Semi-Coupled Dictionary Learning with Applications to Image Super-Resolution and Photo-Sketch Synthesis

Shenlong Wang[1,2] Lei Zhang[2,*] Yan Liang[1] Quan Pan[1]
[1]Northwestern Polytechnical University, [2]The Hong Kong Polytechnic University
shenlong.wang@gmail.com, cslzhang@comp.polyu.edu.hk, liangyan@nwpu.edu.cn, quanpan@nwpu.edu.cn

Abstract

In various computer vision applications, often we need to convert an image in one style into another style for better visualization, interpretation and recognition; for examples, up-convert a low resolution image to a high resolution one, and convert a face sketch into a photo for matching, etc. A semi-coupled dictionary learning (SCDL) model is proposed in this paper to solve such cross-style image synthesis problems. Under SCDL, a pair of dictionaries and a mapping function will be simultaneously learned. The dictionary pair can well characterize the structural domains of the two styles of images, while the mapping function can reveal the intrinsic relationship between the two styles' domains. In SCDL, the two dictionaries will not be fully coupled, and hence much flexibility can be given to the mapping function for an accurate conversion across styles. Moreover, clustering and image nonlocal redundancy are introduced to enhance the robustness of SCDL. The proposed SCDL model is applied to image super-resolution and photo-sketch synthesis, and the experimental results validated its generality and effectiveness in cross-style image synthesis.

1. Introduction

In many computer vision and pattern recognition applications, people often have images of the same scene but obtained from different sources, and consequently the conversion between the images of different styles are required. For example, in law enforcement we may need to compare mug-shot photos to a sketch drawn by an artist based on the verbal description of the suspect. In addition, a low resolution image/video captured by low-end devices often needs to be up-converted to a higher resolution for better visualization and interpretation. Researches on such cross-style image synthesis problems can not only benefit the practical applications (e.g., public security) but also help people understand how the human visual system perceives the distinctive information of the same scene across different sources.

In the past decades, image cross-style synthesis and recognition have been attracting much attention. Since the images under different styles, even describing the same scene, can be very different, how to reveal the underlying relations between the two styles is the key issue to be studied. In order to predict the unknown images in one style from their counterparts in another style, statistical learning approaches can be adopted to learn the underlying mapping from example image pairs. Many image processing and computer vision tasks can be considered as a cross-style image synthesis problem, such as image super-resolution [12, 15, 29, 21], artistic rendering [11, 8, 16], photo-sketch synthesis [24, 26] and multi-modal biometrics [13, 22, 26], *etc*. Various methods have been proposed to solve one of the above mentioned tasks by using patch-based matching [11, 26], coupled subspace learning [13, 16] and coupled dictionary learning [27] techniques, *etc*. However, most of these methods are limited in finding the complex mapping function between styles, as well as limited in reconstructing the style-specific local structures in the conversion process.

In this paper, we propose a simple yet more general model to solve the cross-style image synthesis problems. Specifically, we learn a dictionary pair and a mapping function simultaneously. The pair of dictionaries aims to characterize the two structural domains of the two styles, and the mapping function is to reveal the intrinsic relationship between the two styles for synthesis. In the learning process, the two dictionaries will not be fully coupled, allowing the mapping function much flexibility for accurate synthesis across styles. We call the proposed model semi-coupled dictionary learning (SCDL), and apply it to image super-resolution and photo-sketch synthesis to validate its performance.

In real-world data, the mappings between different styles can be complex, spatial-variant and nonlinear. It is not sufficient to use a single mapping to describe the complex relationship between different image styles. In order to improve the robustness and stability of SCDL, we propose a new model selection (clustering) method and integrate it in-

*Corresponding author. Email: cslzhang@comp.polyu.edu.hk

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

to SCDL. The model selection can effectively separate data into different clusters so that in each cluster a stable linear mapping between the two styles can be learned. Different from the previous methods which do clustering in the signal domain, the proposed model selection performs clustering in the style-specific sparse domains, aiming at enhancing the style conversion capability.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 presents the SCDL framework. Section 4 presents the algorithm. Section 5 conducts experiments and Section 6 concludes the paper.

2. Related Works

Various cross-style image synthesis problems, such as image analogies [11], texture synthesis [8, 16], and multimodal face recognition [26, 13, 22], have been proposed and studied. In this paper, we focus on the problems of image super-resolution and photo-sketch synthesis, and thus we mainly review the methods on these two applications.

Image super-resolution aims to reconstruct a high resolution (HR) image from its low resolution (LR) counterpart. There are mainly two categories of super-resolution methods. In the first category, the LR image is down-sampled from a blurred version of the HR image [10]. The blurring kernel is known (or can be estimated) and used in the HR image reconstruction process. This is basically an inverse problem with an imaging model available. In the second category, often the LR image is modeled as the directly down-sampled version of the HR image. We consider the second case in this paper, and the super-resolution problem can be viewed as an image interpolation problem [12, 29, 15, 21]. Many image interpolation methods, including the classical bi-cubic interpolator [12] and the edge guided interpolators [15, 29], interpolate the missing HR pixel as the weighted average of its local neighbors. The difference between these methods lies in how the weights are determined. In [29] the autoregressive model is used to exploit the image local correlation for effective image interpolation. In [21], a series of linear inverse estimators of HR image are computed based on different priors on the image regularity. These estimators are then mixed in a frame over patches of coefficients, providing a sparse signal representation under l_1-norm minimization weighted by the signal regularity in each patch.

