignment algorithm [39].[6] Specifically, a pixel in the sketch image finds its counterpart in the photo image using a simple warping based on the landmarks. Note that the pairing is performed after sampling so that local structures are not deformed by the warping.

CITE Descriptor. The image is divided into 7 × 5 local regions with equal size, and a histogram of the codes is computed in each region. Then the local histograms are concatenated to form a histogram representation of the image, i.e. the CITE descriptor.

Classifier. We use a simple PCA+LDA classifier[7] [3, 66] to compute the dissimilarity between a photo and a sketch. By

[6] According to our observation, a general face alignment algorithm trained on commonly used face photo data sets is actually also effective for sketch alignment. We did not separately train a face alignment algorithm for sketches.

[7] A small regularization parameter is added to the diagonal elements of the within-class matrix of LDA to avoid singularity.

learning a linear projection matrix on the training set, it projects CITE descriptors into a low-dimensional space. Note that the descriptors are centered, i.e. the mean of the training CITE descriptors is subtracted from them. Then each projected CITE descriptor is normalized to a unit L_2-norm and the Euclidean distance between the normalized low-dimensional representation of a photo and a sketch is computed as their dissimilarity.

Fusion. We use a linear SVM to fuse dissimilarities by different CITE descriptors. The different CITE descriptors can be obtained by running the randomized CITP tree algorithm repeatedly. To train the one-class SVM, we select all the intrapersonal pairs and the same number of interpersonal pairs with smallest dissimilarities.

5.4 Experiments

In this section, we study the performance of our CITE descriptors and CITP trees on face photo-sketch recognition task. We first compare the performance of our CITE descriptor, with a single sampling pattern and single tree, to popular facial features, including LBP [1] and SIFT [45]. The classifier is not

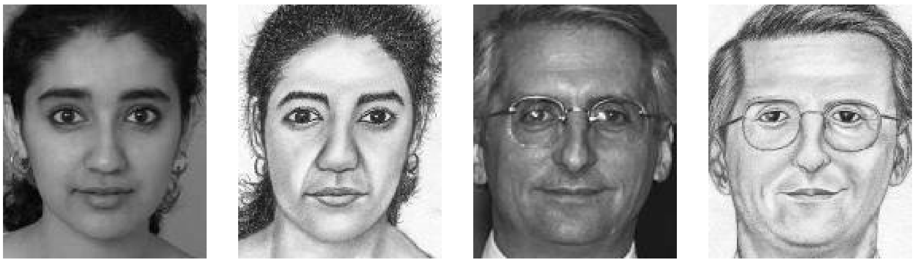


Figure 5.4: Examples of photos from the CUFSF database and corresponding sketches drawn by the artist.

used in this part to clearly show their difference. Then we investigate the effect of various free parameters on the performance of the system. Finally we show that our method is superior to the state-of-the-art.

Datasets. The CUHK Face Sketch FERET Database (CUFSF) is used for the experiments. There are $1,194$ people with lighting variations in the set. Each person has a photo and a sketch with shape exaggeration drawn by an artist. Some examples are shown in Fig. 5.4. The CUFS database [69] is also used as a benchmark. This dataset consists of 606 persons, each of which has a photo-sketch pair. The sketches were drawn without exaggeration by an artist when viewing the photo.

On the CUFSF database, 500 persons are randomly selected as the training set, and the remaining 694 persons form the testing set. On the CUFS database, 306 persons are in the

