Multi-view Nonparametric Discriminant Analysis for Image Retrieval and Recognition

Guanqun Cao, Alexandros Iosifidis, *Senior Member, IEEE*, Moncef Gabbouj, *Fellow, IEEE*

***Abstract*—A novel multi-view nonparametric discriminant analysis method is proposed for the application of cross-modal image retrieval and zero-shot recognition. We exploit the class boundary structure and discrepancy information of the available views in order to formulate an optimization criterion which is automatically adjusted to the multi-view class structures. The proposed method allows for multiple projection directions, by relaxing the Gaussian distribution assumption of related methods. The experiments demonstrate that the proposed method can achieve superior results comparing to several existing methods.**

***Index Terms*—Multi-view learning, subspace learning, image retrieval**

I. Introduction

We have entered a world of multimedia big data. Multimedia contents also become increasingly diverse in their representation and exist in different modalities. It urges the research community to dive into the heterogeneous data to find the desired content across modalities or classify them into the right category from many views. For example, thanks to the available text-image datasets from the collaborative content creations in Wikipedia, matching textual description with their corresponding images becomes a hot-button issue. People start to revisit the image retrieval problem not only in the conventional way of retrieving the best matching image using the query text, but generating human understandable sentences given an image [1]. A visual object can also be observed in various domains in terms of illumination, noise level, viewing angle, and self deformation. Integrating the knowledge obtained from multiple views/modalities contributes to improving the task of object recognition [2].

Subspace learning has proved to be successful among the techniques in multi-view learning for multimedia analysis [3]. It finds a common latent space from different input modalities by fitting an optimization criterion. Among unsupervised methods, Canonical Correlation Analysis (CCA) [4] has been widely used to establish a correlation between views [5], [6]. On the other hand, Multi-view Discriminant Analysis (MvDA) [2] as a supervised algorithm is a direct extension of Linear Discriminant Analysis (LDA) [7], [8]. It seeks for the most discriminant features by maximizing the determinant of the between-class scatters while minimizing

The authors are with the Laboratory of Signal Processing, Tampere University of Technology, Finland. A. Iosifidis is also with the Dept. of Engineering, Electrical and Computer Engineering, Aarhus University, DK-8200, Aarhus N, Denmark.

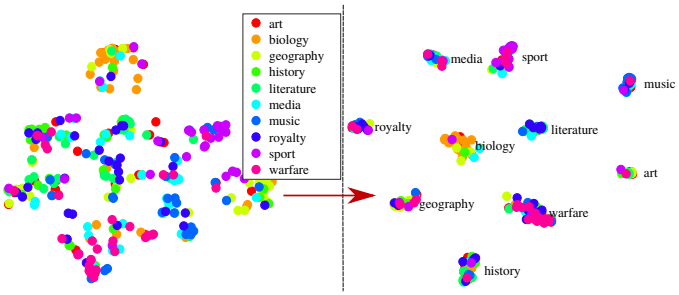


Fig. 1: t-SNE visualization of word2vec represenations before and after applying the proposed method. The samples are grouped together automatically, and each class label indicates the majority class in its group, which matches the corresponding test class.

that of the within-class scatters regardless of view origins. This method can be further extented to nonlinear cases by using (approximate) kernel mappings [9], [10], or integrating with neural nets [11], [6]. Generalized Multi-view Analysis (GMA) [12] was proposed as a framework for numerous techniques to maximize the intra-view discriminant information.

MvDA has certain limitations originating from LDA [13], which is developed upon the assumption that data in each class follow a Gaussian distribution. Only class centers are considered when calculating the between-class scatter matrix and within-class matrix. These parametric methods also suffer performance degradation when the data is non-Gaussian. Several nonparametric techniques [14], [15] were thereby developed to design alternative between-class scatters by exploiting the distances of the data close to the class boundary. However, these techniques are applied in the single-view cases, and view discrepancies should not be overlooked using direct extensions in the multi-view learning.

