Robust Feature Learning for View-Unknown Image Classification

Zhengming Ding
Adivisor: Yun Fu
Northeastern University, Boston, MA, USA
Email: allanding@ece.neu.edu

***Abstract*— Multi-view images are of great abundance in real-world applications, since various view-points and multiple sensors desire to represent the image data in a better way. Conventional multi-view learning methods aimed to learn multiple view-specific transformations meanwhile assumed the view knowledge of training and test data were available in advance. However, they would fail when we do not have any prior knowledge for the probe data's view information, since the correct view-specific projections cannot be utilized to extract effective feature representations. In my research, I manage to develop a Common Low-Rank Subspace (CLRS) algorithm to deal with this problem in view-unknown image classification, which attempts to mitigate the semantic gap across multiple views through seeking a view-shared low-rank projection shared by multiple view-specific transformations. The common low-rank subspace makes our algorithm more flexible when addressing the challenging issue without any prior knowledge of the probe data's view information. To that end, two different settings of experiments on several multi-view benchmarks are designed to evaluate the proposed approach.**

I. Introduction

Multi-view image classification has attracted a great deal of attention recently, since multi-view images are frequently seen in reality [1], [2], [3], [4]. Take face image as an example. Face images captured with cameras located in various viewpoints would have pose variations while different devices would generate different modalities, e.g., low-resolution face taken by a cellphone or even collected with near-infrared sensor. Such image data with large pose or modality divergence would result in a challenging classification problem. Here, we consider cross-pose image and multi-modal image as multi-view image. In general, different views can be treated as different domains drawn from different distributions. Therefore, it is the key to adapt one view to another view to minimize the distribution divergences across them [1], [5].

Conventional multi-view subspace methods [1], [2] were developed to seek many view-specific projections, which transform different views into a common view-free space. Along this line, Canonical Correlation Analysis (CCA) [6] was the most representative one, which learned two projections, each for one view, to align two-view data into the shared space, respectively. Further, multi-view CCA [7] was proposed and extended to multiple view cases based on CCA. Following this, Kan et al. designed a Multi-view Discriminant Analysis (MvDA) algorithm [1], which sought an effective shared space by jointing multiple view-specific linear projections learning and Fisher constraint in a unified framework. One common drawback is that those previous researches [1], [2], [7] mainly dealt with the multi-view learning tasks by applying one labeled view to predict another unlabeled view. Hence, we have to know the view knowledge of training and test data ahead of time, since they only learn multiple view-specific projections. Only with view-information at hand can the view-specific projections be adopted to the exact views, therefore, we need a lot of prior knowledge in real-world multi-view learning scenarios.

Unfortunately, we cannot always obtain the test data's view information in advance at many real-world scenarios, since the test data are always accessible during evaluation. For example, a face image could be captured at running time with view-unknown camera so that we cannot get its exact view knowledge. In such cases, conventional multi-view learning methods cannot work, since they only built multiple view-specific projections during training stage, which are not helpful for each view-known test data. Another phenomenon is that the test images can be in the same distribution with the training data or totally different distributions from the training data. This leads to two scenarios: "traditional multi-view learning" [1], [2], [7] and "multi-view transfer learning"[8], [4] . When fighting off the target multi-view data with no prior knowledge either view information or label knowledge or both, we can ask help from an auxiliary multi-view sources to facilitate the target learning problem. In this scenario, transfer learning has shown appealing performance in dealing with limited data and challenge no labeled data [9]. Along this line, feature adaption is a popular strategy in transfer learning, which aims to extract effective domain invariant features to reduce the domain shift so that the source knowledge could be transferred to the target [10], [8].

II. Summary

So far, we have developed a multi-view learning algorithm, named Collective Low-Rank Subspace (CLRS), to deal with the challenge where the view knowledge of the test data are unavailable during the learning task (Fig. 1), which is extended by my ICDM-14 work [5]. Following conventional multi-view subspace learning algorithms, we also learn the view-specific transformations for view-known training data to project the data into a latent view-free space in the training stage. Since we do not know the probe data's view information, we need to find a surrogate to preserve as much class information as possible, meanwhile reducing the

