Survey on Mining Semantically Consistent Patterns for Cross-View Data

Mis. Ujwala S. Vanve
Department of Computer Engineering,
PVPIT Bavdhan, Savitribai Phule Pune University,
Maharashtra, India,
ujwalasvanve@gmail.com

Abstract- We often face the situation that the similar information is represented by different views with different backgrounds, in some real world applications such as Information Retrieval and Data classification. So it becomes necessary for those applications to obtain a certain Semantically Consistent Patterns (SCP) for cross-view data, which embeds the complementary information from different views. However, eliminating heterogeneity among cross-view representations is a significant challenge in mining the SCP. This paper reviews the research work on a general framework to discover the SCP for cross-view data web crawling algorithms used on searching a general framework to discover the SCP for cross-view data.

Keywords:*Semantically Consistent Patterns; Cross View Data; Canonical correlation analysis; Correlation-based Joint Feature Learning*

*****

I. Introduction

The accelerated growth of the Information Technology makes cross-view data widely available in the real world. Cross-View data can be defined as multiple similar small-scale or aggregate entities represented with different forms, backgrounds or modalities. It will be more helpful, if we create some semantically consistent pattern with the complementary information from different views for those applications which are based on information retrieval. This paper proposes two models namely, Isomorphic Relevant Redundant Information (IRRT) and Correlation-based Joint Feature Learning (CJFL).

In this technique of retrieving (mining) semantically consistent patterns for cross view data, IRRT model is used to linearly map multiple heterogeneous low-level feature spaces to high-dimensional redundant feature spaces, to build mid-level isomorphic feature space. But, some redundant data and noise still remains in it. To remove redundant information and noise from mid- level feature space, CJFL algorithm is used.

II. Literature survey

Proposed by Hotelling in 1936, Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It finds two bases, one for each variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. In an attempt to increase the flexibility of the feature selection, kernelization of CCA (KCCA) has been applied to map the hypotheses to a higher-dimensional feature space. KCCA has been applied in some preliminary work by Fyfe and Lai (2001) and Akaho (2001) and the recent Vinokourov, Shawe-Taylor, and Cristianini (2002) with improved results.

During recent years, there has been a vast increase in the amount of multimedia content available both off-line and online, though we are unable to access or make use of these data unless they are organized in such a way as to allow efficient browsing. To enable content-based retrieval with no reference to labeling, we attempt to learn the semantic representation of images and their associated text. We present a general approach using KCCA that can be used for content-based retrieval (Hardoon & Shawe-Taylor, 2003) and mate-based retrieval (Vinokourov, Hardoon, & Shawe-Taylor, 2003; Hardoon & Shawe-Taylor, 2003).

The purpose of the generalization is to extend the canonical correlation as an associativity measure between two set of variables to more than two sets, while preserving most of its properties. The generalization starts with the optimization problem formulation of canonical correlation. By changing the objective function, we will arrive at the multiset problem. Applying similar constraint sets in the optimization problems, we find that the feasible solutions are singular vectors of matrices, which are derived in the same way for the original and generalized problem.

Proposed by Hotelling in 1936, Canonical correlation analysis can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized. Correlation analysis is dependent on the coordinate system in which the variables are described, so even if there is a very strong linear relationship between two sets of multidimensional variables, depending on the coordinate system used, this relationship might not be visible as a correlation. Canonical correlation analysis seeks a pair of linear transformations, one for each of the sets of variables, such that when the set of variables is transformed, the corresponding coordinates are maximally correlated.

In existing system a general framework to discover the SCP for cross-view data, aiming at building a feature-isomorphic space among different views, a novel Isomorphic Relevant Redundant Transformation (IRRT) is first proposed. The IRRT linearly maps multiple heterogeneous low-level feature spaces to a high-dimensional redundant feature-isomorphic one, which we name as mid-level space. Thus, much more complementary information from different views can be captured. Furthermore, to mine the semantic consistency among the isomorphic representations in the mid-level space, system proposed a new Correlation-based Joint Feature Learning (CJFL) model to extract a unique high-level semantic subspace shared across the feature-isomorphic data. Consequently, the SCP for cross-view data can be obtained.

