Joint-Domain Unsupervised Stylization for Portraits

Saboya Yang, Jiaying Liu*, Shuai Yang, Wenhan Yang and Zongming Guo
Institute of Computer Science and Technology, Peking University, Beijing, P. R. China, 100871

***Abstract*—People wish to own a portrait painting of themselves by Da Vinci. Unfortunately, it is impossible to make this dream come true; nevertheless, it may give us an opportunity by transferring some artistic features from one single reference painting. To address this issue, we propose a joint-domain image stylization approach, particularly for portrait oil paintings. From the view of artistic appreciation, we analyze an amount of oil painting artworks and summarize three critical factors to depict the figure, *i.e.* color, structure and texture. First, the tone of the input image is recolored based on semantic regions corresponding to the reference. Those semantic regions are segmented automatically via the color swatch, by considering the constraints of colors and positions. Then, we exploit sparse representation to reconstruct the layout by acquiring the structure from the reference. The paired training set for sparse dictionary learning is built with the guidance of edge features. Third, considering that texture is usually locally stochastic but regularly repetitive in global, a coarse-to-fine texture synthesis is used to enhance the detail pattern. Subjective results demonstrate the proposed method achieves desirable results compared with state-of-art methods while keeping consistent with artist's style.**

I. Introduction

Nowadays it has been a quite popular trend to share photos through the social media. At the same time, most people prefer uploading photos with special styles generated by apps such as Facebook and Instagram instead of the original ones. This kind of technology, so called *Image Stylization*, makes a more dramatic impression and inspires new creativity.

Image stylization has been a hot spot in both the academia and the industry. In 1998, Hertzmann [4] first utilized strokes to represent image features and incrementally composed virtual strokes to transfer images into artistic styles. On the basis of this stroke-based method, Zhao and Zhu [12] built an abstract painter to simulate brush strokes of oil paintings and reproduced an image in the oil painting style. However, these methods only take advantage of the predefined general information and apply it to the whole image uniformly, which cannot reflect features of multiple techniques in one painting.

To solve a more general case of image stylization, Jia *et al.* [5] came up with doing mappings in cross-style feature spaces. Wang and Tang [11] decomposed images into patches and learned a joint photo-sketch model by a multi-scale Markov Random Fields (MRF) model. Wang *et al.* [10] presented a semi-coupled dictionary learning method to reconstruct the sketch under sparsity constraints. But these learning based methods are on account of a paired training set. Thus, it is hard to apply these methods for handling the case of only one reference, called *unsupervised style transfer*, which is common when taking art works as reference.

Therefore, recent works aim to jointly represent styles and build the mappings across domains for the case of the unsupervised style transfer, with only one reference stylish image. Sunkavalli *et al.* [9] utilized a multi-scale technique to transfer the appearance of one image to another and composited a harmonized image. But this method needed unified masks for harmonization, which are hard to get and lead to the limitation of application scenarios. There is also a new implementation on neural networks to simulate artistic styles [3]. It obtained representations by convolution networks, which may work inefficiently when the reference does not have obvious streamline textures.

In this paper, we first separately represent an art work in three domains: *structure*, *texture* and *color*. Then, considering the intrinsic properties of these three domains and following an artistic creation route, we propose the corresponding approaches to model and map the features in these domains jointly: 1) color is usually region-consistent, thus a local color transfer method is employed to adjust auto-segmented semantic regions; 2) the main structures of the input image are usually deterministic and contain salient features, such as corners or sharp edges, thus it is maintained by sparse reconstruction; 3) texture is usually locally stochastic but regularly repetitive in global, thus a coarse-to-fine texture synthesis is used separately to bring out more brushwork.

In conclusion, the contributions of this paper are as follows:

- Motivated by the artistic appreciation, portrait oil paintings are presented by features in three domains: *color*, *structure* and *texture*. We propose to follow an art creation route and jointly transfer the features in these domains.
- For color domain, we propose an automatic swatch-based color adjustment that transfers colors between semantic swatches with spatial information.
- For structure domain, fundamental structures acquired from the reference are reconstructed by sparse representation, where the coupled dictionary is trained on the coupled patch set created by edge feature correspondences.

The rest of this paper is organized as follows. The proposed multi-scale portrait oil painting stylization approach is elaborated in Section II. Experimental results are presented in Section III. Concluding remarks are given in Section IV.

