Bridging Music and Image: A Preliminary Study with Multiple Ranking CCA Learning

Xixuan Wu[1,2], Yu Qiao[2], Xiaogang Wang[3] and Xiaoou Tang[1]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Shenzhen key lab of CVPR, Shenzhen Institutes of Advanced Technology, CAS, China
[3]Department of Electronic Engineering, The Chinese University of Hong Kong
{wxx010,xtang}@ie.cuhk.edu.hk,yu.qiao@siat.ac.cn,xgwang@ee.cuhk.edu.hk

ABSTRACT

Human perception of music and image are highly correlated. Both of them can inspire human sensation like emotion, power etc. This paper preliminarily investigates how to model the relationship between music and image using 47,888 music-image pairs extracted from music videos. We have two basic observations for this relationship: 1) music space exhibits simpler cluster structure than image space, and 2) the relationship between the two spaces is complex and nonlinear. Based on these observations, we develop Multiple Ranking Canonical Correlation Analysis (MR-CCA) to learn such relationship. MR-CCA clusters the music-image pairs according to their music parts, and then conducts Ranking CCA (R-CCA) for each cluster. Compared with classical CCA, R-CCA takes account of the pairwise ranking information available in our dataset. MR-CCA improves performance and significantly reduce computational cost. Experiment results show that R-CCA outperforms CCA, and MR-CCA has the best performance, a consistency score of 84.52% with human labeling. The proposed method can be generalized to model cross media relationship and has potential applications in video generation, background music recommendation etc.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Music,Image,Human Factors

Keywords

Music-image similarity, Ranking Canonical Correlation Analysis, Combination of Clusters

1. INTRODUCTION

Music and image are two popular forms of media, one in audio the other in vision. It is well known that human perceptions of music and image show strong correlations with each other. Both of them can inspire similar human sensation like emotion, power, weight etc. Psychology and cognition studies indicate that brain information processing of visual and audio are related[8]. Osborne [9] propose that music can stimulate of visual imagery. Juslin et al.[6] argue that visual imagery is an important mechanism by which music brings emotion. For example, people may react with both positive emotions when listening to melodic movement as 'upward' or facing a beautiful nature scene. In movie and TV program, music and image (video) often appear in parallel as complement to each other. Meyer [7] discussed that "it seems probable that ... image processes play a role of great importance in the musical affective experiences".

Many works model the relationship between text and image [10] or text and music [13]. But few work directly analyzes the semantic relationship between image and music. Zhang et al.[15] explores the similarity between audio (NOT music) and image, such as that between a picture of bird and chirps. The audio is clearly categorized, which makes their method not suitable for us. Chao et al.[3] is the closest one to ours which can be seen as an application for our proposed techniques. But they use tags (text) as middle media instead. To our best knowledge, our work is the first one to investigate the matching degree between image and music.

The objective of this paper is to preliminarily study this challenging problem. To bridge music and image is hard. Firstly, image and music have different feature representations. Secondly, both image space and music space exhibit complex structure, and the relationship between them is nonlinear. This paper develops Multiple Ranking Canonical Correlation Analysis (MR-CCA) to deal with this challenging problem. MR-CCA clusters music-image pairs according to their music sides, and makes use of Ranking CCA to model the local relationship for each cluster. It is noted that the proposed method is general and can be applied to other problems.

To examine the effectiveness of the proposed method, we collected a set of 47,888 music-image pairs from more than 1,500 music videos. We chose half of these pairs and asked labelers to compare their matching degree. The labelers largely agree with each other on the annotation, which indicates that human have consensus on this problem. The experimental results of our method show that MR-CCA has the best performance and achieves a consistency score of 84.52% with human annotation.

2. PREPARATIONS

WOODSTOCK '97 El Paso, Texas USA

2.1 Data and annotation

We collect a set of 47,888 music-image pairs from more than 1,500 music videos (MVs), which cover music with various styles and genres, and are from 500 singers. Since a MV contains audio signal with large variation, different audio segments corresponds to different images. So we segment the music with dynamic texture model [1], and each music-image pair is composed by a key frame from the video part and its associated music segment. Since most music videos are created by experts with deliberate consideration, most pairs obtained in this way should be well matched. To further verify the goodness of the dataset, we ask human labelers to compare the matching degree of a pair from MV and a random pair (music is the same but image is selected randomly). All labelers prefer music-image pair from MV with a rate of 89.2%.

