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

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) [9].

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 [2, 10]. The second approach is by extracting modality-invariant features to represent an object [6, 12]. 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 [3, 4, 7, 8, 9]. 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 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 [1, 3, 7] 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).

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 $\mathbf{x}_k$. We have

$$\|\mathbf{W}_{\Omega_{t_i}}\mathbf{x}_i - \mathbf{W}_{\Omega_{t_j}}\mathbf{x}_j\|^2 \leq \|\mathbf{W}_{\Omega_{t_i}}\mathbf{x}_i - \mathbf{W}_{\Omega_{t_k}}\mathbf{x}_k\|^2, \tag{1}$$

where $\mathbf{W}_{\Omega_m}$ is the linear projection matrix for each modality Ω_m. Let $\mathbf{W} = [\mathbf{W}_1\ \mathbf{W}_2\ \ldots \mathbf{W}_M]$, $\mathbf{A} = \mathbf{W}^{\mathrm{T}}\mathbf{W}$, 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 (1) gives

$$\mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,k}) - \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j}) \geq 0 \quad \forall (i, j, k) \in S, \tag{2}$$

where $\mathrm{Tr}(.)$ gives the trace of a matrix, and

$$\mathbf{C}_{i,j} = (\mathbf{A}_{\Omega_{t_i}}\mathbf{x}_i - \mathbf{A}_{\Omega_{t_j}}\mathbf{x}_j)(\mathbf{A}_{\Omega_{t_i}}\mathbf{x}_i - \mathbf{A}_{\Omega_{t_j}}\mathbf{x}_j)^{\mathrm{T}}. \tag{3}$$

Let $\mathbf{C}_{i,j,k} = \mathbf{C}_{i,k} - \mathbf{C}_{i,j}$. By introducing a positive slack variable to each relative proximity constraint (for improving robustness against noise) and a regularization term, learning the projection matrix for each modality can be reformulated into an SVM style [11] energy function, given by

$$\begin{aligned} \min \quad & \|\mathbf{A}\|_{\mathrm{F}}^2 + \gamma \sum \xi_{i,j,k} \\ s.t. \quad & \mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j,k}) \geq 1 - \xi_{i,j,k},\ \ \mathbf{A} \succeq 0 \ \text{and}\ \ \xi_{i,j,k} \geq 0,\ \ \forall (i, j, k) \in S, \end{aligned} \tag{4}$$

where $\|\mathbf{A}\|_{\mathrm{F}}$ is the Frobenius norm of $\mathbf{A}$, and γ controls the relative weights of the regularization and loss terms.

We optimize (4) efficiently by maximizing its dual problem alternatively with eigenvalue decomposition and off-the-shelf first order Newton algorithm such as L-BFGS-B [5]. After getting the optimum $\mathbf{A}^*$, we obtain $\mathbf{W}$ by eigenvalue decomposition.

In conclusion, 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).

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

[2] 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.

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

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

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

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

[7] 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.

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

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

[10] 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.

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

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