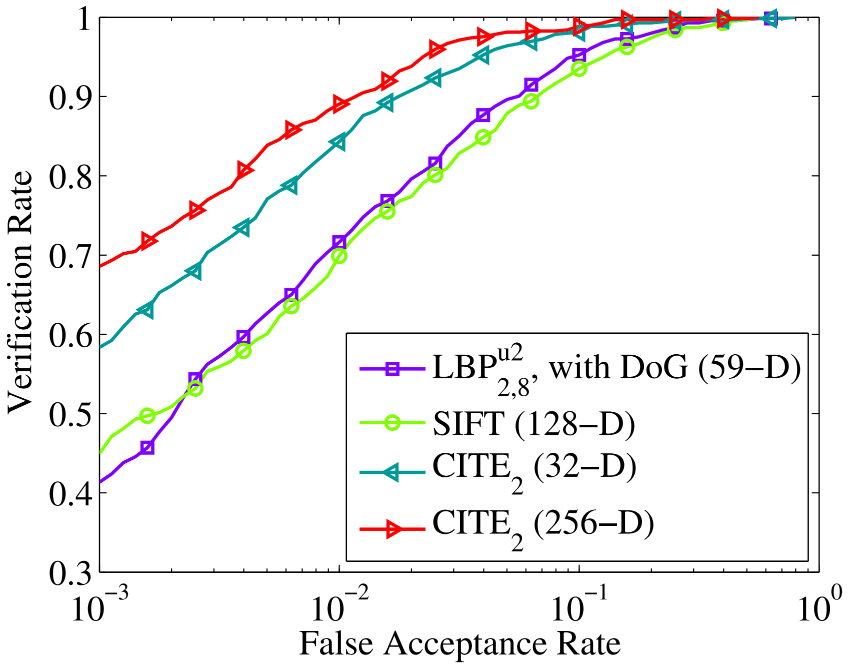


Figure 5.5: Comparison between CITE_2 (single CITP tree), LBP and SIFT. The dissimilarity between a photo and a sketch is computed as the distance between descriptors extracted on them. The χ^2 distance [1] is used for LBP and CITE_2, and Euclidean distance is used for SIFT. For simplicity, we give the length of a local histogram for each descriptor, instead of the length of the whole descriptor, in brackets.

training set and the other 300 persons are in the testing set.

Evaluation metrics. The performance is reported as Verification Rates (VR) at 0.1% False Acceptance Rate (FAR) and Receiving Operator Characteristic (ROC) curves.

5.4.1 Descriptor Comparison

We compare our descriptor with LBP [1] and SIFT [45]. The LBP is computed based on sampling points on a circle. We explore different numbers of sampling points and different radiuses. We find that the LBP descriptors extracted from DoG filtered images perform better than from original images. The 128-dimensional SIFT has 4×4 spatial bins of the same size and 8 orientation bins evenly spaced over $0^\circ - 360^\circ$. The vote of a pixel to the histogram is weighted by its gradient magnitude and a Gaussian window with parameter σ centered at the center of the region. We explore different sizes of the region and different σ. For our CITE descriptor, we use the sampling pattern of a single ring with $r = 2$ as shown in Fig. 5.3. We test on different numbers of leaf nodes (i.e. different sizes of a local histogram).

The ROC curves are shown in Fig. 5.5. Even 32-dimensional CITE$_2$ (please refer to Section 5.3 for this notation) significantly outperforms the 59-dimensional LBP and 128-dimensional SIFT. The 256-dimensional CITE$_2$ (68.58%) beats the best results of LBP (41.35%) and SIFT (44.96%) by 20% on VR at 0.1% FAR.