2.2 Representation of music and image

Two categories of features are considered for music representations. The first is the Echo Nest Song(ENS) features vector, each dimension of which corresponds to a mid-level acoustic characteristic [12] such as tempo, mode etc. The second category consists of a set of semantic features, each of which corresponds to a semantic word description of a song, like happy, relax etc. The semantic words are obtained from famous music sites, such as Google music. A music segment is represented as a vector of posterior probabilities with respect to the pre-defined semantic words, computed with mix hierarchical learning [13]. Mainly following [11], image is represented by three types of features, 1) color features: hue and RGB histogram, color spatialet etc; 2) shape features: Histogram of Gradient, edge orientation histogram etc; 3)texture and appearance features: wavelet, GIST etc. The music and image feature vector has a total dimension of 133 and 2,596, respectively. For each type of music and image feature, we define a distance function for it.

3. MULTIPLE RANKING CANONICAL CORRELATION ANALYSIS

In this section, we introduce Multiple Ranking Canonical Correlation Analysis (MR-CCA). MR-CCA is a learning based approach and its diagram is shown in Fig. 1. Let $V_i = \{I_i, M_i\}$ denote a set of music-image pairs for training. The objective of MR-CCA is to estimate a similarity function $\mathcal{S}(I, M)$ between for image I and music M.

3.1 Cluster and Distance to Reference Transformation

Both music and image space have high dimension and complex structures. To simplify the problem, we cluster music-image pairs at first. One problem is that the relationship between music and image is many-to-many. So it is not a good idea to cluster pairs with both music and image features. Compared with image space, music space includes less diversity and exhibits simpler cluster structure. Thus, we use normalized cut to cluster the music parts, and the images are divided according to the cluster information of their music parts. Let $\{V_1, V_2, ..., V_C\}$ denote the clusters.

Inspired by the success of similarity based representation in [2, 14], we use a Distance to Reference Transformation (DtRT) which converts original feature into a new DtRT representation. For each cluster V_c, we select a set of reference samples $\{I_c^r, M_c^r\}_{r=1}^R$ where R is the size of reference set. For another music I in V_c, we calculate distance $d_I(I, I_c^r)$, then convert the distance to similarity by $s_c^r(I) = \exp\left\{-\frac{d_I(I,I_c^r)^2}{{\sigma_I}^2}\right\}$, where σ_I is a normalization parameter. The new representation for image I is defined as a vector composed by the similarities $x_c(I) = [s_c^1(I), ..., s_c^R(I)]$. In the same way, we can get $y_c(M) = [s_c^1(M), ..., s_c^R(M)]$ for music M. In the next, we use DtRT representation $x_c(I), y_c(M)$ (or x, y for simplicity) instead of I, M, and set similarity function $\mathcal{S}(x, y) = \mathcal{S}(I, M)$. One advantage of using x and y is that both of them have the same dimension, and their components are aligned correspondingly.

3.2 Ranking Canonical Correlation Analysis

This section proposes Ranking CCA (R-CCA) to estimate a similarity function $S_c(x, y)$ for each cluster c. Let $V_c = \{x_i, y_i\}$ denote the set of training pairs in cluster c. Canonical correlation analysis (CCA) [5] aims to find data projections with the largest correlation across two (or more) spaces , which has been successfully used for cross-modal multimedia retrieval [10]. Introduce the project matrices $A = [a_1; a_2; ...; a_J]$, $B = [b_1; b_2; ...; b_J]$. The objective of CCA is $\max_{A,B} \sum_{j=1}^{J} \sum_{i=1}^{N} a_j x_i b_j y_i$. It can be shown that optimal A are composed by the eigenvectors of $\Sigma_{XX}^{-1}\Sigma_{XY}$ $\Sigma_{YY}^{-1}\Sigma_{YX}$ associated with the J largest eigenvalues, while B are composed by the eigenvectors of $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ associated with the J largest eigenvalues, where Σ_{XX}, Σ_{YY} are covariance matrices for x_i, y_i respectively, and Σ_{XY}, Σ_{YX} are covariance matrices between x_i and y_i. With A, B, the similarity function of CCA is defined by,

$$\mathcal{S}_c^{\text{CCA}}(x, y) = \langle Ax, By \rangle, \tag{1}$$

Human annotation described in Section 2.1 yields pairwise ranking information, which means one pair should have higher similarity score than another, i.e., $\mathcal{S}(x_i, y_i) > \mathcal{S}(x_i', y_i')$. Classical CCA cannot handle such kind of pairwise ranking information. Here we develop Ranking Canonical Correlation Analysis (R-CCA) to take account of these information. Mathematically, the objective of R-CCA learning is,

