Image Transformation using Limited Reference with Application to Photo-Sketch Synthesis

Wei Bai, Yanghao Li, Jiaying Liu[1] , Zongming Guo
Institute of computer Science and Technology, Peking University, Beijing, China 100871

***Abstract*—Image transformation refers to transforming images from a source image space to a target image space. Contemporary image transformation methods achieve this by learning coupled dictionaries from a set of paired images. However, in practical use, such paired training images are not easy to get especially when the target image style is not fixed. Thus in most cases, the reference is limited. In this paper, we propose a sparse representation based framework of transforming images with limited reference, which can be used for the typical image transformation application, photo-sketch synthesis. In the learning stage, the edge features are utilized to map patches between different style images, thus building the coupled database for dictionary learning. In the reconstruction stage, sparse representation can well preserve the basic structure of image contents. In addition, a texture synthesis strategy is introduced to enhance target-like textures in the output image. Experimental results show that the performance of our method is comparable to state-of-the-art methods even with limited reference, which is very efficient and less restrictive for practical use.**

***Index Terms*—Image transformation, photo-sketch synthesis, sparse representation, dictionary learning, reconstruction**

I. Introduction

Image transformation has been extensively studied in the last decade. Researchers find that many problems in computer vision can be addressed as transforming images from one style to another [1], *i.e.*, cross-style transformation. The input and output image are visually different, but have the same content and are in registration with each other. It is worth pointing out that in this context ”style” means a form of image instead of a specific image type. Typical applications such as image super-resolution [2], photo-sketch synthesis [3], and artistic rendering [4] all belong to this category. Due to the fact that cross-style image transformation is widely demanded, solving this problem is of great practical importance. For example, photo-sketch synthesis is very popular in the forensic area for suspect identification. Besides, it can also help people understand how we observe the same scene with distinctive information physiologically.

We should note that the visual difference can be very large between two styles of images regardless of the fact that they describe the same scene. Therefore it is important to explore the underlying relations between the two styles. Learning approaches can be adopted to learn the underlying relations from paired images so that we can predict the unknown images of target style from images in another style. The mapping relations may be an explicit function mapping from input to output, or implicitly expressed in the model. Various methods are applied to construct such implicit model, based on which existing image transformation methods can be divided into three categories: Bayesian-based methods, subspace learning methods, and sparse representation methods. Bayesian inference framework [5] estimates the output target image in a maximum a posteriori (MAP) way, which maximizes the likelihood term and the prior term simultaneously. Although it allows the flexibility of prior constraint, the model and model parameters can be very complex. A representative subspace learning work [6] assumes that the sketch and corresponding photo share the same linear combination coefficients. As a matter of fact, most methods assume that both sides of the transformation share the same linear combination of weights. The problem is the number of combinations is usually fixed, which is not true in most cases. Sparse representation based methods [3] adaptively decompose the input image and reconstruct it on the learned coupled dictionaries. Mapping relations are built directly on the sparse domain. The simplicity and effectiveness of sparse representation make it popular for image transformation.

However, all of these methods assume the existence of an external paired dataset for training without considering the fact that paired training images are not easy to get. It is often the case that people see a target style image and they want to transform their own image into the same style with the target image. Thus in most situations, the reference is limited, leaving the aforementioned methods helpless.

In this paper, we develop a general sparse representation based method for image transformation with limited reference and apply it to photo-sketch synthesis. Our contributions are threefold.

1) First, we show that when the reference is very limited, *i.e.*, an external coupled database is inaccessible, we can still build a learning database. The edge features are utilized to map patches between different style images, thus building the coupled database for dictionary learning.
2) Second, we propose a general framework for image transformation with limited reference which can be easily extended to current image transformation methods.
3) Third, we integrate the advantages of both sparse representation-based and example-based methods. Sparse representation can well preserve the basic structure of image contents during reconstruction

[1] Corresponding author.

This work was supported by National Natural Science Foundation of China under contract No.61101078, National High-tech Technology R&D Program (863 Program) of China under Grant 2014AA015205 and Beijing Natural Science Foundation under contract No.4142021.