In law enforcement, we may have to compare mug-shot photos to a sketch drawn by an artist based on the verbal description of the suspect. In addition, since near infrared (NIR) imaging is robust to illumination changes, it is often used in outdoor face image acquisition, and matching face images under NIR and visible lights is necessary. Tang and Wang [24] used eigentransform to learn mappings between different image styles. Their method is based on two important assumptions: transformation between different styles can be approximated as a linear process, and faces can be reconstructed from training samples by PCA. This method works well in face hallucination [25]. However, due to the limitations of PCA, the two assumptions can hardly hold for image styles between which the mappings are highly nonlinear. Another family of cross-style image modeling methods is to construct a hidden subspace [22, 13]. This subspace aims to maximize correlations of different image styles so that images of different styles projected into the subspace are highly correlated. One representative work is canonical correlated analysis, which has been well used in multi-modal face recognition tasks [14]. However, canonical correlated analysis aims at preserving correlation or discriminative information instead of reconstructive information, and it may not lead to highly accurate image reconstruction across styles. To overcome these drawbacks, Lin and Tang [16] proposed a novel coupled subspace learning strategy to learn image mappings between different styles. They first utilized correlative component analysis to find the hidden spaces for each style to preserve correlative information, and then learned a bidirectional transform between two subspaces.

Natural image patches could be sparsely represented by an over-complete dictionary of atoms. Recently, sparse coding (or sparse representation) and dictionary learning have proven to be very effective in image reconstruction [20, 9, 6, 5], while the dictionary plays an important role to successfully accomplish such tasks. Learning a dictionary from example image patches has been attracting much interest, and some representative methods have been proposed, such as K-SVD [1], supervised dictionary learning [19], online dictionary learning [17], *etc*. In [27], Yang *et al*. used a coupled dictionary learning model for image super-resolution. They assumed that there exist coupled dictionaries of HR and LR images, which have the same sparse representation for each pair of HR and LR patches. After learning the coupled dictionary pair, the HR patch is reconstructed on HR dictionary with sparse coefficients coded by LR image patch over the LR dictionary. In our proposed SCDL, this strong regularization of "same sparse representation" is relaxed for cross-style image synthesis, and a more stable cross-style mapping can be learned in the sparse domain.

3. Semi-coupled dictionary learning

3.1. Problem formulation

The image cross-style synthesis problem can be formulated as follows: given an image $\mathbf{x}$ of style s_x, how to recover the associated image $\mathbf{y}$ of style s_y of the same scene? The difficulties of this kind of problems vary with image styles. Suppose that all the images in style s_x form a space $\mathcal{X}$ and images in style s_y form a space $\mathcal{Y}$, and there exists

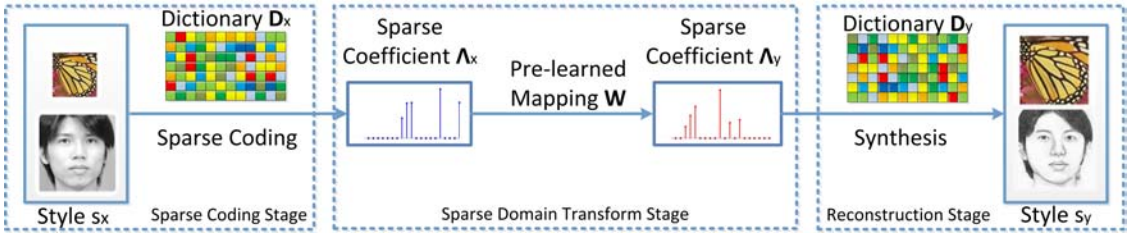

Figure 1. Flowchart of the proposed semi-coupled dictionary learning (SCDL) based image cross-style synthesis.

a mapping $f(\cdot)$ from $\mathcal{X}$ to $\mathcal{Y}$: $\mathbf{y} = f(\mathbf{x})$. If the mapping is invertible and known, we can simply transform between $\mathbf{x}$ and $\mathbf{y}$. Unfortunately, in most cases this kind of transform is invertible and hard to learn directly.

Since each pair of images indicate the same scene, it is reasonable to assume that there exists a hidden space where the styles can be converted to each other. Therefore, some coupled subspace/dictionary learning methods [16, 27] have been proposed, and they assume that in the coupled subspace the representation coefficients of the image pair should be strictly equal. However, this assumption is too strong to address the flexibility of image structures in different styles. In this paper, we relax this assumption and assume that there exists a dictionary pair over which the representations of two styles have a stable mapping. Since the pair of dictionaries is not required to be fully coupled, we call the proposed method semi-coupled dictionary learning (SCDL). In SCDL, we employ dictionaries to seek for the structural hidden spaces and the mapping. Once the dictionary pair and mapping are learned, cross-style image synthesis can be performed, and the synthesis procedures are illustrated in Fig. 1.

