Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Understanding visual-auditory correlation from heterogeneous features for cross-media retrieval*

Hong ZHANG†[1,2], Yan-yun WANG[3], Hong PAN[4], Fei WU[2]

([1]*College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China*)

([2]*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

([3]*School of Elementary Education, Hangzhou Normal University, Hangzhou 310036, China*)

([4]*School of Information Engineering, Hangzhou Normal University, Hangzhou 310036, China*)

†E-mail: zhanghong_zju@yahoo.com.cn

Received Apr. 11, 2007; revision accepted Aug. 12, 2007; published online Jan. 10, 2008

Abstract: Cross-media retrieval is an interesting research topic, which seeks to remove the barriers among different modalities. To enable cross-media retrieval, it is needed to find the correlation measures between heterogeneous low-level features and to judge the semantic similarity. This paper presents a novel approach to learn cross-media correlation between visual features and auditory features for image-audio retrieval. A semi-supervised correlation preserving mapping (SSCPM) method is described to construct the isomorphic SSCPM subspace where canonical correlations between the original visual and auditory features are further preserved. Subspace optimization algorithm is proposed to improve the local image cluster and audio cluster quality in an interactive way. A unique relevance feedback strategy is developed to update the knowledge of cross-media correlation by learning from user behaviors, so retrieval performance is enhanced in a progressive manner. Experimental results show that the performance of our approach is effective.

Key words: Heterogeneity, Cross-media retrieval, Subspace optimization, Dynamic correlation update
doi:10.1631/jzus.A071191 **Document code:** A **CLC number:** TP37; TP391

INTRODUCTION

Multimedia data of different modalities, such as image and audio, carry their own contribution to high-level semantics, and the presence of one has usually a "complementary" effect on the other. For example, when hearing a bird singing we expect the image of a bird, seeing a loudmouthed tiger we expect the presence of its voice, and the image of an explosion usually brings the sound of detonation, etc. It is important and interesting to obtain cross-media retrieval (Wu *et al*., 2006), which returns relevant multimedia data, by submitting a query example of different modalities. In fact, some psychological experiments on cross-media influence have proved the importance of synergistic retrieval of different modalities in the human perception research (McGurk and MacDonald, 1976).

Content-based multimedia retrieval (CBMR) techniques, such as image retrieval (Wang *et al.*, 2004; Lu and Chang, 2007) and audio retrieval (Zhao *et al*., 2002), attempt to provide an effective and efficient tool for searching multimedia data (Zhang and Chen, 2002). For example, in a query-by-example image retrieval system, users can get relevant images by submitting an image as query example. However, most of CBMR approaches focus on single modality retrieval and are not applicable for cross-media retrieval. The basic challenge of cross-media retrieval lies in correlation matching from heterogeneous low-level features. Multimedia data are usually represented with different feature vectors (Zhang *et al*.,

* Project supported by the National Natural Science Foundation of China (Nos. 60533090 and 60773051) and the Natural Science Foundation of Zhejiang Province (No. Y105395), China

2004): images can be represented with visual features of color, texture, and shape, while audio segments can be represented with auditory features like spectral flux and root mean square. So it is not easy to judge the correlation between an image with 200-dimensional visual feature vector and an audio segment with 500-dimensional auditory feature vector, although both may describe relevant high-level semantics, such as bird's picture and bird's sound.