978-1-4799-6139-9/14/$31.00 © 2014 IEEE

and example-based methods provide more target-like textures. Hence, on the basis of sparse representation, we attach more details to the base structure layer with exemplar textures.

At last, the proposed method is applied to photo-sketch synthesis to demonstrate its effectiveness.

The rest of this paper is organized as follows: we address the limited reference problem in Section II, Section III describes how to build a coupled dictionary by mapping between different style images. Then a general framework for image transformation with limited reference is provided in Section IV. Experimental results are shown in Section V. Finally, concluding remarks are given in Section VI.

II. Problem Formulation

Conventional sparse representation based image transformation methods generate a target image y from a corresponding input source x. They are usually composed of two stages, the learning stage and the reconstruction stage. In the learning stage, external image pairs (e.g., low-res and high-res image pairs and photo-sketch pairs) $\{X, Y\}$ are trained based on sparse representation. That is,

$$\begin{aligned} D_x &= \arg\min_{D_x}\{X - D_x\Gamma\}_2^2 + \lambda||\Gamma||_1, \\ D_y &= \arg\min_{D_y}\{Y - D_y\Gamma\}_2^2 + \lambda||\Gamma||_1, \end{aligned} \tag{1}$$

where D_x and D_y are coupled dictionaries, and Γ denotes the sparse coefficients. Specifically, the underlying relations between training image pairs are learned in the sparse domain during training process. Therefore, in the reconstruction stage, an input source image x is first sparse coded in the source style dictionary D_x with coefficient γ, which relates the source style and the target style because they share the same coefficients in the sparse domain. So the output target image corresponding to the input source can be predicted by enforcing the sparse code γ to the target style dictionary D_y, $y = D_y\gamma$.

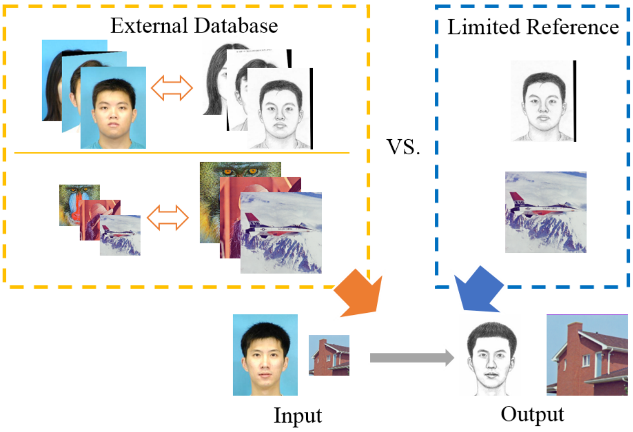

Fig. 1. Simulated scenario vs. real-world scenario. Here we take photo-sketch synthesis and image super-resolution as examples. Regarding the transformation problem itself, most methods assumed existence of an external paired training set. However, we observe that in real-world applications, reference is as limited as in the top-right part of the figure.

As illustrated in Fig.1, the top-left part of the figure describes how external database facilitates the reconstruction process in conventional transformation methods. However, the fact is, in real-world applications, it is often the case that we only have limited reference, *i.e.*, the top-right part in Fig.1. This means if there is a target style image z, we want to transform the input source image x into the same style with z. Frequently, there are no external database like $\{X, Y\}$ to train dictionaries. Plus, z is not necessarily x's counterpart. We know it is critical to build mapping relations between different style images. But in this circumstance, there are no image pairs to train such mapping relations. Thus the problem is how to learn the relation between different styles under the limited-reference situation. We will elaborate on this problem in the following section.

III. How to Match Image Patches of Different Style?

It is very difficult to learn mappings between inconsistent single images unless we can build corresponding training pairs with them. Hence, our solution to the limited-reference problem is creating a paired image patch database from the limited reference. It leaves us to find similar patches between x and z, *i.e.*, the source input image and the target style image. These similar patches, or nearest neighbors (NNs) should meet the following requirement, distributed in different styles but similar in content. For a patch in the source image, we cannot simply search its nearest neighbor in measure of pixel-wise difference because the visual difference can be very large between two styles of images. Therefore we need to develop a feature that can relate the true similar image patches of different style.