II. Joint-Domain Unsupervised Portrait Oil Painting Stylization Method

Some typical features make great art works unique and represent the style of a certain artist. To model those styles, we

*Corresponding author
This work was supported by National Natural Science Foundation of China under contract No.61671025.

978-1-4673-6853-7/17/$31.00 ©2017 IEEE

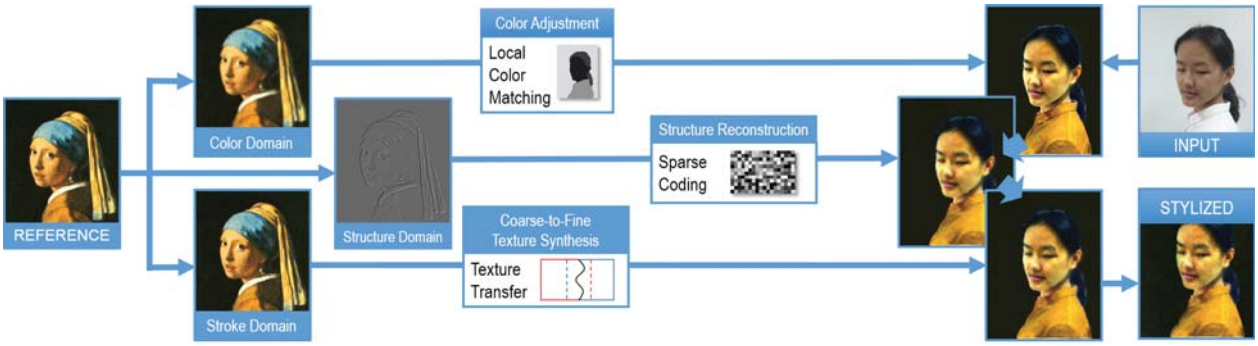

Fig. 1. Framework of the proposed joint-domain portrait oil painting stylization algorithm.

analyze amounts of oil portrait paintings. Through the analysis, three key factors appear to play key roles in affecting art rendering and visual feeling, *i.e.* color, structure and texture. To transfer the desirable styles of the reference portrait oil painting to the input, we propose to decompose the painting into features of these three domains and transfer them jointly to simulate the process of the artistic creation. The framework of the proposed stylization method is illustrated in Fig. 1.

A. Swatch-Based Color Adjustment

To differentiate distinct elements in the frame, artists decide which pigment to use before stroking. The transfer in the color domain is, therefore, the first operation to be applied to the input portrait to adjust the color style of the input to get closer to the reference's tone.

$l\alpha\beta$ color space [6], an orthogonal color space is utilized for color adjustment to avoid distortions. In most cases, the global color transfer method does not work well in this portrait stylization scenario, especially when the person is staying in a complex background in the input. To remedy this issue, we propose a local color transfer method via $l\alpha\beta$ color space.

In order to deal with local regions respectively, segmentation is needed firstly. Hence, GrabCut [7] is conducted to find a binary mask and distinguish the figure from the background. There are usually three main semantic components of the figure in a common portrait: hair, face and clothing, which have to be modified separately to obtain a better-colorized result. However, it is difficult to do partitions directly during color transfer without destroying details. To solve this problem, we utilize the template obtained by nonparametric representation [8] to detect facial landmarks. With these landmarks and the binary mask, spatial relations are provided to obtain three swatches relating to the three regions automatically in Fig. 2. Then the figure in the portrait clusters to those swatches by the normalized feature vector $\mathbf{f} = \{R, G, B, \alpha, \beta\}$ as Equation (1). In the feature vector $\mathbf{f}$, $\{R, G, B\}$ features are utilized to distinguish different colors while $\{\alpha, \beta\}$ features play a role in avoiding abnormities caused by luminance.

$$\min\left(\sum_{x,y}\sqrt{(\mathbf{f}^p_{x,y}-\mathbf{f}^s_{\mathbf{w}_i})^T\Gamma(\mathbf{f}^p_{x,y}-\mathbf{f}^s_{\mathbf{w}_i})}\right), \qquad (1)$$

where $\mathbf{w}_i$ defines the center pixel position of the i-th swatch and pixel (x, y) belongs to the corresponding region Ω_i. $\mathbf{f}^p_{x,y}$ is the feature vector of pixel (x, y) while $\mathbf{f}^s_{\mathbf{w}_i}$ is the mean feature vector of the i-th swatch. Γ controls the balancing weight. According to clustering, the portrait is segmented into three regions besides background: hair, face and clothing.