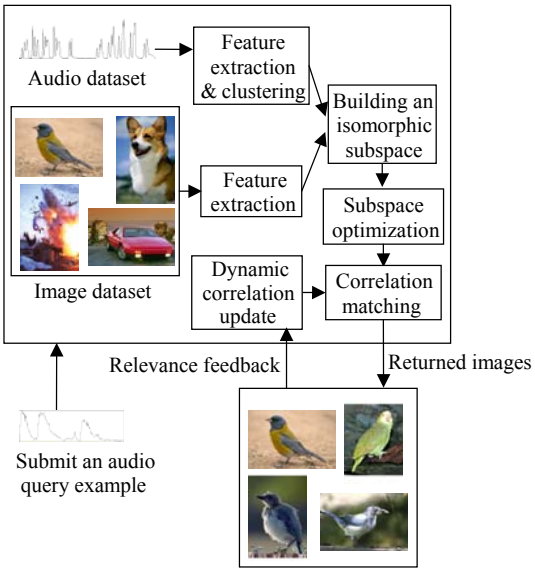

Fig.1 Overview of cross-media retrieval mechanism

This paper focuses on cross-media retrieval between image and audio by exploring content-based visual-auditory correlations. As Fig.1 shows, in our system users can find relevant images by submitting an audio segment as query example, or vice versa. We first build an isomorphic subspace based on canonical correlation analysis of visual-auditory features; and then we optimize the qualities of image cluster and audio cluster by utilizing complementary information; thirdly, users' prior knowledge is melt during relevance feedback to globally refine cross-media correlation matching results.

The organization of this paper is as follows. Section 2 reviews the related works. Section 3 introduces how to discover cross-media correlation and build an isomorphic subspace. Section 4 presents an iterative method to optimize the clustering results. Section 5 presents a correlation update strategy to improve cross-media retrieval through users' interactions. Experimental results are shown in Section 6. The concluding remarks are given in Section 7.

RELATED WORKS

Recently, a number of researches focus on how to learn the correlations between interrelated multimodal data, which is quite similar to cross-media retrieval scenario discussed in this paper. Such research works can be classified into the following groups:

(1) Multi-modal video content analysis. Video content can be represented with different low-level features, such as visual features of key frames and auditory features of video shots. To better understand high-level semantics researchers, refer to multi-modal analysis (Slaney and Covell, 2000; Snoek *et al*., 2005), most of which train a separate classification model for each media track, and then use the weighted-sum rule to fuse a class-prediction decision on semantic events occurred in video streams. Yang and Hauptmann (2004) proposed a statistical learning method to name every individual person appearing in broadcast news videos with names detected from the video transcript. Such research successfully improves video semantic understanding by utilizing different video features synthetically. However, cross-media correlation matching between visual and auditory features is not addressed.

(2) Image-text fusion. To improve content-based image retrieval (CBIR), some researchers resort to different kinds of image attributes, including low-level visual features, surrounding texts, hyperlinks, etc. Chen *et al*.(2001) linearly combined the similarity on textual features measured by dot product and Euclidean distance on visual features with equal weight. Wang *et al*.(2004) treated visual features of images and textual features of surrounding texts as two types of image attributes, and proposed an image-text similarity propagation method to optimize the image clustering results. These methods improve the retrieval performance by using text as complementary information, which itself expresses a certain semantics. However, different from text, there is no direct semantic information for image and audio data. Therefore, image-text fusion techniques cannot solve

the correlation matching problem that cross-media retrieval faces.

(3) Image-text cross. This group focuses on searching across text and image. Image annotation is a hot topic in this area (Duygulu *et al*., 2002; Barnard *et al*., 2003; Jeon *et al*., 2003). Barnard *et al*.(2003) presented a multi-modal modeling approach for predicting words associated with a whole image (auto-annotation) and corresponding to the particular image regions (region naming). Duygulu *et al*.(2002) proposed a model of object recognition as machine translation, and the recognition was a process of annotating image regions with words. Compared with image annotation, cross-media retrieval is relatively new and faced with new challenge of correlation matching among heterogeneous content features.

LINEAR CORRELATION MODELING

In this section, we learn the correlation between visual features and auditory features during dimension reduction, and then build an isomorphic subspace where the correlation is further preserved.

Preprocessing

Image and audio samples can be considered two different representations of semantic concepts. We employ a distinct feature learning method of canonical correlation analysis (CCA) (Hotelling, 1936) to analyze visual and auditory feature spaces simultaneously. The underlying idea of CCA is very intuitive: it looks for basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized.

