Coupled Space Learning of Image Style Transformation

Dahua Lin
Department of Information Engineering
The Chinese University of Hong Kong
dhlin4@ie.cuhk.edu.hk

Xiaoou Tang
Microsoft Research Asia
Beijing, China
xitang@microsoft.com

Abstract

In this paper, we present a new learning framework for image style transforms. Considering that the images in different style representations constitute different vector spaces, we propose a novel framework called Coupled Space Learning to learn the relations between different spaces and use them to infer the images from one style to another style. Observing that for each style, only the components correlated to the space of the target style are useful for inference, we first develop the Correlative Component Analysis to pursue the embedded hidden subspaces that best preserve the inter-space correlation information. Then we develop the Coupled Bidirectional Transform algorithm to estimate the transforms between the two embedded spaces, where the coupling between the forward transform and the backward transform is explicitly taken into account. To enhance the capability of modelling complex data, we further develop the Coupled Gaussian Mixture Model to generalize our framework to a mixture-model architecture. The effectiveness of the framework is demonstrated in the applications including face super-resolution and bidirectional portrait style transforms.

1. Introduction

In recent years, transformation between image style representations becomes an active research topic in computer vision. Representative works on style-transforms include image hallucination[3][1][6] and non-photorealistic rendering[8][4].

Different from conventional approaches where different types of transforms are treated separately. In this paper, we study different transform tasks in a unified perspective and develop a new learning framework to improve the quality of the resultant images.

In statistical learning, each image can be represented by a vector and thus the images in a certain style form a vector space. Under this formulation, the relation between two image styles can be seen as the relation between two vector spaces associated with the two image styles. In the literature, a series of statistical learning approaches have been proposed to model the image space. Among these methods, the most well known one is PCA[10], which finds a principal subspace where the variational energy is maximized. However, PCA is aimed at modelling a single sample space with the goal of best reconstruction, thus cannot be directly applied to model the inter-space dependencies.

Some works have been done to extend the conventional PCA models to learn the dependency between two sample spaces. There are mainly two families of methods: one is to establish a subspace model in the joint space of two vector spaces[9]; the other is to learn the relation between two principal spaces: a representative method in this family is the eigentransformation[8] method, which learns the relationship between photo-space and sketch-space by transferring the synthesis coefficients obtained by PCA. One important drawback in these two families of methods is that some important correlative information, which is not necessarily significant in reconstruction, may be lost in the stage of projecting the sample to the principal subspace learned individually, this is because the learning of these individual spaces does not take the correlation between the two spaces into account. To address the issue, Fernando et al. developed the Asymmetric Coupled Component Analysis(ACCA)[2] where hidden parameter space is made explicit to serve as a bridge coupling the two spaces. In ACCA, though the coupling is explicitly accounted for, however, as shown later, its simple formulation does not fully reflect the essence of coupling and lacks the capability of modelling complex dependencies.

In this paper, we propose a novel framework called *Coupled Space Learning* to learn the dependency between two vector spaces, with each space corresponding to one image style. The core of our framework is to couple the learning process of the forward and the backward transforms. Observing that only the components that are correlated to the other vector space contribute to the inference, we derive the *Maximum Correlation Criteria* and develop a new

Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)
1550-5499/05 $20.00 © 2005 **IEEE**

IEEE COMPUTER SOCIETY

algorithm called *Correlative Component Analysis* to pursue the hidden spaces associated with the two representative spaces so that the correlative information is best preserved. Then the *Coupled Bidirectional Transforms* algorithm is developed to learn the bidirectional transforms between two hidden spaces in a coupled manner where the relation between the transforms in the two opposite directions are explicitly taken into account. The coupling between the forward transform and backward transform are gradually established through repeated information exchange between the two transforms.

To further enhance the framework's capability of modelling complex data, we generalize our framework to a mixture-model architecture, called *Coupled Gaussian Mixture Model*, where GMM for both spaces are jointly trained. The system consists of multiple models, in the training phase each model is adapt to a part of the samples and in the testing phase for a new sample, the results produced by these models are fused together by a weighting scheme using model-posteriori as weights.

Our framework is of broad interest in the realm of computer vision. To illustrate the effectiveness of the framework, we conduct comparative experiments in style-transform applications including face super-resolution and bidirectional transforms between portrait styles.

In the rest of the paper, we first present the theoretical principle and algorithms for Coupled Space Learning in section 2. In section 3, we generalize our model to a mixture-model system with Coupled GMM. Experiments and their results are introduced in section 4. Finally, we conclude the paper and propose future work in section 5.

2. Coupled Space Learning

2.1. Framework of Coupled Modelling

Suppose we have a set of visual objects, denoted as $a_1, a_2, \ldots, a_n$, here n is the number of objects. For each object, it can be expressed by images in different styles, such as photos and nonphotorealistic paintings. In each style, the objects are represented in vectors. The vectors in the first style constitute a sample space $\mathcal{X}$, denoted as $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; likewise, the vector space and vectors in the second style are denoted as $\mathcal{Y}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$. It should be noted that the space dimension for different representation is not necessarily equal, we denote the dimensions of the two representation spaces as d_x and d_y respectively.

Considering that both sample spaces are associated to the same object space, it is reasonable to assume that *there is an intrinsic hidden space $\mathcal{H}$ reflecting the variations which the visual objects inherently bear and the observed spaces are some transformed versions of the hidden space*, which is the fundamental principle in coupled learning.