$$\min_{A,B} \sum_i f(\mathcal{S}_c(x_i, y_i) - \mathcal{S}_c(x_i', y_i')), \tag{2}$$

where similarity function $\mathcal{S}_c(x_i, y_i)$, has the same form as Eq. 1, and f is a hinge penalty function. If $t \leq 0$, $f(t) = -t$; otherwise $f(t) = 0$. The constraints are $\|a_i X\| = \|b_i Y\| = 1$, $a_i^T X^T X a_j = 0$, and $b_i^T Y^T Y b_j = 0$.

To directly optimize Eq. 2 is difficult. This is because $\mathcal{S}$ has a quadratic form of A and B, and both A and B contain a large number of variables. To simplify the problem, we apply classical CCA on matched pairs $\{(x_i, y_i)\}$ to obtain projection matrices A and B to reduce the dimensionality, and introduce the following similarity function for R-CCA,

$$\mathcal{S}_c^{\text{R-CCA}}(x_i, y_i) = x_i^\top A^\top \Sigma B y_i, \tag{3}$$

where Σ is a matrix with size $J \times J$. Since A and B are known, we only need to optimize Σ which minimizes Eq. 2.

Σ contains much less variables than A, B. Moreover, Eq. 3 has a linear form of Σ which makes optimization easier. We can further assume Σ is a diagonal matrix, since the projections obtained by CCA are uncorrelated. Let $W = [w_1, w_2, ..., w_J]$ denote the diagonal of Σ. Introduce variables, $z_i^j = a_j x_i b_j y_i$, $z_i = [z_i^1, z_i^2, ..., z_i^J]$, $z_i^{j\prime} = a_j x_i{}' b_j y_i{}'$,

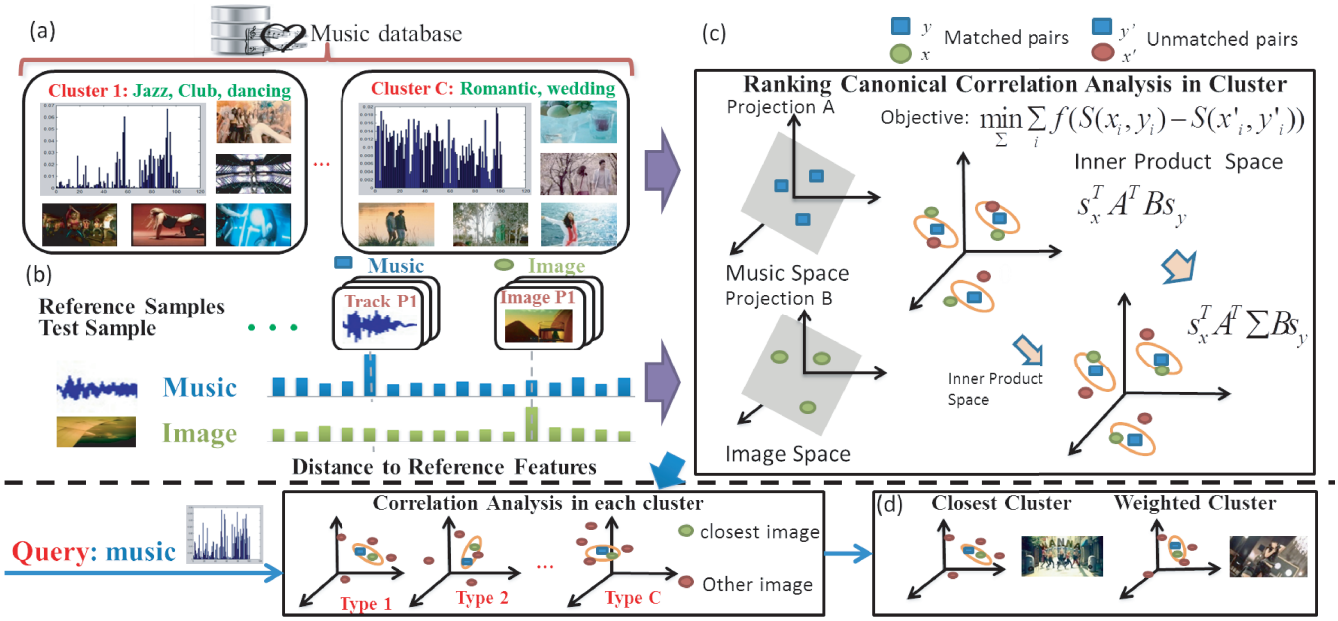

Figure 1: Framework of Multiple Ranking Canonical Correlation Analysis: (a). Cluster in music space; (b). DtRT representation for music and image; (c). Ranking canonical correlation analysis in cluster; (d). Combination of estimation score of clusters.

