Multi-feature Subspace Learning via Sparse Correlation Fusion and Embedding

Hong Zhang[1,2,3] and Yanpeng Zhang[1]

[1] College of Computer Science & Technology, Wuhan University of Science & Technology, China
[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, China
[3] State Key Laboratory of Software Engineering, Wuhan University, 430072, China
zhanghong_wust@163.com

Abstract. Subspace learning is most traditional and important in multimedia analysis. Numerous researches have focused on how to introduce machine learning and statistical methods to multimedia subspace learning for semantic understanding and denoising, and have gained remarkable achievement in different multimedia applications, such as content-based retrieval, data clustering, face recognition, etc. However, most of these researches are based on multimedia data of single modality. Nowadays, with the rapid development of multimedia and information technology, multimedia data of different modalities often coexist, and the presence of one has a complementary effect on the other to some extent. Because different multimedia data are usually represented with heterogeneous low-level features and there exists the well-known semantic gap, it is interesting and challenging to learn multimedia semantics by multi-feature subspace learning of different modalities. In this paper, we analyze sparse canonical correlation between feature matrices of different multimedia data, construct an isomorphic sparse multi-feature subspace; moreover, we propose subspace optimization strategy with correlation fusion, which explores both geometrical-based content correlation and graph-based semantic correlation. Our algorithm has been applied to content-based multimodal retrieval and data classification. Comprehensive experiments have demonstrated the superiority of our method over several existing algorithms.

Keywords: multi-feature subspace, multimedia analysis, correlation fusion.

1 Introduction

Subspace learning plays an important role in many machine learning tasks and applications, such as CBMR (Content-based Multimedia Retrieval) [1][2][3][7][8], data clustering [4][5], face recognition [14][15] and cartoon generation [6][13]. Most of these researches focus on how to analyze low-level multimedia features and find a low-dimensional semantic subspace with minimum noise and underlying semantic correlation discovered. So far considerable subspace learning methods are based on

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 598–607, 2013.

multimedia data of single modality, such as image subspace learning for retrieval and recognition [19][20][21][22][29], audio subspace learning for semantic understanding [23], etc.

Recently, with the rapid development of multimedia and information technology, multimedia data of different modalities usually coexist and the presence of one has a complementary effect on the other to some extent [16]. Some researches have shown that underlying correlations among different modalities help improve the efficiency of multimedia analysis [1][9][10][17][18][24]. It is important and interesting to utilize multi-feature correlation mining of different modalities in the process of subspace learning. However, most of traditional subspace learning methods, which are based on multimedia data of single modality, are hardly usable for different modalities.

In this paper we propose a novel multi-feature subspace learning method via sparse correlation detection and fusion for multimedia analysis. Our method is formulated based on two typical modalities, i.e., image and audio. First, we analyze visual-auditory multi-feature correlation and build the Sparse Multi-feature Subspace (SMFS) based on sparse canonical correlation analysis. Secondly, we further explore underlying content and semantic correlation in the SMFS: the content correlation is explored with geometrical motivated local linear regression model, and semantic correlation is analyzed with graph-based nonlinear learning to bridge the semantic gap. Thirdly, both of above content and semantic learning results are fused into an objective function to calculate a global optimized solution. The efficiency and superiority of our approach is tested and demonstrated with several multimedia applications: content-based multimodal retrieval and data classification.

The rest of this paper is organized as follows. Section 2 describes sparse multi-feature subspace construction for image and audio data. Section 3 gives details of the subspace optimization with correlation fusion. Section 4 presents experimental results and comparisons. Concluding remarks are in section 5.

2 Sparse Multi-feature Subspace

Since multimedia data of different modalities are initially represented with heterogeneous low-level features, in this section we construct a Sparse Multi-feature Subspace (SMFS) where different multimedia data all reside and canonical correlation among original features are furthest preserved.

In our previous work, Canonical Correlation Analysis (CCA) was used to find a map to the isomorphic subspace [10]. CCA is a classical method to explore statistical correlation between two sets of variables. The underlying ideas of CCA are as follows: it looks for two basis vectors for two sets of variables such that the correlation between the projections onto the basis vectors is mutually maximized. However, when the dimension of low-level features is very high it is important to preserve meaningful correlations instead of all of them. Therefore, in this section, we focus on how to find sparse mapping vectors during feature correlation analysis. In the following description our method is formulated based on two typical modalities: image and audio.