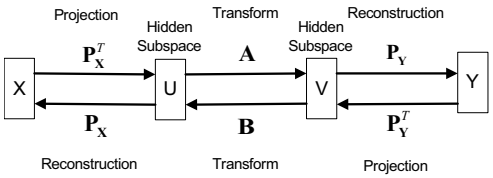

Figure 1. Illustration of Harmony Coupled Learning Framework

Denote the vectors in hidden space as $\mathbf{h}$, the transforms from hidden spaces to observed spaces as $\mathbf{T}_X$ and $\mathbf{T}_Y$

$$\mathbf{x} = \mathbf{T}_X\mathbf{h} + \mathbf{m}_x \qquad \mathbf{y} = \mathbf{T}_Y\mathbf{h} + \mathbf{m}_y. \tag{1}$$

Here, $\mathbf{m}_x$ and $\mathbf{m}_y$ are mean vectors of $\mathbf{x}$ and $\mathbf{y}$. Assume the dimension of hidden space is d, then $\mathbf{T}_X$ is a $d_x \times d$ matrix while $\mathbf{T}_Y$ is a $d_y \times d$ matrix. To investigate the composition of the transform, we perform compact SVD on them as $\mathbf{T}_X = \mathbf{U}_X\mathbf{D}_X\mathbf{V}_X^T$ and $\mathbf{T}_Y = \mathbf{U}_Y\mathbf{D}_Y\mathbf{V}_Y^T$. Here $\mathbf{U}_X$ is a $d_x \times d$ matrix, and $\mathbf{U}_Y$ is a $d_y \times d$ matrix, while $\mathbf{D}_X, \mathbf{D}_Y, \mathbf{V}_X, \mathbf{V}_Y$ are all $d \times d$ matrices. Considering Eq.(1), we have

$$\mathbf{U}_X^T(\mathbf{x} - \mathbf{m}_x) = \mathbf{D}_X\mathbf{V}_X^T\mathbf{h} \quad \mathbf{U}_Y^T(\mathbf{y} - \mathbf{m}_y) = \mathbf{D}_Y\mathbf{V}_Y^T\mathbf{h}. \tag{2}$$

This equation can be interpreted as follows: orthonormal $\mathbf{U}_X$ projects the d_x-dimensional vector $\mathbf{x} - \mathbf{m}_x$ to a subspace of equal dimension to $\mathcal{H}$, which is actually $\mathcal{H}$'s rotated and scaled version, denoted as $\mathcal{U}$. Similar interpretation can be applied to $\mathbf{y}$, where the embedded subspace is denoted as $\mathcal{V}$. Base on this interpretation, we see that there are two d-dimensional embedded spaces associated with $\mathcal{X}$ and $\mathcal{Y}$, which are related to the $\mathcal{H}$ with rotation and scaling. It further follows that the two embedded subspaces are related to each other with rotations and scaling. To clearly emphasize the projection role of $\mathbf{U}_X$ and $\mathbf{U}_Y$, we denote them as $\mathbf{P}_X$ and $\mathbf{P}_Y$.

Based on this rationale, we design a three-level framework for *Coupled Space Learning (CSL)* as illustrated in Figure.1. where the connection between $\mathcal{U}$ and $\mathcal{V}$ is established through $d \times d$ transform matrices $\mathbf{A}$ and $\mathbf{B}$. In mathematics, the whole transform procedure can be represented as follows:

$$\mathbf{y} - \mathbf{m}_y = \mathbf{P}_Y\mathbf{A}\mathbf{P}_X^T(\mathbf{x} - \mathbf{m}_x), \tag{3}$$

$$\mathbf{x} - \mathbf{m}_x = \mathbf{P}_X\mathbf{B}\mathbf{P}_Y^T(\mathbf{y} - \mathbf{m}_y). \tag{4}$$

It is worthwhile to emphasize the following points concerning the design of the framework:

1. Under this formulation, given the relation between $\mathbf{x}$ and $\mathbf{y}$ as linear, why don't we directly use the form $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{x} = \mathbf{B}\mathbf{y}$? As mentioned above, the fundamental concept in coupled learning is the hidden space, and the $\mathbf{P}_X$

IEEE COMPUTER SOCIETY

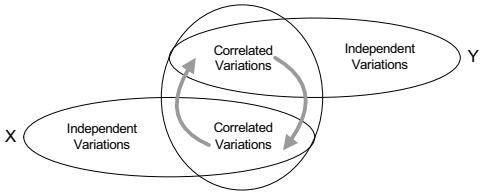

Figure 2. Illustration of Components Decomposition

and $\mathbf{P}_Y$ in Eq.(3) and Eq.(4) indeed embody this concept and enforce it as a structural constraint in transforms.

2. Why don't we directly focus on $\mathbf{T}_X$, $\mathbf{T}_Y$ and $\mathcal{H}$, but decompose the process into $\mathbf{P}_X, \mathbf{P}_Y, \mathbf{A}$, and $\mathbf{B}$? This is mainly for the sake of computational efficiency. Because directly solving transforms between two spaces of different dimensions is difficult and unstable, especially when d_x and d_y is large, it is desirable to first learn a projection to project the vectors to subspaces with equal dimensions and than solve the transform within the subspaces.

2.2. Correlative Component Analysis

As we know, the essence of coupling comes from the statistical dependencies between two sample spaces, which is also the foundation for bidirectional inference. In a linear model, under Gaussian assumption, statistical dependency is equivalent to correlation, where uncorrelated components provide no information for prediction of each other. Concretely, as illustrated in Figure.2, for each space, it can be decomposed into two subspaces, one preserves correlative information for intra-space communication, while the other only captures independent variations special to the space itself. Only the former contributes to the inter-space inference.