and $z'_i = [z_i^{1'}, z_i^{2'}, ..., z_i^{J'}]$. Then Eq. 3 can be written into,

$$\mathcal{S}_c^{\text{R-CCA}}(x_i, y_i) = \sum_j w_j a_j x_i b_j y_i = W^\top z_i. \quad (4)$$

Then our objective reduces to optimize W with

$$\min_W \sum_i f(W^\top z_i - W^\top z'_i). \quad (5)$$

This is in spirit the same as the optimization for ordinal SVM [4], whose due problem is defined as

$$\min_W \|W\|^2 + \sum \xi_i, \quad (6)$$

subject to,

$$\xi_i \geq 0, \quad (7)$$

$$W^T z_i - W^T z'_i \geq 1 - \xi_i. \quad (8)$$

The above problem is a quadratic programming which can be solved by Lagrangian multipliers.

3.3 Similarity score ensemble

In previous subsection, we have estimated a similarity function for each cluster by CCA (Eq. 1) or R-CCA (Eq. 3). For a testing pair (x, y), we need to combine these similarity functions to obtain a final similarity score. The simple idea is to determine which cluster (x, y) belongs to. Let $c*$ denote the index of the nearest cluster. Then

$$\mathcal{S}(x, y) = \mathcal{S}_{c*}(x, y). \quad (9)$$

Eq. 9 is simple. But it cannot deal well with the samples near cluster boundary. To overcome this problem, we introduce a 'soft' similarity function with using softmax function as weights. Let d_c denote the distance between pair (x, y) and the center of c-th cluster. The weighted similarity function is,

$$\mathcal{S}(x, y) = \sum_{c=1}^{C} \frac{\exp(-d_c/\sigma^2)}{\sum_{j=1}^{C} \exp(-d_j/\sigma^2)} \mathcal{S}_c(x, y) \quad (10)$$

where C is the number of cluster, and σ is a normalization parameter. Eq. 10 can model the nonlinear relation between x and y, since weights depend on input pair (x, y).

4. EXPERIMENT

This section presents our experimental evaluation. We randomly select 20,000 of all 22,632 music segments for clustering and training, use the rest for testing. We randomly select a set of samples from each cluster as its reference set. We explore the effect of two parameters, the number of cluster K and the proportion of reference set size to cluster size $1/d$. K is changed from 3 to 20. d is changed from 2 to 20, which means $1/d$-th samples in each cluster are used as the corresponding reference set. The labelers' preference are set as ground truth. The consistency rates (precision) with human labeling are used as evaluation criterion. We repeat each experiment 10 times.

Comparison of Methods. We compared three similarity estimation methods, closest CCA (C-CCA, Eq. 1+Eq. 9), closest Ranking CCA (CR-CCA, Eq. 4+Eq. 9), and Multiple Ranking CCA (MR-CCA, Eq. 4+Eq. 10). Since there is no previous study on this problem, closest CCA can be seen as a baseline method. Firstly, we fix the proportion of reference set $1/d$ as 1/6 and, change the number of cluster K. Experimental results are shown in Fig. 2A. The precisions of all three methods generally increase with the number of cluster when $K \leq 10$. This is partly because too fewer clusters cannot model the whole space precisely. When $K > 10$, the precisions increase very few. The highest precision we obtained is 81.23%, when the number of cluster is 13 and MR-CCA is used. We also make an experiment that only uses some human-labeled pairs to test. It shows that our methods performs better on average. The highest preci-

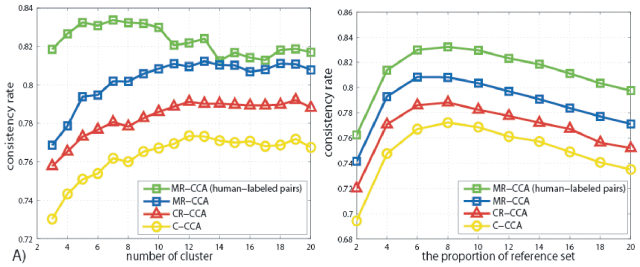

Figure 2: A) Precisions using C-CCA, CR-CCA, MR-CCA and MR-CCA with human-labeled pairs when changing the number of cluster; B) Precisions using C-CCA, CR-CCA, MR-CCA and MR-CCA with human-labeled pairs when changing the proportion of reference set to training set, x axis is the inverse of the proportion.

