Relatively-Paired Space Analysis

Zhanghui Kuang
http://i.cs.hku.hk/~zhkuang/
Kenneth K.Y. Wong
http://i.cs.hku.hk/~kykwong/

Department of Computer Science
The University of Hong Kong
Hong Kong

Abstract

Discovering a latent common space between different modalities plays an important role in cross-modality pattern recognition. Existing techniques often require absolutely-paired observations as training data, and are incapable of capturing more general semantic relationships between cross-modality observations. This greatly limits their applications. In this paper, we propose a general framework for learning a latent common space from relatively-paired observations (i.e., two observations from different modalities are more-likely-paired than another two). Relative-pairing information is encoded using relative proximities of observations in the latent common space. By building a discriminative model and maximizing a distance margin, a projection function that maps observations into the latent common space is learned for each modality. Cross-modality pattern recognition can then be carried out in the latent common space. To evaluate its performance, the proposed framework has been applied to cross-pose face recognition and feature fusion. Experimental results demonstrate that the proposed framework outperforms other state-of-the-art approaches.

1 Introduction

It is very common that an object can have very different presentations in different modalities. For instance, printed and hand-written forms of the same character can look very different, so are face photo and face sketch of the same person. Humans have little problem in recognizing objects across different modalities (e.g., matching face sketches to face photos). In contrast, conventional machine learning methods, such as k-NN classifiers, perform poorly in cross-modality pattern recognition since they assume both the training data and test patterns are randomly sampled from the same distribution (which is not the case in cross-modality pattern recognition) [31].

There exist a number of research studies in the literature targeting at cross-modality pattern recognition, which can be roughly classified into one of the three main approaches. The first approach consists of transforming one modality into another in a preprocessing step [3, 34]. The second approach is by extracting modality-invariant features to represent an object [17, 28]. A major limitation of these two approaches is that methods based on these approaches are usually tailor-made for each different modality pair involved in different recognition tasks. The third approach is to find an underlying latent common space shared between different modalities [4, 14, 20, 30, 31]. Unlike the first two approaches, the third approach does not depend on task-dependent knowledge. Methods based on the third

approach are therefore general frameworks that can be applied to different applications. Existing methods of the third approach often require absolutely-paired observations as training data. We refer to them as *Absolutely-Paired Space Analysis* (APSA). These methods assume the projections of paired observations being dependent in the latent space, and can only represent a binary relationship between observations (i.e., either paired observations or non-paired observations).

In many application scenarios, however, it is more suitable to consider relatively-paired observations (i.e., two observations from different modalities are more-likely-paired than another two) than absolutely-paired observations. For instance, given an input text query, an image search-engine (such as Google) will return a list of most probable images. The images clicked by the user are not absolutely-paired with the input text, but instead are more-likely-paired with the input text than other returned images. In fact, relative-pairing is a general pairing relationship that also covers absolute-pairing. One can safely consider two observations that are absolutely-paired being more-likely-paired than other non-paired observations. Another advantage of considering relatively-paired training data is that label information of the observations can be easily integrated to boost recognition performance. It is reasonable to assume observations with the same label being more-likely-paired than those with different labels. This strategy can be used to reduce within-class scatter while maximizing between-class scatter in the latent common space, as well as increase the minimium distance between observations with different labels in the latent common space.

In this paper, we propose a general framework named *Relatively-Paired Space Analysis* (RPSA) which works on relatively-paired observations. Note that RPSA is *not* a trivial extension of APSA as they are based on completely different models. APSA methods are often based on generative models [2, 4, 20] which either explicitly or implicitly assume the distributions of model parameters and noise (e.g., Gaussian distribution). The final estimation will be unreliable when real data do not fit the assumption. As opposed to APSA, our method is based on a discriminative model that has no distribution assumption. Besides, APSA methods learn a projection function for each modality by exploring the statistics dependence of the projections of absolutely-paired observations in the latent common space. This one-to-one absolute-pairing requirement makes them not suitable for relatively-paired observations. In our proposed framework, we compute the projection functions by preserving the relative proximities of observations in the latent common space (i.e., if observations a and b are more-likely-paired than observations a and c, then the distance between the projections of a and c in the latent common space is assumed to be longer than that between a and b). We validate our RPSA framework by applying it to cross-pose face recognition and feature fusion. Experimental results demonstrate that our proposed framework outperforms other state-of-the-art approaches. The main contributions of this paper are

1. We propose a general framework for automatically learning a latent common space between different modalities from relatively-paired observations, which, to the best of our knowledge, has not been explored before.

2. We apply our proposed RPSA framework to cross-pose face recognition and feature fusion, and achieve significant improvement in recognition performance compared with other state-of-the-art methods.