Formally, let $X \in R^{n \times p1}$ denote image feature matrix, and $Y \in R^{n \times p2}$ denote audio feature matrix. We assume that the columns of X and Y have been standardized to have mean zero and standard deviation one. CCA calculates linear combinations of the variables in X and Y that are maximally correlated with each other.

Let $u \in R^{p1}$ and $v \in R^{p2}$ denote canonical vectors which maximize the correlation between Xu and Yv, CCA is to solve the following extremum problem:

$$\begin{aligned} &\max\ u^T X^T Y v \\ &s.t.\ \ u^T X^T X u = 1,\ \ v^T Y^T Y v = 1 \end{aligned} \tag{1}$$

The main challenging issue of CCA is that it is not appropriate when $p_1, p_2 \approx n$ or $p_1, p_2 \gg n$. Then Sparse Canonical Correlation Analysis (SCCA) was proposed to address the problem [12], and its objective function takes the following form:

$$\begin{aligned} &\max\ u^T X^T Y v \\ &s.t.\ \ \|u\|_2 \le 1,\ \|v\|_2 \le 1,\ P_1(u) \le c_1,\ P_2(v) \le c_2 \end{aligned} \tag{2}$$

where P_1, P_2 are convex and non-smooth sparsity-inducing penalty functions that yield sparse u, v. And the constraints $\|u\|_2 \le 1,\ \|v\|_2 \le 1$ are convex relaxations of the quality constraints. Paper [12] studied two specific forms of the penalty P_1, P_2 with structure of L1-norm penalty and the chain-structured fused lasso penalty, which resulted in unique canonical vectors, even when $p_1, p_2 \gg n$. With u fixed, the criterion is convex in v, and with v fixed, it is convex in u.

To solve image canonical vector u and audio canonical vector v, the objective function is used the same as function (2). Here canonical vectors u and v define a linear combination of visual features in X that is correlated with a linear combination of audio features in Y. Elements of u and that equal zero indicate features in X and Y that are not involved in the linear combinations. In this paper we impose sparsity on v. For the ease of illustration, we assume:

$$P_1(u) = \|u\|_1 \tag{3}$$

$$P_2(v) = \sum_j |v_j| + \sum_j |v_j - v_{j-1}| \tag{4}$$

where $P_1(u)$ is an L_1 penalty and $P_2(v)$ is a fused lasso penalty. To substitute Eq.(3) and Eq.(4) for $P_1(u)$ and $P_2(v)$ in the objective function (2), the optimized sparse canonical vectors u and v can be calculated with the algorithm proposed in paper [12]. Accordingly, image feature matrix and audio feature matrix can be mapped to the Sparse Multi-feature Subspace (SMFS).

3 Subspace Optimization by Correlation Fusion

Although image and audio samples are represented with isomorphic dimensionality in the SMFS where sparse canonical correlation is preserved, the SMFS isn't well consistent with high-level semantics because of the well-known semantic gap. Thus, in this section, we propose subspace optimization strategy by correlation fusion, which explores content correlation with local linear regression and utilizes semantic correlation based on a weighted k-nearest neighbor graph. Moreover, both content and semantic correlation we learned are fused into an overall objective function; the optimum solution of this function is therefore the Optimized Sparse Multi-feature Subspace (OSMFS)

In the following descriptions, $z_i, (i \in [1,2n])$ denote a sample (which could be image or audio) in the SMFS, and $m_i, (i \in [1,2n])$ denote the corresponding coordinate vector in the OSMFS after optimization, $Z = \{z_1, z_2, ..., z_{2n}\}$ denote all of the samples in the SMFS.

3.1 Content Correlation Analysis with Local Linear Regression

To predict the value of m_i, we use a local linear regression model of $\psi_i^T z_i + \xi_i$ where ψ_i is a matrix and ξ_i is a vector of bias term. The definition of local means k-nearest neighbors of z_i (including z_i itself), denoted as $\mathbb{N}^k(z_i)$. The regression parameters ψ_i and ξ_i are common for all the samples in $\mathbb{N}^k(z_i)$. Based on the local linear regression model, we define a prediction error for each z_i as:

$$\sigma(z_i) = \left\| \psi_i^T z_i + \xi_i - m_i \right\|_F^2 \tag{5}$$

Then, by summing up all prediction errors for samples in $\mathbb{N}^k(z_i)$ we obtain the local prediction error as below:

$$\sigma_L = \sum_{z_j \in \mathbb{N}^k(z_i)} \left(\left\| \psi_i^T z_j + \xi_i - m_j \right\|_F^2 + \alpha \left\| \psi_i \right\|_F^2 \right) \tag{6}$$

where the second term is added as a regularizer to avoid overfitting. We minimize the local error and get the objective function

$$\arg\min_{m_i, \psi_i, \xi_i} \left(\left\| \psi_i^T Z_i + \xi_i \xi_i^T - M_i \right\|_F^2 + \alpha \left\| \psi_i \right\|_F^2 \right) \tag{7}$$

where $Z_i = [z_i, z_i^1, z_i^2, ..., z_i^k]$ is a feature matrix of samples in $\mathbb{N}^k(z_i)$, and $M_i = [m_i, m_i^1, m_i^2, ..., m_i^k]$ is a matrix of the coordinate vectors for the samples in $\mathbb{N}^k(z_i)$. It is easy to find that the global projection error is to sum (7) on all the samples in training set. Therefore the objective function can be rewritten as:

$$\arg\min \sum_{i=1}^{n} \left(\left\| \psi_i^T Z_i + \xi_i \xi_i^T - M_i \right\|_F^2 + \alpha \left\| \psi_i \right\|_F^2 \right) \tag{8}$$

By setting the derivatives to be zero with respect to ψ_i and ξ_i, we have:

$$\begin{cases} \psi_i = (Z_i H Z_i^T + \alpha I)^{-1} Z_i H M_i \\ \xi_i = (k+1)^{-1} (M_i \theta_{k+1} - Z_i^T \psi_i \theta_{K+1}) \end{cases}, H = I - (k+1)^{-1} \theta_{k+1} \theta_{k+1}^T \in \mathbb{R}^{(k+1)\times(k+1)} \tag{9}$$

where $\theta_{k+1} \in \mathbb{R}^{k+1}$ is a column vector with all ones, and H is the centering matrix. With (8)(9), the objective function becomes:

$$\arg\min_{M_i} \sum_{i=1}^{n} M_i L_i M_i^T \tag{10}$$

$$L_i = H - HZ_i^T (Z_i H Z_i^T + \alpha I)^{-1} Z_i H \quad , \quad L_i \in \mathbb{R}^{n\times(k+1)} \tag{11}$$

We define $M_i = MS_i$ where matrix M consists of all coordinate vectors in the OSMFS and $S_i \in \mathbb{R}^{n\times(k+1)}$ is a selecting matrix made up of 1 and 0. To be specific, the cell value of $S_i(r,t) = 1$ when both z_r and z_t are in the k-nearest neighbor set of $\mathbb{N}^k(z_i)$, and otherwise $S_i(r,t) = 0$. Therefore, the objective function can finally be written as:

$$\arg\min_{M} tr(MLM^T) \tag{12}$$

where $L = [S_1, S_2, ..., S_n] \begin{bmatrix} L_1 & & \\ & ... & \\ & & L_n \end{bmatrix} [S_1, S_2, ..., S_n]^T$ is name as content correlation Laplacian matrix.

3.2 Semantic Correlation Analysis Based on Graph Model

When the query example r is inside training set, the system finds its k-nearest neighbors in the MSS, ranks them by distance in ascending order, and returns to the user as query results. On the other hand, when the query example r is out of training set it needs to be mapped into the MS, then the retrieval process is the same as before. Moreover, we use weighted k-nearest neighbor graph to explore underlying semantic correlation. Since it is difficult to find large amount of labeled image and audio samples for supervised analysis, we utilize both labeled and unlabeled datasets for semantic exploration. Formally, let $G_s(V,E)$ denote a weighted k-nearest neighbor graph with its vertex set V being the set of Z, and the corresponding semantic weight matrix $A = [a_{ij}]$ is defined by

$$a_{ij} = \begin{cases} 1, & if\ z_i \in \mathbb{N}^k(z_j)\ or\ z_j \in \mathbb{N}^k(z_i) \\ 0, & otherwise \end{cases} \tag{13}$$

where $\mathbb{N}^k(z_j)$ denotes the set of k-nearest neighbors of z_j. Note that: if z_i and z_j are both labeled and belong to the same class then definitely there is