We propose a new formulation for multi-view discriminant analysis which successfully exploits the boundary structure of the classes on data from different sources, as well as the view discrepancy for balancing the contribution of each view in the overall optimization process. Following the graph embedding framework [16], we design the intrinsic and penalty graphs characterizing the within-class compactness and between-class separability, while encoding both intra-view and inter-view discrimination simultaneously. Class compactness is encoded using a k_1-nearest neighbor graph connecting neighboring samples from the same class with the same view origin, while class discrimination is modeled using another k_2-nearest neighbor graph connecting nearest sample pairs from the same

view but belonging to different classes. We also enhance the class discrimination of each node in the penalty graph by weighting the contribution of neighboring pairs based on their proximity to the class boundary. Moreover, global class discrimination is combined to the adaptive local graph to better adjust to the properties of heterogeneous classes.

We outline the strength of the proposed method as follows: 1) It allows for a larger number of projection directions than MvDA, and makes use of all the samples when developing the intrinsic and penalty graphs, while MvDA merely uses the class centers. 2) It assumes that each class is formed by multiple subclasses, denoted by the different views. In this way, it relaxes the assumption of MvDA in that each class is formed by samples drawn from a multi-dimensional Gaussian distribution, independent from the view they come from. 3) By exploiting both the between-class and within-class margins in the same view, we obtain a better class discrimination in the penalty graph and compactness in the intrinsic graph, and result in an improved performance. 4) Multi-view extension of Marginal Fisher Analysis (MFA) under the GMA framework [12] only considers the intra-view discriminant information, while MvNDA also takes into account of the inter-view discrimination.

The rest of the letter is organized as follows. In Section II, we will present our multi-view nonparametric discriminant analysis in detailed after describing the previous work on MvDA [2]. In Section III, we present quantitative results in cross-modal image retrieval on the Wikipedia dataset and zero-shot recognition on the Animal with Attribute (AwA) dataset. Finally, Section IV concludes the letter.

II. Approach

We denote the data matrix by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, $\mathbf{x}_i \in \mathbb{R}^D$, where N is the number of samples and D is the feature dimension. In the multi-view case, we define $\mathbf{X}_v \in \mathbb{R}^{D_v \times N}, v = 1, \ldots, V$ for the feature vectors of the vth view. The dimensionality of the various feature spaces D_v can vary across the views. $\mathbf{W} = [\mathbf{W}_1^\top, \mathbf{W}_2^\top, \ldots, \mathbf{W}_V^\top]^\top$, where $\mathbf{W}_v \in \mathbb{R}^{D_v \times d}, v = 1, \ldots, V$ is the projection matrix in view v, d is the number of dimensions in the latent (common) space. For multi-class learning problems, the class label of the sample $\mathbf{x}_i$ is defined as $c_i \in \{1, 2, \ldots, C\}$, where C is the number of classes. We also denote the index set of the cth class by π_c.

We use the graph embedding notation, where we define by $\mathbf{G} = \{\mathbf{X}, \mathbf{V}\}$ an undirected weighted graph with vertex set $\mathbf{X}$ and simlarity matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$. The diagonal matrix $\mathbf{D}$ and the Laplacian matrix $\mathbf{L}$ of a graph $\mathbf{G}$ in the vth view are denoted as $\mathbf{L}_v = \mathbf{D}_v - \mathbf{V}_v$, $\mathbf{D}_{ii}^v = \sum_{j \neq i} \mathbf{V}_{ij}^v$, $\forall i$.

A. Multi-view Discriminant Analysis (MvDA)

MvDA [2] is the multi-view version of parametric LDA which maximizes the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix. Mathematically, it is written as

$$\mathcal{J}_{\text{MvDA}}(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{S}_B^P)}{\text{Tr}(\mathbf{S}_W^P)}, \tag{1}$$

where the between-class scatter matrix is

$$\mathbf{S}_B^P = \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \underbrace{\left(\sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top - \frac{1}{N} \mathbf{e}\,\mathbf{e}^\top \right)}_{\mathbf{L}_B^P} \mathbf{X}_j^\top \mathbf{W}_j \tag{2}$$

and the within-class scatter matrix is

$$\mathbf{S}_W^P = \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \underbrace{\left(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top \right)}_{\mathbf{L}_W^P} \mathbf{X}_j^\top \mathbf{W}_j \tag{3}$$

$\mathbf{L}_B^P$ and $\mathbf{L}_W^P$ are the between-class Laplacian matrix and within-class Laplacian matrix, respectively [17]. Both the single-view and multi-view linear discriminant analysis are parametric methods under the assumption that the data of each class follows a Gaussian distribution. Their performance degrades when the data distribution is non-Gaussian. Moreover, since the rank of the between-class matrix is at most $C-1$ in the vth view, the number of the final MvDA feature is at most $(C-1) \times V$. The classification performance is constrained by the limited number of dimensionality in the subspace.