2 Related Work

There exist a large number of research studies on cross-modality pattern recognition in the literature. Due to page limitation, however, we focus our discussion only on those most relevant work that automatically learn a latent common space between different modalities. In [4], Borga *et al.* proposed the Canonical Correlation Analysis (CCA) which finds a latent common space by maximizing the correlation of the projections of cross-modality observations. Sun *et al*. [30] extended CCA by maximizing the within-class correlations and minimizing between-class correlations. In [8], Torre and Black developed the Asymmetric Coupled Component Analysis (ACCA) to explicitly learn the dependence of projections in a latent common space. Similarly, Lin and Tang [13] explored the coupled space by alternatively maximizing the correlation of projections of cross-modality observations and finding the relations between these projections. Different from CCA, Partial Least Square (PLS) [20, 22] chooses linear mappings such that the covariance between projections of cross-modality observations in the latent common space is maximized. Bilinear Models (BLM) was proposed in [31] to separate style and content. Besides, researchers have proposed advanced nonlinear methods based on GPLVM [9, 18, 28]. All the above methods require absolutely-paired observations as training data. Recently, Lampert and Krömer [12] learned a latent space based on weakly-paired data (i.e., subsets of observations of one modality are paired with those of another modality) by alternatively finding element pairs and maximizing covariance of projections of cross-modality observations. Sharma *et al*. [25] proposed a General Multi-view Analysis (GMA) approach which is solved as a generalized eigenvalue problem. Different from previous work, our proposed framework depends on neither prior distribution assumptions nor statistics computations, and learns a latent common space by preserving relative proximities of the relatively-paired training data in the latent common space.

Metric learning can be interpreted as finding a latent space for a single-modality observation space by linear projection. In [35], Xing *et al.* proposed to minimize the distances between samples from a similar set while keeping the distances of those from a dissimilar set above a threshold. Goldberger *et al.* [10] directly maximized a stochastic variant of the leave-one-out k-NN score on the training set. Since then, many other methods [7, 26, 33] were proposed to achieve a similar goal. However, these methods only focus on a single modality. In the recent work of Quadrianto and Lampert [21], they extended metric learning to multiple modalities by explicitly modeling linear projections. Their objective function is non-convex and thus the final optimum obtained depends on initialization. Moreover, their method requires the dimension of the latent common space to be known a priori. As opposed to their method, our model is convex which guarantees a global optimum, and can find a latent common space with any dimension in a single optimization.

Exploiting latent spaces can be also found in related research studies, such as local metric learning [1], hashing [5], multi-task learning [19], domain adaption [23] and ranking [32]. However, their goals are very different from the one in this paper.

3 Relatively-Paired Space Analysis

In this section, we describe our RPSA framework for learning a latent common space from relatively-paired observations. The goal is to find linear mappings that project observations from different modalities into a latent common space in which the relative proximities of the

relatively-paired observations are preserved.

3.1 The RPSA Model

Consider a set of M modalities $\{\Omega_1, \Omega_2, \ldots, \Omega_M\}$ with dimensions $\{d_1, d_2, \ldots d_M\}$ respectively, and a training dataset of N observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ with a corresponding flag set $\{t_1, t_2, \ldots, t_N\}$ such that $t_i \in \{1, \ldots, M\}$ indicates that $\mathbf{x}_i$ comes from Ω_{t_i}. Let the relative-pairing knowledge of the observations be represented by a set of triplets $S = \{(i, j, k)\}$, where each triplet (i, j, k) encodes that $\mathbf{x}_i$ and $\mathbf{x}_j$ are more-likely-paired than $\mathbf{x}_i$ and $\mathbf{x}_k$. Note that $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$ can come from either the same or different modalities. When they are from the same modality, "being more-likely-paired" means "being more similar".

To learn a latent common space Z with dimension d_z, we seek a $d_z \times d_m$ linear projection matrix $\mathbf{W}_{\Omega_m}$ for each modality Ω_m such that the relative proximities of the projections of the relatively-paired observations are preserved in Z, i.e.,

$$d(i, j) \leq d(i, k) \quad \forall (i, j, k) \in S, \tag{1}$$

where

$$d(i, j) = \|\mathbf{W}_{\Omega_{t_i}} \mathbf{x}_i - \mathbf{W}_{\Omega_{t_j}} \mathbf{x}_j\|^2 \tag{2}$$

denotes the squared Euclidean distance between the projections of $\mathbf{x}_i$ and $\mathbf{x}_j$ in Z. Let $\mathbf{W} = [\mathbf{W}_1\ \mathbf{W}_2\ \ldots \mathbf{W}_M]$ and $\mathbf{A}_{\Omega_m}$ be a $\sum d_n \times d_m$ matrix with all elements being zero except for row $\sum_{n<m} d_n + 1$ to row $\sum_{n \leq m} d_n$ being an identity matrix, such that $\mathbf{W}_{\Omega_m} = \mathbf{W}\mathbf{A}_{\Omega_m}$. Substituting this into (2) gives