Suppose $\mathbf{x} \sim \mathcal{N}(\mathbf{m}_x, \mathbf{C}_x)$ and $\mathbf{y} \sim \mathcal{N}(\mathbf{m}_y, \mathbf{C}_y)$, for a component in $\mathcal{X}$ characterized by projection direction $\mathbf{p}_x$ and a component in $\mathcal{Y}$ projected by $\mathbf{p}_y$, their correlation can be measured in terms of covariance as

$$E\left[(\mathbf{p}_x^T(\mathbf{x}-\mathbf{m}_x))(\mathbf{p}_y^T(\mathbf{y}-\mathbf{m}_y))^T\right] = \mathbf{p}_x^T\mathbf{C}_{xy}\mathbf{p}_y. \quad (5)$$

Here, $\mathbf{C}_{xy} = E\left[(\mathbf{x}-\mathbf{m}_x)(\mathbf{y}-\mathbf{m}_y)^T\right]$ is the covariance matrix between $\mathbf{x}$ and $\mathbf{y}$.

Considering that $\mathbf{C}_{xy}$ is not a semidefinite matrix, thus the value of the covariance matrices can be negative, however, only the magnitude but not the sign of the value represents the intensity of the correlation. For mathematical tractability, we use the square of covariance value as *Correlation Intensity*:

$$CI(\mathbf{p}_x, \mathbf{p}_y) = (\mathbf{p}_x^T\mathbf{C}_{xy}.\mathbf{p}_y)^2. \quad (6)$$

For a set of components obtained by projection matrices $\mathbf{P}_x$ and $\mathbf{P}_y$, their covariance is a generalization of Eq.5 as

$$E\left[(\mathbf{P}_X^T(\mathbf{x}-\mathbf{m}_x))(\mathbf{P}_Y^T(\mathbf{y}-\mathbf{m}_y))^T\right] = \mathbf{P}_X^T\mathbf{C}_{xy}\mathbf{P}_Y. \quad (7)$$

By taking all components as a whole, the total correlation intensity can be derived as

$$\begin{aligned} CI(\mathbf{P}_X, \mathbf{P}_Y) &= \mathrm{tr}\left((\mathbf{P}_X^T\mathbf{C}_{xy}\mathbf{P}_Y)(\mathbf{P}_X^T\mathbf{C}_{xy}\mathbf{P}_Y)^T\right) \\ &= \mathrm{tr}\left(\mathbf{P}_X^T\mathbf{C}_{xy}\mathbf{P}_Y\mathbf{P}_Y^T\mathbf{C}_{yx}\mathbf{P}_X\right) \quad (8) \\ &= \mathrm{tr}\left(\mathbf{P}_Y^T\mathbf{C}_{yx}\mathbf{P}_X\mathbf{P}_X^T\mathbf{C}_{xy}\mathbf{P}_Y\right). \quad (9) \end{aligned}$$

Here, $\mathbf{C}_{yx} = E\left[(\mathbf{y}-\mathbf{m}_y)(\mathbf{x}-\mathbf{m}_x)^T\right] = \mathbf{C}_{xy}^T$ is the covariance matrix between $\mathbf{y}$ and $\mathbf{x}$.

Given training sets as $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and $[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$, and denote their sample mean as $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$, then we can arrange the mean-offset samples into two matrices as $\widetilde{\mathbf{X}} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \mathbf{x}_2 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}]$ and $\widetilde{\mathbf{Y}} = [\mathbf{y}_1 - \bar{\mathbf{y}}, \mathbf{y}_2 - \bar{\mathbf{y}}, \ldots, \mathbf{y}_n - \bar{\mathbf{y}}]$, thus the maximum likelihood estimation of covariance matrices[5] can be written as $\mathbf{C}_{xy} = \frac{1}{n}\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T$ and $\mathbf{C}_{yx} = \frac{1}{n}\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T$

To facilitate further analysis of the relation between the two spaces, it is desirable to pursue the subspaces which best preserve the correlative information, hence we derive the *Maximum Correlation Criteria* for learning two correlative subspaces as follows:

$$(\mathbf{P}_X, \mathbf{P}_Y) = \underset{\mathbf{P}_X, \mathbf{P}_Y}{\mathrm{argmax}}\, CI(\mathbf{P}_X, \mathbf{P}_Y). \quad (10)$$

Here,

$$\begin{aligned} CI(\mathbf{P}_X, \mathbf{P}_Y) &= \mathrm{tr}\left(\mathbf{P}_X^T\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T\mathbf{P}_Y\mathbf{P}_Y^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T\mathbf{P}_X\right), \quad (11) \\ &= \mathrm{tr}\left(\mathbf{P}_Y^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T\mathbf{P}_X\mathbf{P}_X^T\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T\mathbf{P}_Y\right). \quad (12) \end{aligned}$$

To optimize the Maximum Correlation Criteria, we develop an algorithm called *Correlative Component Analysis (CCA)*, which pursues optimal $\mathbf{P}_X$ and $\mathbf{P}_Y$ alternately. The procedure is described in Table.1:

1. Initialize $\mathbf{P}_X^{(0)}$ and $\mathbf{P}_Y^{(0)}$ to be identity matrices.
2. Repeat the following steps, at the t-th step:
  (a) Compute $\mathbf{S}_X^{(t)} = \widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T\mathbf{P}_Y^{(t-1)}\mathbf{P}_Y^{(t-1)T}\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T$
  (b) Update $\mathbf{P}_X$ by $\mathbf{P}_X^{(t)} = \mathrm{argmax}_{\mathbf{P}_X}\, \mathbf{P}_X^T\mathbf{S}_X^{(t)}\mathbf{P}_X$
  (c) Compute $\mathbf{S}_Y^{(t)} = \widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T\mathbf{P}_X^{(t)}\mathbf{P}_X^{(t)T}\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T$
  (d) Update $\mathbf{P}_Y$ by $\mathbf{P}_Y^{(t)} = \mathrm{argmax}_{\mathbf{P}_Y}\, \mathbf{P}_Y^T\mathbf{S}_Y^{(t)}\mathbf{P}_Y$
  (e) Compute the objective function $C^{(t)}$ by Eq.10.
3. Stop and exit when $C^{(t)} - C^{(t-1)} < \varepsilon$.