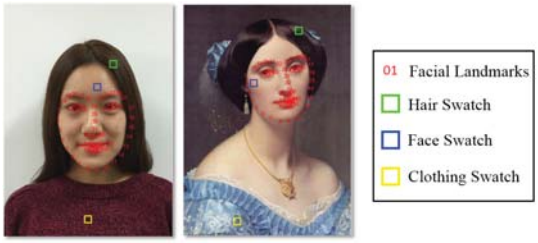

Fig. 2. The swatches in the input and the reference are obtained based on the binary mask and facial landmarks. Different colors are used to represent different swatch pairs.

Since relevant regions have to be mapped between the input I^S and the reference I^T to transfer the corresponding color styles, the segmentation is executed on both images. Then, with statistics of three cluster pairs, the color of each pixel $I^S_{x,y}$ is shifted in the $l\alpha\beta$ color space depending on its nearest cluster centers and the tone of the reference is therefore transferred to the input.

B. Structure Reconstruction via Sparse Representation

With the desirable colors, artists paint the fundamental structure of the painting to determine the basic layout. The features in the structure domain are utilized here to maintain the structure of the input portrait.

Traditional sparse representation based methods are relied on a paired training set, several paired images in homologous uniform styles of the input and the reference. However, in this unsupervised case, the paired training set is absent. Therefore, we need to build mappings in the structure domain between the input and the reference for dictionary learning. As the input and the reference describes different people, the cross-style mappings cannot be built directly. The edge feature [1], which represents structure information, is style-invariant upon most occasions and manipulated to relate the corresponding patches together and build the training set. We take advantage of the pyramid to offer edge features.

To be more specific, the pyramid is constructed by filtering the image with a set of linear filters, which refers to Haar filter

here. The i-th subband B_i^C and B_i^T of the colorized input I^C and the reference I^T are calculated by the i-th linear filter. The established pyramid with n levels has three subbands at each level, and then leaves a low-pass residue B_{3n+1}^T. We notice that subtracting the residue from the original image produces the edge map $(I^T - B_{3n+1}^T)$ so that edges patches ready for matching are isolated in Fig. 3.

Edge patches e^C and e^T from the input edge map $(I^C - B_{3n+1}^C)$ and the reference edge map $(I^T - B_{3n+1}^T)$ are matched together based on the patch similarity $\delta(e^C, e^T)$, which is evaluated by both the intensity and structure similarities.

$$\delta(e^C, e^T) = \|e^C - e^T\|_2^2 + \tau\|\nabla e^C - \nabla e^T\|_2^2, \tag{2}$$

where τ defines a weighting parameter and set as 0.5 in this paper. ∇ is the gradient operator.

To summarize, with the guidance of edge features, the paired training set $\{C, T\}$ is acquired. Then inspired by [10], sparse coding is managed to learn the ultimate mapping relations. Traditional sparse representation based method assumes that the coupled dictionaries D^C and D^T strictly share the same sparse coefficients α for each patch pair. However, this assumption is too strong for cross-style patches, and we loose the assumption to admit that there exists a stable linear mapping W between the corresponding sparse coefficients α^C and α^T. The sparse dictionary learning problem is formulated as the following ridge regression problem in Eq. (3).

$$\begin{aligned} &\min_{\{D^C, D^T, W\}} \|C - D^C\alpha^C\|_F^2 + \|T - D^T\alpha^T\|_F^2 \\ &+\varphi\|\alpha^T - W\alpha^C\|_F^2 + \lambda^C\|\alpha^C\|_1 + \lambda^T\|\alpha^T\|_1 + \lambda^W\|W\|_F^2, \\ &\qquad s.t. \quad \|d_j^C\|_2^2 \le 1, \|d_j^T\|_2^2 \le 1, \forall j, \end{aligned} \tag{3}$$

where d_j^C and d_j^T are the j-th dictionary atoms of the coupled dictionaries D^C and D^T. φ, λ^C, λ^T and λ^W refer to regularization parameters to balance different terms. Then the cross-style mapping W and the coupled dictionaries D^C and D^T are learned iteratively to optimize the variables alternatively.