$$\begin{aligned} d(i, j) &= (\mathbf{A}_{\Omega_{t_i}} \mathbf{x}_i - \mathbf{A}_{\Omega_{t_j}} \mathbf{x}_j)^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \mathbf{W} (\mathbf{A}_{\Omega_{t_i}} \mathbf{x}_i - \mathbf{A}_{\Omega_{t_j}} \mathbf{x}_j) \\ &= \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j}), \end{aligned} \tag{3}$$

where $\mathrm{Tr}(.)$ gives the trace of a matrix, $\mathbf{A} = \mathbf{W}^{\mathrm{T}}\mathbf{W}$, and

$$\mathbf{C}_{i,j} = (\mathbf{A}_{\Omega_{t_i}} \mathbf{x}_i - \mathbf{A}_{\Omega_{t_j}} \mathbf{x}_j)(\mathbf{A}_{\Omega_{t_i}} \mathbf{x}_i - \mathbf{A}_{\Omega_{t_j}} \mathbf{x}_j)^{\mathrm{T}}. \tag{4}$$

Substituting (3) into (1) gives

$$\mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,k}) - \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j}) \geq 0 \quad \forall (i, j, k) \in S. \tag{5}$$

(5) defines the relative proximity constraints on $\mathbf{A}$ which encodes $\mathbf{W}$ (i.e., the set of projection matrices). Since t_i, t_j, $t_k \in \{1, \ldots, M\}$, there are M^3 possible modality configurations for a triplet (i, j, k). When $t_i = t_j = t_k$, $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$ are from the same modality, and (5) provides constraints in one modality which is the same as metric learning. Now to learn the latent common space, we find a positive-semidefinite matrix $\mathbf{A}$ (i.e., $\mathbf{A} \succeq 0$.) which fulfills constraints (5). Note that if $\mathbf{A}^*$ is a solution, multiplying $\mathbf{A}^*$ by any arbitrary positive scalar will also give a solution. To specify a unique solution, we let $\mathrm{Tr}(\mathbf{A}) = 1$ and maximize the distance margin, i.e.,

$$\begin{aligned} &\max \quad \varphi \\ &s.t. \quad \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j,k}) \geq \varphi, \ \ \mathbf{A} \succeq 0 \ \text{ and } \ \mathrm{Tr}(\mathbf{A}) = 1, \ \ \forall (i, j, k) \in S, \end{aligned} \tag{6}$$

where $\mathbf{C}_{i,j,k} = \mathbf{C}_{i,k} - \mathbf{C}_{i,j}$. By introducing a positive slack variable to each relative proximity constraint (for improving robustness against noise), (6) can be reformulated into an SVM

style [36] energy function, given by

$$\begin{array}{ll} \min & \mathrm{Tr}(\mathbf{A})^2 + \gamma_1 \sum \xi_{i,j,k} \\ s.t. & \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j,k}) \geq 1 - \xi_{i,j,k},\ \mathbf{A} \succeq 0 \text{ and } \xi_{i,j,k} \geq 0,\ \forall (i,j,k) \in S, \end{array} \tag{7}$$

where γ_1 controls the relative weights of the regularization and loss terms. The regularization term $\mathrm{Tr}(\mathbf{A})^2$ not only regularizes the ambiguous problem in (5), but also forces $\mathbf{A}$ to be a low rank matrix. Inspired by [27], we employ Frobenius norm of $\mathbf{A}$ as the regularization term for the sake of simplicity and scalability, and get

$$\begin{array}{ll} \min & \|\mathbf{A}\|_{\mathrm{F}}^2 + \gamma_2 \sum \xi_{i,j,k} \\ s.t. & \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j,k}) \geq 1 - \xi_{i,j,k},\ \mathbf{A} \succeq 0 \text{ and } \xi_{i,j,k} \geq 0,\ \forall (i,j,k) \in S, \end{array} \tag{8}$$

where γ_2 is set to 1 in all our experiments. Although using Frobenius norm regularization does not theoretically guarantee a low rank optimum $\mathbf{A}^*$, all our experiments validate that a low rank approximation of $\mathbf{A}^*$ is sufficient in practice (e.g., a rank-25 approximation of $\mathbf{A}^*$ with a dimension of 3840×3840 performs well in cross-pose face recognition). (8) can be efficiently optimized by maximizing its dual problem alternatively with eigenvalue decomposition and off-the-shelf first order Newton algorithm such as L-BFGS-B [15].