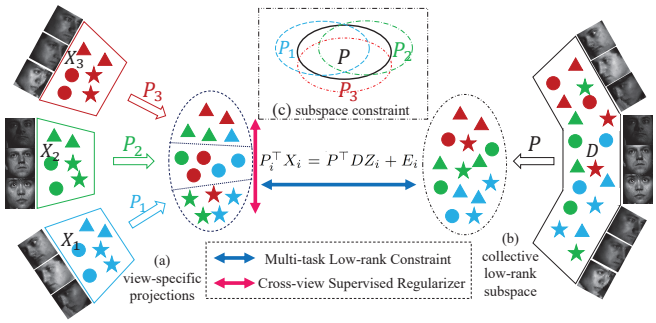

Fig. 1. Framework of the proposed Collective Low-Rank Subspace (CLRS) algorithm. Here we show three views (poses) and the same color represents the same view, while each view consists of three classes (the same shape denotes the data with same label). (a) We still adopt multiple view-specific projections $\{\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3\}$ for three views $\{\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3\}$ during the training stage. (b) Aiming to address the view-unknown testing data, we seek a surrogate by learning a low-rank shared transformation P for data D with three mixed views. (c) To further uncover more shared knowledge across multiple views, we adopt low-rank sparse decomposition to make the common P capture more shared information across those view-specific ones ($\mathrm{P_i} = \mathrm{P} + \mathrm{S_i}$, where $\mathrm{S_i}$ is the sparse residue of the i-th view projection).

impact of view divergence for mixed view-unknown test data, either in the same distribution or different distributions. On the account that the multiple view-specific projections all preserve the within-class knowledge for its specific view. In other words, those view-specific projections should have the similar discriminability for classification in different views. In other words, it is essential to find the consistent knowledge across multiple view-specific projections for view-unknown test data. To seek a more effective projection for view-unknown test data, we employ a collective low-rank projection to uncover most of the compatible structure across multiple view-specific projections, which are decomposed into the common part and sparse unique parts. Thus, the proposed algorithm is more flexible to solve real-world multi-view problems when we cannot have the view or even label information for the probe data at hand.

A. The Proposed Algorithm

Assume we have k-view training data as $\mathrm{X} = [\mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{k}]$, and each view $\mathrm{X_i} \in \mathbb{R}^{d\times m}$ contains the same c classes with m data samples. The view-specific transformation $\bar{\mathbf{P}}_\mathrm{i} \in \mathbb{R}^{d\times d}$ would be learned for the i-th view $\mathrm{X_i}$ following the conventional multi-view learning. Hence, each $\bar{\mathbf{P}}_\mathrm{i}$ represents the basis to expand the space of each view $\mathrm{X_i}$, i.e., $\bar{\mathbf{P}}_\mathrm{i} = \mathrm{X_i A_i}$, where $\mathrm{A_i}$ is weight matrix. As discussed before, multiple view-specific projections have the similar discriminability in their own view so that they should have a lot of shared knowledge. Then, we aim to seek as many common bases as possible across multi-view data so that such common basis can be generalized to view-unseen test data. To this end, we adopt a collective low-rank transformation $\bar{\mathbf{P}} \in \mathbb{R}^{d\times d}$ to uncover such consistent knowledge that it can be extended to work for view-unknown test data. Specifically, we exploit low-rank sparse decomposition by assuming each $\bar{\mathbf{P}}_\mathrm{i}$ is combined of $\bar{\mathbf{P}}$ and their unique sparse residue $\bar{\mathbf{S}}_\mathrm{i} \in \mathbb{R}^{d\times d}$, so more common knowledge could be uncovered.

Since $\bar{\mathbf{P}}$ is low-rank, there are many bases very similar, resulting in much redundant information within $\bar{\mathbf{P}}$. Assume the rank of $\bar{\mathbf{P}}$ is p ($p \ll d$), hence, we can adopt the p bases to extract effective features from multi-view data, which could help well deal with the *curse of dimensionality*. Specifcially, we have $\mathrm{P} \in \mathbb{R}^{d\times p}, \mathrm{P_i} \in \mathbb{R}^{d\times p}, \mathrm{S_i} \in \mathbb{R}^{d\times p}$ are p columns of $\bar{\mathbf{P}}, \bar{\mathbf{P}}_\mathrm{i}, \bar{\mathbf{S}}_\mathrm{i}$, respectively. And we further add an orthogonal constraint $\mathrm{P^\top P} = \mathrm{I}_p$ ($\mathrm{I}_p \in \mathbb{R}^{p\times p}$ is an identity matrix) to make the P with the full rank of p.

