Discriminative Sparse Generalized Canonical Correlation Analysis (DSGCCA)

Chenfeng Guo

School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
E-mail: cfguo@hust.edu.cn

Dongrui Wu, *Senior Member, IEEE*

School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
E-mail: drwu@hust.edu.cn

***Abstract*—Multi-view learning (MVL) is a strategy for fusing multi-view data, which has better generalization performance than single-view learning algorithms. Canonical correlation analysis (CCA) is a representative multi-view subspace learning approach, which plays an important role in MVL classification and information retrieval. Traditional CCA can only be used to calculate the correlation of two views, and the learned features are usually dense. Moreover, it is unsupervised, and hence wastes label information in supervised learning. To overcome these limitations, this paper proposes discriminative sparse generalized CCA (DSGCCA), which integrates generalized CCA to handle more than two views, and supervised discriminative sparse principal component analysis to make use of the label information. DSGCCA can handle small multi-view datasets with high feature dimensionality and any number of views. Experiments on four classification datasets demonstrated that DSGCCA outperformed several other representative CCA-based MVL approaches.**

***Index Terms*—Canonical correlation analysis, classification, multi-view learning, principal components analysis**

I. Introduction

Many real-world datasets can be described by multiple feature sets. For instance, a web page can be represented by texts and images above, a video can be represented by visual and audio features, etc. Multi-view learning (MVL) improves the learning performance by exploiting the consensual and complementary information among multiple views [1]–[4].

As summarized in [2], MVL approaches can be divided into three major categories:

1) *Co-training* [5], [6], which exchanges information between different views by training multiple models alternatively.
2) *Multi-kernel learning* [7], [8], which fuses different features that are projected with different kernels.
3) *Subspace learning* [9], [10], which assumes that there exists a shared latent space from which all views are generated, and tries to identify it.

Hotelling [9] proposed Canonical correlation analysis (CCA) in 1936. As a representative subspace learning approach, it projects two different views onto a shared correlated subspace to maximize the correlation between. It is widely used in multi-view clustering [11], [12], multi-view regression [13], multi-view classification [14], [15], and so on [16], [17].

The traditional CCA has several limitations, and many extensions have been proposed to accommodate them in the past few decades. First, CCA cannot be used to calculate the correlations of more than two views. Carroll [18] in 1968 proposed generalized CCA (GCCA) to maximize the correlations of multiple views by finding a common latent correlated space. In addition to estimating the correlations in pairs [19]–[21], Luo *et al.* [22] analyzed a high-order covariance tensor to directly maximize the correlations among multiple views. Second, CCA is unsupervised, which completely ignores the label information in supervised scenarios. Many works [23]–[25] utilized the discriminative label information by taking the inter-class and intra-class similarities of different views into consideration. Third, the projections obtained from CCA are usually very dense. Witten *et al.* [26] proposed a sparse CCA by imposing LASSO-constraints on the canonical vectors.

CCA can be regarded as a dimensionality reduction method [27], [28] for multi-view data. For single-view data, a representative dimensionality reduction approach is principal component analysis (PCA), which maps all instances onto a lower-dimensional uncorrelated linear space and maximizes their variance in that space. Traditional PCA has the same problem as CCA, i.e., it ignores the label information and is not sparse. Recently, Feng *et al.* [29] proposed supervised discriminative sparse PCA (SDSPCA), by imposing additional constraints on the learned linear space. Inspired by SDSPCA, we propose discriminative sparse generalized CCA (DSGCCA) in this paper, which integrates GCCA and SDSPCA.

Our DSGCCA MVL approach has three desirable properties:

1) It can handle any number of views.
2) It can extract more discriminative information by using label information.
3) It can handle small multi-view datasets with high feature dimensionality.

The remainder of this paper is organized as follows: We first reviews related works in section II, and then introduces the proposed DSGCCA algorithm. Section III presents experimental results on several multi-view classification datasets. Section IV draws conclusions.

978-1-7281-4094-0/19/$31.00 ©2019 IEEE

II. The DSGCCA Algorithm

This section first briefly reviews the traditional CCA, GCCA, and SDSPCA algorithms, and then introduces our proposed DSGCCA algorithm.