After getting the optimum $\mathbf{A}^*$, we obtain $\mathbf{W}$ by minimizing $\left\|\mathbf{A}^* - \mathbf{W}^{\mathrm{T}}\mathbf{W}\right\|_{\mathrm{F}}$. Suppose the rows of $\mathbf{W}$ are orthogonal to each other, $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ will then be a positive-semidefinite matrix with rank d_z (i.e., the dimension of the latent common space Z). According to Eckart-Young theorem [29], $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ will be the rank-d_z approximation of $\mathbf{A}^*$. We perform eigenvalue decomposition over the positive-semidefinite matrix $\mathbf{A}^*$, getting $\mathbf{A}^* = \mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}$ with $\mathbf{U}$ being an orthogonal matrix and Σ a real diagonal matrix with decreasing singular values $\sigma_1 \geq \ldots \geq \sigma_{\sum d_m}$. We obtain $\mathbf{W} = \Sigma'\mathbf{U}^{\mathrm{T}}$ with Σ' being a diagonal matrix with decreasing diagonal values $\sqrt{\sigma_1}, \sqrt{\sigma_2}, \ldots, \sqrt{\sigma_{d_z}}, 0, \ldots, 0$. Linear projections $\mathbf{W}_{\Omega_m}$ for different dimensions of Z can be obtained after optimizing (8) and one eigenvalue decomposition. Note that the appropriate latent common space dimension d_z is application dependent, and is determined by cross validation in this paper.

3.2 Time Complexity

To solve problem (8), eigenvalue decomposition of $\mathbf{A}$ dominates the computation complexity at each iteration if the number of constraints is not far more than the dimensionality of A. The optimization algorithm can converge in a small number of iterations. After getting the optimum $\mathbf{A}^*$, one more eigenvalue decomposition is performed to obtain $\mathbf{W}$. In this case, the overall time complexity is $\mathcal{O}(t \cdot D^3)$ with $D = \sum d_m$ and t being around 10 in our experiments.

3.3 Discussion

We would like to discuss the differences between our model and some related work with exploiting latent spaces. Parameswaran and Weinberger [19] proposed multi-task metric learning which learns a metric for each task and a common metric for multi-task. Each metric is regularized separately, and only one task gets involved in each training triplet. Bronstein and Bronstein [5] developed multi-modality hashing method in the Adaboost framework. It has no regularization term and only uses similarity or dissimilarity pairs as constraints. Wang *et al*. [32] learned a ranking function for each modality. It has no cross-modality comparison and jointly regularizes ranking functions by $l_{(2,1)}$ norm. In [23], visual category

models are adapted to new domains with information-theoretic regularization and similarity or dissimilarity pairs as constraints. Therefore, Both the regularization and loss terms in [5, 19, 23, 32] are different from the proposed RPSA. Frome *et al*. [1] proposed globally-consistent local metric learning. It only learns a weight vector for each instance while a projection matrix is learned for each modality in our model. Our final optimization problem is semi-definite programming while theirs not, and thus the optimization methods are also different. We stress the problems solved in all the above work are very different from the one in this paper.

4 Experiments

We evaluated the performance of our proposed RPSA framework by applying it to cross-pose face recognition and feature fusion.

4.1 Training Triplets

Training triplets (i, j, k) can be generated in an unsupervised or supervised fashion. This kind of relatively-paired data can be collected from clickthrough data of search engines. It can also be generated from labels based on the principle that observations with the same label are expected to be more-likely-paired than those with different labels. Let l_i denote the label of an observation $\mathbf{x}_i$. Given a pair of cross-modality observations $(\mathbf{x}_p, \mathbf{x}_q)$ (where $t_p \neq t_q$) for an object, we define four types of triplets to describe the relative-pairing knowledge (see Table 1). Each triplet (i, j, k) suggests that $\mathbf{x}_i$ is more-likely-paired with $\mathbf{x}_j$ than with $\mathbf{x}_k$. Euclidean distance between two observations is used in defining nearest neighbor in Table 1. Figure 1 gives a graphical illustration for these four types of triplets. If the numbers of these four types of triplets are n_1, n_2, n_3 and n_4, respectively, for each given pair $(\mathbf{x}_p, \mathbf{x}_q)$, we say that the training data have a structure of (n_1, n_2, n_3, n_4). The total number of triplets is therefore $(n_1 + n_2 + n_3 + n_4) \times N_p$, where N_p is the number of pairs.

Table 1: Four types of triplets defined for describing relative-pairing information of a given pair of observations ($\mathbf{x}_p$, $\mathbf{x}_q$).