Denote by $\mathbf{X}$ and $\mathbf{Y}$ the training datasets formed by the image patch pairs of styles s_x and s_y. We propose to minimize the energy function below to find the desired semi-coupled dictionaries as well as the desired mapping:

$$\begin{aligned}&\min_{\{\mathbf{D}_x,\mathbf{D}_y,f(\cdot)\}} E_{data}(\mathbf{D}_x,\mathbf{X}) + E_{data}(\mathbf{D}_y,\mathbf{Y})\\ &+\gamma E_{map}(f(\mathbf{\Lambda}_x),\mathbf{\Lambda}_y) + \lambda E_{reg}(\mathbf{\Lambda}_x,\mathbf{\Lambda}_y,f(\cdot),\mathbf{D}_x,\mathbf{D}_y)\end{aligned} \tag{1}$$

where $E_{data}(\cdot,\cdot)$ is the data fidelity term to represent data description error, $E_{map}(\cdot,\cdot)$ is the mapping fidelity term to represent the mapping error between the coding coefficients of two styles, and E_{reg} is the regularization term to regularize the coding coefficients and mapping. Note that in the proposed model, the coding coefficients of $\mathbf{X}$ and $\mathbf{Y}$ over $\mathbf{D}_x$ and $\mathbf{D}_y$ will be related by a mapping $f(\cdot)$. The two dictionaries ($\mathbf{D}_x$ and $\mathbf{D}_y$) and the mapping function $f(\cdot)$ will be jointly optimized.

One special but important case is that the mapping $f(\cdot)$ is linear, and then the framework in Eq. 1 can be turned into the following dictionary learning and ridge regression problem:

$$\begin{aligned}&\min_{\{\mathbf{D}_x,\mathbf{D}_y,\mathbf{W}\}} \|\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x\|_F^2 + \|\mathbf{Y}-\mathbf{D}_y\mathbf{\Lambda}_y\|_F^2\\ &+\gamma\|\mathbf{\Lambda}_y-\mathbf{W}\mathbf{\Lambda}_x\|_F^2 + \lambda_x\|\mathbf{\Lambda}_x\|_1 + \lambda_y\|\mathbf{\Lambda}_y\|_1 + \lambda_W\|\mathbf{W}\|_F^2\\ &\text{s.t. } \|\mathbf{d}_{x,i}\|_{l_2}\le 1, \|\mathbf{d}_{y,i}\|_{l_2}\le 1, \forall i\end{aligned} \tag{2}$$

where $\gamma, \lambda_x, \lambda_y, \lambda_W$ are regularization parameters to balance the terms in the objective function and $\mathbf{d}_{x,i}, \mathbf{d}_{y,i}$ are the atoms of $\mathbf{D}_x$ and $\mathbf{D}_y$, respectively. The objective function in Eq. 2 is not jointly convex to $\mathbf{D}_x, \mathbf{D}_y, \mathbf{W}$. However, it is convex w.r.t. each of them if others are fixed. Therefore, we can design an iterative algorithm to alternatively optimize the variables. In [27], the mapping transform $\mathbf{W}$ is predefined as an identity matrix and the coding coefficients $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ are assumed the same. This model actually approximates $f(\cdot)$ as a conformal mapping on the coupled dictionaries. However, for complex data with invertible mapping, this model is limited to reconstruct the image structures across different styles. In comparison, our proposed SCDL model relaxes the coupling of dictionaries by allowing mapping errors between coding coefficients.

3.2. Training

To tackle the energy-minimization in Eq. 2, we separate the objective function into 3 sub-problems, namely sparse coding for training samples, dictionary updating and mapping updating. First, we need to initialize the mapping $\mathbf{W}$ and dictionary pair. $\mathbf{W}$ can be simply initialized as the identity matrix. There are many ways to initialize $\mathbf{D}_x$ and $\mathbf{D}_y$ such as random matrix, PCA basis, DCT basis, *etc*. Using l_1-minimization, the sparse codes $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ can then be obtained. Note that mapping by $\mathbf{W}$ is assumed to be linear, and the bidirectional transform learning strategy can be adopted to learn transforms from $\mathbf{\Lambda}_x$ to $\mathbf{\Lambda}_y$ and from $\mathbf{\Lambda}_y$ to $\mathbf{\Lambda}_x$ simultaneously.

With some initialization of W and dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$, we can calculate the sparse coding coefficients $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ as follows:

$$\begin{aligned}&\min_{\{\mathbf{\Lambda}_x\}} \|\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x\|_F^2 + \gamma\|\mathbf{\Lambda}_y-\mathbf{W}_x\mathbf{\Lambda}_x\|_F^2 + \lambda_x\|\mathbf{\Lambda}_x\|_1\\ &\min_{\{\mathbf{\Lambda}_y\}} \|\mathbf{Y}-\mathbf{D}_y\mathbf{\Lambda}_y\|_F^2 + \gamma\|\mathbf{\Lambda}_x-\mathbf{W}_y\mathbf{\Lambda}_y\|_F^2 + \lambda_y\|\mathbf{\Lambda}_y\|_1\end{aligned} \tag{3}$$

Eq. 3 is a multi-task lasso problem. Many l_1-optimization algorithms can solve it effectively, such as FISTA [2], LARS [7], *etc*. In this paper, we choose LARS [7] as the l_1-optimization method for its efficiency and stability.