A. CCA

Let $\left\{X_j \in \mathbb{R}^{d_j \times N}\right\}_{j=1}^{J}$ be a dataset containing J mean-zero views with N instances, where d_j is the feature dimensionality of View j. CCA maximizes the correlation between two views ($J = 2$). It seeks K linear projections for each view, $W_1 \in \mathbb{R}^{d_1 \times K}$ and $W_2 \in \mathbb{R}^{d_2 \times K}$, called canonical vectors, to maximize the correlation between $W_1^T X_1$ and $W_2^T X_2$.

The objective function of CCA is:

$$\begin{aligned} \max_{W_1, W_2} \; & W_1^T X_1 X_2^T W_2 \\ s.t. \; & W_1^T X_1 X_1^T W_1 = I, W_2^T X_2 X_2^T W_2 = I. \end{aligned} \tag{1}$$

By solving a singular value decomposition problem [30] or a generalized eigen decomposition problem [31], the canonical vectors, W_1 and W_2, is obtained.

B. GCCA

GCCA [18] is a representative approach for dealing with more than two views ($J > 2$). It assumes that there exists a set of multivariate latent variables $G = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N]^T \in \mathbb{R}^{N \times K}$, from which each view is generated.

The objective function of GCCA is:

$$\min_{G, \{W_j\}_{j=1}^{J}} \sum_{j=1}^{J} \left\| W_j^T X_j - G^T \right\|_F^2 \quad s.t. \; G^T G = I, \tag{2}$$

where $W_j \in \mathbb{R}^{d_j \times K}$ contains the canonical vectors of View j, $\| \cdot \|_F$ denotes the Frobenius-norm of a matrix, and the constraint $G^T G = I$ guarantees that G is an orthonormal matrix.

The matrices G and $\{W_j\}_{j=1}^{J}$ in (2) can be solved by the approaches described in [32], [33].

C. PCA

PCA [34] is a representative single-view dimensionality reduction algorithm. It maps the instances onto a lower-dimensional uncorrelated linear space, in which the variance of the instances is maximized.

Let $X \in \mathbb{R}^{d \times N}$ be the single-view input data with d features. Then, the objective function of PCA is:

$$\min_{W, G} \; \|X - W G^T\|_F^2 \quad s.t. \; W^T W = I, \tag{3}$$

where each column of $G \in \mathbb{R}^{N \times K}$ is a principal component, and each column of $W \in \mathbb{R}^{d \times K}$ is a principal direction. The constraint $W^T W = I$ guarantees that the principal directions are orthonormal.

Some works [29], [35] modified the traditional PCA by replacing $W^T W = I$ with $G^T G = I$, i.e., (3) becomes:

$$\min_{W, G} \; \|X - W G^T\|_F^2 \quad s.t. \; G^T G = I. \tag{4}$$

D. SDSPCA

In supervised classification, we have label information, which should be taken into consideration to help extract more discriminative principal components. Additionally, the principal components calculated from PCA are usually dense, whereas sparse features are usually preferred in machine learning. For these reasons, Feng *et al.* [29] proposed SDSPCA by imposing label correlated constraints and sparse constraints on G.

The objective function of SDSPCA is:

$$\begin{aligned} \min_{W, G} \; & \|X - W G^T\|_F^2 + \alpha \|B - A G^T\|_F^2 + \beta \|G\|_{2,1} \\ s.t. \; & G^T G = I, \end{aligned} \tag{5}$$

where $B \in \mathbb{R}^{c \times N}$ is the labels' one-hot coding matrix (c denotes the number of classes), and α and β denote trade-off parameters. Initialize the matrix $A \in \mathbb{R}^{c \times K}$ randomly. $\|G\|_{2,1}$ is the $L_{2,1}$ norm of G, i.e., $\|G\|_{2,1} = \sum_{i=1}^{N} \|\mathbf{g}_i\|_2$, where $\|\cdot\|_2$ is the 2-norm of a vector.

E. DSGCCA

Inspired by GCCA, we further extend SDSPCA to DSGCCA to deal with more than two views. Similar to GCCA, we assume that each view is generated from a set of multivariate latent variables $G = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N]^T \in \mathbb{R}^{N \times K}$, which maximize the correlations of all views. Similar to SDSPCA, we add extra constraints on G to take the sparsity and label information into account.