Type	Form	Num.	Remark
1	(p, q, q_1)	n_1	$\mathbf{x}_{q_1}$ is the kth ($k \leq n_1$) nearest neighbor of $\mathbf{x}_q$ s.t. $t_p \neq t_q \wedge t_q = t_{q_1} \wedge l_p = l_q \wedge l_q \neq l_{q_1}$
2	(q, p, p_1)	n_2	$\mathbf{x}_{p_1}$ is the kth ($k \leq n_2$) nearest neighbor of $\mathbf{x}_p$ s.t. $t_q \neq t_p \wedge t_p = t_{p_1} \wedge l_q = l_p \wedge l_p \neq l_{p_1}$
3	(p, p_1, p_2)	n_3	$\mathbf{x}_{p_1}$ is the kth ($k \leq n_3$) nearest neighbor of $\mathbf{x}_p$ s.t. $t_p = t_{p_1} \wedge l_p = l_{p_1}$ $\mathbf{x}_{p_2}$ is the kth ($k \leq n_3$) nearest neighbor of $\mathbf{x}_p$ s.t. $t_p = t_{p_2} \wedge l_p \neq l_{p_2}$
4	(q, q_1, q_2)	n_4	$\mathbf{x}_{q_1}$ is the kth ($k \leq n_4$) nearest neighbor of $\mathbf{x}_q$ s.t. $t_q = t_{q_1} \wedge l_q = l_{q_1}$ $\mathbf{x}_{q_2}$ is the kth ($k \leq n_4$) nearest neighbor of $\mathbf{x}_q$ s.t. $t_q = t_{q_2} \wedge l_q \neq l_{q_2}$

4.2 Cross-Pose Face Recognition

Faces observed under a particular pose can be considered as being sampled from one modality, and therefore faces observed under different poses correspond to different modalities. RPSA can be used to recognize faces under different poses, in which gallery faces are in one pose while probe faces are in another pose. Note that our method requires knowing the rough pose of each photo (i.e., to which modality it belongs) as in [24]. CMU PIE face

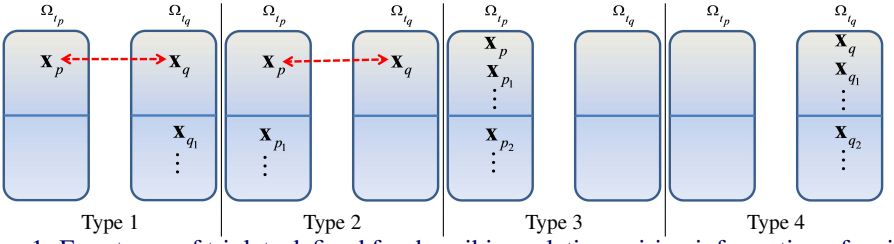


Figure 1: Four types of triplets defined for describing relative-pairing information of a given pair of observations $(\mathbf{x}_p, \mathbf{x}_q)$. $\mathbf{x}_p \leftarrow\cdots\rightarrow \mathbf{x}_q$ means $\mathbf{x}_p$ and $\mathbf{x}_q$ are paired observations from different modalities. Grids on the same horizontal line contain cross-modality observations with the same label.

database [1] was used in our experiments. This data set consists of 68 subjects, each of which has face photos in 13 different poses (indexed by c27/05/29/37/11/07/09/02/14/22/34/25/31). Photos in the same pose were aligned by the eyes and mouth. All photos were cropped and down-sampled to 48×40. Each photo was then reshaped into a column vector giving an observation $\mathbf{x}_i$. In our experiments, subject 1 to 34 were used as training data, while the rest were used as testing data. In the training phase, we derived a set of triplets for each training pair with a structure of $(5,5,0,0)$. We therefore had 10×34 triplets in total. The learned latent common space had a dimension of 25. In the testing phase, the nearest gallery face of each probe face was found in the learned latent common space, and the recognition rates were reported.

In Table 2, we compare our method with those using frontal faces (photos indexed by c27 in CMU PIE dataset) as gallery, in terms of mean recognition rates over different subsets of probe poses. It can be seen that RPSA is only slightly worse than TFA [20], but outperforms all the others. Note that TFA requires 14 user-elaborately-clicked points for photo alignment and Gabor filter for extracting complex features, whereas our method only needs 3 points (eyes and mouth) for photo alignment and directly employs the face image as a feature vector. We also compare our method with PLS [24] which, to the best of our knowledge, reports the best performance in the recent literature. It can be seen that our method is better than PLS when frontal faces are used as gallery and faces with other poses as probes.

Table 2: Mean recognition rates for frontal faces (c27) gallery.

Gallery	Probe	Method	Accuracy	Method	Accuracy
c27	c05/37/25/22/29/11/14/34	PGFR [16]	0.86	RPSA	**0.94**
c27	c05/22	TFA [20]	**0.95**	RPSA	0.93
c27	c05/29/37/11/07/09	LLR [8]	0.95	RPSA	**1.00**
c27	c05/29/37/11/07/09	ELF [11]	0.90	RPSA	**1.00**
c27	c05/29/37/11/07/09/02/14/22/34/25/31	PLS [24]	0.94	RPSA	**0.95**

4.3 Feature Fusion

For classifying patterns with different kinds of features stemming from different sources, a critical issue is to efficiently utilize these cross-modality features. A common solution is feature fusion by first projecting cross-modality features into a latent common space to

[1] http://vasc.ri.cmu.edu/idb/html/face/

reduce dimension and suppress noise, and then adding the paired projections together as a final feature vector. The fused feature for two modalities [30, 37] is usually given by