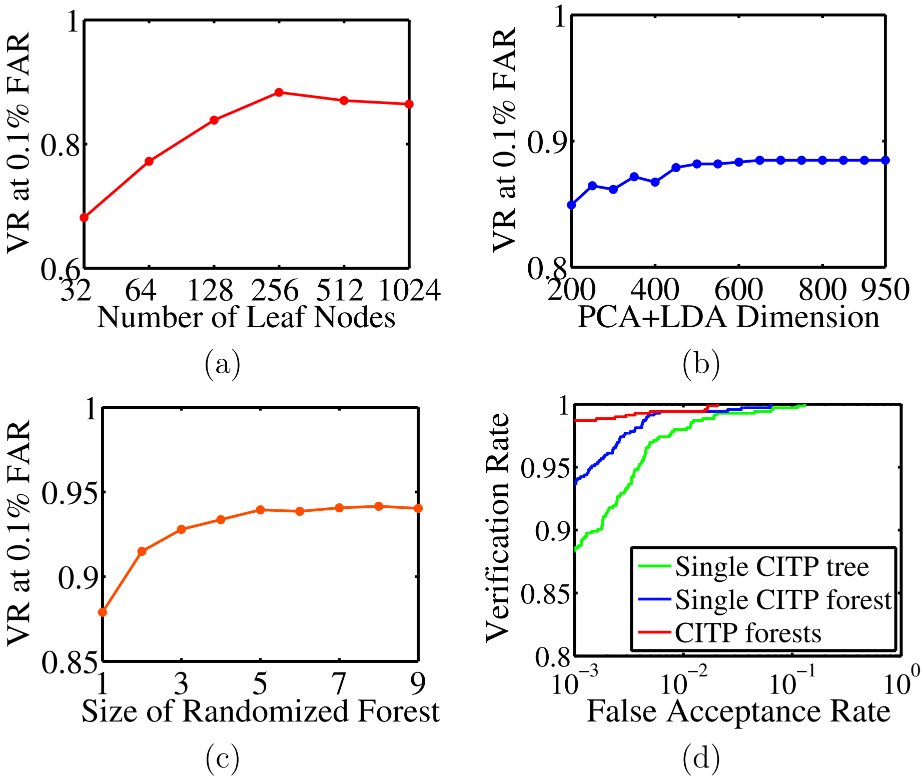
5.4.2 Parameter Exploration

We investigate the effect of various free parameters on the performance of the system, including the number of leaf nodes, the projected dimension by PCA+LDA, the size of randomized forest and the effect of using different sampling patterns. We fix the other factors when investigating one parameter.

Number of Leaf Nodes. We compare the effect of using different numbers of leaf nodes in a CITP tree. The number is extensively studied from 32 (2^5) to 1024 (2^{10}). As shown in Fig. 5.6(a), the VR initially increases, and does not increase when the number is larger than 256. Due to small performance gain and high computational cost of a large leaf node number, we choose 256 leaf nodes as our default setting.

PCA+LDA Dimension. The reduced dimension is an important parameter of PCA+LDA. The VR has a fairly large stable region and varies less than 1% from 500 to 950 (see Fig. 5.6(b)). We choose 600 PCA+LDA dimensions in our final system.

Size of Randomized Forest. We vary the number of randomized trees in the CITP forest from 1 to 9. Fig. 5.6(c) shows that increasing the number of trees from 1 to 5 increases the VR from 87.90% to 93.95%, with little improvement beyond this. Hence, we fix the number of randomized trees in a CITP forest to be 5.



(a)

(b)

(c)

(d)

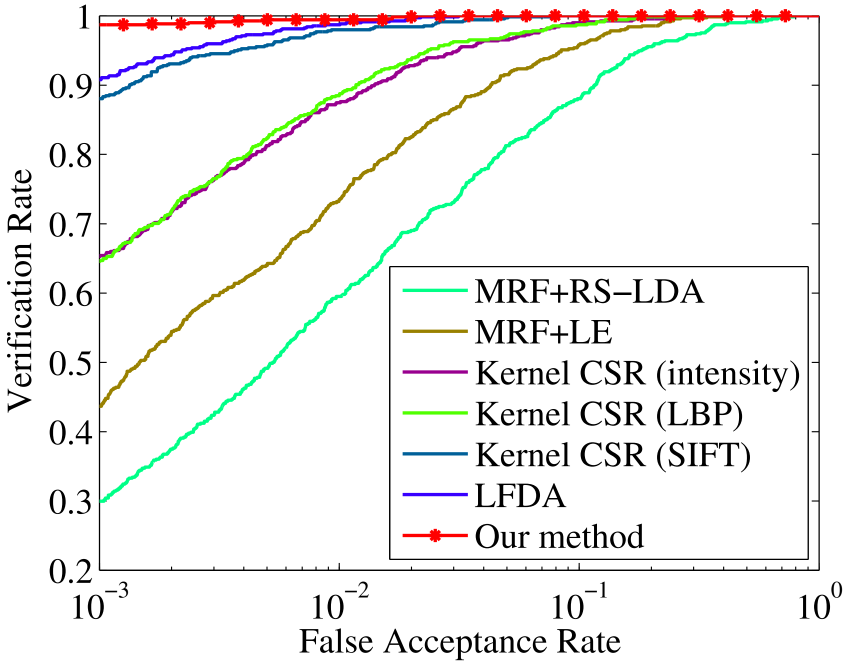
Figure 5.6: VR at 0.1% FAR vs. (a) number of leaf nodes; (b) PCA+LDA dimension; (c) size of randomized forest; (d) comparison of ensemble of forests with different sampling patterns and the forest with a single sampling pattern. In (a)–(c), The descriptor is CITE$_2$. In (a), the descriptors are compressed to 600 dimensional using PCA+LDA, and a single CITP tree is used. In (b), we use 256 leaf nodes and a single CITP tree. In (c) and (d), we use 256 leaf nodes and 600 PCA+LDA dimensions.