The objective function of DSGCCA is:

$$\begin{aligned} \min_{G, \{W_j\}_{j=1}^{J}} \; & \frac{1}{J} \sum_{j=1}^{J} \|X_j - W_j G^T\|_F^2 \\ & + \alpha \|B - A G^T\|_F^2 + \beta \|G\|_{2,1} \\ s.t. \; & G^T G = I, \end{aligned} \tag{6}$$

where $\{W_j\}_{j=1}^{J}$ contains the canonical vectors for View j.

The optimization of (6) is similar to SDSPCA [29]. Let

$$\ell = \frac{1}{J} \sum_{j=1}^{J} \|X_j - W_j G^T\|_F^2 + \alpha \|B - A G^T\|_F^2 + \beta \|G\|_{2,1}. \tag{7}$$

The partial derivative of ℓ w.r.t. $\{W_j\}_{j=1}^{J}$ is:

$$\frac{\partial \ell}{\partial W_j} = -2 \left(X_j - W_j G^T\right) G, \quad j = 1, \ldots, J. \tag{8}$$

By setting $\frac{\partial \ell(W_j)}{\partial W_j} = 0$ and noting that G is orthonormal, we have

$$W_j = X_j G, \quad j = 1, \ldots, J. \tag{9}$$

Similarly, taking the partial derivative of ℓ w.r.t. A and setting it to zero, we have

$$A = BG. \tag{10}$$

$\|G\|_{2,1}$ in (7) can be transformed into $\mathrm{tr}(G^T V G)$, where $V \in \mathbb{R}^{N \times N}$ is a diagonal matrix with its i-th diagonal element

be $V_{ii} = \frac{1}{2\|\mathbf{g}_i\|_2}$ if $\mathbf{g}_i \neq \mathbf{0}$. Substituting (9) and (10) into (7), and noting that $G^TG = I$, it follows that:

$$\begin{aligned}
\ell =& \frac{1}{J}\sum_{j=1}^{J} \operatorname{tr}\left(X_j - X_jGG^T\right)\left(X_j - X_jGG^T\right)^T \qquad (11)\\
&+ \alpha \operatorname{tr}\left(B - BGG^T\right)\left(B - BGG^T\right)^T + \beta \operatorname{tr}\left(G^TVG\right)\\
=& \frac{1}{J}\sum_{j=1}^{J}\left(-\operatorname{tr}\left(G^TX_j^TX_jG\right) + \|X_j\|_F^2\right)\\
&- \alpha \operatorname{tr}\left(G^TB^TBG\right) + \alpha\|B\|_F^2 + \beta \operatorname{Tr}\left(G^TVG\right)\\
=& -\operatorname{tr}(G^T(\frac{1}{J}\sum_{j=1}^{J}X_j^TX_j)G) - \alpha \operatorname{tr}\left(G^TB^TBG\right)\\
&+ \beta \operatorname{tr}\left(G^TVG\right) + \frac{1}{J}\sum_{j=1}^{J}\|X_j\|_F^2 + \alpha\|B\|_F^2\\
=& \operatorname{tr}(G^T(-(\frac{1}{J}\sum_{j=1}^{J}X_j^TX_j) - \alpha B^TB + \beta V)G)\\
&+ \frac{1}{J}\sum_{j=1}^{J}\|X_j\|_F^2 + \alpha\|B\|_F^2.
\end{aligned}$$

(6) can be transformed into a convex optimization problem:

$$\begin{aligned}
\min_{G}\ & \operatorname{tr}\left[G^T\left(-\frac{1}{J}\sum_{j=1}^{J}X_j^TX_j - \alpha B^TB + \beta V\right)G\right]\\
&+ \frac{1}{J}\sum_{j=1}^{J}\|X_j\|_F^2 + \alpha\|B\|_F^2 \qquad (12)\\
s.t.\ & G^TG = I.
\end{aligned}$$

We use the Lagrange multiplier to obtain G, which is equivalent to the K leading eigenvectors of $T = \frac{1}{J}\sum_{j=1}^{J}X_j^TX_j + \alpha B^TB - \beta V$. Once G is obtained from the training set, the canonical vectors $\{W_j^T\}_{j=1}^J$ are calculated using (9). On the test set, $\{W_j^TX_j\}_{j=1}^J$ are the inputs to a classifier.