Some redundant and noisy components inevitably co-exist with the requisite component in the midlevel high-dimensional space. In order to eliminate the redundant and noisy information in the mid-level space, CJFL will extract a unique high-level low-dimensional semantic subspace shared across the feature-isomorphic data obtained by IRRT.

CJFL achieves much better classification performance than IRRT without the involvement of semantic information. Which shows the existing IRRT is not effective in combination with the CJFL.

Methods for learning the feature space:

1. Principle Component Analysis (PCA) [2]:
PCA can be defined as a technique which involves a transformation of a number of possibly correlated variables into a smaller number of uncorrelated variables known as principle components, which is helpful to compression and classification of data. PCA only makes use of the training inputs while making no use of the labels [2].

2. Independent Component Analysis (ICA) [2]:
Similar to PCA, ICA also find new components (new space) that are mutually independent in complete statistical sense, but it has additional feature that ICA considers higher order statistics to reduce dependencies. PCA considers second order statistics only. ICA gives good performance in pattern recognition, noise reduction and data reduction. We can say that, ICA is generalization of PCA. However, ICA is not yet often used by statisticians.

3. Partial Least Squares (PLS) [2]:
It is a method similar to canonical correlation analysis (CCA) [2], CCA is used to find directions of maximum correlation where PLS used to find directions of maximum covariance. PLS able to model multiple dependent as well as independent variables, it is suitable for the datasets which are small, suffer from multi-colinearity, missing values or where the distribution is unknown as it minimizes adverse effects of these conditions.

4. Canonical Correlation Analysis (CCA) [2]:
CCA is a method of correlating linear relationships between two multidimensional variables [2]. CCA can be seen as a using complex labels as a way of guiding feature selection towards the underlying semantics [2]. CCA uses two views of the same semantic object to extract the features of the semantics and permits to summarize the relationships between multidimensional variables into lesser number of statistics with preserving important features of the relationships. The main difference between CCA and other three methods is that CCA is closely related to mutual information and it is invariant with respect to affine transformations (a class of 2-D geometric transformations) of the variables. Hence CCA has wide area of acceptance e.g. Economics, medical sciences, meteorology, classification of malt whisky as it closely related to mutual information

III. CONCEPTS HELPFUL TO MINE CONSISTENT PATTERNS FOR CROSS VIEW DATA

The existing system uses the IRRT model is used to build a feature isomorphic space among different views. IRRT creates a mid-level high dimensional redundant feature isomorphic space. In this project work we are going to implement the canonical correlation analysis (CCA) algorithm to achieve better performance with the CJFL algorithm.

Consider a multivariate random vector of the form (X, Y). Suppose we are given a sample of instances,

$$S = \big((x_1, y_1), \ldots (x_n, \ldots \ldots y_n)\big) \text{ Of } (x, y)$$

Let, S_x denote $(x_n, \ldots \ldots x_n)$ and S_x denote $(x_n, \ldots, x_n)$.We can consider defining a new coordinate for X by choosing a direction w_x and projecting X onto that direction,

$$X \rightarrow \langle W_x, X\rangle$$

If we do the same for Y by choosing a direction w_y , we obtain a sample of the new X coordinate. Let

$$S_{x,w_x} = (\langle w_x, x_1\rangle, \ldots, \langle w_x, x_n\rangle)$$

With the corresponding values of the new y coordinate being

$$S_{y,w_y} = \big(\langle w_y, y_1\rangle, \ldots, \langle w_y, y_n\rangle\big)$$

The first stage of canonical correlation is to choose w_x and w_y to maximize the correlation between the two vectors. In other words, the function's result to be maximized is

$$\rho = \max_{w_x w_y} corr\big(S_x w_x, S_y w_y\big)$$

$$= \max_{w_x w_y} \frac{\langle S_x w_x, S_y w_y\rangle}{\|S_x w_x\| \|S_y w_y\|} \quad (1)$$

If we use $\hat{\mathbb{E}}[f(x, y)]$ to denote the empirical expectation of the function f (x, y), where,

$$\hat{\mathbb{E}}[f(x, y)] = \frac{1}{m} \sum_{i=1}^{m} f\,(x_i, Y_i) \quad (2)$$

Where, m = number of samples in x and y.