Table 1: Local reference Vs Global reference

Reference Type	Consistency Rate	Training Time Cost
Local reference	80.35%	46.3999s
Global reference	79.24%	869.2618s

sion is 84.52% with MR-CCA when $K = 5$ and $d = 5$. This means our methods are comparable with human annotation performance on this task.

Secondly, we fix the number of cluster K as 10 and change the proportion of reference set $1/d$. The results are depicted in Fig. 2B. The best $1/d$ is around $1/6 \sim 1/8$. This is because DtRT features cannot convey enough information when too fewer references are used, while too many reference can lead to redundant and noisy representations.

In all our experiments, we find MR-CCA archives the best performance among the three methods, and CR-CCA always outperforms C-CCA. This indicates that 1)Ranking CCA is more effective than classical CCA due to the fact that it can use pairwise ranking information in training, and 2) the weighted summation of similarity functions (Eq. 10) is better than the single closet similarity function (Eq. 9).

Comparison of Time Efficiency. Another issue is how to select reference samples. In spite of selecting local references for each cluster, one may suggest to select global reference set from the whole training set. In the next, we make comparison between local reference and global reference. To be fair for both kinds of reference, we fixed the proportion of reference set as 1/10. Cluster number K is chosen as 10. We also conduct experiments with other d and K. The tendency is similar, thus we omit these results due to space limitation. The precisions and training time cost are shown in Table 1. All the experiments were run on a machine with 2.80GHz CPU and 16GB of RAM. It can be seen that local reference has slight better performance and is much faster, which is more scalable for larger or distributed database. That's mainly because the DtRT feature representations x, y has lower dimension with local reference.

5. CONCLUSIONS

This paper develops multiple ranking CCA which considers both cluster structure and pairwise ranking information to learn similarity between music and image. Our experiments on 47,888 pairs show the effectiveness of the proposed methods, and MR-CCA achieves a consistency score of 84.52% with human labelers on comparing music-image pairs. It is noted that to model the relationship between music and image is a challenging and general problem. This paper is more of an attempt to introduce the topic and present methods with inspiring results on data from music video. Although the discussions are limited to music-image pairs, our analysis and methods can be generalized to other types of multimodal relationship. The proposed methods have potential applications in video generation, etc.

6. ACKNOWLEDGMENTS

This work is partly supported by National Natural Science Foundation of China (61002042), Shenzhen Basic Research Program for Distinguished Young Scholar (JC201005270350A), 100 Talents Programme of Chinese Academy of Sciences, and Introduced Innovative R&D Team of Guangdong Province "Robot and Intelligent Information Technology".

7. REFERENCES

[1] L. Barrington, A. Chan, and G. Lanckriet. Modeling music as a dynamic texture. *IEEE Trans. on ASLP*, 2010.

[2] E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *BMVC*, 2005.

[3] J. Chao, H. Wang, W. Zhou, W. Zhang, and Y. Yu. Tunesensor: A semantic-driven music recommendation service for digital photo albums. 2011.

[4] R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In *ICANN*, 1999.

[5] R. Johnson and D. Wichern. *Applied multivariate statistical analysis*. 1992.

[6] P. Juslin and D. Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 2008.

[7] L. Meyer. *Emotion and meaning in music*. University of Chicago Press, 1961.

[8] I. Olson, J. Gatenby, and J. Gore. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research*, 2002.

[9] J. Osborne. The mapping of thoughts, emotions, sensations, and images as responses to music. *Journal of Mental Imagery*, 1981.

[10] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM-MM*, 2010.

[11] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang. Intentsearch:capturing user intention for one-click internet image search. *IEEE Trans. on PAMI*, 2011.

[12] D. Tingle, Y. Kim, and D. Turnbull. Exploring automatic music annotation with "acoustically-objective" tags. In *ACM-MIR*, 2010.

[13] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. on ASLP*, 2008.

[14] X. Wang, K. Liu, and X. Tang. Query-specific visual semantic spaces for web image re-ranking. In *CVPR*, 2011.

[15] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *ACM-MM*, 2007.