B. Proposed Multi-view Nonparametric Discriminant Analysis (MvNDA)

We propose a new criterion to learn a mapping from the multiple feature spaces defined over the various views to a common space as follows,

$$\mathcal{J}_{\text{MvNDA}}(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{S}_B^N)}{\text{Tr}(\mathbf{S}_W^N)}, \tag{4}$$

where $\mathbf{W}$ is the projection matrix containing the eigenvectors of $\mathbf{S} = {\mathbf{S}_W^N}^{-1}\mathbf{S}_B^N$ associated with the top d eigenvalues λ, and can be solved efficiently from the generalized eigenvalue problem as in [2], [6]. We define the within-class scatter matrix $\mathbf{S}_W^N$ and between-class scatter matrix $\mathbf{S}_B^N$ as follows. In the latent space, we enforce the samples from the same class of the same view to be close to each other. Therefore, the intrinsic graph is designed to strengthen the intra-view class compactness from these subclasses, and the within-class scatter matrix is

$$\mathbf{S}_W^N = \sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i (\mathbf{D}_W - \mathbf{V}_W) \mathbf{X}_i^\top \mathbf{W}_i \tag{5}$$

where $\mathbf{L}_W^N = \mathbf{D}_W - \mathbf{V}_W$ is the within-class Laplacian matrix and the intrinsic graph $\mathbf{V}_W$ is defined as

$$\mathbf{V}_{pq}^W = \begin{cases} 1, & \text{if } p \in \text{NN}_{k_1}(q) \text{ or } q \in \text{NN}_{k_1}(p) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$\text{NN}_{k_1}(p)$ denotes the index set of the k_1 nearest neighbors of the sample $\mathbf{x}_p$ in the same class.

We also design a view-specific penalty graph to push apart the marginal samples from different classes of the same view with the following between-class scatter matrix:

$$\mathbf{S}_B^{\text{VS}} = \sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i [\mathbf{Q} \circ (\mathbf{D}_B - \mathbf{V}_B)] \mathbf{X}_i^\top \mathbf{W}_i, \tag{7}$$

where $\mathbf{L}_B^{\text{VS}} = \mathbf{D}_B - \mathbf{V}_B$ is the between-class view-specific

Laplacian matrix, and its intrinsic graph is characterized as:

$$\mathbf{V}^B_{pq} = \begin{cases} 1, & \text{if } (p,q) \in \text{NP}_{k_2}(c_p) \text{ or } (p,q) \in \text{NP}_{k_2}(c_q) \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

$\text{NP}_{k_2}(c)$ is a set of data pairs which contains the k_2 nearest pairs in the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$. The weight matrix $\mathbf{Q}$ aims to highlight the importance of the samples on the classification boundary. Specifically, the value in $\mathbf{Q}$ goes to 0.5 if the sample falls close to the boundary, but reduces to 0 otherwise. $d(p,q)$ is the Euclidean distance between two vectors p and q. $\mathbf{Q}$ is mathematically described below,

$$\mathbf{Q}_{pq} = \begin{cases} \frac{\min\{d(p,q), d(p,\text{NN}_{k_2}(p))\}}{d(p,q)+d(p,\text{NN}_{k_2}(p))} & \text{if } (p,q) \in \text{NP}_{k_2}(c_p) \\ & \text{or } (p,q) \in \text{NP}_{k_2}(c_q) \\ 0 & \text{otherwise.} \end{cases}$$

In order to enforce both inter-view and intra-view class discrimination, our penalty term is based on the linear combination of $\mathbf{S}^P_B$ of MvDA (2) and $\mathbf{S}^{\text{VS}}_B$ of (7) as follows

$$\mathbf{S}^N_B = \alpha\, \mathbf{S}^P_B + (1-\alpha)\, \mathbf{S}^{\text{VS}}_B, \quad (9)$$

where $\alpha \in [0,1]$ is a weighting factor which is set close to 1 if the training data has a Gaussian distribution, and some other value if the data distribution is unknown.

