Boosting Cross-media Retrieval via Visual-Auditory Feature Analysis and Relevance Feedback

Hong Zhang[1,2,3], Junsong Yuan[3], Xingyu Gao[4,5], Zhenyu Chen[4]
[1]College of Computer Science & Technology, Wuhan University of Science and Technology, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing & Real-time Industrial System, China
[3]School of EEE, Nanyang Technological University, Singapore
[4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[5]School of Computer Engineering, Nanyang Technological University, Singapore
zhanghong_wust@163.com, jsyuan@ntu.edu.sg

ABSTRACT
Different types of multimedia data express high-level semantics from different aspects. How to learn comprehensive high-level semantics from different types of data and enable efficient cross-media retrieval becomes an emerging hot issue. There are abundant correlations among heterogeneous low-level media content, which makes it challenging to query cross-media data effectively. In this paper, we propose a new cross-media retrieval method based on short-term and long-term relevance feedback. Our method mainly focuses on two typical types of media data, i.e. image and audio. First, we build multimodal representation via statistical canonical correlation between image and audio feature matrices, and define cross-media distance metric for similarity measure; then we propose optimization strategy based on relevance feedback, which fuses short-term learning results and long-term accumulated knowledge into the objective function. Experiments on image-audio dataset have demonstrated the superiority of our method over several existing algorithms.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms
Algorithms, Design.

Keywords
Cross-media retrieval; feature analysis; relevance feedback

1. INTRODUCTION
Cross-media retrieval is emerging as a new search paradigm that enables seamless information processing from different types of data. Plenty of work has been done on cross-media retrieval [1][5][7][14], and these methods do provide effective ways to better manage multimodal data. However, there are two issues mostly remain uncovered.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654975

The first issue is: most of the existing works focused on web images and their tagging texts, while the analysis of images and other types of multimedia data was mostly ignored. For example, [2] proposed supervised coupled dictionary learning with group structures for Wiki Text-Image data; [3] proposed effective multimodal stacked auto-encoders for retrieval among web images and associated tags. Very limited research efforts were devoted to content-based cross-media retrieval among other types of media data, such as image-audio database [9][14]. In fact, audio is an important kind of sensory information, which affects a lot on human perception [4]. It is interesting and challenging to learn cross-media semantics for retrieval from image and audio data simultaneously.

The other issue is that most related works do not pay enough attention to long-term optimization of cross-media retrieval results. Generally, these works share a similar two-step processing strategy [3][5]. First, they learn a set of mapping functions to project high-dimensional features into a common low-dimensional latent space. Second, a multi-dimensional index in the metric space is built for applications like efficient retrieval [3][9], classification [7], event detection [8], etc. User interactions in relevance feedback provide useful prior knowledge that could be used to narrow the semantic gap [1]. However, it turns out to be a great challenge to mine accumulated prior knowledge with long-term relevance feedback because there are abundant correlations among heterogeneous media data.

Regarding above two issues, we propose a new cross-media retrieval and optimization method for image and audio data based on relevance feedback. For example, users can query images of an animal by submitting either its images or a sound of its roaring. Specifically, we first analyze underlying statistical correlation between visual feature matrix of images and auditory feature matrix of audios, find a correlation-preserved mapping to a low-dimension isomorphic space where the cross-media distance is define for similarity retrieval. Furthermore, we propose short-term learning within a single query session and long-term learning over the course of many query sessions for performance optimization. For long-term learning, we classify positive and negative feedbacks into pair-wise constraints, based on which two discriminative functions are defined. The short-term optimizing results are fused into the objective function together with the long-term learning results to improve multimodal data representation. Our approach not only explores global statistical cross-media correlation between image and audio data, but also optimizes data representation via relevance feedback.

2. MULTIMODAL REPRESENTATION

In this section, we analyze statistical correlation between image features and audio features to construct multimodal representation and estimate cross-media distance, based on which cross-media retrieval is generated.