With $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ fixed, dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$ can be updated as follows:

$$\begin{aligned} \min_{\{\mathbf{D}_x,\mathbf{D}_y\}} \|\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x\|_F^2 + \|\mathbf{Y}-\mathbf{D}_y\mathbf{\Lambda}_y\|_F^2 \\ \text{s.t. } \forall i, \|\mathbf{d}_{x,i}\|_{l_2} \le 1, \|\mathbf{d}_{y,i}\|_{l_2} \le 1 \end{aligned} \tag{4}$$

Eq. 4 is a quadratically constrained quadratic program problem (QCQP) and we adopt a one-by-one update strategy [28] to solve it.

With dictionary and coding coefficients fixed, we can then update the mapping $\mathbf{W}$:

$$\min_{\{\mathbf{W}\}} \|\mathbf{\Lambda}_y - \mathbf{W}\mathbf{\Lambda}_x\|_F^2 + (\lambda_W/\gamma)\cdot\|\mathbf{W}\|_F^2 \tag{5}$$

Eq. 5 is a ridge regression problem and the solution can be analytically derived as:

$$\mathbf{W} = \mathbf{\Lambda}_y\mathbf{\Lambda}_x^T(\mathbf{\Lambda}_x\mathbf{\Lambda}_x^T + (\lambda_W/\gamma)\cdot\mathbf{I})^{-1} \tag{6}$$

where $\mathbf{I}$ is an identity matrix.

With SCDL, we can learn the dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$ on which the sparse coding coefficients of two styles have stable bidirectional linear transformations. In Section 4 we can further enhance its stability by clustering samples into several clusters and exploiting the image nonlocal redundancy of patches.

3.3. Synthesis

After learning the dictionaries $\mathbf{D}_x$ and $\mathbf{D}_y$ and the linear mapping $\mathbf{W}$, for a given image $\mathbf{x}$ in style s_x, we can easily convert it into an image $\mathbf{y}$ of style s_y by solving the following optimization:

$$\begin{aligned} \min_{\{\boldsymbol{\alpha}_{x,i},\boldsymbol{\alpha}_{y,i}\}} \|\mathbf{x}_i - \mathbf{D}_x\boldsymbol{\alpha}_{x,i}\|_F^2 + \|\mathbf{y}_i - \mathbf{D}_y\boldsymbol{\alpha}_{y,i}\|_F^2 \\ +\gamma\|\boldsymbol{\alpha}_{y,i} - \mathbf{W}\boldsymbol{\alpha}_{x,i}\|_F^2 + \lambda_x\|\boldsymbol{\alpha}_{x,i}\|_1 + \lambda_y\|\boldsymbol{\alpha}_{y,i}\|_1 \end{aligned} \tag{7}$$

where $\mathbf{x}_i$ is a patch of $\mathbf{x}$ and $\mathbf{y}_i$ is the corresponding patch in the intermediate estimate of $\mathbf{y}$ to be synthesized. Eq. 7 can be solved by alternatively updating $\boldsymbol{\alpha}_{x,i}$ and $\boldsymbol{\alpha}_{y,i}$. Finally, each patch of $\mathbf{y}$ can be reconstructed as:

$$\hat{\mathbf{y}}_i = \mathbf{D}_y\hat{\boldsymbol{\alpha}}_{y,i} \tag{8}$$

After all the patches are estimated, the estimation of the desired image $\mathbf{y}$ can then be obtained.

In our synthesis method, an initial estimation of $\mathbf{y}$ is needed. Depending on the problem, different strategies can be adopted to initialize $\mathbf{y}$. For example, in the problem of image super-resolution, $\mathbf{y}$ can be simply initialized by bicubic interpolation. In the problem of photo-sketch synthesis, we can first code $\mathbf{x}_i$ on $\mathbf{D}_x$ for coding vector $\boldsymbol{\alpha}_{x,i}$, and then initialize $\mathbf{y}_i$ as $\mathbf{D}_y\mathbf{W}\boldsymbol{\alpha}_{x,i}$.

4. Enhanced algorithm

4.1. Clustering and model selection

Due to the complex structures in images of different styles, learning only one pair of dictionaries and an associated linear mapping function is often not enough to cover all variations of image cross-style synthesis. For example, in face sketch-photo synthesis the mapping may vary significantly in different facial regions. Therefore multi-model should be learned to enhance the robustness. Intuitively, pre-clustering could be conducted to separate training data into several groups so that the linear mapping in each group can be more stably learned. Lin and Tang [16] proposed a Coupled Gaussian Mixture Model to tackle this coupled data clustering problem. They dealt with sample pairs as a whole in the joint spaces and modeled them as mixtures of several cluster centers. The objective function of the clustering algorithm is [16]:

$$\max_{\{\mathbf{M},\mathbf{c}\}} \prod_{i=1}^{n} P(\mathbf{u}_i, \mathbf{v}_i|\mathbf{M}_{\mathbf{c}(i)}) \tag{9}$$

where

$$\mathbf{c}(i) = \arg\max_{\{k\}} P(\mathbf{u}_i, \mathbf{v}_i|\mathbf{M}_k) \tag{10}$$

and $\mathbf{M}_k$ indicates a coupled Gaussian model $\mathbf{u} \sim N(\mathbf{m}_{u,k}, \mathbf{\Sigma}_{u,k})$ and $\mathbf{v} \sim N(\mathbf{m}_{v,k}, \mathbf{\Sigma}_{v,k})$.