Formally, consider a multivariate random vector of the form (x, y), and a sample of instances $S=\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, where $\boldsymbol{x}_i=(x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{y}_i=(y_{i1}, \ldots, y_{iq})$. Let $\boldsymbol{S}_x$ denote $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{S}_y$ denote $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. Project x onto a direction $\boldsymbol{W}_x$, and y onto a direction $\boldsymbol{W}_y$:

$$\begin{cases} \boldsymbol{S}_x \xrightarrow{\boldsymbol{W}_x} \boldsymbol{L} =< \boldsymbol{x}'_1, ..., \boldsymbol{x}'_n >,\ \boldsymbol{x}'_i = (x'_{i1}, ..., x'_{im}), \\ \boldsymbol{S}_y \xrightarrow{\boldsymbol{W}_y} \boldsymbol{M} =< \boldsymbol{y}'_1, ..., \boldsymbol{y}'_n >,\ \boldsymbol{y}'_i = (y'_{i1}, ..., y'_{im}). \end{cases} \tag{1}$$

Then the problem of correlation preserving boils down to finding the optimal $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$, which makes the correlation between $\boldsymbol{S}_x'$ and $\boldsymbol{S}_y'$ be maximally in accordance with that between $\boldsymbol{S}_x$ and $\boldsymbol{S}_y$. In other words, the function to be maximized is

$$\begin{aligned} \rho &= \max_{\boldsymbol{W}_x, \boldsymbol{W}_y} corr(\boldsymbol{S}_x\boldsymbol{W}_x, \boldsymbol{S}_y\boldsymbol{W}_y) \\ &= \max_{\boldsymbol{W}_x, \boldsymbol{W}_y} \frac{(\boldsymbol{S}_x\boldsymbol{W}_x, \boldsymbol{S}_y\boldsymbol{W}_y)}{\|\boldsymbol{S}_x\boldsymbol{W}_x\| \cdot \|\boldsymbol{S}_y\boldsymbol{W}_y\|} \\ &= \max_{\boldsymbol{W}_x, \boldsymbol{W}_y} \frac{\boldsymbol{W}_x'\boldsymbol{C}_{xy}\boldsymbol{W}_y}{\sqrt{\boldsymbol{W}_x'\boldsymbol{C}_{xx}\boldsymbol{W}_x\boldsymbol{W}_y'\boldsymbol{C}_{yy}\boldsymbol{W}_y}}, \end{aligned} \tag{2}$$

where $\boldsymbol{C}$ is a covariance matrix. Since the solution of Eq.(2) is not affected by re-scaling $\boldsymbol{W}_x$ or $\boldsymbol{W}_y$ either together or independently, the optimization of ρ is equivalent to maximizing the numerator subject to $\boldsymbol{W}_x'\boldsymbol{C}_{xx}\boldsymbol{W}_x=1$ and $\boldsymbol{W}_y'\boldsymbol{C}_{yy}\boldsymbol{W}_y=1$. Then with Lagrange multiplier method we can get

$$\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{W}_x = \lambda^2\boldsymbol{C}_{xx}\boldsymbol{W}_x, \tag{3}$$

which is a generalized Eigen problem of the form $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$. And the sequence of $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ can be obtained by solving the generalized eigenvectors.

Simi-supervised correlation preserving mapping

We propose a semi-supervised correlation preserving mapping (SSCPM) method, which needs only partially labelled data, to analyze canonical correlation between image and audio that have similar semantics. Given unlabelled image and audio data, and suppose be belonging to Z semantic categories, then SSCPM is described below:

(1) Semi-supervised clustering. (i) For each semantic category Z_i (e.g., Z_1 represents "tiger" category), manually select several image examples Ω_i which describe the same semantics as Z_i (e.g., tiger image), calculate centroid C_i for the images in Ω_i, and label C_i with Z_i; (ii) Run K-means clustering algorithm (Xing *et al*., 2003) on the whole image dataset with C_i (i=1, …, Z) as the start point and Z as the number of clusters; (iii) Let C_i^* (i=1, …, Z) denote the output centroid after K-means clustering, label every image cluster by comparing the distance between C_i^* and the start point C_i, label the ith image cluster with Z_i if C_j is the nearest to C_i^*; (iv) Implement the above steps on the audio dataset, and let I_i^*, A_i^* (i=1, …, Z)

denote image clusters and audio clusters respectively after semi-supervised clustering.