Since different types of multimedia data are heterogeneous in low-level features, cross-media distance in feature level is not directly measurable. Therefore, we first analyze canonical correlation between image feature vectors and audio feature vectors to generate a correlation-based unified multimodal representation for all training samples. Formally, let I denote image feature matrix, A denote audio feature matrix. Motivated by CCA (Canonical Correlation Analysis) which has been used in many data mining applications[12][14], we calculate basis vectors for image and audio samples such that the correlation between the projections of the two sets of samples onto the basis vectors are mutually maximized, that is:

$$\max_{W_1,W_2} corr(IW_1, AW_2) = \max_{W_1,W_2} \frac{(IW_1, AW_2)}{\|IW_1\|\|AW_2\|} = \max_{W_1,W_2} \frac{W_1' C_{IA} W_2}{\sqrt{W_1' C_{II} W_1 W_2' C_{AA} W_2}} \quad (1)$$

where C is covariance matrix, W_1 and W_2 are the transformation matrices. Since the solution of equation (1) is not affected by re-scaling W_1 or W_2 either together or independently, the optimization of (1) is equivalent to maximizing the numerator subject to $W_1' C_{II} W_1 = 1$ and $W_2' C_{AA} W_2 = 1$. Then with Lagrange multiplier method we can find solutions for W_1 and W_2 by $C_{IA} C_{AA}^{-1} C_{AI} W_1 = \lambda^2 C_{II} W_1$, which is a generalized Eigenproblem. By choosing the same number of eigenvectors, we not only maximize canonical correlation between image and audio feature matrices, but also transfer them into the same dimensions.

Let $X = \{x_1, x_2, ..., x_n\}' \in \mathrm{R}^{n \times d}$ denote the isomorphic multimodal representation of image and audio data with the total number of n where d is the dimension after CCA process. We calculate the Euclidean distance between any two samples $x_i, x_j, (i, j \in [1, n], i \neq j)$, and obtain the corresponding distance matrix $\mathrm{H} \in \mathrm{R}^{n \times n}$.

We further analyze multimodal geometrical structure of all samples based on the matrix $\mathrm{H} \in \mathrm{R}^{n \times n}$. Specifically, we construct a multimodal adjacency graph $G(V, E)$: for any sample x_i, there is a corresponding vertex $v_i \in V$ and for any two vertices $v_i, v_j \in V$, we put an edge between them with distance c_{ij} which is defined as follows:

$$c_{ij} = \begin{cases} \mathrm{H}_{ij}, & if\ \mathrm{H}_{ij} < \varepsilon \\ \infty, & otherwise \end{cases} \quad (2)$$

where ε is a distance reflecting the measurement of locality. To model the global geometrical structure, we define the length of a path as the sum of all pair-wise distances along the path, and replace the value of c_{ij} with the length of the shortest path between v_i, v_j. So far, cross-media retrieval can be generated based on the geodesic distance on the multimodal graph.

3. RELEVANCE FEEDBACK

Although in recent years, relevance feedback has been studied from the perspective of machine learning, most of which are used for application of content-based image retrieval [13]. However, in cross-media retrieval that we target at, relevance feedback covers not only images but also audios. Besides, most learning methods only take into account current query session and the knowledge obtained from the past user interactions with the system is forgotten. Therefore, in this section, we propose both short-term and long-term strategies to further explore prior knowledge over the course of many query sessions so as to improve cross-media retrieval performance.

3.1 Short-term Refinement

Here we present a simple way to update distance c_{ij} gradually. Intuitively, the samples marked by the user as positive examples in a query session share some common semantics. Therefore, in short-term relevance feedback, we shorten the distance between them. Similarly, we can lengthen the distances between the positive examples and negative examples, as follows:

$$c_{ij} = c_{ij} / \alpha \ \ (if\ x_i, x_j \in \mathrm{K}^+) \quad (3)$$

$$c_{ij} = c_{ij} \cdot \beta \ \ (if\ x_i \in \mathrm{K}^+ \ \&\ x_j \in \mathrm{K}^-) \quad (4)$$

where $\alpha > 1$ and $\beta > 1$. K^+ and K^- represent positive feedbacks and negative feedbacks respectively. Then we define the weight of edges on the multimodal graph as below:

$$w_{ij} = \begin{cases} \exp(-c_{ij} / t), & if\ \ c_{ij} < \rho \\ 0, & otherwise \end{cases} \quad (5)$$

where t, ρ are suitable constants.

Let $Y = \{y_1, y_2, ..., y_n\}' \in \mathrm{R}^{n \times k}$ denote the optimized multimodal representation. Then the objective of the whole relevance feedback strategy is to learn a mapping from $X = \{x_1, x_2, ..., x_n\}'$ to $Y = \{y_1, y_2, ..., y_n\}'$. That is, for any $x_i \in X (1 \le i \le n)$ we have:

$$y_i = M^T x_i \quad (6)$$

where $M \in \mathrm{R}^{d \times k}$ is the projection matrix.