Table 1. Training process of CCA

Note that for a positive semidefinite matrix $\mathbf{S}$, $\mathrm{argmax}_{\mathbf{P}}\, \mathrm{tr}(\mathbf{P}^T\mathbf{S}\mathbf{P})$ can be obtained by performing eigenvalue-eigenvector analysis on $\mathbf{S}$, and takes the d eigenvectors associated with largest eigenvalues as the column vectors of $\mathbf{P}$.

Discussion

1. As in Eq.11 and Eq.12, their equivalence elegantly reflects the duality of the two spaces.

IEEE COMPUTER SOCIETY

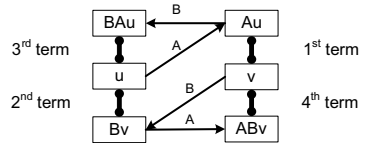

Figure 3. Illustration of Objectives of Bidirectional Transform

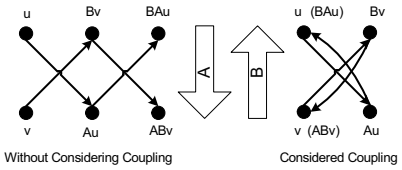

Figure 4. Illustration of Significance of Coupling

2. When $\mathbf{P}_X$ or $\mathbf{P}_Y$ is fixed, the objective function is convex w.r.t $\mathbf{P}_Y$ or $\mathbf{P}_X$, thus the updating achieves global optimal $\mathbf{P}_Y$ or $\mathbf{P}_X$ with the other matrix fixed. Moreover, we can see that due to the equivalence of two forms of the objective function, step.2.2 and step.2.4 is actually optimizing the same objective, and the procedure is thus guaranteed to converge.

3. Intuitively, the Correlative Component Analysis algorithm embodies a *Negotiation Mechanism*: in each iteration, both spaces convey the information of themselves through the projection matrices, and adjust their subspace projection to cater for the other part's need. In this procedure the commonality between two subspaces is gradually amplified via continuous conversation between the two parts.

2.3. Coupled Bidirectional Transform

When the two hidden subspaces are established by Correlative Component Analysis, we can learn the bidirectional transform between the two spaces. Here we denote the vectors in the two hidden spaces as $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ and $[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ respectively, which can be computed as follows

$$\mathbf{u}_i = \mathbf{P}_X^T(\mathbf{x}_i - \bar{\mathbf{x}}), \tag{13}$$

$$\mathbf{v}_i = \mathbf{P}_Y^T(\mathbf{y}_i - \bar{\mathbf{y}}). \tag{14}$$

Before we derive the objective function, it is worthwhile to analyze the goal of learning. Different from conventional unidirectional model where transform accuracy is the chief aim, in a bidirectional model the relation between forward transform and backward transform should be taken into account. In ideal case, they should be inverse process of each other. Therefore, there are two goals in learning the pair of transforms: the first goal is the accuracy of transform as in unidirectional models, while the second goal is their coupling relationship, which can be measured in terms of fidelity of reconstruction. Based on this rationale, the objective function to be minimized can be written as

$$J(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{n} \big(||\mathbf{v}_i - \mathbf{A}\mathbf{u}_i||^2 + ||\mathbf{u}_i - \mathbf{B}\mathbf{v}_i||^2 + ||\mathbf{u}_i - \mathbf{B}\mathbf{A}\mathbf{u}_i||^2 + ||\mathbf{v}_i - \mathbf{A}\mathbf{B}\mathbf{v}_i||^2 \big). \tag{15}$$

Denote $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$, then Eq.15 can be rewritten as

$$J(\mathbf{A},\mathbf{B}) = ||\mathbf{V} - \mathbf{A}\mathbf{U}||_F^2 + ||\mathbf{U} - \mathbf{B}\mathbf{V}||_F^2 + ||\mathbf{U} - \mathbf{B}\mathbf{A}\mathbf{U}||_F^2 + ||\mathbf{V} - \mathbf{A}\mathbf{B}\mathbf{V}||_F^2. \tag{16}$$

As illustrated in Figure 3, the 1st term and the 2nd term of Eq.16 correspond to the goal of transform accuracy, while the other 2 terms correspond to the fidelity of coupling.