The clustering in [16] is actually performed according to the concentration of data points. The objective function simply assumes that joint data assembling closely in vector space share the same linear mapping between the two styles. In this paper, we propose to conduct clustering in the sparse domains spanned by the two dictionaries. In this way, a linear mapping between the sparse codes of two image styles can be more stably and accurately learned than in the non-sparse original signal spaces. For easy calculation and modeling, we suppose that the model prediction error is Gaussian distributed. Based on the above discussions, we integrate a novel model selection procedure into the proposed SCDL framework by optimizing the following objective function:

$$\begin{aligned} &\max_{\{\mathbf{W},\mathbf{c}\}} \prod_{i=1}^{n} P(\boldsymbol{\alpha}_{x,i}, \boldsymbol{\alpha}_{y,i}|\mathbf{W}_{\mathbf{c}(i)}) \\ &= \min_{\{\mathbf{W},\mathbf{c}\}} \sum_{i=1}^{n} \|\boldsymbol{\alpha}_{x,i} - \mathbf{W}_{c(i)}\boldsymbol{\alpha}_{y,i}\|_2 \end{aligned} \tag{11}$$

where $\mathbf{c}$ are model indices for samples and $\mathbf{W}$ are style mappings in each cluster.

Our method focuses on concentration around superplanes in the sparse coding domains instead of centroids in the non-sparse original signal domains. Eq. 11 can be alternatively optimized by fixing one of the two sets of variables, $\mathbf{c}$ or $\mathbf{W}$. Therefore, a heuristic strategy which is similar to K-Means clustering can be integrated in our SCDL framework. In each iteration we update the clustering index of

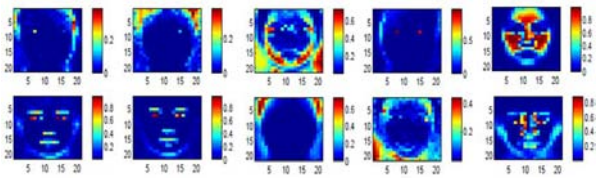

Figure 2. Examples of cluster distributions in photo-sketch synthesis. Each sub-figure shows the distribution of a cluster, while the stronger color represents higher frequency.

each training sample based on model fitting error, and update linear mappings according to the current clusters.

After integrating clustering into the learning of SCDL, multiple dictionary pairs and mappings are learned. In the synthesis stage, a model must be selected for a local image patch. However, we only have the image in style s_x, and coupled clustering for model seeking cannot be conducted directly. To solve this problem, we can initialize $\mathbf{y}$, and then determine the initial cluster index $\mathbf{c}(i)$ of each patch by:

$$\min_{\{c\}} \|\boldsymbol{\alpha}_{x,i} - \mathbf{W}_{\mathbf{c}(i)}\hat{\boldsymbol{\alpha}}_{y,i}\|_2 \tag{12}$$

where $\boldsymbol{\alpha}_{x,i}$ and $\hat{\boldsymbol{\alpha}}_{y,i}$ are the sparse coding coefficients of source style image patch $\mathbf{x}_i$ and initial guess of target style image patch $\mathbf{y}_i$. In image super-resolution this model selection is effective because $\mathbf{y}$ can be well initialized by bicubic interpolator. However when dealing with cross-style face synthesis, it is difficult to get a good enough initialization of $\mathbf{y}$. Based on the structure of face images, we found that patches in different clusters have distinctive spatial distribution, as shown in Fig. 2. We see that the patches in different clusters concentrate at different spatial locations. The strong color means high frequency.

With this observation, we can have an empirical estimation of the spatial distribution of each cluster in the face images. Initial model selection can then be transformed into a MAP problem:

$$\max_{\{c\}} P(\boldsymbol{\alpha}_{x,i}, \hat{\boldsymbol{\alpha}}_{y,i}|\mathbf{W}_c)P(\mathbf{L}_i|c) \tag{13}$$

where $\mathbf{L}_i = (row_i, col_i)^T$ are coordinates of patches $\mathbf{x}_i$ in spatial domain and distribution $P(\mathbf{L}_i|c)$ is the prior probability from empirical observation on training data.The MAP problem in Eq. 13 is a weighted distance minimum problem that can be easily solved.

4.2. Exploiting nonlocal self-similarities

Recently many works have shown that the nonlocal redundancies existing in natural images are very useful for image restoration and a good combination of local sparsity and nonlocal redundancy can greatly enhance the performance of image reconstruction [3, 4, 18, 23]. Our synthesis framework can also be enhanced by integrating nonlocal similarities. For each local patch $\mathbf{y}_i$, we can search for its similar patches in the whole image, and then predict this patch as: $\hat{\mathbf{y}}_i = \sum_{i=1}^{L} b_i^l \mathbf{y}_i^l$, where $\mathbf{y}_i^l$ is the lth most similar patch to $\mathbf{y}_i$ and b_i^l is the nonlocal weight as defined in [3]. Consequently, the nonlocal based cross-style image synthesis can be performed by:

$$\arg\min_{\{\mathbf{y}_i\}} E_{SCDL} + \delta\|\mathbf{y}_i - \sum\nolimits_{l=1}^{L} b_i^l \mathbf{y}_i^l\|_2^2 \tag{14}$$

where E_{SCDL} is the energy function defined in Eq. 7 and δ is the balancing parameter.