Ensemble of Randomized Forests with Different Sampling Patterns. Although the performance increases slowly when the number of randomized trees is more than 5, using ensemble of randomized forests with different sampling patterns can further boost the performance. Different sampling patterns can capture rich information across multiple scales. Fig. 5.6(d) shows that using five sampling patterns improves the VR at 0.1% FAR from 93.95% to 98.70%.

5.4.3 Experiments on Benchmarks

We compare our algorithm with the following state-of-the-art approaches on the CUFSF database. The algorithms are tuned to the best settings according to their paper.

- MRF-based synthesis [69]. Pseudo photos are synthesized from query sketches, and random sampling LDA (RS-LDA) [68] is used to match them to gallery photos. In addition, we test LE [8] on matching pseudo photos and gallery photos.
- Kernel CSR [38]. The CSR model is trained to seek for a common discriminative subspace, based on intensities, LBP and SIFT feature vectors separately.



VR at 0.1% FAR		
MRF+RS-LDA	MRF+LE	LFDA
29.54%	43.66%	90.78%
Kernel CSR (LBP)	Kernel CSR (SIFT)	Ours
64.55%	88.18%	98.70%

Figure 5.7: Comparison of the state-of-the-art approaches and our method on the CUFSF database. ROC curves and VR at 0.1% FAR are shown.

- LFDA [35]. It fuses the LBP features with four different radiuses and the SIFT features with a discriminative model. For each feature, multiple projection vectors are learnt.

Fig. 5.7 shows that our method significantly outperforms the state-of-the-art approaches. MRF-based synthesis requires that there is no significant shape distortion between photos and sketches in the training set, and also that training photos are taken under similar lighting conditions. This method does not work well in this new data set because the drawing style of the artist involves large shape exaggeration and the photos in the FERET database are taken under different lightings with large variations. Therefore, the pseudo photos by MRF-based synthesis have artifacts such as distortions. Such artifacts degrade the performance of state-of-the-art face photo recognition algorithms including RS-LDA and LE. The results of Kernel CSR on different features verify that the inappropriate selection of features will reduce the discriminative power of the classifier. SIFT features have better results than LBP on the photo-sketch recognition task. LFDA achieves a good result by fusing five different kinds of features with two different spatial partitions. However, its error rate (9.22%) is much higher than ours (1.30%) for 0.1% FAR.

Our method also has superior performance on the CUFS database, a standard benchmark for face photo-sketch recognition, as shown in Table 5.1. Apparently, this dataset is now an easy one for the state-of-the-art methods.

Table 5.1: Rank-1 recognition rates on the CUFS database. The recognition rates are averaged over five random splits of 306 training persons and 300 testing persons. We test our method with the same configuration of training and testing splits as [69, 35].

MRF+RS-LDA [69]	LFDA [35]	Ours
96.30%	99.47%	99.87%

5.5 Conclusions

We proposed a coupled information-theoretic encoding based descriptor for face photo-sketch recognition. We introduced coupled information-theoretic projection forest to maximize the mutual information between the encoded photo and encoded sketch of the same subject. Our system significantly outperforms the state-of-the-art approaches. In the future work, we would like to further investigate the system with more cross-modality recognition problems.

□ **End of chapter.**

Chapter 6

Conclusion

In many domains of computer vision, data can be represented in multiple modalities. Different modalities of the same object or scene are generated by different processes. In the previous chapters we have been through several topics in inter-modality image synthesis and recognition: face sketch synthesis, example-based image stylization, and face photo-sketch recognition.