Following the idea of low-rank subspace learning, we desire to exploit low-rank representation to build a bridge across the view-specific features and the shared features (Fig. 1). Hence, knowledge across multiple view-specific transformations could be transferred to the common subspace. Due to the real-world data are always noisy, we design a sparse error term to figure out the noise or outliers. Finally, the objective function can be achieved by integrating common subspace and low-rank reconstruction into a unified framework as:

$$\begin{aligned}\min_{\substack{\mathrm{P},\mathrm{Z_i},\mathrm{E_i},\\ \mathrm{S_i},\mathrm{P_i},\mathrm{Z}}} \sum_{i=1}^{k} \left(\mathrm{rank}(\mathrm{Z_i}) + \lambda_0\|\mathrm{S_i}\|_1 + \lambda_1\|\mathrm{E_i}\|_{2,1}\right) + \lambda_2\Omega(\mathrm{P},\mathrm{Z}) \\ \text{s.t.}\ \ \mathrm{P_i^\top X_i} = \mathrm{P^\top DZ_i} + \mathrm{E_i},\ \mathrm{P_i} = \mathrm{P} + \mathrm{S_i}, \\ i = 1, \cdots, k,\ \ \mathrm{P^\top P} = \mathrm{I}_p,\end{aligned} \tag{1}$$

where $\mathrm{Z_i} \in \mathbb{R}^{\bar{m}\times m}$ is the i-th low-rank reconstruction coefficient. $\mathrm{E_i} \in \mathbb{R}^{p\times m}$ is the error term and $\|\cdot\|_{2,1}$ is the $L_{2,1}$-norm, i.e., $\|\mathrm{E_i}\|_{2,1} = \sum_{k=1}^{p}\sqrt{\sum_{j=1}^{m}([\mathrm{E_i}]_{kj})^2}$, which aims to detect and remove outliers. And λ_1 and λ_2 are two trade-offs to balance three parts. $\Omega(\mathrm{P},\mathrm{Z})$ is a regularizer to preserve more discriminative information across multiple views.

In the above objective function, $\mathrm{D} \in \mathbb{R}^{d\times\bar{m}}$ denotes the data with mixed k views, which has different definitions in different scenarios. In feature learning setting, D means the dictionary ($\bar{m}$ is the atom size of dictionary), which usually adopts the data itself X for simplicity. In this paper, we also directly use X as the basis. Whilst in transfer learning setting, D denotes the unlabeled target domain and X represents the well-labeled source domain. We can easily understand that we are dealing with an unlabeled multi-view dataset by borrowing the knowledge from a well-learned source domain. Objective function (1) would help facilitate the target learning with the the view/label knowledge of source domain.

To better utilize the label information in the training stage, we employ a supervised graph regularizer to align cross-view data within the same class. Model (1) only utilizes the view-information of the training data so that it works in a weakly supervised fashion. Moreover, model (1) exploits a multi-task scheme, that is, data from each view are reconstructed by the commonly projected data in an individual manner. It is very important to align different views to make the learned collective subspace more discriminative. We first denote the projected low-dimensional data of each view $\mathrm{Y}_i = \mathrm{P}_i^\top\mathrm{X}_i$ ($\mathrm{P^\top DZ}_i \in \mathbb{R}^{p\times m}$ can be treated as its clean version), so the multi-view projected data $\mathrm{Y} = [\mathrm{Y}_1, \cdots, \mathrm{Y}_k] \approx \mathrm{P^\top DZ} = \mathrm{P^\top D}[\mathrm{Z}_1, \cdots, \mathrm{Z}_k] \in \mathbb{R}^{p\times km}$. Based on the reconstructed

TABLE I

RECOGNITION PERFORMANCE (%) OF 10 ALGORITHMS ON THE **ORIGINAL IMAGES** FROM CMU-PIE FACE DATASET, IN WHICH CASE 1: {C02, C14}, CASE 2: {C02, C27}, CASE 3: {C14, C27}, CASE 4: {C05, C07, C29}, CASE 5: {C05, C14, C29, C34}, CASE 6: {C02, C05, C14, C29, C31}