4.3. Summary of algorithms

The proposed SCDL approach involves two algorithms: the dictionary and mapping learning algorithm and the image synthesis algorithm, which are summarized in the following Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Semi-Coupled Dictionary Learning

Input: Training datasets $\mathbf{X}$ and $\mathbf{Y}$ of two image styles. Each corresponding pair indicates the same object. Initial dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$, and initial mapping $\mathbf{W}_x$ and $\mathbf{W}_y$.
For each iteration **Until** convergence:
For each cluster

1. Fix other variables, update $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$ by sparse coding in Eq. 3.
2. Fix other variables, update $\mathbf{D}_x$ and $\mathbf{D}_y$ in Eq. 4.
3. Fix other variables, update $\mathbf{W}_x$ and $\mathbf{W}_y$ in Eq. 5.

Update clustering index of each pair by Eq. 13.
Output: $\mathbf{D}_x$, $\mathbf{D}_y$, $\mathbf{W}_x$ and $\mathbf{W}_y$

Algorithm 2 Cross-style Image Synthesis

Input: Test image $\mathbf{x}$, well trained dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$, the learnt mapping $\mathbf{W}_x$ and $\mathbf{W}_y$ for two styles.
Initialization: Initialize $\mathbf{y}$ as discussions in 3.3. Initialize clustering index of each patch according to Eq. 13.
For each iteration **Until** convergence:

1. Update $\mathbf{y}$ by the nonlocal based cross-style synthesis in Eq. 14.
2. Update clustering index of each patch according to Eq. 13.

Output: Synthesized image $\mathbf{y}$.

5. Experiment results

The proposed SCDL model is simple yet general. It can be adapted to solve various cross-style image synthe-

sis problems. In this paper, we apply it to image super-resolution and photo-sketch synthesis to verify its effectiveness.

It is crucial to select appropriate parameters for different applications. In this paper, a combined line-search strategy was used to select parameters for each application according to the minimal energy after converge. The parameters selected in this way include those regularization parameters $\lambda_x, \lambda_x, \lambda_x, \gamma, \delta$. For the number of clusters in pre-clustering, image patch size and the number of atoms in the dictionary, we empirically set them by experience. The specific values of these parameters will be given in the following experiments.

Due to the page limit, only partial experimental results are shown. More results and the MATLAB source codes of this paper can be found at http://www.comp.polyu.edu.hk/~cslzhang/SCDL.htm.

5.1. Image Super-resolution

As we discussed in Section 2, we consider the image super-resolution problem where the low-resolution (LR) image is directly down-sampled from the high-resolution (HR) image. Since there is no blurring (or we can say that the blur kernel is the Dirac delta function) before down-sampling, the missing pixels have no direct connection with the sampled pixels, making the super-resolution a highly ill-posed problem. Fortunately, natural images have a rich amount of local and nonlocal redundancy, and we can assume that there exists a piecewise linear mapping between the HR and LR image patches in the domains spanned by an HR dictionary $\mathbf{D}_h$ and an LR dictionary $\mathbf{D}_l$. With a training set $\mathbf{Y}$ of HR patches and the associated training set $\mathbf{X}$ of LR patches, the model in Eq. 2 can be adapted to learn the mapping $\mathbf{W}$.

In our experiments, 500 thousand training patch pairs are extracted from the Kodak PhotoCD dataset*, which has no relation with the testing images used in the experiments. The patch size is 5×5. Pre-clustering is conducted and cluster number is set to be 64. We choose nine widely used testing images in the experiments. The regularization parameter $\lambda_x, \lambda_x, \lambda_x, \gamma, \delta$, are set to be 0.01, 0.01, 0.1, 0.1 and 0.25, respectively. The number of atoms in the learned dictionary is set as 256 for each cluster. In the reconstruction (i.e., synthesis) stage, bicubic interpolation is used for the initialization of HR image. The representative and state-of-the-art image super-resolution methods, including bicubic [12], SAI [29], SME [21] and ScSR [27], are employed to compare with the proposed SCDL method. All the codes are downloaded from the authors' websites. Note that in the implementation of ScSR, the Matlab function "`imresize`" is used to generate the LR image, which actually involves a smooth filtering before down-sampling.

*http://r0k.us/graphics/kodak/

We first do image super-resolution with scaling factor 2. The PSNR results are listed in Table 1, while an example (Butterfly) is shown in Fig. 3. For color images, we only calculate PSNR measures for the luminance channel. From Table 1 we can see that our proposed method outperforms state-of-the-arts in most cases, and its PSNR is in average 0.26dB higher than SAI, which is the second best among all competing methods. In particular, from Fig. 3 we can see that although the SAI method can preserve well the image edges, it will also over-smooth the edges to some extent. For example, some fine structures in the wing of the Butterfly are smoothed out by SAI, but interestingly such fine structures can be partially preserved by the proposed SCDL method.

We then do image super-resolution with scaling factor 3. Since the codes of SAI and SME can only do super-resolution with scaling factor 2^n, where n is an integer, we only compare SCDL with bicubic and ScSR in this experiment. The PSNR results are listed in Table 2, and an example (Leaves) is shown in Fig. 4. Again, SCDL performs much better than ScSR in terms of both PSNR and visual perception quality.