We can rewrite the correlation expression as

$$\rho = \max_{w_x w_y} \frac{\hat{E}[\langle W_x, X\rangle \langle W_y, Y\rangle]}{\sqrt{\hat{E}[\langle W_x, X\rangle^2] \hat{E}[\langle W_x, x\rangle^2]}}$$

$$= \max_{w_x w_y} \frac{\hat{E}[W'_x X Y' W_y]}{\sqrt{\hat{E}[W'_x X X' W_x] \hat{E}[W'_y Y Y' W_y]}} \quad (3)$$

It follows that,

$$\rho = \max_{w_x w_y} \frac{W'_x \hat{E}[XY'] W_y}{\sqrt{W'_x \hat{E}[XX'] W_x W'_y \hat{E}[YY'] W_y}} \quad (4)$$

Now observe that the covariance matrix of (x, y) is

$$C(x, y) = \hat{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}'\right] = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} = C \quad (5)$$

The total covariance matrix C is a block matrix where the within-sets covariance matrices are C_{XX} and C_{YY} and the

ISSN: 2321-8169
2307 – 2312

between-sets covariance matrices are C_{XY} and C_{YY}, although equation is the covariance matrix only in the zero-mean case.

Hence, we can rewrite the function ρ as

$$\rho = \max_{w_x w_y} \frac{W'_x C_{XY} W_y}{\sqrt{W'_x C_{XX} W_x W'_y C_{YY} W_y}} \quad (6)$$

The maximum canonical correlation is the maximum of ρ with respect to w_x and w_y.

CCA Algorithm:

Observe that the solution of equation (6) is not affected by rescaling wx or wy either together or independently, so that, for example, replacing wx by αwx gives the quotient

$$\frac{\alpha W'_x C_{XY} W_y}{\sqrt{\alpha^2 W'_x C_{XX} W_x W'_y C_{YY} W_y}} = \frac{W'_x C_{XY} W_y}{\sqrt{W'_x C_{XX} W_x W'_y C_{YY} W_y}} \quad (7)$$

Since the choice of rescaling is therefore arbitrary, the CCA optimization problem formulated in equation (6) is equivalent to maximizing the numerator subject to

$$W'_x C_{XX} W_x = 1$$

$$W'_y C_{YY} W_y = 1 \quad (8)$$

The corresponding Lagrangian is strategy for finding the local maxima and minima of a function subject to equality constraints

$$L(\lambda, W_X, W_Y) = \lambda_x W'_y C_{YY} W_y - \lambda_X W'_x C_{XX} W_x \quad (9)$$

This together with the constraints implies that, Assuming C_{YY} is invertible, we have

$$W_y = \frac{C_{YY}^{-1} C_{YX} W_x}{\lambda} \quad (10)$$

Using above formula we have,

$$C_{XY} C_{YY}^{-1} C_{YX} W_x = \lambda^2 C_{XX} W_x \quad (11)$$

We are left with a generalized Eigen problem of the form $Ax = \lambda Bx$. We can therefore find the coordinate system that optimizes the correlation between corresponding coordinates by first solving for the generalized eigenvectors of equation (11) to obtain the sequence of W_x ‘s and then using equation (10) to find the corresponding W_y ‘s.