We provide a qualitative illustration of the intrinsic and penalty graph in Fig. 2. The intrinsic graph shows the within-class compactness by connecting a sample to its k_1-nearest-neighbors of the same class and view. The between-class separability is characterized by both the connected marginal point pairs from the same view but of different classes, and the distance of different class centers.

We also follow the standard kernel-based learning approach to define non-linear multi-view mappings. Each input space is then mapped to the so-called kernel space $\mathcal{F}_v$ using a non-linear function ϕ, i.e. $\mathbf{X}_v \in \mathbb{R}^{D_v \times N} \xrightarrow{\mathbf{\Phi}(\cdot)} \mathbf{\Phi}(\mathbf{X}_v) \in \mathbb{R}^{|\mathcal{F}_v| \times N}$. In $\mathcal{F}_v$, following the Representer Theorem [18], [19], a linear projection can be expressed as $\mathbf{W}_v = \mathbf{\Phi}(\mathbf{X}_v)\mathbf{A}_v$ and dot products between data pairs can be expressed using the kernel matrix $\mathbf{K}_v = \mathbf{\Phi}(\mathbf{X}_v)^\top \mathbf{\Phi}(\mathbf{X}_v)$ [20]. Then,

$$\mathcal{J}_{\text{MvNDA}}(\mathbf{A}) = \arg\max_{\mathbf{A}} \frac{\text{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}^N_B\, \mathbf{K}\mathbf{A})}{\text{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}^N_W\, \mathbf{K}\mathbf{A}\,)}, \quad (10)$$

where the between-class Laplacian matrix $\mathbf{L}^N_B = \alpha \mathbf{L}^P_B + (1-\alpha)\mathbf{L}^{\text{VS}}_B$, and $\mathbf{K} = \text{diag}(\mathbf{K}_1, \cdots, \mathbf{K}_V)$. For the cases where the direct solution of (10) is impractical, due to the training data size, we employ the approximate kernel mapping proposed in [10] followed by the linear mapping defined in (4).

III. Experiments

A. Wikipedia dataset

The cross-modal retrieval dataset named "Wikipedia" was collected from the "Wikipedia featured articles" [1]. The dataset has 10 generic classes and is composed of $2,866$ documents. Each document is a short paragraph with a median text length of 200 words, and is coupled with a single image.

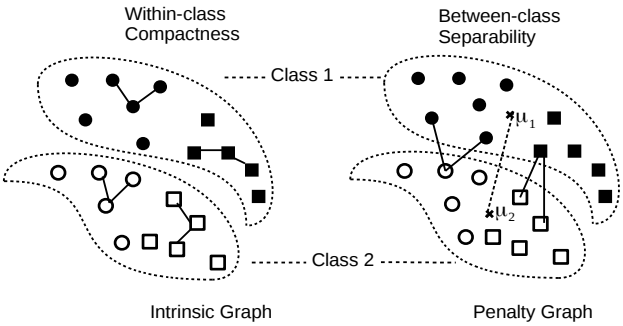


Fig. 2: The adjacency relationship of the intrinsic and penalty graphs of the proposed MvNDA. The circular and rectangle dots indicate samples from different views. We illustrate the 2-nearest adjacencies (i.e. $k_1 = k_2 = 2$) of one sample in each class per view origin for clarity.

We follow the train/test split in [1] using $2,173$ training and 693 test pairs of images and documents. Furthermore, a validation set is held out by 20% of the training image/text pairs. We perform PCA beforehand and reduce the dimensionality of input features to 100. We set the dimensionality of the latent space d to 50 for all methods, and the maximal number of dimesional by MvDA is used. We set $\alpha = 0.5$ and $k_1 = k_2 = 20$ in all experiments based on the validation set.

Here, we briefly describe the feaures extracted from each view in this dataset. For images, two off-the-shelf CNNs models are used to produce the visual features. VGGNet provides the *view 1* feature using the output from the $fc8$ layer in VGGNet with 16 weight layers [21]. We also use the GoogleNet outputs as the *view 3* features. *View 2* feature is extracted from the Wikipedia paragraphs surrounding the images using a pre-trained skip-thoughts model [22]. An additional *view 4* feature is the regression outputs from the Word2Vec by mapping the visual feature to the word feature [23]. The same set of features has been adopted and detailed description can be found in [6].