It is reasonable that a "good" mapping should minimize the objective function $\sum (y_i - y_j)^2 w_{ij}$, which incurs a heavy penalty if neighboring data i and j are mapped far apart. Then we have:

$$\sum_{i,j} (y_i - y_j)^2 w_{ij} = 2 \sum_i M^T x_i D_{ii} x_i^T M - 2 \sum_{i,j} M^T x_i w_{ij} x_i^T M = 2 M^T \mathrm{X}(\mathrm{D} - W) X^T M \quad (7)$$

where $W = [w_{ij}]$, $L = D - W$ is a Laplacian matrix and D is a diagonal matrix defined as $D_{ii} = \sum_j w_{ji}$. Thus the minimization problem is:

$$\min_{M^T M = I} tr(M^T X L X^T M) \quad (8)$$

3.2 Long-term Relevance Feedback

Formally, we define two sets of data pairs P_t and N_t based on the positive and negative sample sets the user labels in the *t*-th round of relevance feedback. In P_t each pair of data is of the same labels, and in N_t each pair of data is of different labels. For long-term relevance feedback we define:

$$P = \bigcup_t P_t, \quad N = \bigcup_t N_t \quad (9)$$

In this way, data pairs in P are semantically similar to each other and those in N are semantically dissimilar to each other, which can be used as pairwise constraints. The sum of squared distances of data pairs $(x_i, x_j) \in P$ can be calculated as follows:

$$\sum_{(x_i, x_j) \in P} (M^T x_i - M^T x_j)^T (M^T x_i - M^T x_j) = \sum_{(x_i, x_j) \in P} \mathrm{tr}[M^T (x_i - x_j)(x_i - x_j)^T M] = tr(M^T S_w M) \quad (10)$$

where $S_w = \sum_{(x_i,x_j)\in P} (x_i - x_j)(x_i - x_j)^T$ and tr is the trace operator. The distances among data pairs $(x_i, x_j) \in P$ should be as small as possible. Thus we have the following function:

$$\min_{M^T M = I} tr(M^T S_w M) \tag{11}$$

in which the constraint $M^T M = I$ is imposed to prevent arbitrary scaling of the projection. Similarly, the distances between data pairs from N should be as large as possible. Thus we have:

$$\max_{M^T M = I} tr(M^T S_b M) \tag{12}$$

where $S_b = \sum_{(x_i,x_j)\in N} (x_i - x_j)(x_i - x_j)^T$.

3.3 Combining Short- & Long-term Learning

So far, the relevance feedback strategy should satisfy not only (8) but also (11)(12) in order to preserve refined locality structure and benefit from accumulated pairwise constraints. Motivated by [6] which proposed an image classification method via cluster distance optimization, we merge (8)(11)(12) into an overall discriminative objective function as below:

$$\max_{M^T M = I} tr[M^T S_b M (M^T S_t M + \delta M^T X L X^T M)^{-1}] \tag{13}$$

where $S_t = S_b + S_w$ denotes the total scatter matrix and the coefficient δ balances the model complexity and the empirical loss. Then the solution of the optimal matrix M is given by the maximum eigenvalues to the generalized eigenvalue problem in form of $XLX^T M = \lambda X(I + \delta L)X^T M$. Based on above discussions, the optimization process is listed below.

Algorithm 1. Relevance feedback algorithm

Input: distance matrix $C = [cij]$;
similar set P and dissimilar set N;

Output: optimized multimodal representation Y.

1. Calculate the weight matrix $W = [w_{ij}]$ in (5), and obtain the Laplacian matrix L in (8);
2. Calculate data sets P and N in Eq.(9) based on long-term relevance feedback;
3. Calculate pairwise constraint matrices S_t, S_b;
4. Compute $M = [m_1, m_2, ..., m_c]$, in which $m_1, m_2, ..., m_c$ are the eigenvectors corresponding to the c largest non-zero eigenvalues of (13), and map x_i to the optimized multimodal subspace with $y_i = M^T x_i \ (1 \le i \le n)$.

4. EXPERIMENTS

We test our proposed algorithm for cross-media retrieval between image and audio data, and provide an extensive performance study of our algorithm in comparison with the state-of-the-art methods. Precision is defined as the percentage of correctly retrieved samples in the top-k-returned results.