Algorithm 1 shows the pseudo-code of DSGCCA. Performing eigen decomposition of matrix T is the most time-consuming step in solving DSGCCA, which depends on the sample size N but is independent of the feature dimensionality $\{d_j\}_{j=1}^J$. This suggests that DSGCCA is more suitable for handling small datasets with high feature dimensionality.

III. Experiments

This section evaluates DSGCCA algorithm on several multi-view datasets with two or three views.

A. Datasets

Table I summarizes the four datasets used in our experiments, which are:

1) *Corel5k*: This dataset contains 5,000 images with 50 semantic topics, with 100 images per topic. We selected three topics (topic tag 1000, 10000, and 12000), with a total of 300 instances.
2) *MIR Flickr*: This dataset has 25,000 images in 38 categories. We selected three categories (animals, food, and night), with a total of 362 images.
3) *NUS-WIDE-LITE* [36]: This dataset contains 48,615 images with a total of 81 concepts. We selected 915 single-label images from two concepts (lake and snow). Six descriptors were provided, and we selected color histogram (CH), wavelet textures (WT), and block-wise color moments (CM55) as the views.
4) *Wiki Text-Image* [37]: This dataset contains two views of images and text with ten categories and a total of 2,866 instances. The 10-D text features were extracted with latent Dirichlet allocation, and 128-D SIFT features were extracted from the images. We selected five categories (art, biology, geography, history, and literature), with a total of 1,472 instances.

All datasets were z-normalized. For datasets with three views, the first two views were used to evaluate the two-view CCA approaches.

B. Experimental Setup and Model Parameters

We included three baselines in our experiments:

1) *Combined-view baseline*, which concatenated the features from all views and then trained an SVM classifier.
2) *PCA*, where PCA was used to reduce the dimensionality of the concatenated features (according to the 95% variance threshold) from all views, and then an SVM classifier was trained.
3) *SDSPCA*, where SDSPCA was used on concatenated features from all views, and then an SVM classifier was trained.

We also compared DSGCCA with several representative CCA-based MVL approaches for two and more views:

1) *Sparse CCA (sCCA)* [26]. As in [26], sparse CCA obtains sparsity by imposing LASSO-constraints on the canonical vectors. The sparse coefficients c_1 and c_2 were set to $c\sqrt{d_1}$ and $c\sqrt{d_2}$, respectively, where $c = 0.2$.
2) *Kernel CCA (KCCA)* [38], which projects data onto a higher dimensional space using kernel functions. The Gaussian RBF kernel, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right)$, was used, and the bandwidth parameter σ was set as the mean distance of the ten nearest neighbors for all instances.
3) *Random CCA (RCCA)* [39], which deals with nonlinearity between views by projecting data randomly. For simplicity, all views had the same number of nonlinear random features, which was selected from $\{100, 120, \ldots, 300\}$.
4) *Discriminative CCA (DisCCA)* [23], which takes label information into consideration by minimizing the within-class similarity and maximizing the between-class similarity.
5) *Multi-view linear discriminant analysis (MLDA)* [24], which utilizes the label information by integrating linear discriminant analysis [40] and CCA. The trade-off parameter was selected from $\{1, 5, 10, 15, 20\}$.

Algorithm 1: Discriminative sparse generalized canonical correlation analysis (DSGCCA).

Input: J views $\{X_j \in \mathbb{R}^{d_j \times N}\}_{j=1}^{J}$ with N instances, where d_j is the feature dimensionality of View j;
$B \in \mathbb{R}^{c \times N}$, the labels' one-hot coding matrix, where c denotes the number of classes;
K, the number of canonical vectors;
α, β, the trade-off parameters;

Output: The canonical vectors $\{W_j\}_{j=1}^{J}$.