The objective function is nonlinear with respect to $\mathbf{A}$ and $\mathbf{B}$, and it has no analytic solution. One approach is to employ gradient-based numerical optimization, the derivatives of the objective w.r.t $\mathbf{A}$ and $\mathbf{B}$ are deduced as follows

$$\frac{\partial J}{\partial \mathbf{A}} = -2\mathbf{V}\mathbf{U}^T + 2\mathbf{A}\mathbf{U}\mathbf{U}^T - 2\mathbf{B}^T\mathbf{U}\mathbf{U}^T - 2\mathbf{V}\mathbf{V}^T\mathbf{B}^T + 2\mathbf{B}^T\mathbf{B}\mathbf{A}\mathbf{U}\mathbf{U}^T + 2\mathbf{A}\mathbf{B}\mathbf{V}\mathbf{V}^T\mathbf{B}^T, \tag{17}$$

$$\frac{\partial J}{\partial \mathbf{B}} = -2\mathbf{U}\mathbf{V}^T + 2\mathbf{B}\mathbf{V}\mathbf{V}^T - 2\mathbf{A}^T\mathbf{V}\mathbf{V}^T - 2\mathbf{U}\mathbf{U}^T\mathbf{A}^T + 2\mathbf{A}^T\mathbf{A}\mathbf{B}\mathbf{V}\mathbf{V}^T + 2\mathbf{B}\mathbf{A}\mathbf{U}\mathbf{U}^T\mathbf{A}^T. \tag{18}$$

A drawback of traditional optimization method is that it is computationally expensive and the convergence is slow. To enhance the efficiency of optimization, we develop a novel algorithm to learn the *Coupled Bidirectional Transform (CBT)* as described in Table.2

1. Initialize $\mathbf{A}$ and $\mathbf{B}$ by linear regression:

$$\mathbf{A}^{(0)} = \underset{\mathbf{A}}{\operatorname{argmin}} ||\mathbf{V} - \mathbf{A}\mathbf{U}||_F^2 = (\mathbf{V}\mathbf{U}^T)(\mathbf{U}\mathbf{U}^T)^{-1}, \tag{19}$$

$$\mathbf{B}^{(0)} = \underset{\mathbf{B}}{\operatorname{argmin}} ||\mathbf{U} - \mathbf{B}\mathbf{V}||_F^2 = (\mathbf{U}\mathbf{V}^T)(\mathbf{V}\mathbf{V}^T)^{-1}. \tag{20}$$

2. Iterate the following steps until the change of objective is below some specified threshold:

(a) Backward transform variables $[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ by $\mathbf{B}^{(t-1)}$. Then we can form the augmented sample matrices $\mathbf{U}_{aug} = [\mathbf{U}, \mathbf{B}^{(t-1)}\mathbf{V}]$, and $\mathbf{V}_{aug} = [\mathbf{V}, \mathbf{V}]$, then update $\mathbf{A}$ as $\mathbf{A}^{(t)} = \operatorname{argmin}_{\mathbf{A}} ||\mathbf{V}_{aug} - \mathbf{A}\mathbf{U}_{aug}||_F^2$

(b) Forward transform variables $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ by $\mathbf{A}^{(t)}$. Then we can form the augmented sample matrices $\mathbf{U}_{aug} = [\mathbf{U}, \mathbf{U}]$, and $\mathbf{V}_{aug} = [\mathbf{V}, \mathbf{A}^{(t)}\mathbf{U}]$, then update $\mathbf{B}$ as $\mathbf{B}^{(t)} = \operatorname{argmin}_{\mathbf{B}} ||\mathbf{U}_{aug} - \mathbf{B}\mathbf{V}_{aug}||_F^2$

Table 2. Training process of CBT

IEEE COMPUTER SOCIETY

Discussion

1. The goal of the transform accuracy and that of the reconstruction fidelity are not equivalent. This is clearly illustrated in Figure.4, it can be easily seen that although the transform accuracy of the left one equals the right one, however the right one achieves much higher reconstruction fidelity and thus is more preferable, which indicates the significance of explicit considering reconstruction fidelity in learning bidirectional transforms.

2. As in correlative component analysis, the process of learning the coupled bidirectional transforms also reflects the "negotiation mechanism", where the forward transform $\mathbf{A}$ conveys the information about it by augmenting the training set of $\mathbf{B}$ with the forward transformed variables $\mathbf{AU}$, so that the optimization of $\mathbf{B}$ will take the information from $\mathbf{A}$ into account. The principle is similar for backward transform. Therefore, the "augmented set" plays an crucial role for exchanging information between the two transforms, and through repeated communication, the two transforms are finally coupled together and a well equilibrium is achieved between transform accuracy and reconstruction fidelity.

2.4. Procedure of Coupled Space Learning

Here, we summarize the whole procedure for coupled space learning. In training stage, the process can be briefly described as follows:

1. Compute the mean vectors $\mathbf{m}_x$ and $\mathbf{m}_y$, covariance matrices $\mathbf{C}_x$, $\mathbf{C}_y$, $\mathbf{C}_{xy}$ and $\mathbf{C}_{yx}$ for both spaces.
2. Learn the two hidden subspaces $\mathbf{P}_X$ and $\mathbf{P}_Y$ by Correlative Component Analysis.
3. Project the mean-offset samples $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ onto hidden spaces to obtain $\mathbf{U}$ and $\mathbf{V}$.
4. Learn the bidirectional transforms between two $\mathcal{U}$ and $\mathcal{V}$ using Coupled Bidirectional Transforms algorithm.

Table 3. The whole training procedure of CSL

In testing stage, for arbitrary new sample $\mathbf{u}$ or $\mathbf{v}$, we can infer corresponding $\mathbf{v}$ or $\mathbf{u}$ following Eq.3 or Eq.4.

3. Generalization to Mixture Model

3.1. Coupled Gaussian Mixture Model

Due to the complexity of the real data, one single linear model is often not enough to capture all the aspects of variations and dependencies. Motivated by the successful application of Gaussian Mixture Model(GMM)[13] in many practical problems, we develop *Coupled Gaussian Mixture Model (CGMM)* which effectively integrates GMM and Coupled Space Learning.