The first part of the thesis focuses on real-world face sketch synthesis. Automatic face sketch synthesis has important applications in law enforcement and digital entertainment. Although great progress has been made in recent years, previous methods only work under well controlled conditions and often fail when there are variations of lighting and pose. In Chapter 3, we propose a robust algorithm for synthesizing a face sketch from a face photo taken under a different lighting condition and in a different pose from the training set. It synthesizes local sketch patches using a multiscale Markov Random Field (MRF) model. The robustness to lighting and pose variations is achieved in three steps. Firstly, shape priors specific to facial components are introduced to reduce artifacts and distortions caused by variations of lighting and pose. Secondly, new patch descriptors and metrics which are more robust to lighting variations are used to find candidates of sketch patches given a photo patch. Lastly, a smoothing term measuring both intensity compatibility and gradient compatibility is used to match neighboring sketch patches on the MRF network more effectively. The proposed approach significantly improves the performance of the state-of-the-art method. Its effectiveness is shown through experiments on the CUHK face sketch database and celebrity photos collected from the web.

Then we explore unsupervised style transfer, i.e., transferring the artistic style from a template image to photos. It is a more general problem than face sketch synthesis, and has wide applications of making artistic effects for images and videos. However, most existing methods do not consider the content and style separately. In Chapter 4, we propose a style transfer algorithm via a novel frequency band decomposition approach, based on techniques from image analysis. First, an image is decomposed into the low-frequency (LF), mid-frequency (MF), and high-frequency(HF) components, which describe the content, main style, and information along the boundaries. Then the style is transferred from the template image to the source photo in the MF and HF components. Style transfer is also formulated as a global optimization problem by using MRF, and a coarse-to-fine belief propagation algorithm is used to solve the optimization problem. To combine the LF component of the source photo and the obtained style information, the final artistic result can be achieved via a reconstruction step. Compared to the other algorithms, our method not only synthesizes the style, but also preserves the image content well. We extend our algorithm to personalized artwork. The results indicate that our approach performs excellently in image stylization and the extended application.

Finally we consider the problem of inter-modality face recognition through proposing a new feature representation. We studied face photo-sketch recognition, which is a typical and challenging inter-modality face recognition problem. Recent research has focused on transforming photos and sketches into the same modality for matching or developing advanced classification algorithms to reduce the modality gap between features extracted from photos and sketches. In Chapter 5, we propose a new inter-modality face recognition approach by reducing the modality gap at the feature extraction stage. A new face descriptor based on coupled information-theoretic encoding is used to capture discriminative local face structures and to effectively match photos and sketches. Guided by maximizing the mutual information between photos and sketches in the quantized feature spaces, the coupled encoding is achieved by the proposed coupled information-theoretic projection tree, which is extended to the randomized forest to further boost the performance. We create the largest face sketch database including sketches of $1,194$ people from the FERET database. Experiments on this large scale dataset show that our approach significantly outperforms the state-of-the-art methods.

Although our work on inter-modality image synthesis achieved state-of-the-art performance, it should be noted that there is still a long way to go for simulating arbitrary artistic styles with automatic algorithms. Our work does not work well on styles with large shape exaggeration or high abstraction, because patch-based methods require good alignment between modalities and do not model semantic information. To our best knowledge, there is barely any existing research on such styles. It is worthy to explore along this direction.

For the future work, we also plan to explore more innovative feature representations for inter-modality matching and more applications of the proposed methods. We will apply the coupled information-theoretic encoding based descriptor to more potential applications, such as optical-infrared face recognition, cross-age face recognition and cross-pose face recognition. We will utilize the proposed descriptor in our face sketch synthesis algorithm. We will test the face sketch synthesis algorithm for more sketch styles and other artistic styles. Some of these extensions are straight-forward, but we believe many exciting problems can be found in this process and it will benefit the future development of the multi-modality computer vision.

□ **End of chapter.**

Bibliography

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 28(12):2037, 2006.

[2] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

[3] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 19(7):711–720, 2002.

[4] P. Benson and D. Perrett. Perception and recognition of photographic quality facial caricatures: Implications for the

recognition of natural images. *European Journal of Cognitive Psychology*, 3(1):105–135, 1991.

[5] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. On matching sketches with digital face images. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[6] M. Blaschko and C. Lampert. Correlational spectral clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] C. Cao, S. Chen, W. Zhang, and X. Tang. Automatic motion-guided video stylization and personalization. In *ACM International Conference on Multimedia*, 2011.

[8] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[9] L. Chang, M. Zhou, Y. Han, and X. Deng. Face sketch synthesis via sparse representation. In *Proc. International Conference on Pattern Recognition*, 2010.