Intuitively, we come up with the edge feature to relate two style images. It origins from the observation that edges rarely decay between styles since they are the most important features for visual cognitive tasks. In other words, edges are style-invariant for most instances. Another reason is relative to the a recent work [7], which suggests that it is possible to build a dictionary for edge patches as opposed to intensity patches. It improves the computational speed and memory requirement without noticeable loss of performance.

These facts motivate us to build a dictionary for coupled edge patches with the limited reference. To build edge patch based dictionary, edge preserving filters are applied on reference images. The smoothed image is subsequently subtracted from the original image to obtain an edge image which captures the edges well. Therefore edge features can be used to map patches between different style images for coupled dictionary learning. In this work we suggest using the guided image filter [8]. Let g be the image to be filtered and h be the filtered output. Using the input image as guidance image I, the filtering output at a pixel i can be expressed as a weighted average:

$$h_i = \Sigma_j W_{i,j}(I) g_j, \tag{2}$$

where i and j represent pixel indexes. The filter kernel $W_{i,j}$

is a function of the guidance image, taking the form as:

$$W_{i,j} = \frac{1}{|\omega|^2} \sum_{k:(i,j)\in\omega_k} (1 + \frac{(I_i-\mu_k)(I_j-\mu_k)}{\sigma_k^2+\epsilon}) \tag{3}$$

where μ_k and σ_k denote the mean and the variance of I in a $r \times r$ window ω_k. Pixel k is the center of window ω_k and $|\omega|$ stands for the number of pixels in this window. ϵ is a smoothness parameter. To demonstrate how the filter kernel preserve edges of I in the output, we take a 1-D step edge for example. If I_j and I_i are on the same side of an edge, $(I_i - \mu_k)(I_j - \mu_k)$ in Eq.(3) should have a positive sign. Otherwise, it will be negative. Thus the term $1+\frac{(I_i-\mu_k)(I_j-\mu_k)}{\sigma_k^2+\epsilon}$ in Eq.(3), is large for pixel pairs on the same side of the edge and small for pixel pairs on the different side of the edge. Hence, edges are distinguished by different pixel weights.

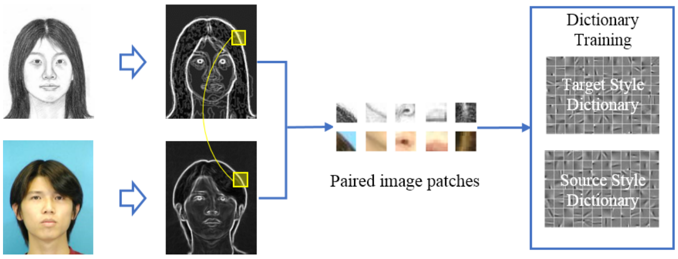

Fig. 2. Edge features are used to map patches between different style images for coupled dictionary learning.

For an input source image x, we use Eq.(2) to get a filtered output x_f. Afterwards, it is subtracted from the original image to obtain an edge image, $x_e = x - x_f$, which captures the edges well. In the same way, we can get the edge image of the target image, z_e.

With the edge images, we can implement the patch matching on them. Mean squared error (MSE) is a widely used distance measure to evaluate similarity of image patches. But the pixel-wise similarity measure is not able to reflect intrinsic image structures, which is a main defect especially when the reference images are just rough edge maps. To solve this problem, we utilize the gradient mean square error, or GMSE for short. Let p be a patch in x_e, and q be a patch in z_e. The similarity criterion $D(p,q)$ is defined as:

$$D(p,q) = ||p-q||_2^2 + \eta||\nabla p - \nabla q||_2^2. \tag{4}$$

where ∇ is the gradient operator and η is a weighting parameter. It emphasizes on both intensity and structure similarity, which is important for our content-oriented matching process. As can be seen from Fig.2, even with single and inconsistent image reference, we are still able to build corresponding image pairs.