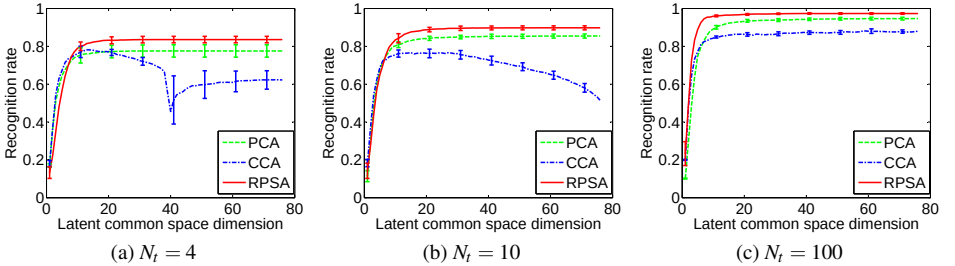
$$\mathbf{y} = \mathbf{W}_{\Omega_{t_i}}\mathbf{x}_i + \mathbf{W}_{\Omega_{t_j}}\mathbf{x}_j, \tag{9}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two feature vectors in different modalities for one object (i.e., $t_i \neq t_j$).

The proposed method was used to fuse features of UCI Multiple Features dataset[2]. This dataset consists of 2000 instances of ten hand-written numerals ('0'-'9'). Each instance has six features, namely Fou, Fac, Kar, Pix, Zer and Mor, with dimensions 76, 216, 64, 240, 47, and 6 respectively. We considered each feature as one modality. In our experiment, any two kinds of features were selected to fuse, and we had $C_2^6 = 15$ combination pairs. In the training phase, for each feature pair, the number of training data for each digit (N_t) was set to 4, 10 or 100. We derived a set of relatively-paired observations with a structure of $(4,4,3,3)$. Therefore, the total number of triplets is $(14 \times N_t \times 10)$. The latent common space had a dimension of 25, except for feature pairs involving Mor where it had a dimension of 6. In the testing phase, we find the nearest training fused feature with label for each testing fused feature. The experiment was repeated 10 times by randomly selecting fixed number of training data (i.e., $N_t \times 10$, here 10 is the number of digit categories). We evaluated our method by mean recognition rates and standard deviations.

Table 3: Mean recognition rates and standard deviations on UCI Multiple Features dataset.

Pairs		$N_t = 4$			$N_t = 10$			$N_t = 100$		
		PCA	CCA	RPSA	PCA	CCA	RPSA	PCA	CCA	RPSA
Fac	Fou	0.78±0.03	0.72±0.02	**0.83±0.02**	0.84±0.01	0.76±0.02	**0.89±0.01**	0.94±0.01	0.86±0.01	**0.97±0.00**
Fac	Kar	0.79±0.03	0.75±0.02	**0.82±0.02**	0.86±0.01	0.82±0.02	**0.89±0.01**	0.94±0.01	0.95±0.00	**0.98±0.01**
Fac	Pix	0.78±0.04	0.73±0.03	**0.83±0.03**	0.86±0.03	0.84±0.02	**0.91±0.01**	0.94±0.01	0.93±0.00	**0.98±0.00**
Fac	Zer	0.79±0.03	0.58±0.05	**0.83±0.03**	0.86±0.01	0.69±0.02	**0.90±0.01**	0.95±0.01	0.84±0.01	**0.97±0.01**
Fac	Mor	0.76±0.03	0.32±0.28	**0.80±0.04**	0.82±0.02	0.27±0.26	**0.84±0.02**	0.88±0.01	0.74±0.01	**0.92±0.01**
Fou	Kar	0.61±0.07	0.57±0.05	**0.77±0.03**	0.72±0.04	0.55±0.03	**0.87±0.01**	0.85±0.02	0.88±0.01	**0.97±0.00**
Fou	Pix	0.75±0.03	0.58±0.04	**0.84±0.03**	0.85±0.02	0.61±0.03	**0.92±0.01**	0.95±0.01	0.77±0.01	**0.98±0.00**
Fou	Zer	0.50±0.03	0.48±0.04	**0.71±0.03**	0.65±0.02	0.46±0.03	**0.77±0.01**	0.77±0.01	0.80±0.01	**0.85±0.01**
Fou	Mor	0.52±0.05	0.30±0.25	**0.59±0.04**	0.53±0.02	0.52±0.02	**0.63±0.02**	0.63±0.01	0.75±0.01	**0.74±0.01**
Kar	Pix	0.76±0.03	0.67±0.04	**0.80±0.03**	0.86±0.01	0.80±0.02	**0.90±0.01**	0.96±0.01	0.95±0.01	**0.97±0.00**
Kar	Zer	0.64±0.05	0.52±0.04	**0.77±0.04**	0.76±0.02	0.60±0.02	**0.86±0.01**	0.90±0.01	0.88±0.01	**0.94±0.01**
Kar	Mor	0.58±0.05	0.28±0.24	**0.64±0.04**	0.69±0.05	0.57±0.03	**0.71±0.03**	0.77±0.01	0.77±0.02	**0.84±0.02**
Pix	Zer	0.76±0.03	0.56±0.05	**0.82±0.03**	0.86±0.01	0.67±0.02	**0.89±0.01**	0.95±0.01	0.80±0.01	**0.95±0.00**
Pix	Mor	0.74±0.04	0.31±0.26	**0.77±0.04**	0.82±0.02	0.27±0.25	**0.82±0.03**	0.87±0.01	0.71±0.01	**0.88±0.02**
Zer	Mor	0.49±0.05	0.26±0.13	**0.63±0.04**	0.54±0.03	0.62±0.02	**0.68±0.01**	0.61±0.01	**0.73±0.01**	0.72±0.02