[10] J. Chen, D. Yi, J. Yang, G. Zhao, S. Li, and M. Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[11] S. Chen, Y. Tian, F. Wen, Y. Xu, and X. Tang. EasyToon: an easy and quick tool to personalize a cartoon storyboard using face photos. In *ACM International Conference on Multimedia*, 2008.

[12] L. Cheng, S. Vishwanathan, and X. Zhang. Consistent image analogies using semi-supervised learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[13] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[14] T. Cover and J. Thomas. *Elements of information theory.* John Wiley and Sons, 2006.

[15] Z. Cui, S. Shan, X. Chen, and L. Zhang. Sparsely encoded local descriptor for face recognition. In *IEEE International*

Conference on Automatic Face and Gesture Recognition and Workshops, 2011.

[16] I. Drori, D. Cohen-Or, and H. Yeshurun. Example-based style synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[17] A. A. Efros and W. Freeman. Image quilting for texture synthesis and transfer. In *Proc. annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, 2001.

[18] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[19] W. T. Freeman, J. B. Tenenbaum, and E. C. Pasztor. Learning style translation for the lines of a drawing. *ACM Transactions on Graphics*, 22(1):33–46, 2003.

[20] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *Advances in Neural Information Processing Systems*, 2007.

[21] X. Gao, N. Wang, D. Tao, and X. Li. Face sketch-photo synthesis and retrieval using sparse representation. *IEEE*

Transactions on Circuits and Systems for Video Technology, pages –, 2012.

[22] X. Gao, J. Zhong, J. Li, and C. Tian. Face sketch synthesis algorithm based on E-HMM and selective ensemble. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(4):487–496, 2008.

[23] L. Gibson and D. F. Mills. *Faces of Evil: Kidnappers, Murderers, Rapists and the Forensic Artist Who Puts Them Behind Bars.* Liberty Corner, NJ: New Horizon Press, 2006.

[24] B. Gooch and A. Gooch. *Non-photorealistic rendering.* AK Peters, Ltd. Natick, MA, USA, 2001.

[25] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer, second edition, 2009.

[26] A. Hertzmann. A survey of stroke-based rendering. *IEEE Computer Graphics and Applications*, 23(4):70–81, July–Aug. 2003.

[27] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proc. annual confer-*

ence on Computer graphics and interactive techniques (SIGGRAPH), 2001.

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

[29] G. Hua and A. Akbarzadeh. A robust elastic and partial matching metric for face recognition. In *Proc. International Conference on Computer Vision*, 2009.

[30] T. Huang, S. Chen, J. Liu, and X. Tang. Image inpainting by global structure and texture propagation. In *ACM International Conference on Multimedia*, 2007.

[31] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[32] N. Ji, X. Chai, S. Shan, and X. Chen. Local regression model for automatic face sketch generation. In *International Conference on Image and Graphics (ICIG)*, 2011.

[33] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. International Conference on Computer Vision*, 2005.

[34] T. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions Pattern Analysis and Machine Intelligence*, pages 1415–1428, 2008.

[35] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mugshot photos. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[36] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami. On KANSEI facial image processing for computerized facialcaricaturing system PICASSO. In *Proc. IEEE International Conf. on Systems, Man, and Cybernetics*, 1999.

[37] V. Kwatra, A. Schodl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: image and video synthesis using graph cuts. In *Proc. annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, 2003.

[38] Z. Lei and S. Li. Coupled spectral regression for matching heterogeneous faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[39] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *Proc. European Conference on Computer Vision*, 2008.

[40] D. Lin and X. Tang. Coupled space learning of image style transformation. In *Proc. International Conference on Computer Vision*, 2005.

[41] D. Lin and X. Tang. Inter-modality face recognition. In *Proc. European Conference on Computer Vision*, 2006.

[42] C. Liu, H. Shum, and W. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[43] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[44] W. Liu, X. Tang, and J. Liu. Bayesian tensor inference for sketch-based facial photo hallucination. In *International Joint Conferences on Artificial Intelligence*, 2007.

[45] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[46] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[47] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Advances in Neural Information Processing Systems*, 2007.

[48] U. Park, Y. Tong, and A. Jain. Age-invariant face recognition. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 32(5):947–954, 2010.

[49] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2002.