The cross-modal retrieval is conducted in both ways by

TABLE I: MAP Score (%) on the Wikipedia Dataset

	Linear methods			Kernel methods		
Method	img. query	txt. query	Avg.	img. query	txt. query	Avg.
2 views						
MvCCA [6]	36.92	34.96	35.94	44.78	41.83	43.31
MvPLS [6]	42.49	40.42	41.46	42.94	40.46	41.70
GMA [12]	41.91	38.55	40.23	45.65	36.97	41.31
MvDA [2]	39.73	37.14	38.44	44.16	37.82	40.99
MvNDA	43.51	40.72	**42.12**	48.41	41.97	**45.19**
3 views						
MvCCA [6]	36.40	34.51	35.46	44.06	41.41	42.74
MvPLS [6]	41.29	39.34	40.31	42.03	39.40	40.71
GMA [12]	42.26	38.66	40.46	43.96	36.06	40.01
MvDA [2]	39.34	35.04	37.19	41.25	34.58	37.92
MvNDA	43.21	40.81	**42.01**	48.17	42.67	**45.42**
4 views						
MvCCA [6]	40.50	37.91	39.21	45.13	41.66	43.40
MvPLS [6]	41.86	39.74	40.80	41.94	38.84	40.39
GMA [12]	42.26	38.67	40.47	43.30	35.95	39.63
MvDA [2]	41.07	39.21	40.14	41.31	37.16	39.24
MvNDA	43.44	40.63	**42.04**	48.00	42.43	**45.21**

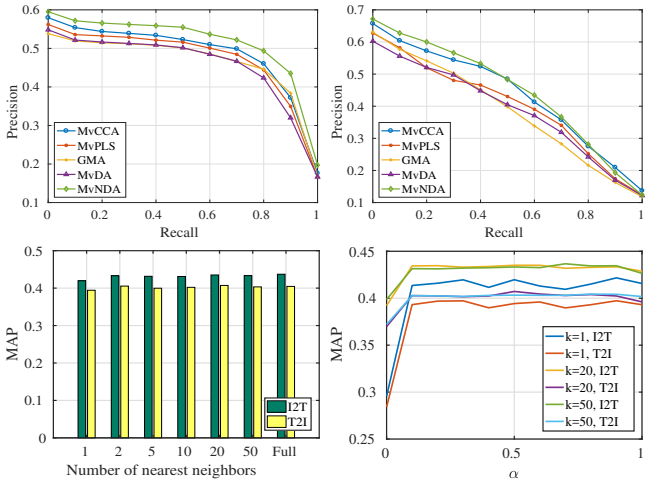


Fig. 3: Clockwise from top left: The precision-recall curve by querying images for text annotations, the retrieval performance of matching text to images, the MAP scores with various α under different fixed numbers of nearest neighors k, (here $k = k_1 = k_2$), and the MAP scores with the different k nearest neighbors and a fixed $\alpha = 0.5$. The legends in the figures in the first row indicate the method producing the PR curve, and we denote querying images for texts by "I2T", and querying texts by images by "T2I" in the figure in the bottom row. k is the number of nearest neighbors.

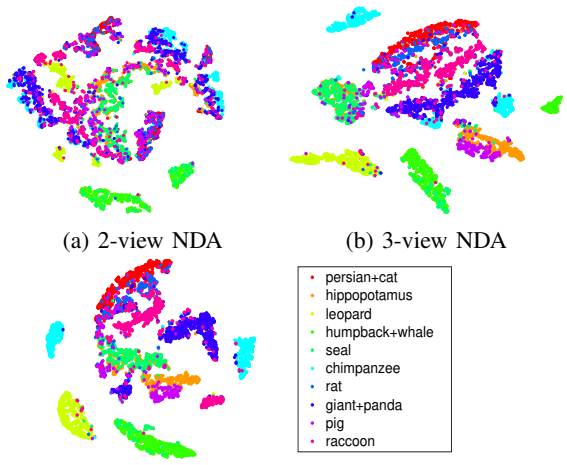
querying every test image and searching for the most relevant texts in the test set, and vice versa. The Mean Average Precision (MAP) is used to evaluate the retrieval performance based on the position of all retrieved images/annotations. We compare the retrieval performance using the features in the subspace of the proposed MvNDA with that of numerous methods in the literature. Both matching images (*view 1*) to text (*view 2*) and text to images are tested. Additional views are projected to the latent space to show more results. In Table I, we see MvNDA outperforms the previous methods in all scenarios using different numbers of views. The further results are confirmed by the Precision-Recall curves in Fig. 3, which shows the retrieval results by the proposed MvNDA are among the leading group in both querying images for text and using text to seek relevant images. We also analyze the effects of different numbers of nearest neighbors and the weight factors α in Fig. 3. It shows the consistent retrieval performance with the different values of k or α, while only using the view-specific discrimination ($\alpha = 0$) degrades the MAP score. We also show the word embedding in its original feature space and the projected latent space in Fig. 1.