Fig.2 depicts the dictionary learning process with the mapped patches. First, the target style image (*e.g.*, a sketch photo) and the source input image are used to generate edge features with guided image filtering. Then, similar patches of two style images are mapped in measure of gradient MSE on edge features. At last, coupled dictionaries are learned on the paired image patches.

IV. The General Framework for Image Transformation

After we trained coupled dictionaries with limited reference, we show how our proposed algorithm can be modified to work with conventional cross-style transformation methods. Apart from the dictionary learning part, the remaining reconstruction stage has been explored by various methods [2], [3]. These methods devoted in finding the complex mapping function between styles. Undoubtedly, more accurate mappings lead to better reconstruction performance. So the basic structures of the reconstructed images are well preserved. However, we can not deny that the sparse representation process is compromised of some details for approximate solution. It explains the fact that sparse representation based methods smoothed more details than texture synthesis methods.

Considering this issue, we propose a general framework for image transformation, which integrates the advantages of both sparse representation based methods and synthesis based methods.

Fig.3 is the work-flow of the proposed framework. As can be seen, two kinds of reconstruction contribute to the final result. On the basis of the learned coupled dictionaries, the input source image is sparse coded to get a sparse coefficient vector. Then the coefficient vector is multiplied by the target style dictionary, recovering the structural layer of the to-be-transformed image. On the other hand, texture features are extracted from the target style image and then synthesized to the structure layer. In this way, we incorporate the benefits of both methods.

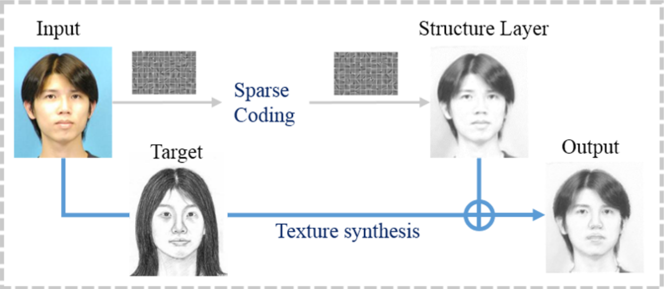

Fig. 3. Two-step image reconstruction of the general image transformation framework.

Specifically, we just use the plain sparse representation framework here. But the fact is either the semi-coupled dictionary learning framework in [3] or the coupled dictionary based framework [2] can replace the sparse representation part and work well. We even test linear embedding method framework on this framework and it turns out that our method is very adaptive and flexible.

As for the synthesis method, we use a simple texture transfer method in [9]. Texture transfer, as it indicates, transfer the sample texture to the target image. This is done by considering how well a block corresponds to the target image. In terms of the Markov model, this means adding more information to each state (location in the synthetic image). In addition