$z_i \in \mathbb{N}^k(z_j)$, $z_j \in \mathbb{N}^k(z_i)$; if one of them is unlabeled or both of them are unlabeled we use the Euclidean distance in the SMFS to calculate whether they are k-nearest neighbors to each other. The normalized semantic correlation Laplacian matrix is then defined as

$$L_s = I - B^{-1/2} A B^{1/2} \tag{14}$$

where I is a $n \times n$ identity matrix and B is an $n \times n$ diagonal matrix with its i-th diagonal element being the sum of the i-th row of A . Then, similar to above subsection, we have the following objective function according to graph-based semantic learning result of L_s :

$$\arg \min_{M^T M = I} tr(M L_s M^T) \tag{15}$$

So far, we obtain two objective functions (12) and (15). Therefore, the overall objective function to calculate a good map to the OSMFS is defined as:

$$\arg \min_{M^T M = I} tr(M(L + L_s)M^T) \tag{16}$$

Therefore, the optimum solution of (16) can be obtained by eigen-decomposition of $(L + L_s)$. In summary, the construction of SMFS and its optimization via content and semantic correlation fusion are stated below:

Step 1. Extract low-level features of images and audio clips, and formulate visual feature matrix $X \in R^{n \times p_1}$ and auditory feature matrix $Y \in R^{n \times p_2}$;

Step 2. Analyze sparse canonical correlation between matrices X and Y , compute SMFS where the correlation learned is furthest preserved, and map all samples into the SMFS;

Step 3. Compute k-nearest neighbors $\mathbb{N}^k(z_i)$ for each sample z_i in the SMFS, calculate sub-matrices $S_i, L_i \ (i \in [1, n])$ in (11), then get content correlation Laplacian matrix L in (12);

Step 4. Construct the k-nearest neighbor graph $G_s(V, E)$ with labeled and unlabeled data, calculate corresponding semantic weight matrix $A = [a_{ij}]$ in (13), then get the normalized semantic correlation Laplacian matrix L_s in (14);

Step 5. Compute $V = [v^1, v^2, ...v^c]$ in which $v^1, v^2, ...v^c$ are eigenvectors obtained from c minimum non-zero eigenvalues of (16), then for sample z_i its coordinate vector in the OSMFS is calculated by $m_i = (v_i^1, v_i^2, ..., v_i^c)$ where v_i^j is the ith entry of eigenvector v^j .

Fig. 1. The construction and optimization of SMFS

4 Experiments

4.1 Dataset

To evaluate the effectiveness of the proposed approach, we experimented with an image-audio dataset consisting of 12 semantic categories, including dog, car, bird, explosion, tiger, train, dolphin, drum, piano, plane, zither, lightning. There are 2000 images and 1200 audio clips in total, which are divided equally into the 20 categories. 85% of the image-audio dataset are used as labeled data and the rest 15% are used as unlabeled. We test our proposed algorithm with content-based multimodal retrieval and data classification.

The extracted visual features include Color Histogram, CCV, and Tamura Texture. Auditory features are made up of Centroid, Rolloff, Spectral Flux and Root Mean Square. Since audio is a kind of time series data, the dimensionalities of combined auditory feature vectors are inconsistent. We require collected audio clips not exceed 7 seconds and employ Fuzzy Clustering [10] on auditory features in preprocessing to get index vectors.

4.2 Performance Comparison

Content-based multimodal retrieval is performed based on the Euclidean distance in the subspace. In our experiments, if a returned result and the query example are in the same semantic category, it is regarded as a correct result. Precision is defined as the percentage of correctly retrieved samples in the top-k-returned results.

To evaluate the efficiency of our method, k-nearest neighbors, which could be image and audio, are calculated for each query example according to the Euclidean distance in the SMFS and in the OSMFS respectively. The SMFS is obtained with SCCA which explores original multi-feature canonical correlation, while the OSMFS is obtained with further optimization of the SMFS. Therefore, our experiments compare the retrieval performance before and after subspace optimization, together with the correlation ranking algorithm in [11]. In all the following figures, the precision is the averaged precision values from all the query samples.