(a) $N_t = 4$ (b) $N_t = 10$ (c) $N_t = 100$

Figure 2: Mean recognition rates and standard deviations of PCA, CCA, and RPSA under different latent common space dimensions.

We first compared our method with PCA which reduces dimension independently in each modality, and Canonical Correlation Analysis (CCA) [4] which is a representative multi-feature fusion approach. In our experiments, both PCA and CCA kept the same latent common space dimension as RPSA. It can be seen from Table 3 that our method outperforms

[2] http://archive.ics.uci.edu/ml/datasets/Multiple+Features

Table 4: Recognition rates on Multiple Features Dataset.

Pairs		DCCA	bgCCA	bgDCCA	bsCCA	bsDCCA	PLS	RCE	RPSA
Fac	Fou	0.89	0.86	0.89	0.84	0.88	0.94	0.95	**0.97**
Fac	Kar	0.98	0.95	0.98	0.93	0.98	0.94	**0.98**	**0.98**
Fac	Pix	0.97	0.86	0.97	0.86	0.97	0.94	0.95	**0.98**
Fac	Zer	0.88	0.86	0.88	0.84	0.87	0.96	**0.97**	**0.97**
Fac	Mor	0.82	0.75	0.82	0.74	0.81	0.88	0.88	**0.92**
Fou	Kar	0.90	0.90	0.90	0.88	0.89	**0.97**	0.96	**0.97**
Fou	Pix	0.89	0.77	0.89	0.74	0.87	**0.98**	0.95	**0.98**
Fou	Zer	0.83	0.82	0.83	0.80	0.81	0.81	**0.85**	**0.85**
Fou	Mor	0.77	0.75	0.77	0.74	0.76	0.44	**0.80**	0.74
Kar	Pix	0.95	0.94	0.95	0.93	0.94	**0.98**	0.96	0.97
Kar	Zer	0.88	0.90	0.88	0.89	0.86	0.83	**0.96**	0.94
Kar	Mor	0.80	0.77	0.80	0.76	0.79	0.62	**0.86**	0.84
Pix	Zer	0.87	0.83	0.87	0.80	0.86	0.84	0.94	**0.95**
Pix	Mor	0.79	0.73	0.79	0.71	0.77	0.71	0.84	**0.88**
Zer	Mor	0.75	0.72	0.75	0.70	0.74	0.72	**0.77**	0.72

both PCA and CCA with impressive margin for all the cases, except for Zer-Mor feature pair with $N_t = 100$. Besides, RPSA performs reasonably well even when having only 4 training instances for each digit (i.e., $N_t = 4$). This is expected since RPSA is a discriminative model and it does not need any distribution assumption which usually needs many training samples for a good estimation or fitting. Figure 2 shows the recognition rates under different latent common space dimensions for Fac-Fou feature pair. For $N_t = 4$ and $N_t = 10$, RPSA is slightly poorer than PCA and CCA when the dimension of the latent common space is less than 10 (see Figure 2 (a) and (b)). In fact, both PCA and CCA also have very low recognition rates under such settings, and it is therefore impractical to learn an extremely low dimension latent common space for pattern recognition. Nonetheless, this problem can be avoided by optimizing (7) with trace regularization. When the number of training data increases, our method performs much better than PCA and CCA under all dimensions of the latent common space (see Figure 2 (c)).

The proposed method was also compared with Discriminative Canonical Correlation Analysis (DCCA) [30], Partial Least Squares (PLS), bgCCA, bgDCCA, bsCCA, bsDCCA (two kinds of variants of CCA and DCCA) and Random Correlation Ensemble (RCE) [37]. For fair comparison, all the methods employ nearest neighbor method as classifier. The results of competitors are from Table 2 in [37]. From Table 4, we see that RPSA is superior to DCCA, bgCCA, bgDCCA, bsCCA, bsDCCA and PLS in most cases. RPSA shows advantages compared with RCE, although RCE is a sophisticated method which first finds random cross-view correlations between within-class examples and then boosts performance by ensemble learning. Similarly, our method can also be further improved by ensemble learning.