Moreover, the image can be reconstructed with the learned dictionary. Taking k-th patch p_k^C in the colorized input portrait I^C as an example, the corresponding patch p_k^R in the structure-maintained output I^R can be reconstructed through the following optimization equation. The parameter φ, λ^C and λ^T are correspondingly set as 0.05, 0.01 and 0.001.

$$\begin{aligned} &\min_{\{\alpha_k^C, \alpha_k^R\}} \|p_k^C - D^C\alpha_k^C\|_F^2 + \|p_k^R - D^T\alpha_k^R\|_F^2 \\ &+\varphi\|\alpha_k^R - W\alpha_k^C\|_F^2 + \lambda^C\|\alpha_k^C\|_1 + \lambda^T\|\alpha_k^R\|_1. \end{aligned} \tag{4}$$

To solve the equation, we first sparsely code patch p_k^C on the dictionary D^C to obtain sparse coefficients α_k^C. Then patch p_k^R is initialized by enforcing sparse coefficients α_k^R with the mapping W to the reference dictionary D^T as $D^T W\alpha_k^C$. In the end, the structure-maintained output image I^R is predicted by the traversal and overlap of all patches p_k^R as $p_k^R = D^T\alpha_k^R$.

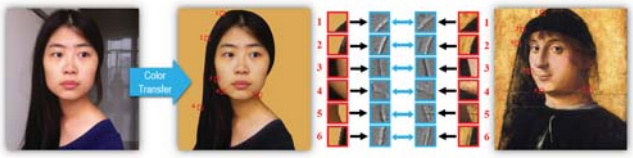

Fig. 3. Edge features are used to map similar patches between different styles for coupled dictionary learning.

C. Coarse-to-Fine Texture Synthesis

After settling down the fundamental structures of the painting, artists draw with various brushes and creative strokes. Therefore, we analyze the features in the texture domain of the reference to accomplish the stroke texture synthesis in the structure-reconstructed input.

In this paper, we tend to manipulate the input portrait by patches in raster scanning [2]. For each input patch, we search for a set of suitable reference patches considering the intensity similarity and the normalized distance. With the candidate patch set, we seek a candidate patch which can fit in with its neighbors best to paste into the result. Then the simulated image with stroke textures I^E is blended with the structure-maintained image I^R through alpha matting to obtain the stroke-confluent result I^K.

At the same time, it is noticed that when artists are drawing with brushes on the drawing surface, there exist some fine-grained textures from the surface which cannot be completely covered by strokes. Enlightened by Sunkavalli's work [9], these fine-grained textures are regarded as noises and imitated by pyramid matching. After pyramid matching, it comes out that the fine-grained textures induced by the surface have been injected to the input.

III. Experimental Results

To evaluate the effectiveness of the proposed method, we conduct experiments on several test image pairs. In experiments, we import an original portrait photo as the input to be stylized and a portrait oil painting as the reference. The reference portrait oil painting set is collected from the Internet. We also find and take some portrait photos as the input to test the algorithm. The patch size is 5×5.

In the meanwhile, the background also plays an important role in the mood of a portrait. But it turns out that artists usually design the background instead of painting the authentic one in real life when drawing portraits. Thus in experiments, the previously obtained mask is utilized to extract the background and directly replace the background of the synthetic portrait through Poisson image editing. When the reference mask and the input mask cannot overlap perfectly, image inpainting is utilized to extrapolate the missing area.

We compare the proposed algorithm with Sunkavalli's method [9] and Gatys' method [3]. Then subjective results are illustrated in Fig. 4. In Fig. 4, our method transforms the reference style to the input containing colors and coarse-to-fine textures while preserving the original structures. Sunkavalli's method [9] harmonizes the figure into the reference seamlessly, but also leads to the darkness of the figure and the loss of some