If C_{XX} is invertible, we are able to formulate equation (11) as a standard Eigen problem of the form $B - 1Ax = \lambda x$, although to ensure a symmetric standard Eigen problem, we do the following. As the covariance matrices C_{XX} and C_{YY} are symmetric positive definite, we are able to decompose them using a complete Cholesky decomposition,

$$C_{XX} = R_{XX} \cdot R'_{XX} \quad (12)$$

Where, R_{XX} is a lower triangular matrix. If we let $u_x = R'_{XX} \cdot$ Wx, we are able to rewrite equation (11) as follows:

$$C_{XY} C_{YY}^{-1} C_{YX} R_{XX}^{-1} u_x = \lambda^2 R_{XX} u_x$$

$$R_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX} R_{XX}^{-1'} u_x = \lambda^2 u_x \quad (13)$$

We are therefore left with a symmetric standard Eigen problem of the form, $Ax = \lambda x$

Exploiting the distance problem, we can give a generalization of the canonical correlation for more than two known samples. Let us give a set of samples in matrix form,

$\{S^{(1)}, \ldots, S^{(k)}\}$ With dimension $m \times n_1, \ldots, m \times n_k$.

We are looking for the linear combinations of the columns of these matrices. In the matrix form $W^{(1)}, \ldots, W^{(k)}$ such that they give the optimum solution of the problem:

$$\min_{W^{(1)}, \ldots W^{(k)}} \sum_{k,l=1,k\neq l}^{K} \left\| S^{(k)} W^{(k)} - S^{(l)} W^{(l)} \right\|_F$$

$$\text{s.t.} \quad W^{(k)T} C_{kk} W^{(k)} = I,$$

$$W_i^{(k)T} C_{kl} W_j^{(l)} = 0. \quad (14)$$

Unfolding the objective function, we have the sum of the squared Euclidean distances between all of the pairs of the column vectors of the matrices, $S^{(k)} W^{(k)}$, k = 1, . . . ,K. One can show this problem can be solved by using singular value decomposition for arbitrary K.

The CJFL Model:

In above Section, we have built a mid-level high dimensional redundant feature-isomorphic space for correlating different views, in which the embedded requisite component tends to be exact, clear, complete, and objective. However, as shown in Fig.2, some redundant and noisy components inevitably co-exist with the requisite one in the space. These factors can seriously affect the performance of the mid-level data representations.

Therefore, it is essential to extract a unique high-level low-dimensional semantic subspace shared across the feature-isomorphic data to eliminate both the redundant and noisy information in the mid-level high-dimensional redundant space. Recently, some trace ratio algorithms such as Linear Discriminant Analysis (LDA) [11], Semantic Subspace Projection (SSP) [12], and Trace Ratio Optimization Problem (TROP) [13] have been proven to be effective in redundancy and noise reduction. For the purpose of finding a projection matrix W to simultaneously minimize the within-class distance while maximizing the between-class distance, a trace ratio optimization problem is formulated as follows [13]:

$$\max_{W^T W = I} \frac{tr(W^T H W)}{tr(W^T G W)}$$

(15)

Where, H and G denote the between-class and within class scatter matrices, respectively. However, since these methods were originally developed for handling single view problems, they do not fully take into account the correlation across isomorphic representations. Thereby, unlike some previous supervised shared subspace learning methods based on least squares and matrix factorization techniques, we use a new trace

 ISSN: 2321-8169
2307 – 2312

ratio based shared subspace learning algorithm, called Correlation-based Joint Feature Learning (CJFL) model. By exploiting the correlations across isomorphic representations, CJFL could extract a unique high-level semantically shared subspace. In this subspace, the requisite component will be maintained to a large extent without the redundant and noisy information being remained. Correspondingly, the SCP for cross-view data can be obtained.