5 Conclusion and Future work

In this paper, we have proposed a framework called Relatively-Paired Space Analysis (RPSA) which can automatically learn a latent common space between multiple modalities from relatively-paired observations. Relative-pairing can explore more general semantic relationships between observations than absolute-pairing, and allows easy integration of label information. Theoretically, RPSA is a discriminative model which does not assume any parameter

or noise distribution, and is a general framework which can be used in any cross-modality pattern recognition. We have evaluated the performance of RPSA by applying it to cross-pose face recognition and feature fusion. Experimental results show that RPSA outperforms other state-of-the-art techniques, some of which are tailored for the particular problems. We have made the code available online (http://i.cs.hku.hk/∼zhkuang/Software.html). In future work, we would like to extend RPSA to a nonlinear version.

References

[1] F. Andrea, S. Yoram, F. Sha, and M. Jitendra. Learning globally-consistent local distance functions for shape-based image retrieval. In *ICCV*, pages 1–8, 2007.

[2] F.R. Bach and M.I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005.

[3] V. Blanz, P. Grother, P.J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *CVPR*, volume 2, pages 454–461, 2005.

[4] M. Borga, H. Knutsson, and T. Landelius. Learning canonical correlations. In *SCIA*, volume 1, pages 1–8, 1997.

[5] M.M. Bronstein and A.M. Bronstein. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2010.

[6] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose-invariant face recognition. *TIP*, 16(7):1716–1725, 2007.

[7] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information theoretic metric learning. In *ICML*, pages 209–216, 2007.

[8] F. De la Torre and M.J. Black. Dynamic coupled component analysis. In *CVPR*, volume 2, pages 643–650, 2001.

[9] C.H. Ek, J. Rihan, P.H.S. Torr, G. Rogez, and N.D. Lawrence. Ambiguity modeling in latent spaces. In *MLMI*, 2008.

[10] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, pages 513–520, 2005.

[11] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *PAMI*, 26(4):449–465, 2004.

[12] C. Lampert and O. Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *ECCV*, pages 566–579, 2010.

[13] D. Lin and X. Tang. Coupled space learning of image style transformation. In *ICCV*, volume 2, pages 1699–1706, 2005.

[14] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, pages 13–26, 2006.

[15] D.C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[16] X. Liu and T. Chen. Pose-robust face recognition using geometry assisted probabilistic modeling. In *CVPR*, volume 1, pages 502–509, 2005.

[17] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2): 91–110, 2004.

[18] R. Navaratnam, A.W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *ICCV*, 2007.

[19] S. Parameswaran and K. Weinberger. Large margin multi-task metric learning. In *NIPS*, pages 1–9, 2010.

[20] S.J.D. Prince, J. Warrell, J.H. Elder, and F.M. Felisberti. Tied factor analysis for face recognition across large pose differences. *PAMI*, 30(6):970–984, 2008.

[21] N. Quadrianto and C. Lampert. Learning multi-view neighborhood preserving projections. In *ICML*, pages 425–432, 2011.

[22] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*, pages 34–51, 2006.

[23] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 1–14, 2010.

[24] A. Sharma and D.W. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *CVPR*, pages 593–600, 2011.

[25] A. Sharma, A. Kumar, H. Daume, and D.W. Jacobs. Generalized multiview analysis: a discriminative latent space. In *CVPR*, pages 2160–2167, 2012.

[26] C. Shen, J. Kim, L. Wang, and A. Hengel. Positive semidefinite metric learning with boosting. In *NIPS*, pages 1651–1659, 2009.

[27] C. Shen, J. Kim, and L. Wang. A scalable dual approach to semidefinite metric learning. In *CVPR*, pages 2601–2608, 2011.

[28] A.P. Shon, K. Grochow, A. Hertzmann, and R.P.N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In *NIPS*, 2006.

[29] G.W. Stewart. On the early history of the singular value decomposition. *SIAM review*, pages 551–566, 1993.

[30] T. Sun, S. Chen, J. Yang, and P. Shi. A novel method of combined feature extraction for recognition. In *ICDM*, pages 1043–1048, 2008.

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

[32] B. Wang, J. Tang, W. Fan, S. Chen, Z. Yang, and Y. Liu. Heterogeneous cross domain ranking in latent space categories and subject descriptors. In *CIKM*, 2009.

[33] K.Q. Weinberger, J. Blitzer, and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[34] B. Xiao, X. Gao, D. Tao, Y. Yuan, and J. Li. Photo-sketch synthesis and recognition based on subspace learning. *Neurocomputing*, 73(4-6):840–852, 2010.

[35] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, volume 15, pages 505–512, 2002.

[36] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In *NIPS*, pages 2214–2222, 2009.

[37] J. Zhang and D. Zhang. A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples. *Pattern Recognition*, 44(6):1162 – 1171, 2011.

[38] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520, 2011.