	PCA[11]	LDA[12]	LPP[13]	TFRR[14]	SRRS[15]	LRCS [5]	MvDA[1]	RMSL[3]	Ours
Case 1	69.03±0.08	70.46±0.05	57.25±0.06	77.92±0.03	78.27±0.04	87.78±0.22	85.23±0.05	**88.15±0.06**	87.24±0.03
Case 2	69.21±0.08	71.32±0.02	58.83±0.07	76.24±0.12	78.74±0.23	86.67±0.09	85.81±0.09	**87.05±0.07**	86.82±0.11
Case 3	68.52±0.12	63.51±0.75	59.25±0.56	75.29±0.07	77.45±0.02	87.38±0.39	86.12±0.12	87.40±0.17	**87.97±0.09**
Case 4	52.65±0.04	56.53±0.02	43.56±0.08	69.74±0.05	71.44±0.03	74.84±0.04	**75.36±0.18**	75.16±0.12	72.97±0.03
Case 5	34.94±0.08	24.07±0.25	19.67±0.05	33.91±0.12	38.86±0.02	44.48±0.03	**54.13±0.16**	44.93±0.11	45.92±0.06
Case 6	29.09±0.01	07.06±0.01	13.11±0.01	28.36±0.04	30.16±0.02	36.17±0.11	**47.67±0.18**	37.14±0.08	39.17±0.08

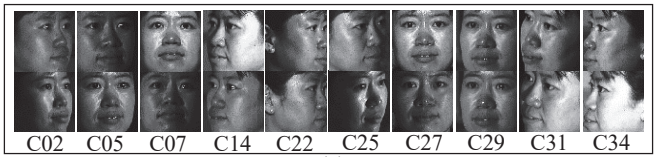

Fig. 2. Face samples from different views of one individual in CMU-PIE cross-pose face dataset.

view-invariant features, we build a Fisher regularizer $\Omega(\mathrm{P}, \mathrm{Z})$ to keep the within-class compactness while preserving the between-class discrimination.

B. Experiments

1) Datasets & Experimental Setting: **CMU-PIE Face dataset** totally consists of 68 subjects with different poses (Fig. 2). There is 21 different illumination variations for the samples of each subject. Specifically, we adopt such different poses, which show large view variances within the same subject across different poses. In the experiment, we select different numbers of views to build various evaluation scenarios. For each pose, we randomly choose 10 samples for training while the left for testing. Furthermore, we crop faces into size of 64×64 and adopt the gray-scale value as the input.

In our experiment, we address the challenging problem where the view knowledge of the probe data is unavailable. Thus, conventional multi-view methods would fail. Therefore, we mainly compare with PCA [11], LDA [12], LPP [13], TFRR [14], SRRS [15], RMSL [3] and LRCS [5]. Specifically, LDA, RSR, SRRS, RMSL and ours are five supervised algorithms; and PCA, LPP, TFRR and LRCS are four unsupervised methods. Furthermore, we compare with one conventional multi-view subspace learning algorithm, MvDA [1], by providing it extra view knowledge of the probe data to show the effectiveness of our algorithm.

The nearest neighbor classifier (NNC) is adopted to testify the final classification results. We choose ten images per individual per pose to build the training set, and the remaining data are used for testing. We do 5 random selections and report the average performance. Table I represents recognition performance on the original images.

III. FUTURE PLANS AND CHALLENGES

Currently, we develop a linear algorithm, which is not effective enough in feature learning compared with deep learning models. Thus, deep learning based algorithm will be proposed to effectively extract view-invariant features. Moreover, our current model cannot well deal with large-scale data, as the complexity is $O(m^3)$, and therefore, more efficient alternative will be developed to well handle large-scale dataset. Furthermore, our current model assumes the dimensionality of each view is same, which may be not satisfied in reality. Thus, a more general model will be designed to fight off this shortcoming.

Specifically, I will focus on the last two challenges by designing a more efficient and general model [*from Feb 2017 to June 2017*]. After that, I will explore deep learning for view-unseen image classification [*from July 2017 to May 2018*].

REFERENCES

[1] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proceedings of European Conference on Computer Vision*. Springer, 2012, pp. 808–821.

[2] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou, "Regularized latent least square regression for cross pose face recognition," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 1247–1253.

[3] Z. Ding and Y. Fu, "Robust multi-view subspace learning through dual low-rank decompositions," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1181–1187.

[4] P. Yang and W. Gao, "Multi-view discriminant transfer learning," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1848–1854.

[5] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *IEEE International Conference on Data Mining*, 2014.

[6] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[7] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Conference on Data Mining and Data Warehouses*, 2010, pp. 1–4.

[8] Z. Ding, M. Shao, and Y. Fu, "Incomplete multisource transfer learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–14, 2016.

[9] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[10] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *International Journal of Computer Vision*, pp. 1–20, 2014.

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[12] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[13] X. He and P. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, p. 153.

[14] R. Liu, Z. Lin, F. De la Torre, and Z. Su, "Fixed-rank representation for unsupervised visual learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 598–605.

[15] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2160–2173, 2016.