Specifically, let (A*, B*) be the optimal solutions of the above problem [1]. Then we have the sets of relevant redundant representations

J= $\{a_i$ = A*T $X_i\}n_i$ =1 and R= $\{b_i$ = B*T $Y_i\}n_i$ =1

Let, C_X^t and C_Y^t be the sample sets of t-th class from J and R, respectively. We define

$$S_X^t = \{(a_i, a_j) | a_i a_j \in C_X^t, i \neq j\},$$

$$S_Y^t = \{(b_i, b_j) | b_i b_j \in C_Y^t, i \neq j\},$$

$$D_X^{tk} = \{(a_i, a_j) | a_i \in C_X^t \wedge a_j \in C_X^k, i \neq j, t \neq k\},$$

$$D_Y^{tk} = \{(b_i, b_j) | b_i \in C_Y^t \wedge b_j \in C_Y^k, i \neq j, t \neq k\},$$

Let,

$S_X = \cup_t S_X^t$ and $S_Y = \cup_t S_Y^t$

$D_X = \cup_t \cup_k D_X^{tk}$ and $D_Y = \cup_t \cup_k D_Y^{tk}$ (16)

Obviously, each pair of data from S_X or S_Y is semantically similar to each other and the one from D_X or D_Y is semantically dissimilar to each other. To eliminate the redundant and noisy information in the mid-level high-dimensional space, we need to learn a linear transformation θ ϵ Ṙp˟k with prior knowledge (class information in our case) to parameterize the semantically shared subspace, where K is the dimensionality of the subspace. Mathematically, we would like to minimize the within-class distance as follows:

$$= tr(\theta^T (J_S + R_S)\theta)$$

Where

$$J_S = \sum_{\forall (a_i, a_j) \in S_Y} (a_i - a_j)(a_i - a_j)^T,$$

$$R_S = \sum_{\forall (b_i, b_j) \in S_Y} (b_i - b_j)(b_i - b_j)^T$$

And J_S+ R_S are a joint within-class scatter matrix from both J and R. Meanwhile, we also expect to maximize the between-class distance as follows:

$$= tr(\theta^T (J_D + R_D)\theta)$$

Where

$$J_D = \sum_{\forall (a_i, a_j) \in D_X} (a_i - a_j)(a_i - a_j)^T,$$

$$R_D = \sum_{\forall (b_i, b_j) \in D_Y} (b_i - b_j)(b_i - b_j)^T$$

(18)

And J_D+ R_D is a joint between-class scatter matrix from both J and R. To simultaneously minimize the within-class distance while maximizing the between class distance, it is straightforward to formulate the above problem as a trace ratio optimization problem:

$$\Omega_1 = \max_{\theta^T\theta = I} \frac{tr(\theta^T (J_D + R_D)\theta)}{tr(\theta^T (J_S + R_S)\theta)}$$

(19)

Where, the orthogonal constraint for θ is used to eliminate the redundant information in the mid-level space, which takes high relativity with the requisite component. Unlike the scatter matrices in LAD [11], SSP [12], and TROP [13], both the joint within class and between-class scatter matrices JS + RS and

J_D+ R_D make a full use of the identity of sample distributions from different views in the mid-level feature-isomorphic space. On the other hand, the complementarity across isomorphic representations should be well preserved. Thus, we can redefine the formulation Ω1 by:

$$\Omega_2 = \max_{\theta^T\theta = I} \frac{tr(\theta^T (J_D + R_D)\theta)}{tr(\theta^T (J_S + R_S)\theta) + \alpha \|J\theta - R\theta\|_F^2 + \beta \|\theta\|_F^2}$$

(20)

Where the term ||Jθ – Rθ||F2 denotes the correlation based residual to avoid violating the intrinsic structure of the coupled representations, the regularization term ||θ||F2 controls the complexity of the model, and α, β > 0 are the regularization parameters.