[50] S. Pizer, E. Amburn, J. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. Romeny, J. Zimmerman, and K. Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[51] G. Ramanarayanan and K. Bala. Constrained texture synthesis via energy minimization. *IEEE Transactions on Visualization and Computer Graphics*, 13(1):167–178, 2007.

[52] R. Resales, K. Achan, and B. Frey. Unsupervised image translation. In *Proc. International Conference on Computer Vision*, 2003.

[53] A. Sharma and D. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[54] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[55] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. International Conference on Computer Vision*, 2003.

[56] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[57] J. Suo, S. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 32(3):385–401, 2010.

[58] X. Tan and B. Triggs. Fusing Gabor and LBP feature sets for kernel-based face recognition. In *Proc. International Conference on Analysis and Modeling of Faces and Gestures*, 2007.

[59] X. Tang and X. Wang. Face photo recognition using sketch. In *ICIP*, 2002.

[60] X. Tang and X. Wang. Face sketch synthesis and recognition. In *Proc. International Conference on Computer Vision*, 2003.

[61] X. Tang and X. Wang. Face sketch recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):50–57, 2004.

[62] J. Tenenbaum and W. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[63] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[64] B. Wang, W. Wang, H. Yang, and J. Sun. Efficient example-based painting and synthesis of 2D directional texture. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):266–277, 2004.

[65] X. Wang and X. Tang. Unified subspace analysis for face recognition. In *Proc. International Conference on Computer Vision*, 2003.

[66] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 26(9):1222–1228, 2004.

[67] X. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 35(3):425–434, 2005.

[68] X. Wang and X. Tang. Random sampling for subspace face recognition. *International Journal of Computer Vision*, 70(1):91–104, 2006.

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

[70] X. Wang, C. Zhang, and Z. Zhang. Boosted multi-task learning for face verification with applications to web image and video search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[71] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*, 2008.

[72] J. Wright and G. Hua. Implicit elastic matching with random projections for pose-variant face recognition. In *Proc.*

IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[73] R. Xiao, H. Zhu, H. Sun, and X. Tang. Dynamic cascades for face detection. In *Proc. International Conference on Computer Vision*, 2007.

[74] J. Yang, D. Zhang, and J. Yang. Is ICA significantly better than PCA for face recognition? In *Proc. International Conference on Computer Vision*, 2005.

[75] M. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[76] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[77] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li. Face matching between near infrared and visible light images. *Advances in Biometrics*, pages 523–530, 2007.

[78] B. Zhang, S. Shan, X. Chen, and W. Gao. Histogram of gabor phase patterns (HGPP): a novel object representation

approach for face recognition. *IEEE Transactions Image Processing*, 16(1):57–68, 2006.

[79] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Li. Face detection based on multi-block LBP representation. *Advances in Biometrics*, pages 11–18, 2007.

[80] W. Zhang, S. Chen, J. Liu, and X. Tang. Style transfer via frequency band decomposition. Technical report, The Chinese University of Hong Kong, 2010.

[81] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In *Proc. International Conference on Computer Vision*, 2005.

[82] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *Proc. European Conference on Computer Vision*, 2010.

[83] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In

Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[84] Y. Zhang, S. Ellyson, A. Zone, P. Gangam, J. Sullins, C. McCullough, S. Canavan, and L. Yin. Recognizing face sketches by a large number of human subjects: A perception-based study for facial distinctiveness. In *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, 2011.

[85] Y. Zhang, C. McCullough, J. Sullins, and C. Ross. Hand-drawn face sketch recognition by humans and a pca-based algorithm for forensic applications. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 40(3):475–485, 2010.

[86] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[87] H. Zhou, Z. Kuang, and K.-Y. K. Wong. Markov weight fields for face sketch synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[88] K. Zhou, P. Du, L. Wang, Y. Matsushita, J. Shi, B. Guo, and H. Shum. Decorating surfaces with bidirectional texture functions. *IEEE Transactions on Visualization and Computer Graphics*, 11(5):519–528, 2005.

[89] Y. Zhou, L. Gu, and H. J. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via Bayesian inference. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[90] J. Zou, Q. Ji, and G. Nagy. A comparative study of local matching approach for face recognition. *IEEE Transactions Image Processing*, 16(10):2617–2628, 2007.