4.1 Datasets

The image-audio dataset is collected from Corel image gallery, Webpages, etc. There are 4800 image and audio samples in total, which are grouped into 12 categories, such as zither, dog, dolphin, bird, elephant, tiger, explosion, car, train etc. In our experiments, if a returned result and the query example are in the same category, it is regarded as a correct result. Two types of visual features are extracted and normalized, including color correlogram in Hue Saturation Value color space and Tamura Texture. Auditory features include Mel-frequency Cepstral Coefficients (MFCC), Centroid, Rolloff and Spectral Flux. Since audio is a kind of time series data, we employ Fuzzy Clustering method [10] on original auditory features to get isomorphic indexes.

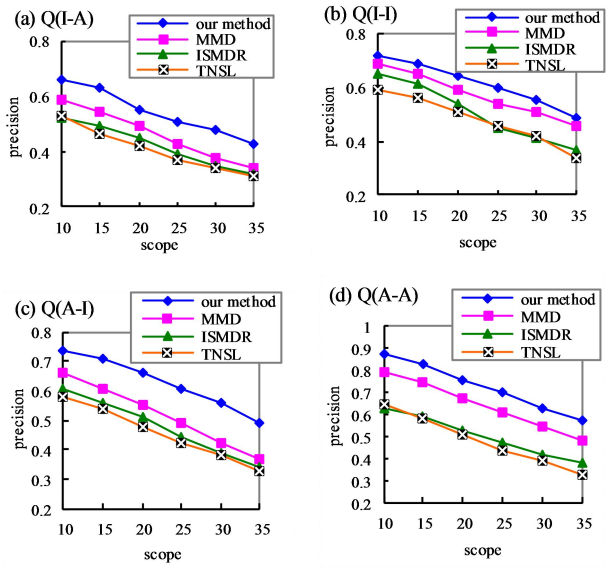

Figure 1(a)(b)(c)(d). Retrieval performance comparison of different algorithms in different retrieval scenarios

4.2 Performance Comparison

We compare our method with three content-based cross-media retrieval and relevance feedback methods: the Multimedia Document (MMD) method in [9], the Isomorphic and Sparse Multimodal Data Representation (ISMDR) method in [14] and Tagging-based Nonnegative Subspace Learning (TNSL) method in [11]. We randomly choose 20% of images and audios in each category as query examples to perform cross-media retrieval, and for each query, we select at most 3 positive and 3 negative multimedia objects, which could be images or audios, from returned results. The user's relevance feedbacks are used for both sort-term and long-term learning.

Since query examples and retrieval results could be images or audios, we generate four types of queries to give detailed comparisons, namely *Q(I-A)*, *Q(I-I)*, *Q(A-I)*, *Q(A-A)*. For instance, *Q(I-A)* searches relevant audio samples given an image example. Figure 1 shows the average *precision-scope* curves of different algorithms after relevance feedback. We have the following observations from Figure 1: our method outperforms the other three methods in four retrieval scenarios. For example, in Figure 2(c) Q(A-I), when the number of returned images is 15 the precision is 0.712 with our method, while the MMD, ISMDR and TNSL methods achieve precision of 0.611, 0.564 and 0.543 respectively. The performance gain of our method is probably attributed to the following reasons:

(1) ISMDR, TNSL and MMD methods optimize cross-media similarity with the help of surrounding texts or strictly selected auxiliary tags which express semantic information more directly than audios do; (2) our algorithm utilizes feature-level image-audio statistical correlation and optimizes multimodal data representation in relevance feedback, which are ignored in the other three methods; (3) the MMD method takes a global view on

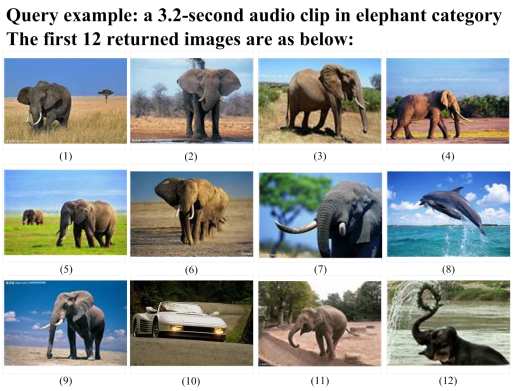

Figure. 2 An example of *Q(A-I)*

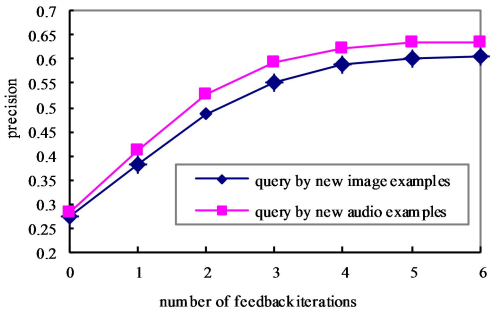

Figure. 3. Average results of querying by new examples

different objects in the multimedia documents, thus it is more flexible and capable to different retrieval scenarios compared to ISMDR and TNSL methods.