Algorithm 3: Correlation-based Joint Feature Learning (CJFL)

Input: an arbitrary columnly orthogonal matrix Θ_0, the matrices J, R, J_D, R_D, J_S and R_S, a positive integer h, two positive numbers α and ß, and max- iter

Output: Θ *

1: for t= 0,1,…..max-iter do

2: Compute

$$\eta_t = \frac{tr(\Theta_t^T (J_D + R_D)\Theta_t)}{tr(\Theta_t^T (J_S + R_S)\Theta_t) + \alpha \|J\Theta_t - R\Theta_t\|_F^2 + \beta \|\Theta_t\|_F^2}$$

(21)

3: Perform Eigen-decomposition of the matrix

$J_D + R_D - \eta_t (J_S + R_S + \alpha (J^T J - 2J^T R +$
$RTR+\beta Ias P\Lambda PT.$

4: Θ_{t+1} is given by the column vectors of the matrix P corresponding to the h largest Eigen value.

5: end-for

6: Set $\Theta^* = \Theta_{t+1}$

Semantically Consistent Patterns:

Let (A*, B*) be the optimal solution of the problem and be the optimal one of the problem θ* Then, for the i-th couple of heterogeneous representations (xi, yi), we can obtain their own isomorphic relevant representations with the optimal (A*, B*) and θ* as follows:

$$T_{xi} = \theta * T\,A * T_{xi}$$
$$T_{yi} = \theta * T\,A * T_{yi} \quad (22)$$

In addition, we can exploit the consistent representation τi of different views, i.e., the Semantically Consistent Patterns (SCP) for the cross-view data in the semantically shared subspace based on τ_{xi} and τ_{yi} :

$$\tau_i = (\tau_{xi} + \tau_{yi})/2 \quad (23)$$

IV. System Implmentation

In this section, we will discuss the experimental requirements of proposed framework for cross view data.

Datasets

Our experiments will be conducted on publicly available cross view dataset namely COREL 5K [7]. The statistics of the datasets are given in Table 1.

- Corel 5K dataset

It contains 260 different categories of images with different content ranging from animals to vehicles, which are enclosed in fifteen different feature sets. Each set of features is stored in a separate file. Each category contains a number of pictures under different natural scenarios belonging to the same semantic class. We also select the DenseHue and HarisHue feature sets. This data set contains 5000 Corel images. There are 374 words in total in the vocabulary (labels) and each image has 4-5 keywords. Images are segmented using Normalized Cuts. Only regions larger than a threshold are used, and there are typically 5-10 regions for each image. Regions are then clustered into 499 blobs using k-means, which are used to describe the final image. Dense Hue and Harris Hue are features provided by Gillanium et al [7]. Both features use local SIFT features and local hue histograms.

TABLE 1
Statistics of cross view datasets

Dataset	Total Attributes	Total Classes	Total Samples
COREL 5K	37152	260	4999

TABLE 2
Brief Descriptions of the feature sets

Dataset	Feature Set	Total Attributes	Total Labels	Total Instances
COREL 5K	DenseHue (V_x)	100	260	4999
	HarrisHue (V_y)	100	260	4999

Corel images data set			
Type	Multi label	Origin	Real world
Features	499	(Real / Integer / Nominal)	(0 / 0 / 499)
Instances	5000	Classes	374
Missing values?			No

Fig. 3 Corel 5K dataset Description

Experimental Setup

To carry out the experiment we need to see that all the data are normalized to unit length. Each dataset is randomly separated into a training samples account for 80 percent of each original dataset, remaining will act as the test data. Such a partition of each dataset will be repeated five times and the average performance will be reported.

The k-nearest (k=5) neighbor classifier serves as the benchmark for the tasks of classification. In the case of retrieval, we use the Euclidean distance and the Mean average Precision (MAP) score to evaluate the retrieval performance.

ISSN: 2321-8169
2307 – 2312

V. Conclusion

The main objective of the review paper was to throw some light on the existing methods for learning the feature space. We also discussed the existing frameworks for cross-view data retrieval and their strengths and weaknesses associated. We believe that all algorithms discussed are effective, but our approach will be more effective to obtain SCP, in terms of complexity of algorithm as compared to IRRT and CJFL. Thus we have used CCA and CJFL for semantic consistent patterns from cross view data representation.