B. Animal with Attributes (AwA)

We also demonstrate the effectiveness of multi-view embeddings in tackling the domain shift problem for zero-shot recognition [24]. The Animal with Attribute (AwA) dataset has 50 animal classes with $30,475$ images, and 85 class-level attributes. We follow the experimental protocol in [6] by splitting 40 classes ($24,295$ images) to train the recognition model while the other 10 classes with $6,180$ images for testing

TABLE II: Recognition accuracy (%) on the AwA dataset

	Linear methods			Approxmiate kernel methods		
Method	2 views	3 views	4 views	2 views	3 views	4 views
MvCCA [6]	55.86	75.88	82.01	43.93	47.33	49.51
MvPLS [6]	**58.52**	73.59	77.09	45.37	**47.50**	**52.10**
MvDA [2]	49.95	68.55	70.00	36.65	42.73	42.72
GMA [12]	52.12	73.49	78.46	42.42	44.81	46.84
TMV-HLP [24]	-	73.50	80.50	-	-	-
MvNDA	56.16	**77.16**	**82.78**	**48.78**	46.74	47.56



(a) 2-view NDA (b) 3-view NDA

(c) 4-view NDA

Fig. 4: t-SNE Embedding of Latent Feature Representation: We visualize the embeddings from different numbers of views using the proposed method.

the zero-shot recognition. Each animal class contains more than one positive attribute, and the attributes are shared across classes which enables zero-shot recognition. The detailed class labels and attributes are provided in [25]. Besides the visual features (*view 1,4*) and the class label encoding (*view 3*) generated in the same way as the Wikipedia dataset, a new attribute encoding is added as *view 2* by mapping the visual feature to the attribute probabilities of the animal classes [25].

Table II shows the quantitative results in zero-shot recognition. α, k_1, k_2 are determined based on the grid search using the held-out set. By integrating all available views, we see that recognition accuracy improves with more input views. Due to the size of the training set, we adopt the Nystriöm method for the approximate kernel mapping [10]. MvNDA produces the leading results in all linear cases. We can also observe that the performance of nonlinear methods is inferior compared to the linear ones, which can be explained by the high-dimensionality of the input representations and the use of approximate kernel-based learning. We also graphically show in Fig. 4 that with more available views, the embedded features are grouped into the correct animal classes using the proposed method.

IV. Conclusion

We proposed a novel multi-view nonparametric discriminant analysis technique for the problem of cross-modal image retrieval and recognition. This method has several advantages in exploiting the view difference and class boundary structure information, providing more available projection directions, and achieving better class discrimination in different tasks on both Wikipedia and AwA dataset.

REFERENCES

[1] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 521–535, 2014.

[2] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.

[3] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[4] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

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

[6] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Transactions on Cybernetics*, 2017, doi: 10.1109/TCYB.2017.2742705.

[7] 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 (T-PAMI)*, vol. 19, no. 7, pp. 711–720, Jul 1997.

[8] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1491–1497, Sept 2013.

[9] ——, "Kernel reference discriminant analysis," *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.

[10] A. Iosifidis and M. Gabbouj, "Nyström-based approximate kernel subspace learning," *Pattern Recognition*, vol. 57, pp. 190–197, 2016.

[11] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855.

[12] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.

[13] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[14] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, no. 6, pp. 671–678, 1983.

[15] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 31, no. 4, pp. 755–761, 2009.

[16] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.

[17] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 27, no. 3, pp. 328–340, 2005.

[18] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[19] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of Annual Conference of Computational Learning Theory,*. Springer, Heidelberg, Germany, 2001, pp. 416–426.

[20] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[22] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3111–3119.

[24] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 11, pp. 2332–2345, Nov 2015.

[25] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 453–465, 2014.