Figure 2 shows the comparison results. From Figure 2 we have the following observations: after subspace optimization the retrieval performance is improved, and the SMFS slightly outperform the correlation ranking method in [11]. The phenomenon is attributed to the following reasons:

(1) Sparse correlation is explored from a global perspective and preserved in the SMFS, while in the OSMFS the knowledge learned from both content correlation analysis and semantic correlation analysis further helps bridge the semantics gap.

(2) Paper [11] estimated multimodal similarity based on correlation ranking in the text-image-audio graph, in which text data worked as a bridge to propagate ranking scores among image and audio; when text data is not included the correlation ranking method in [11] doesn't work well, while our multi-feature subspace learning method are based on intrinsic sparse canonical correlation learned from low-level visual and auditory features.

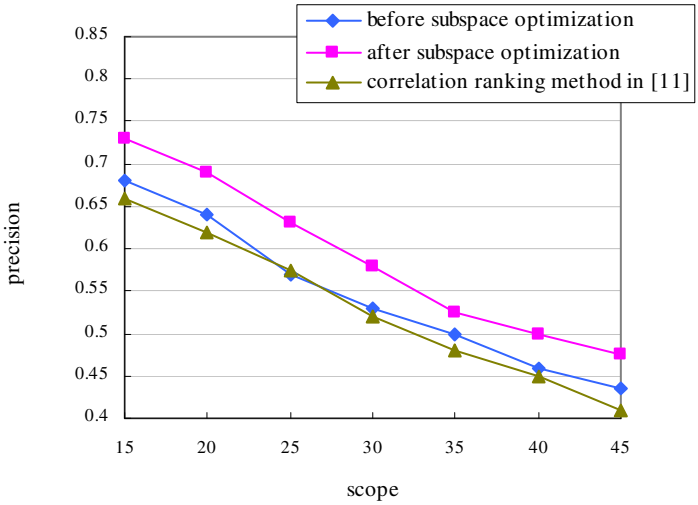

Fig. 2. Comparison of before and after subspace optimization and the correlation ranking algorithm in [11] for multimodal retrieval

To further evaluate the efficiency of our algorithm, data classification is performed in the OSMFS. We use images and audio clips from all 12 categories for performance evaluation respectively. Table1 shows ACC (Accuracy) and AUC(Area Under Curve) results for image classification and audio classification.

Table 1. ACC & AUC performance for image clustering and audio clustering

	Image clustering		Audio clustering	
	ACC	AUC	ACC	AUC
Dog	0.412	0.512	0.397	0.503
Car	0.426	0.533	0.422	0.522
Bird	0.453	0.549	0.389	0.512
Explosion	0.445	0.538	0.431	0.534
Tiger	0.493	0.563	0.418	0.518
Train	0.408	0.522	0.392	0.508
Dolphin	0.447	0.557	0.442	0.537
Drum	0.398	0.515	0.429	0.521
Piano	0.456	0.573	0.425	0.528
Plane	0.461	0.562	0.413	0.514
Zither	0.458	0.519	0.399	0.509
Lightning	0.418	0.487	0.408	0.503

In our experiments KNN (k-nearest neighbor) method is used to execute classification process on image and audio data located in the OSMFS space. We observe that our algorithm works well on the whole. This demonstrates that the OSMFS space learns underlying correlation among image and audio data, and is basically in accordance with high-level semantics.

5 Conclusions

Different from most existing subspace learning methods which focus on single modality feature analysis and denoising, this paper proposes a novel multi-feature subspace learning method, which not only maps multimedia data of different modalities into the sparse multi-feature subspace where original canonical correlation is furthest preserved, but also optimizes the subspace with correlation fusion which explores content correlation with local linear regression and utilizes semantic correlation based on a weighted k-nearest neighbor graph. The experimental results are promising and advantageous; also show that our approach is effective on several multimedia applications. To the best of our knowledge, it is hard to find a public multimodal database as a benchmark. The collected image-audio database is relatively small. Therefore, future work includes further study on large-scale multimodal data analysis and semantic understanding.

Acknowledgment. This work is supported by National Natural Science Foundation of China (No.61003127), State Key Laboratory of Software Engineering (SKLSE2012-09-31).