References

[1] Lei Zhang, Yao Zhao_, Senior Member, IEEE, Zhenfeng Zhu, Shikui Wei,and Xindong Wu, Fellow, IEEE" Mining Semantically Consistent Patterns for Cross-View Data" DOI 10.1109/TKDE.2014.2313866, IEEE Transactions on Knowledge and Data Engineering.

[2] David R. Hardoon ,Sandor Szedmak,John Shawe-Taylor "Canonical Correlation Analysis: An Overview with Application to Learning Methods "Neural Computation 16, 2639–2664 (2004).

[3] Lu, Z. and Chi, Z. and Siu, W.C.,"Extraction and optimization of B-spline PBD templates for recognition of connected handwritten digit strings," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.24, no.1, pp.132-139, 2002.

[4] Boutell, M.R. and Luo, J. and iShen, X. and Brown, C.M., "Learning multi-label scene classification," Elsevier. Pattern recognition, vol.37, no.9, pp.1757-1771, 2004.

[5] Yang, Y. and Nie, F. and Xu, D. and Luo, J. and Zhuang, Y. and Pan, Y., "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.34, no.4, pp.723-742, 2012.

[6] Zhuang, Y. and Yang, Y. and Wu, F. and Pan, Y., "Manifold learning based cross-media retrieval"

[7] to media object complementary nature," Springer. The Journal of VLSI Signal Processing, vol.46, no.2, pp.153-164, 2007.

[8] Robert P.W. Duin., "UCI Repository of Machine Learning Databases", http://archive.ics.uci.edu/ml/datasets/Multiple+Features.

[9] Guillaumin, Matthieu and Verbeek, Jakob and Schmid, Cordelia,"Multiple instance metric learning from automatically labeled bags of faces", Proc. European Conference on ComputerVision, vol.37, no.9, pp.634-647, 2010.

[10] Rasiwasia, N. and Costa Pereira, J. and Coviello, E. and Doyle, G. and Lanckriet, Gert RG. and Levy, R. and Vasconcelos, N., "A new approach to cross-modal multimedia retrieval," Proc. ACM. International Conference on Multimedia, pp. 251-260, 2010.

[11] Cheng, Qiang and Zhou, Hongbo and Cheng, Jie, "The Fisher- Markov Selector Fast Selecting Maximally Separable Feature Subset for Multiclass Classification with Applications to High- Dimensional Data," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.33, no.6, pp.1217-1233, 2011.

[12] Keinosuke Fukunaga, Introduction to statistical pattern recogni- tion, Academic Press, 1991.

[13] Yu, J. and Tian, Q.,Learning image manifolds by semantic subspace projection," Proc. ACM. International Conference on Multimedia, pp.297-306, 2006.

[14] Wang, H. and Yan, S. and Xu, D. and Tang, X. and Huang, T.,"Trace Ratio vs. Ratio Trace for Dimensionality Reduction," Proc. IEEE. Computer Vision and Pattern Recognition, pp.457-464, 2007.

[15] Akaho, S. (2001). A kernel method for canonical correlation analysis. In International Meeting of Psychometric Society.

[16] Osaka, Japan. Bach, F., & Jordan, M. (2002). Kernel independent component analysis. Journal of Machine Leaning Research, 3, 1–48.

[17] Borga, M. (1999). Canonical correlation. Online tutorial. Available at: http://www.imt.liu.se/~magnus/cca/tutorial.

[18] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press.

[19] Cristianini, N., Shawe-Taylor, J., & Lodhi, H. (2001). Latent semantic kernels. In C. Brodley & A. Danyluk (Eds.), Proceedings of ICML-01, 18th International Conference on Machine Learning (pp. 66–73). San Francisco:Morgan Kaufmann.

[20] Fyfe, C., & Lai, P. L. (2001). Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems, 10, 365–374.

[21] Gifi, A. (1990). Nonlinear multivariate analysis. New York: Wiley.

[22] Golub, G. H., & Loan, C. F. V. (1983). Matrix computations. Baltimore: Johns Hopkins University Press.