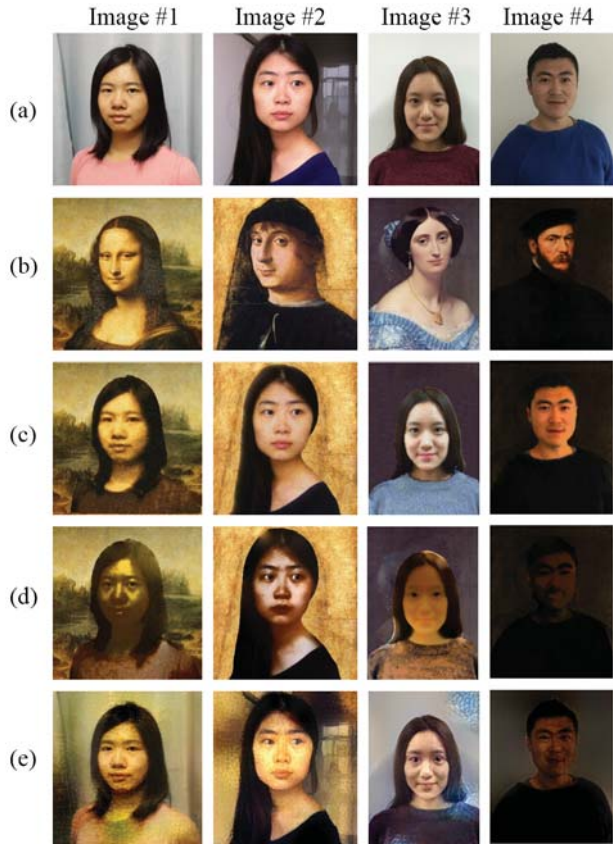

Fig. 4. Subjective experimental results. (a) Original input stylish image. (b) Reference image. (c) Stylized image by the proposed method. (d) Stylized image by Sunkavalli's method [9]. (e) Stylized image by Gatys' method [3].

details. In the meanwhile, neural networks based method [3] works sensitively to conspicuous patterns but is inefficient on relatively smooth oil paintings.

In the meanwhile, to ensure the credibility of our method, we invited 40 observers with diversity to fill in our questionnaires. The number of female observers is 20 while the age varies from 18 to 63. The distinct professions are computer, design, finance and so on. To verify the fairness, the orders of results randomly change every round. Observers are asked to score the similarity of styles between the reference and the stylized image from 5 to 1, from the most similar to the least similar. According to the statistics in Table I, our method acquires the highest score in each round of Fig. 4. We also ask observers to choose the best method which transfers the most similar style every round. 75% observers in average agree that our method works the best in Fig. 5. It signifies the proposed method produces fairly similar and visually desirable stylized images. It could not only transfer styles from references but also keep consistent contents with the given portrait.

TABLE I
SCORES OF DIFFERENT METHODS

Images	Proposed	Sunkavalli's	Gatys'
Image #1	**4.38**	2.00	3.18
Image #2	**3.95**	2.95	3.35
Image #3	**3.73**	1.60	2.85
Image #4	**3.60**	2.08	3.18
Average	**3.92**	2.16	3.14

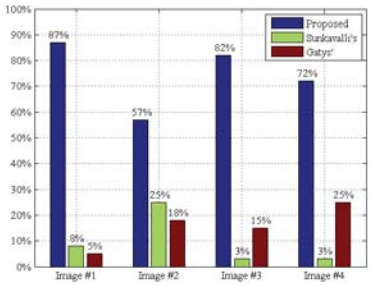

Fig. 5. Statistics of votes for the method produces the most similar style.

IV. CONCLUSIONS

In this paper, we propose a joint-domain method to stylize portraits into a customized oil painting style. Each of these three unique factors, color, structure, and texture, plays an important role in the expression of the painting and is applied to the input in order to imitate the reference's style. Hence, the color style and coarse-to-fine textures are blended to the input while the structure of the input maintains. Experimental results indicate that our proposed method outperforms state-of-art algorithms in most people's eyes.

REFERENCES

[1] 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*, 23(1):356–365, 2014. 2

[2] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 341–346. ACM, 2001. 3

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 3, 4

[4] A. Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, pages 453–460. ACM, 1998. 1

[5] K. Jia, X. Wang, and X. Tang. Image transformation based on learning dictionaries across image spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):367–380, 2013. 1

[6] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 2

[7] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004. 2

[8] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *IEEE 12th International Conference on Computer Vision*, pages 1034–1041. IEEE, 2009. 2

[9] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics*, 29(4):125, 2010. 1, 3, 4

[10] 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*, pages 2216–2223. IEEE, 2012. 1, 3

[11] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. 1

[12] M. Zhao and S. C. Zhu. Sisley the abstract painter. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, pages 99–107. ACM, 2010. 1