The fundamental difference between CGMM and GMM is that the pair of samples $\mathbf{u}_i$ and $\mathbf{v}_i$ should be dealt with as a whole instead of being handled individually. In CGMM, suppose we have K models, denoted as $M_1, M_2, \ldots, M_K$; then the probability of a sample-pair $(\mathbf{u}_i, \mathbf{v}_i)$ conditioning on the k-th model is

$$p(\mathbf{u}_i, \mathbf{v}_i|M_k) = p(\mathbf{u}_i|\mathbf{m}_{uk}, \mathbf{\Sigma}_{uk})p(\mathbf{v}_i|\mathbf{m}_{vk}, \mathbf{\Sigma}_{vk}). \quad (21)$$

Here $\mathbf{m}_{uk}, \mathbf{\Sigma}_{uk}$ and $\mathbf{m}_{vk}, \mathbf{\Sigma}_{vk}$ are the mean vectors and covariance matrices of the samples belonging to M_k in hidden space $\mathcal{U}$ and $\mathcal{V}$ respectively.

3.2. Optimization by EM Algorithm

With the definition of coupled conditional probability, the CGMM can be learned by Expectation-Maximization algorithm similar to that in GMM[13]. The procedure is described as follows:

1. Initialize CGMM by Random Clustering:
(a) Randomly select K pairs of samples as cluster centers, denoted as $\mathbf{m}_{u1}^{(0)}, \ldots, \mathbf{m}_{uK}^{(0)}$ and $\mathbf{m}_{v1}^{(0)}, \ldots, \mathbf{m}_{vK}^{(0)}$, which are also the initial estimation of mean vectors for CGMM.
(b) For each pair of samples$(\mathbf{u}_i, \mathbf{v}_i)$, categorize it to the cluster where the cluster center is closest. The distance is simply defined as $d_{ik} = ||\mathbf{u}_i - \mathbf{m}_{uk}^{(0)}||^2 + ||\mathbf{v}_i - \mathbf{m}_{vk}^{(0)}||^2$
(c) Compute the covariance matrices in clusters $\mathbf{\Sigma}_{u1}^{(0)}, \ldots, \mathbf{\Sigma}_{uK}^{(0)}$ and $\mathbf{\Sigma}_{v1}^{(0)}, \ldots, \mathbf{\Sigma}_{vK}^{(0)}$ as initial estimation of covariance matrices.
(d) Initialize the prior probability for all models to be the same, i.e. $P^{(0)}(M_1) = \cdots = P^{(0)}(M_K) = \frac{1}{K}$

2. Update the CGMM by iterating the following steps:
(a) Compute the probability of every training sample-pair belonging to the k-th model as

$$w_{ik} = \frac{P^{(t-1)}(M_k)p(\mathbf{u}_i, \mathbf{v}_i|M_k)}{\sum_{j=1}^{K} P^{(t-1)}(M_j)p(\mathbf{u}_i, \mathbf{v}_i|M_j)}. \quad (22)$$

The calculation of conditional probability follows Eq.21 using the mean vectors and covariance matrices computed in the $(t-1)$-th step.
(b) Update the priori of models as

$$P(M_k) = \frac{1}{n}\sum_{i=1}^{n} w_{ik}. \quad (23)$$

(c) Update the mean vectors and covariance matrices as follows:

$$\mathbf{m}_{uk}^{(t)} = \frac{1}{nP(M_k)}\sum_{i=1}^{n} w_{ik}\mathbf{u}_i, \quad (24)$$

$$\mathbf{m}_{vk}^{(t)} = \frac{1}{nP(M_k)}\sum_{i=1}^{n} w_{ik}\mathbf{v}_i, \quad (25)$$

$$\mathbf{\Sigma}_{uk}^{(t)} = \frac{1}{nP(M_k)}\sum_{i=1}^{n} w_{ij}(\mathbf{u}_i - \mathbf{m}_{uk})(\mathbf{u}_i - \mathbf{m}_{uk})^T, \quad (26)$$

$$\mathbf{\Sigma}_{vk}^{(t)} = \frac{1}{nP(M_k)}\sum_{i=1}^{n} w_{ij}(\mathbf{v}_i - \mathbf{m}_{vk})(\mathbf{v}_i - \mathbf{m}_{vk})^T. \quad (27)$$

Training process of CGMM

IEEE COMPUTER SOCIETY

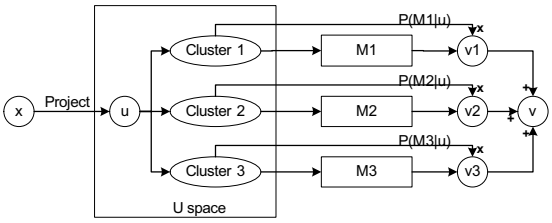

Figure 5. Illustration of the CGMM Inference Process

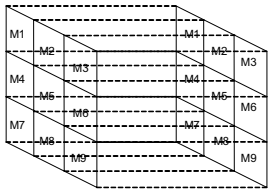

Figure 6. Illustration of Patch-based Models

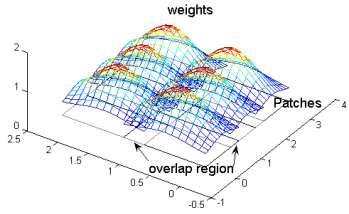

Figure 7. Overlapped Patch Partition and Weighted Pixel Synthesis

The CGMM is trained on $\mathcal{U}$ and $\mathcal{V}$ after the hidden subspaces are obtained. After the CGMM is trained, the bidirectional transforms are learned for every model, denoted as $\mathbf{A}_1, \ldots, \mathbf{A}_K$ and $\mathbf{B}_1, \ldots, \mathbf{B}_K$. For a new sample $\mathbf{x}$ it is first projected to $\mathcal{U}$ to obtain $\mathbf{u}$, then $\mathbf{v}$ can be computed as

$$\mathbf{v} = \sum_{k=1}^{K} P(M_k|\mathbf{u})(\mathbf{A}_k\mathbf{u}). \tag{28}$$

Here the posteriori can be calculated as

$$P(M_k|\mathbf{u}) = \frac{p(\mathbf{u}|M_k)p(M_k)}{\sum_{j=1}^{K} p(\mathbf{u}|M_j)p(M_j)}. \tag{29}$$

This inference process is illustrated in figure 5.