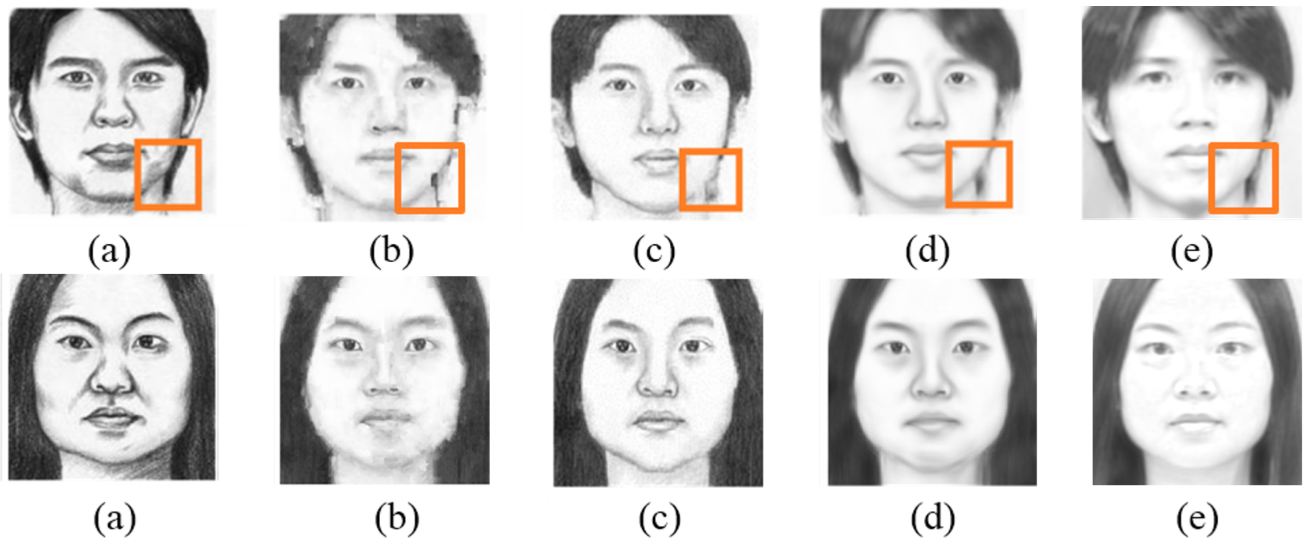

Fig. 4. Subjective experimental results. (a) Ground truth. (b) Wang et al.'s method [5] without MRF. (c) Wang et al.'s method [5] with MRF. (d) SCDL. (e) The proposed method.

to finding blocks that matches well in the region of overlap, blocks must also have high correspondence with the target image. In this way, textures are attached to the structural layer.

V. Experimental Results

To evaluate the efficiency of the our method, the proposed framework is applied to the typical cross-style transformation application, photo-sketch synthesis. Here we conduct photo-sketch face synthesis on the CUFS Database [1]. Sketches which are often drawn by artists have significantly different appearance from the original photos.

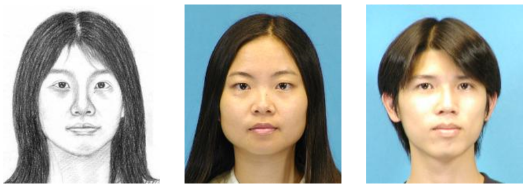

Fig. 5. Target image and test input source images from the CUFS database.

We test the proposed method on photo images with a sketch image as target (see Fig.5, the left one serves as target and the right ones are input images). For each input, a coupled edge patch based dictionary is learned separately. The patch size is 7×7 and the overlap between patches is [5, 5]. We compare our method with two state-of-the-art methods, Wang et al.'s method [5] and SCDL [3].

Fig.4 shows subjective results on two test photo images. Note the highlighted part in the comparing images. In contrast with the fully referenced methods, our algorithm achieves similar results with SCDL and outperforms Wang *et al.*'s method as well. Note that, owing to the texture transfer process, the reconstructed image by the proposed method presents more target-like details.

VI. Conclusion

In this paper, we proposed a sparse representation based framework of transforming images using limited reference and applied it to photo-sketch synthesis. In the learning stage, we managed to build the coupled database for dictionary learning by mapping patches between different style images. In the reconstruction stage, a texture synthesis strategy is introduced to improve the basic structure image layer. Experimental results show that the performance of our method is comparable to state-of-the-art methods even with limited reference.

References

[1] K. Jia, X. Wang, and X. Tang, "Image transformation based on learning dictionaries across image spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 367–380, Feb. 2013.

[2] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[3] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super resolution and photo-sketch synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Sep. 2012, pp. 2216–2223.

[4] B. Wang, W. Wang, and H. Yang, "Efficient example-based painting and synthesis of 2d directional texture," *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, no. 3, pp. 266–277, 2004.

[5] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[6] Q. Liu and X. Tang, "A nonlinear approach for face sketch synthesis and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2005, pp. 1005–1010.

[7] H. Bhujle and S. Chaudhuri, "Novel speed-up strategies for non-local means denoising with patch and edge patch based dictionaries," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 356–365, Jan. 2014.

[8] K. He, J. Sun, and X. Tang, "Guided image filtering," in *European Conference on Computer Vision*. Springer, 2010, pp. 1–14.

[9] A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of SIGGRAPH*, 2001, pp. 341–346.

[1] http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html