Figure 2 gives an example of Q(A-I). We submit an example of a 3.2-second audio clip selected from elephant category, and the first 12 returned images are shown in Figure 2. It can be seen that 10 images are corrected returned.

4.3 Query Example outside the Database

Above results are obtained when the query examples are in the database. If a query example is out of the database, we call it a new sample. We perform retrieval with new samples to evaluate the generalization capability of our algorithm. 360 images and 360 audios are used as queries outside the training database for testing. Although the testing set have no optimized multimodal representations, we can use the transformation matrices learned in equation (1) for initial feature mapping, and then calculate the Euclidean distances between testing set and training set, based on which cross-media retrieval could be generated. Further, we make use of relevance feedback to narrow the semantic gap. We classify cross-media retrieval into querying by new audio examples and querying by new image examples.

Figure 3 shows precision performance by looking at the top 20 returns. It can be seen that the performances of querying by new examples are good overall. Although when no relevance feedback is involved, some new images and audios are well recognized, the precision performances climb to 0.592 and 0.553 for querying with new audios and querying with new images respectively at the third round of relevance feedback. As users interact with the system, querying with new images and new audios can be further optimized with our long-term learning model.

5. CONCLUSIONS

In this paper, we propose a new cross-media retrieval method based short-term and long-term relevance feedback. Our approach first learns a multimodal representation via statistical correlation analysis between image and audio features, and generates flexible cross-media retrieval; more importantly, by fusing both accumulated discriminative knowledge and local data structure into the objective function of relevance feedback, we improve multimodal representation for retrieval performance optimization. Experiment results of cross medial retrieval on image-audio dataset confirm the improvements of our method over previous works in search accuracy.

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (No.61373109, No.61003127), State Key Laboratory of Software Engineering (SKLSE2012-09-31).

REFERENCES

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

[2] Zhuang Y.T., Wang Y.F., Wu F., et al. 2013. Supervised Coupled Dictionary Learning with Group Structures for Multi-modal Retrieval. *Proceeding of the 27th Conference on Artificial Intelligence.*1070-1076.

[3] Wang W., Ooi B.C., Yang X.Y., et al. 2014. Effective multi-modal retrieval based on stacked auto-encoders. *International Conference on Very Large Data Bases.*

[4] McGurk H., MacDonald J. 1976. Hearing Lips and Seeing Voices. *Nature*, 264: 746-748.

[5] Yang Y., Ma Z.G., Hauptmann A., et al. 2013. Feature Selection for Multimedia Analysis by Sharing Information among Multiple Tasks. *IEEE Transactions on Multimedia.* 15(3):661-669.

[6] Cai D., He X.F., Han J.W. 2007. Semi-supervised Discriminant Analysis. *International Conference on Computer Vision.* 1-7.

[7] Srivastava N., Salakhutdinov R. 2012. Multimodal Learning with Deep Boltzmann Machines. *In Neural Information Processing Systems*. 2231-2239.

[8] Yuan J.S., Luo J.B. and Wu Y.. 2010. Mining Compositional Features from GPS and Visual Cues for Event Recognition in Photo Collections, *IEEE Trans. on Multimedia*, 12(7):705-716.

[9] Yang Y., Zhuang Y.T. and Wang W.H. 2008. Heterogeneous multimedia data semantics mining using content and location context. *ACM Multimedia Conference*. 655-658.

[10] Zhang H., Liu J. and Ma Z.G. 2013. Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval. *Neurocomputing* vol.119, 10-16.

[11] Gupta S.K., Phung D., Adams B. et al. 2010. Nonnegative shared subspace learning and its application to social media retrieval. *ACM International Conference on Knowledge Discovery and Data Mining*. 1169-1178.

[12] Sun T., Chen S. 2007. Locality preserving CCA with applications to data visualization and pose estimation. *Image and Vision Computing*. Vol.25, 531-543.

[13] Yu J., and Tian Q. 2006. Learning Image Manifolds by Semantic Subspace Projection. *ACM Multimedia Conference.* 297-306.

[14] Zhang H., Chen L. 2013. Isomorphic and Sparse Multimodal Data Representation based on Correlation Analysis. *International Conference on Image Processing*. 3959-3962.