(2) Correlation preserving mapping. (i) Rank all image clusters I_i^* so as to match all audio clusters in semantics; (ii) Extract visual feature matrix $\boldsymbol{S}_i^x$ for image cluster I_i^*, and auditory feature matrix $\boldsymbol{S}_i^y$ for audio cluster A_i^*, and find optimal $\boldsymbol{W}_i^x$ and $\boldsymbol{W}_i^y$ for $\boldsymbol{S}_i^x$ and $\boldsymbol{S}_i^y$ with the method presented in Section 3.1; (iii) Construct SSCPM subspace S^m by $(\boldsymbol{S}_i^x)'=\boldsymbol{S}_i^x\boldsymbol{W}_i^x$ and $(\boldsymbol{S}_i^y)'=\boldsymbol{S}_i^y\boldsymbol{W}_i^y$.

In this way, visual features are analyzed together with auditory features, which is a kind of "interaction" process. For example, tiger image cluster is analyzed together with tiger audio cluster for correlation detection, therefore tiger image affects the location of tiger audio in subspace S^m to a certain extent, and vice versa. Moreover, bird image cluster is mapped with bird audio cluster, and bird audio differs from tiger audio, so bird image could be located differently from tiger image in S^m. Therefore, the cross-media correlations are discovered and preserved in the mapping process.

General distance function

With $\boldsymbol{W}_i^x$ and $\boldsymbol{W}_i^y$ we can map heterogeneous visual feature vectors and audio feature vectors into an m-dimensional subspace S^m, where canonical correlations between initial visual and auditory features are further preserved. We define a general similarity measure for all images and audio objects in S^m.

After SSCPM, large amounts of complex numbers occur. Let $\boldsymbol{x}_i'=(x_{i1}', \ldots, x_{im}')$ $(x_{ik}'=a+b\cdot i,\ a,\ b\in\mathbb{R})$ denote the coordinates of an image or an audio object in S^m. We represent x_{ik}' in polar coordinates as

$$x_{ik}'=(\beta_{ik},|x_{ik}'|), \tag{4}$$

where $\beta_{ik}=\arctan(b/a)$, $|x_{ik}'|=\sqrt{a^2+b^2}$, $k\in[1,m]$.

And define the general distance function as

$$Crodis(\boldsymbol{x}_i',\boldsymbol{x}_j')=\mathrm{sqrt}\left\{\sum_{k=1}^{m}\left(|x_{ik}'|^2+|x_{jk}'|^2-2|x_{ik}'||x_{jk}'|\cos(|\beta_{ik}-\beta_{jk}|)\right)\right\}. \tag{5}$$

We name Eq.(4) as SSCPM coordinates of $\boldsymbol{x}_i'$. Since $\boldsymbol{x}_i'$ and $\boldsymbol{x}_j'$ can represent all media objects in S^m, Eq.(5) is a cross-media distance function for image and audio.

OPTIMIZATION STRATEGY

Because of the well-known semantic gap (Zhao and Grosky, 2002), the results of image clustering and audio clustering in S^m are not always consistent with human perception, which would cause bad effects on cross-media retrieval. It has been proved that relationships among different types of objects are effective on improving cluster quality (Ye and Li, 2005; Zhang *et al*., 2007). Enlightened by this mechanism, we optimize image and audio cluster quality in subspace S^m by interactional correlation propagation to bridge the semantic gap.

Overview of the approach

The basic idea of interactional optimization is that image similarity and audio similarity can mutually influence each other through the bridge of cross-media correlation, and in this way the cluster quality in SSCPM subspace S^m can be improved. Let P and N denote image objects and audio objects in S^m, respectively. As the values of image visual feature differ greatly from those of auditory feature in threshold, we assume P and N occupy different areas in S^m. Let P_i represent a specific image object and N_j a specific audio object.