Initialize $V = I_{N \times N}$, $\theta = \infty$, $\ell' = 0$, and $A \in \mathbb{R}^{c \times K}$ randomly;
while $|\theta| > 10^{-5}$ **do**
  Construct G as the K leading eigenvectors of matrix $T = \frac{1}{J}\sum_{j=1}^{J} X_j^T X_j + \alpha B^T B - \beta V$;
  $W_j = X_j G,\ j = 1, \ldots, J$;
  $A = BG$;
  $V_{ii} = \frac{1}{2\|\mathbf{g}_i\|_2}$, where $\mathbf{g}_i$ is the i-th row of G;
  $\ell = \frac{1}{J}\sum_{j=1}^{J} \|X_j - W_j G^T\|_F^2 + \alpha\|B - AG^T\|_F^2 + \beta\|G\|_{2,1}$;
  $\theta = \ell - \ell'$;
  $\ell' = \ell$;
end
return $\{W_j\}_{j=1}^{J}$

TABLE I
SUMMARY OF THE FOUR CLASSIFICATION DATASETS.

Dataset	No. of instances	No. of views	No. of features per view	No. of classes
Corel5k[1]	300	3	512/100/100	3
MIR Flickr[1]	362	3	512/100/100	3
NUS-WIDE-LITE[2]	915	3	64/128/225	2
Wiki Text-Image[3]	1472	2	128/10	5

[1] http://lear.inrialpes.fr/people/guillaumin/data.php
[2] https://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[3] http://www.svcl.ucsd.edu/projects/crossmodal/

6) *Generalized CCA (GCCA)* [18], which calculates a rank-m approximation of each view, where m was selected from $\{50, 100, \ldots, \widetilde{m}\}$, in which $\widetilde{m} = \min(200, d_1, d_2, d_3)$.
7) *Least squares CCA (LS-CCA)* [21], which minimizes the distances between the instances projected from different views.
8) *Tensor CCA (TCCA)* [22], which analyzes a high-order covariance tensor to directly maximize the correlations among multiple views. The final feature dimensionality, K, was selected from $\{50, 100, \ldots, 300\}$.

For DSGCCA, as in [29], the trade-off parameters α and β were selected from $\{10^{-20}, 10^{-19}, \ldots, 10^{20}\}$. The number of canonical variables, K, was selected from $\{10, 20, 30\}$.

For each dataset, a random one-fourth of the training set was used for parameter search. To simplify the experiments, the covariance regularization parameters were all set to 0.01. The number of canonical variables, K, was selected from $\{10, 20, \ldots, \widetilde{d}\}$, where $\widetilde{d} = \min(100, d_1, d_2)$ for two views, and $\widetilde{d} = \min(100, d_1, d_2, d_3)$ for three views.

For each view, the canonical variables were input to an SVM classifier separately to compute the confidence for each class, and then the confidences from all views were averaged to give the final classification. All experiments were repeated six times with 5-fold cross-validation.

C. Experimental Results

Because the classes in every dataset were balanced, we computed the raw classification accuracy on the test set as the performance measure. Tables II and III shows the means and standard deviations, for the 2-view and 3-view datasets, respectively. We can observe that:

TABLE II
AVERAGE CLASSIFICATION ACCURACIES (%) ON THE 2-VIEW DATASETS [MEAN±STANDARD DEVIATION].

Algorithm	Corel5k	MIR-Flickr	NUS-WIDE-LITE	Wiki Text-Image	Average
Combined-view	91.67±3.13	60.87±5.54	69.89±2.40	68.59±2.33	72.61±3.35
PCA	91.72±2.64	57.88±5.63	72.75±2.61	70.15±2.03	72.48±3.23
CCA [19]	83.22±5.67	55.01±5.48	67.05±3.64	75.83±1.88	72.47±4.17
sCCA [26]	90.83±3.09	55.66±6.67	70.66±3.14	67.09±3.29	70.17±4.05
KCCA [38]	92.28±2.68	57.46±7.34	66.43±3.60	63.11±2.98	68.99±4.15
RCCA [39]	84.83±4.36	54.47±5.81	69.11±2.44	75.76±2.11	72.71±3.68
DisCCA [23]	80.83±6.31	54.73±5.03	67.38±3.32	74.04±1.71	70.91±4.09
MLDA [24]	81.67±5.59	51.69±7.40	64.68±3.59	75.15±2.00	70.91±4.65
SDSPCA	88.78±4.26	61.79±6.26	**73.72±3.04**	73.99±2.04	74.64±3.90
DSGCCA	**92.72±3.80**	**63.17±4.95**	72.73±2.71	**76.46±1.52**	**77.20±3.24**