References

[1] Yang, Y., Nie, F., Xu, D., Luo, J., et al.: A Multimedia Retrieval Framework based on Semi-Supervised Ranking and Relevance Feedback. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(4), 723–742 (2012)

[2] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences and trends of the new age. ACM Computing Surveys 40(2) (2008)

[3] Zhang, R., Zhang, Z.: Effective Image Retrieval based on Hidden Concept Discovery in Image Database. IEEE Transactions on Image Processing 16(2), 562–572 (2007)

[4] Nie, F., Xu, D., Tsang, I., Zhang, C.: Spectral Embedded Clustering. In: International Joint Conference on Artificial Intelligence (IJCAI), California, pp. 1181–1186 (2009)

[5] Ye, J., Zhao, Z., Wu, M.: Discriminative k-means for clustering. Advances in Neural Information Processing Systems 20, 1649–1656 (2008)

[6] Liang, D.W., Liu, Y., Huang, Q.M., et al.: Video2Cartoon: Generating 3D Cartoon from Broad-cast Soccer Video. In: Proceedings of ACM Multimedia (2005)

[7] Typke, R., Wiering, F., Veltkamp, R.: A survey of music information retrieval systems. In: Proceedings of ISMIR, pp. 153–160 (2005)

[8] Lew, M., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state-of-the-art and challenges. ACM Transactions on Multimedia Computing, Communication, and Applications 2(1), 1–19 (2006)

[9] Yang, Y., Xu, D., Nie, F., Luo, J.: Ranking with local regression and global alignment for cross-media retrieval. ACM Multimedia (2009)
[10] Zhang, H., Zhuang, Y., Wu, F.: Cross-modal correlation learning for clustering on image-audio dataset. ACM Multimedia (2007)
[11] Zhang, H., Meng, F.: Multi-modal Correlation Modeling and Ranking for Retrieval. In: IEEE Pacific-Rim Conference on Multimedia, pp. 637–646 (2009)
[12] Witten, D.M., Tibshirani, R.: Extensions of sparse canonical correlation analysis, with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 8(1) (2009)
[13] Yang, Y., Zhuang, Y., Xu, D., Pan, Y., Tao, D., Maybank, S.: Retrieval Based Interactive Cartoon Synthesis via Unsupervised Bi-Distance Metric Learning. ACM Multimedia, 311–320 (2009)
[14] Turk, M.A., Pentland, A.P.: Face Recognition using Eigenface. In: Computer Vision and Pattern Recognition, pp. 586–591 (1991)
[15] Guo, G., Li, S.Z., Chan, K.: Face Recognition by Support Vector Machines. In: IEEE Intl. Conf. on Auto. Face and Gesture Recognition, pp. 196–201 (2000)
[16] McGurk, H., MacDonald, J.: Hearing Lips and Seeing Voices. Nature 264, 746–748 (1976)
[17] Zhang, H., Liu, Y., Ma, Z.: Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval. Neurocomputing (2013), doi:10.1016/j.neucom.2012.03.033
[18] Ma, Q., Akiyo, N., Katsumi, T.: Complementary Information Retrieval for Cross-media News Content. In: Proceedings of Information Systems, vol. 31(7), pp. 659–678 (2006)
[19] Joliffe: Principal component analysis. Springer, New York (1986)
[20] He, X.F., Yan, S.C., Hu, Y.X., et al.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)
[21] Tenenbaum, J.B., Silva, V.D., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290, 2319–2323 (2000)
[22] Hansen, L., Larsen, J., Kolenda, T.: On Independent Component Analysis for Multimedia Signals. In: Multimedia Image and Video Processing, pp. 175–200. CRC Press (2000)
[23] Guo, G., Li, S.Z.: Content-based Audio Classification and Retrieval by Support Vector Machines. IEEE Transactions on Neural Networks 14(1), 209–215 (2003)
[24] Slaney, M., Covell, M.: FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. In: NIPS, pp. 814–820 (2000)
[25] Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
[26] Zhang, H., Yu, J., Wang, M., Liu, Y.: Semi-supervised Distance Metric Learning based on Local Linear Regression for Data Clustering. Neurocomputing 93, 100–105 (2012)
[27] Lovasz, L., Plummer, M.: Matching Theory, pp. 307–349. Akadémiai Kiadó, North Holland (1986)
[28] Cai, D., He, X., Han, J.: Semi-supervised Discriminant Analysis. In: IEEE 11th International Conference on Computer Vision, pp. 1–7 (2007)
[29] Ma, Z., Yang, Y., Nie, F., Uijlings, J., Sebe, N.: Exploiting the entire feature space with sparsity for automatic image annotation. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 283–292