4. Integrated Framework for Image Style Transform

In this section we introduce the framework integrating the coupled space learning algorithm with image analysis and synthesis. Due to the complexity and high dimensionality of image, directly modelling the whole image sample space is very difficult and inefficient. However, the images associated with one visual object such as face maintain a stable global structure over the image, which will not change notably when the style changes. For example, whatever style you employ to express a face, the eyes, nose and mouth remain in all the styles and correspondence can be established between the same facial components in different image styles. Moreover, inter-pixel dependency is believed to consist within a neighboring region but not the whole image. Based on these rationales, we can partition the images into patches, and learn the dependencies within these patches. As illustrated in figure 6, in our patch-based approach, we train different models for patches in different positions. The patch-based strategy brings us two-fold merits: 1)The vector space dimensions for each model is much lower, thus both robustness and efficiency will be enhanced; 2)Since each model focuses on a small region, more subtle details can be captured in the model.

Since patches in different positions are independently modelled, the continuity in the patch-boundary cannot be guaranteed. To enhance the smoothness of the whole image and reduce the artifacts incurred by inter-patch discontinuities, we design a scheme (illustrated in figure 7), where adjacent patches are overlapped and the value of each pixel is weighed sum of values in synthesized patches covering that pixel. The weights of a patch on each pixel is attenuated softly as the distance of the pixel to the center of the patch increases. Specially, we employ an exponential function to describe the attenuation of the weights as $w(r) = exp(-r^2/\sigma^2)$, where r is the distance of a pixel to the patch-center, σ controls the speed of attenuation.

The whole procedure of image-transform can be described as follows:

Step 1. For an input image, divide it into overlapped patches. The partition scheme should follow that in the training stage.

Step 2. For each patch, infer its corresponding part using the Coupled Space Learning Model and Coupled GMM Model as introduced in previous sections.

Step 3. Use the weighed-sum scheme to combine all patches synthesized to form the entire image.

5. Experiments

In this section, we test the framework in two applications: face super-resolution and portrait style transforms.

5.1. Face Super-resolution

Super-resolution technique is to infer the high-resolution image based on a given low-resolution image in order to restore the details of the image. There are mainly two families of super-resolution methods: reconstruction-based and learning-based. Recently, learning-based approaches become more popular due to its capability of utilizing the prior knowledge in the inference, which is shown to play a crucial role, especially for face super-resolution. Baker. et al[1] propose the Gradient Prior Prediction algorithm with an MAP framework incorporated. Liu et al.[6] develop a framework integrating a global parametric linear model and a local patch-based Markov Random Field. Wang and Tang propose to use eigentransformation for inferring high-resolution faces from low-resolution ones [12][11]. Though face super-resolution seems to be a unidirectional process, however due to the fact that the super-resolution process is related to its inverse process: down sampling, thus we believe that it can be benefit from coupled learning.

We conduct our experiments on the FERET database[7], where 1276 images are selected into the training set while another 1272 images are selected into the testing set. Each image is preprocessed by affine transform to fix the positions of eyes and mouth

IEEE COMPUTER SOCIETY

center and cropped to size of 96×120 as high-resolution image. Then each one is low-pass-filtered and down-sampled to size of 24×30 as low-resolution image. Models are then trained on the 1276 pairs of images. Here, we employ the patch-based strategy, each image is divided into 11×11 overlapped patches, the size of patches in the high-resolution images is 16×20 while the size of patches in the low-resolution images is 4×5. In testing, a low-resolution image is input and its high-resolution counter-part is inferred by the algorithms in testing.

We compare the experimental results by our framework and those produced by other state-of-the-art algorithms in figure 8. It can be seen from the results that the quality of high-resolution images obtained by our Coupled Space Learning(CSL) framework is better than other algorithms. Moreover, the Coupled GMM further refines the details of the image and reduces the artifacts.

An objective evaluation of all these algorithms is shown in the following table in terms of mean square error compared to original high resolution image, which shows that the resultant images of Coupled Space Learning approximates the desired high resolution image better.

Algorithm	B-Spline	Baker	Eigen Trans.
MSE	0.00813	0.00290	0.00243
Algorithm	C.Liu	CSL	CSL + CGMM
MSE	0.00152	0.00094	0.00047

Table 4. Comparison of reconstruction errors of different methods

5.2. Portrait Style Transforms

The transforms between different art styles are mainly investigated in Computer Graphics. The most representative work is the image analogies[4] which use graphics techniques to filter the image so that it takes on desirable artistic effect. Tang and Wang[8] consider the problem as a learning problem and derive an eigen-transformation method to transform photos to sketches for recognition.