1) DSGCCA outperformed SDSPCA, suggesting that extending the single-view SDSPCA algorithm to the multi-view DSGCCA algorithm improved the generalization performance. This may be because DSGCCA reduces the noise contained in each view when considering the consensus among different views.
2) On average, DSGCCA performed the best among all CCA-based algorithms. It achieved the highest average classification accuracy and relatively low standard deviation.
3) DSGCCA did not outperform SDSPCA on two of the three 3-view datasets, although their performances were

TABLE III
AVERAGE CLASSIFICATION ACCURACIES (%) ON THE 3-VIEW DATASETS [MEAN±STANDARD DEVIATION].

Algorithm	Corel5k	MIR-Flickr	NUS-WIDE-LITE	Average
Combined-view	92.89±4.08	63.04±4.74	67.36±3.23	74.55±4.02
PCA	94.00±2.79	62.10±6.18	72.04±2.98	76.05±3.98
GCCA [18]	87.72±4.70	55.43±6.59	70.86±2.69	71.33±4.66
LS-CCA [21]	85.89±5.12	53.49±6.00	67.25±2.37	68.88±4.50
TCCA [22]	90.44±3.79	53.27±6.28	69.80±3.13	71.17±4.40
SDSPCA	89.33±4.75	**64.19±5.23**	**74.90±3.10**	76.14±4.36
DSGCCA	**94.28±3.49**	64.05±4.93	74.24±3.32	**77.52±3.92**

close. This may be caused by the fact that the average of the predicted confidences from all views were used in classification. Estimating the correlation of multiple views is more difficult than two views. When different views have different importance, view weights may be considered for better performance.

4) Although the feature dimensionality of several datasets was higher than their sample size, such as Corel5k and MIR-Flickr, DSGCCA still achieved relatively high classification accuracy, which suggests that DSGCCA can deal with high feature dimensionality without significant overfitting.
5) Many CCA-based MVL approaches did not outperform the combined-view baseline, suggesting that blindly choosing a CCA-based MVL algorithm for a multi-view dataset may not always be appropriate. However, the performance of DSGCCA on different datasets was more consistent than other algorithms, suggesting that it is a safer choice among these CCA-based MVL approaches.

IV. CONCLUSIONS AND FUTURE WORK

This paper has proposed a novel MVL approach, DSGCCA, which integrates GCCA and SDSPCA. DSGCCA can handle dataset with any number of views, and takes the label information and the sparsity into consideration. Experiments on four classification datasets demonstrated that DSGCCA outperformed other representative CCA-based MVL approaches.

The proposed DSGCCA also has some limitations, which will be addressed in our future research. First, it can only optimize the linear correlation among different views, whereas the correlation may be nonlinear in real-world datasets. A possible solution is to combine DSGCCA with deep neural networks, or to project the data onto a higher dimensional space using kernel functions. Second, searching for the optimal parameters of DSGCCA is time-consuming. Third, the computational complexity of DSGCCA increases with the sample size, making it unsuitable for big data. A remedy is to use mini-batch instead of the full batch in its optimization, just like how mini-batch is used in training neural networks.

V. ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China (61873321).

REFERENCES

[1] J. Tang, Y. Tian, P. Zhang, and X. Liu, "Multiview privileged support vector machines," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3463–3477, 2018.
[2] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *CoRR*, vol. abs/1304.5634, 2013. [Online]. Available: http://arxiv.org/abs/1304.5634
[3] X. Xue, F. Nie, S. Wang, X. Chang, B. Stantic, and M. Yao, "Multi-view correlated feature learning by uncovering shared component," in *Proc. 31th AAAI Conf. on Artificial Intelligence*, San Francisco, CA, Feb. 2017, pp. 2810–2816.
[4] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 12th Annual Conf. on Computational Learning Theory*, NY, Jul. 1998, pp. 92–100.
[6] S. Abney, "Bootstrapping," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, Jul. 2002, pp. 360–367.
[7] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. 12th IEEE Int'l Conf. on Computer Vision*, Kyoto, Japan, Sep. 2009, pp. 221–228.
[8] M. Gonen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.
[9] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.
[10] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Multiview Fisher discriminant analysis," in *Proc. Neural Information Processing Systems Workshop on Learning from Multiple Sources*, Whistler, BC, Canada, Dec. 2008.
[11] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annual Int'l Conf. on Machine Learning*, Montreal, Quebec, Canada, Jun. 2009, pp. 129–136.
[12] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, Jun. 2008.
[13] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *Proc. 20th Annual Conf. on Learning Theory*, San Diego, CA, Jun. 2007.
[14] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. on Neural Networks*, vol. 17, no. 1, pp. 233–238, 2006.
[15] C. Shan, S. Gong, and P. McOwan, "Fusing gait and face cues for human gender recognition," *Neurocomputing*, vol. 71, no. 10, pp. 1931–1938, 2008.
[16] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.
[17] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7, pp. 2031–2038, 2013.
[18] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *Proc. 76th Annual Convention of the American Psychological Association*, Washington, DC, Sep. 1968, pp. 227–228.
[19] P. Horst, "Generalized canonical correlations and their applications to experimental data," *Journal of Clinical Psychology*, vol. 17, no. 4, pp. 331–347, 1961.
[20] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, p. 433–451, 1971.
[21] J. Vía, I. Santamaría, and J. Pérez, "A learning algorithm for adaptive canonical correlation analysis of several data sets," *Neural Networks*, vol. 20, no. 1, pp. 139–152, 2007.
[22] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, 2015.
[23] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *Proc. 8th IEEE Int'l Conf. on Data Mining*, Washington, DC, Dec. 2008, pp. 1043–1048.
[24] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Trans. on Cybernetics*, vol. 46, no. 12, pp. 3272–3284, 2016.

[25] N. E. D. Elmadany, Y. He, and L. Guan, "Multiview learning via deep discriminative canonical correlation analysis," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016, pp. 2409–2413.
[26] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.
[27] D. P. Foster, R. Johnson, S. M. Kakade, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," Toyota Technical Institute, Chicago, IL, Tech. Rep. TR-2008-4, 2008.
[28] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proc. 14th Conf. of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, Apr. 2014, pp. 462–471.
[29] C. Feng, Y. Xu, J. Liu, Y. Gao, and C. Zheng, "Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data," *IEEE Trans. on Neural Networks and Learning Systems*, pp. 1–12, 2019.
[30] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int'l Conf. on Machine Learning*, Atlanta, GA, Jun. 2013, pp. 1247–1255.
[31] 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.
[32] P. Rastogi, B. V. Durme, and R. Arora, "Multiview LSA: Representation learning via generalized CCA," in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, CO, May 2015, pp. 556–566.
[33] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *Proc. Symp. on Machine Learning in Speech and Language Processing*, Portland, Oregon, Sep. 2012.
[34] S. Wold, K. H. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987.
[35] B. Jiang, C. Ding, B. Luo, and J. Tang, "Graph-laplacian PCA: Closed-form solution and robustness," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, OR, Jun. 2013, pp. 3492–3498.
[36] T.-S. Chua, J. Tang, R. Hong, H. LI, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Int'l Conf. on Image and Video Retrieval*, Santorini, Fira, Greece, Jul. 2009, pp. 48:1–48:9.
[37] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int'l Conf. on Multimedia*, Firenze, Italy, Oct. 2010, pp. 251–260.
[38] S. Akaho, "A kernel method for canonical correlation analysis," *CoRR*, vol. abs/cs/0609071, 2006. [Online]. Available: http://arxiv.org/abs/cs/0609071
[39] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schoelkopf, "Randomized nonlinear component analysis," in *Proc. 31st Int'l Conf. on Machine Learning*, Bejing, China, Jun. 2014, pp. 1359–1367.
[40] R. A. Fisher, "The use of multiple measurements in taxonomicproblems," *Annals of Eugenics*, vol. 7, no. 2, p. 179–188, 1936.