5.2. Face synthesis between sketch and photo

The proposed SCDL can also be used for other applications such as sketch-photo/photo-sketch synthesis, which have potential applications in law enforcement and entertainment. Sketches which are often drawn by artists have significantly different appearance from the original photos. Here we conduct photo-sketch and sketch-photo face synthesis on the CUFS Database [26], which consists of three parts: 188 subjects from CUHK students, 295 subjects from XM2TVS database and 123 subjects from the AR database. Each subject has one photo image and one corresponding sketch image drawn by artists. In our experiments we use the 88 subjects from CUHK students for training and others as testing samples.

As mappings between photo and sketch are highly nonlinear, we do synthesis on image patches. As artists prefer to exaggerate some local structures of human faces, for similar patches in photo their corresponding patches in sketch can be very different. Therefore, we need to pre-cluster patch pairs to learn multiple dictionary pairs and linear mappings to address the complex relationship between photo and sketch. In the synthesis, the initialization of sketch-photo is made as explained in Section 3.3, and each patch pair is clustered by Eq. 13. 50,000 pairs of patches are randomly selected for training. The patch size is 10×10 and the cluster number is 64. The number of atoms in the dictionary is 256. The regularization parameters, $\lambda_x, \lambda_x, \lambda_x, \gamma, \delta$, are set to be 0.015, 0.015, 0.15, 0.1 and 0.4, respectively.

Fig.5 shows the synthesis results for photo-sketch and sketch-photo synthesis, respectively. Wang and Tang's

Figure 3. Experimental results on image super-resolution (scaling factor: 2). From left to right: low resolution image, high resolution ground-truth, and reconstructed images by Bicubic [12], ScSR [27], SAI [29], SME [21] and the proposed SCDL method.

Table 1. PSNR (dB) results on image super-resolution (scaling factor = 2)

Image	*Girl*	*Butterfly*	*Fence*	*Starfish*	*Parthenon*	*House*	*Foreman*	*Cameraman*	*Leaves*	Average
Bicubic[12]	33.83	27.68	24.52	30.22	27.08	32.15	35.56	25.36	26.85	29.25
SAI[29]	34.13	29.17	23.78	30.73	27.10	32.84	**37.68**	25.88	28.72	30.00
SME[21]	34.03	28.65	24.53	30.35	27.13	33.15	37.17	**26.14**	28.21	29.93
ScSR[27]	33.29	28.27	24.05	30.35	26.46	31.78	35.68	25.28	27.52	29.19
Proposed	**34.25**	**29.62**	**24.76**	**30.94**	**27.32**	**33.21**	37.26	26.06	**28.92**	**30.26**

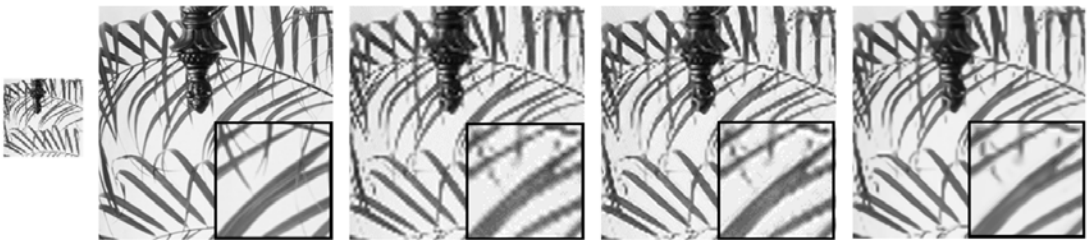

Figure 4. Experimental results on image super-resolution (scaling factor: 3). From left to right: low resolution image, high resolution ground-truth, and reconstructed images by Bicubic [12], ScSR [27] and the proposed SCDL method.

Table 2. PSNR (dB) results on image super-resolution (scaling factor = 3)

Image	*Girl*	*Butterfly*	*Fence*	*Starfish*	*Parthenon*	*House*	*Foreman*	*Cameraman*	*Leaves*	Average
Bicubic[12]	31.24	23.32	20.30	25.97	24.05	28.55	32.00	22.09	21.74	25.47
ScSR[27]	31.10	23.84	20.38	26.08	24.06	28.53	32.29	22.21	21.93	25.60
Proposed	**31.90**	**24.61**	**20.96**	**26.60**	**24.68**	**29.25**	**33.37**	**22.89**	**22.64**	**26.32**

method in [26] is used for comparison. The method in [26] actually has two steps. In the first step, a nearest neighbor searching based method is used to synthesize the photo or sketch patches, as shown in the 2nd row of Fig. 5. In the second step, patches will be optimized with an MRF post-processing framework, as shown in the 3rd row of Fig. 5. The MRF post-processing significantly improves the results of the first step, whereas we can still see some artifacts (highlighted in the last row of Fig. 5) generated in the incorrect patch matching process. Compared with the final synthesis results reported in [26], our result seems over-smoothed, as shown in the 4th row of Fig. 5. However, it should be noted that there is no complex MRF post-processing in our method. We simply use the averaging strategy for fusing overlapped patches. Our results have a large room to improve by coupling with some post-processing techniques.