In our experiments, we apply the CSL framework to learn the transforms between two image styles for portrait. The face images in the FERET database[7] are divided into a training set with 1276 samples and a testing set with 1272 samples. The images are normalized to size of 96×120 and partitioned into 11×11 patches of size 16×20. The relationship between real portraits and PosterEdge-Style images and that between real portraits and HalfTone images are respectively learned in training stage. Figure 9 illustrates the results of forward and backward transforms between real photos and PosterEdge-Style renderings, while figure 10 illustrates the results for real photos and images with halftone effects. The results show the good performance of our framework in the application of style transforms.

6. Conclusion

In this paper, we have proposed a new framework to learn the dependency between two associated vector space. Correlative Component Analysis and Bidirectional Transforms are integrated to learn the relation in a coupled manner. We further develop a Coupled GMM model to enhance the framework's capability of modelling data under complex distribution by adapting each model to a part of samples and fuse them together. Experiments in face super-resolution and portrait style transforms clearly demonstrate the effectiveness of the framework.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region (4190/01E and N_CUHK409/03). The work was conducted at the Chinese University of Hong Kong.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. on PAMI*, pages 1167–1183, 2002.
[2] F. de la Torre and M. J. Black. Dynamic coupled component analysis. *Proc. of CVPR'01*, pages 643–650, 2001.
[3] W. T. Freeman and E. C. Pastor. Learning low-level vision. *Proc. of ICCV'99*, pages 1182–1189, 1999.
[4] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. *Proc. of SIGGRAPH'01*, pages 687–694, 2001.
[5] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, Inc., 5th edition, 2003.
[6] C. Liu, H. Shum, and C. S. Zhang. A two-step approach to hallucinating faces: Global parametric model and local nonparametric model. *Proc. of CVPR'01*, pages 192–198, 2001.
[7] P. J. Philips, H. Moon, S. A. Ryzvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. on PAMI*, 12(10):1090–1104, 2000.
[8] X. Tang and X. Wang. Face sketch synthesis and recognition. *Proc. of ICCV'03*, pages 687–694, 2003.
[9] T.Cootes, G.Edwards, and C.Taylor. Active appearance mode. *Proc. of ECCV'98*, pages 484–498, 1998.
[10] M. Turk and A. Pentland. Face recognition using eigenfaces. *Proc. of CVPR'91*, pages 586–591, 1991.
[11] X. Wang and X. Tang. Face hallucination and recognition. *Proc. of AVBPA'03*, pages 486–494, 2003.
[12] X. Wang and X. Tang. Face hallucination by eigentransformation. *IEEE Trans. on Systems, Man and Cybernetics-Part C, Special issues on Biometrics Systems*, 35(3), Aug 2004.
[13] A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons Ltd., 2nd edition, 2002.

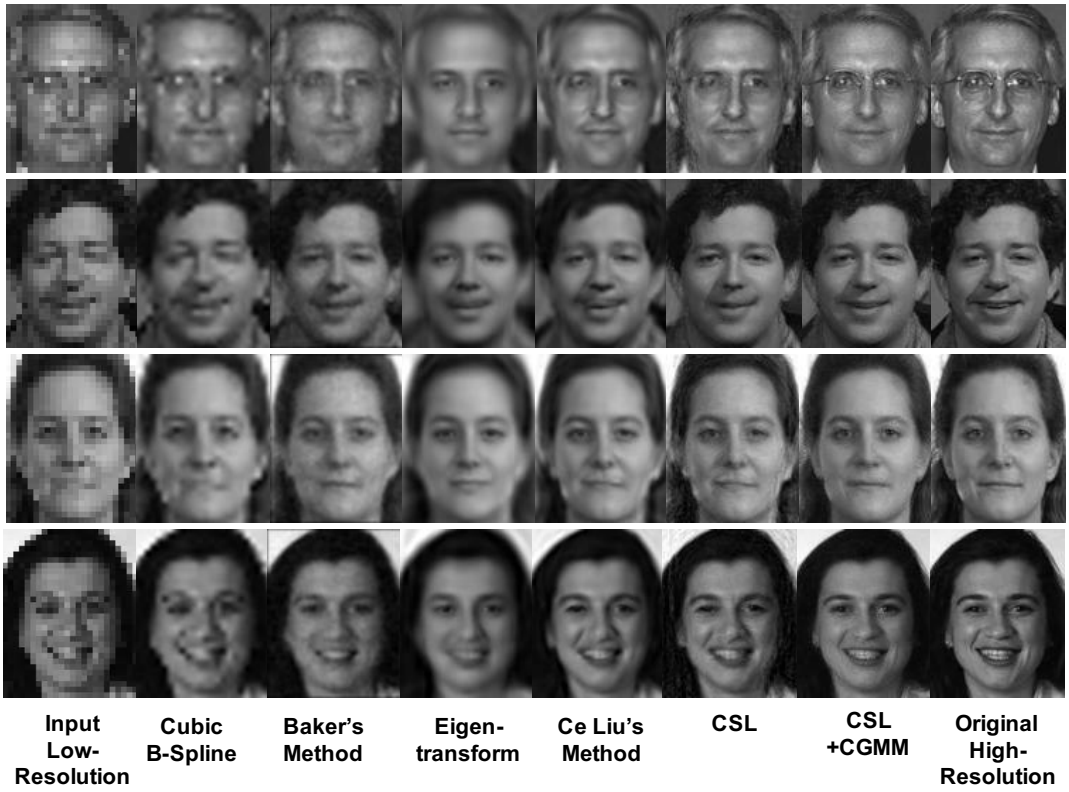

Figure 8. Results of face hallucination

Figure 9. Results of bidirectional transforms between real photos and PosterEdge-Renderings

Figure 10. Results of bidirectional transforms between real photos and Halftone-images

IEEE COMPUTER SOCIETY