6. Conclusions

In this paper, we proposed a novel semi-coupled dictionary learning (SCDL) framework for cross-style image synthesis. SCDL jointly optimizes the dictionary pair and the mapping function in the sparse domain. The learned dictionary pair can not only ensure the style-specific data fidelity but also span the hidden spaces for stable mapping between image styles. The proposed SCDL is adapted to applications of image super-resolution and photo-sketch synthesis, and shows very competitive performance with state-of-the-arts. In the future study, we will adapt SCDL to more types of image synthesis tasks and extend it to cross-style image recognition tasks.

7. Acknowledgements

This work is supported by HK RGC General Research Fund (PolyU 5375/09E) and NSFC Key Project (61135001).

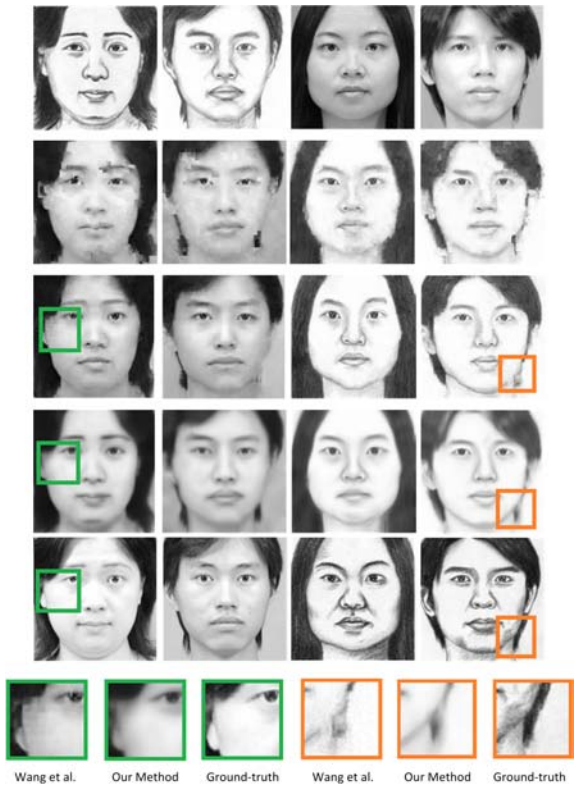

Figure 5. Sketch-photo (left two columns) and photo-sketch (right two columns) synthesis. From top to bottom: input images, results by Wang et al.'s method [26] without MRF post-processing, results by method [26] with MRF post-processing, results by SCDL, ground-truths and zoom-in sub-images.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Trans on*, 54(11):4311–4322, 2006. 2

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 4

[3] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *CVPR*. IEEE, 2005. 5

[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans on IP*, 16(8):2080–2095, 2007. 5

[5] W. Dong, L. Zhang, and G. Shi. Centralized sparse representation for image restoration. In *ICCV*. IEEE, 2011. 2

[6] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Trans on IP*, 20(7):1838–1857, 2011. 2

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. 4

[8] A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, 2011. 1, 2

[9] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans on IP*, 15(12):3736–3745, 2006. 2

[10] J. T. Freeman W.T. and P. E.C. Example-based super-resolution. *Computer Graphics and Applications, IEEE*, 22(2):56–65, 2002. 2

[11] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *SIGGRAPH*, pages 327–340, 2001. 1, 2

[12] R. Keys. Cubic convolution interpolation for digital image processing. *Acoustics, Speech and Signal Processing, IEEE Trans on*, 29(6):1153–1160, 1981. 1, 2, 6, 7

[13] Z. Lei and S. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*. IEEE, 2009. 1, 2

[14] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *CVPR*. IEEE, 2009. 2

[15] X. Li and M. Orchard. New edge-directed interpolation. *IEEE Trans on IP*, 10(10):1521–1527, 2001. 1, 2

[16] D. Lin and X. Tang. Coupled space learning of image style transformation. In *ICCV*. IEEE, 2005. 1, 2, 3, 4

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*. ACM, 2009. 2

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *ICCV*. IEEE, 2009. 5

[19] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *NIPS*, 2009. 2

[20] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans on IP*, 17(1):53–69, 2008. 2

[21] S. Mallat and G. Yu. Super-resolution with sparse mixing estimators. *IEEE Trans on IP*, 19(11):2889–2900, 2010. 1, 2, 6, 7

[22] A. Sharma and D. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR*. IEEE, 2011. 1, 2

[23] J. Sun and M. Tappen. Learning non-local range markov random field for image restoration. In *CVPR*, pages 2745–2752. IEEE, 2011. 5

[24] X. Tang and X. Wang. Face sketch synthesis and recognition. In *ICCV*. IEEE, 2003. 1, 2

[25] X. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE Trans on SMC-C*, 35(3):425–434, 2005. 2

[26] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Trans on PAMI*, pages 1955–1967, 2008. 1, 2, 6, 7, 8

[27] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans on IP*, 19(11):2861–2873, 2010. 1, 2, 3, 6, 7

[28] M. Yang, L. Zhang, J. Yang, and D. Zhang. Metaface learning for sparse representation based face recognition. In *ICIP*, pages 1601–1604. IEEE, 2010. 4

[29] X. Zhang and X. Wu. Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation. *IEEE Trans on IP*, 17(6):887–896, 2008. 1, 2, 6, 7