Fig.2 shows the sketch map of this approach. The dotted lines denote cross-media distance between image and audio in subspace S^m, and the real lines are distances within image and within audio. The length of the real line represents the degree of single-

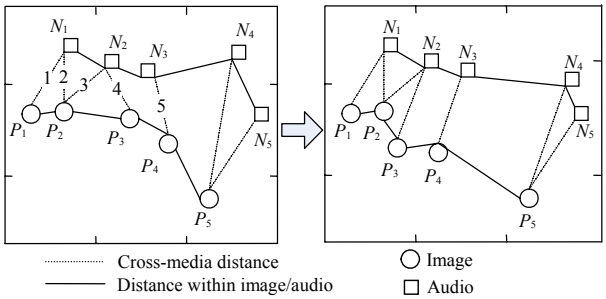

Fig.2 Sketch map of image-audio optimization

modality similarity, and the length of the dotted line represents the degree of cross-media correlation. The left-part shows the original image-audio relationships, i.e., P_1 and P_2 are similar images, but both are dissimilar to P_3 and P_4. Also N_1, N_2, N_3 are similar audio data, but are dissimilar to audio object N_4.

Simultaneously, audios N_1, N_2, N_3 have close correlation among images P_1, P_2, P_3, P_4, i.e., the lengths of dotted lines 1~5 are short and similar. Assume that semantically P_1 and P_2 are similar to P_3 and P_4, but P_5 are dissimilar to P_3 and P_4. Then the similarity among audios N_1, N_2, N_3 can be propagated to images P_1, P_2, P_3, P_4 through the dotted lines 1~5, and accordingly P_1, P_2, P_3, P_4 are clustered together as the right-part shows. On the other hand, although originally N_4 is not quite similar to N_5, they are cross-media correlated to the same image P_5, thus the dissimilarity between N_4 and N_5 is reduced. When this process performs iteratively, image cluster quality and audio cluster quality are improved in an interactional way, so as to better reflect intrinsic similarities between multimedia objects.

Formal description

This interactional optimization method consists of three steps: (1) the construction of image similarity matrix, audio similarity matrix and image-audio cross-media correlation matrix; (2) normalization; (3) iterative similarity propagation till convergence.

According to general distance function in Eq.(5), we can get image similarity matrix $\boldsymbol{L}$ and audio similarity matrix $\boldsymbol{M}$, and also cross-media correlation matrix $\boldsymbol{C}$. Matrices $\boldsymbol{L}$, $\boldsymbol{M}$, $\boldsymbol{C}$ are normalized below:

$$L_{ij}^{*}=\frac{L_{ij}}{\max(L(i,\cdot))},\ M_{ij}^{*}=\frac{M_{ij}}{\max(M(i,\cdot))},\ C_{ij}^{*}=\frac{C_{ij}}{\sum C(i,\cdot)}. \tag{6}$$

Let $\boldsymbol{L}^{*}$ and $\boldsymbol{M}^{*}$ denote the normalized similarity matrices after each iteration step. Then the interactional optimization process can be described as follows:

$$\begin{cases}\boldsymbol{L}^{*}=\alpha\boldsymbol{L}+(1-\alpha)\gamma\boldsymbol{C}\boldsymbol{M}^{*}\boldsymbol{C}',\\ \boldsymbol{M}^{*}=\beta\boldsymbol{M}+(1-\beta)\gamma\boldsymbol{C}'\boldsymbol{L}^{*}\boldsymbol{C},\end{cases}\quad 0<\alpha,\beta,\gamma<1, \tag{7}$$

where α, β are the weights, γ is a decay factor to make sure that the propagated similarities are weaker than the original similarities. The convergence of this iteration process can be proved by calculating the limitation of $\boldsymbol{L}^{*(n)}-\boldsymbol{L}^{*(n-1)}$ as follows:

$$\begin{aligned}\boldsymbol{L}^{*(n)}-\boldsymbol{L}^{*(n-1)}&=[\alpha\boldsymbol{L}+(1-\alpha)\gamma\boldsymbol{C}\boldsymbol{M}^{*(n)}\boldsymbol{C}']-\\&\quad[\alpha\boldsymbol{L}+(1-\alpha)\gamma\boldsymbol{C}\boldsymbol{M}^{*(n-1)}\boldsymbol{C}']\\&=(1-\alpha)\gamma\boldsymbol{C}(\boldsymbol{M}^{*(n)}-\boldsymbol{M}^{*(n-1)})\boldsymbol{C}',\end{aligned} \tag{8}$$

$$\begin{aligned}\boldsymbol{M}^{*(n)}-\boldsymbol{M}^{*(n-1)}&=[\beta\boldsymbol{M}+(1-\beta)\gamma\boldsymbol{C}'\boldsymbol{L}^{*(n-1)}\boldsymbol{C}]-\\&\quad[\beta\boldsymbol{M}+(1-\beta)\gamma\boldsymbol{C}'\boldsymbol{L}^{*(n-2)}\boldsymbol{C}]\\&=(1-\beta)\gamma\boldsymbol{C}'(\boldsymbol{L}^{*(n-1)}-\boldsymbol{L}^{*(n-2)})\boldsymbol{C}.\end{aligned} \tag{9}$$

Then replace $\boldsymbol{M}^{*(n)}-\boldsymbol{M}^{*(n-1)}$ in Eq.(8) with Eq.(9), we obtain

$$\boldsymbol{L}^{*(n)}-\boldsymbol{L}^{*(n-1)}=(1-\alpha)(1-\beta)\gamma^{2}\boldsymbol{C}\boldsymbol{C}'(\boldsymbol{L}^{*(n-1)}-\boldsymbol{L}^{*(n-2)})\boldsymbol{C}\boldsymbol{C}', \tag{10}$$

Let $\rho=(1-\alpha)(1-\beta)\gamma^{2}$, $\chi=\boldsymbol{C}\boldsymbol{C}'$, we have

$$\begin{aligned}\boldsymbol{L}^{*(n)}-\boldsymbol{L}^{*(n-1)}&=\rho\chi(\boldsymbol{L}^{*(n-1)}-\boldsymbol{L}^{*(n-2)})\chi\\&=\ldots=\rho^{n-1}\chi^{n-1}(\boldsymbol{L}^{*(1)}-\boldsymbol{L})\chi^{n-1}.\end{aligned} \tag{11}$$

Since $\boldsymbol{L}^{*(1)}-\boldsymbol{L}$ is a constant matrix, the convergence of $\boldsymbol{L}^{*(n)}-\boldsymbol{L}^{*(n-1)}$ depends on those of ρ^{n-1} and χ^{n-1}. According to Eqs.(12) and (13),

$$\chi_{ij}=\frac{\sum C(i,\cdot)C(j,\cdot)}{\sum C(i,\cdot)\sum C(j,\cdot)}<1\Rightarrow\chi^{n-1}\xrightarrow{n\to\infty}0, \tag{12}$$

$$\rho^{n-1}=[(1-\alpha)(1-\beta)\gamma^{2}]^{n-1}\xrightarrow{n\to\infty}0. \tag{13}$$

We can get $\boldsymbol{L}^{*(n)}-\boldsymbol{L}^{*(n-1)}\xrightarrow{n\to\infty}\boldsymbol{0}$, which proves the convergence of Eq.(7). Another advantage of this optimization method is that, in relevance feedback procedure (see Section 5), users' interactions can be better utilized on a well-clustered dataset.

DYNAMIC CORRELATION UPDATE

After the above optimization process, image similarity matrix $\boldsymbol{L}^{*}$ and audio similarity matrix $\boldsymbol{M}^{*}$ tend to be more consistent with human perception. However, cross-media correlation between image and audio remains unchanged because SSCPM coordinates of image and audio are unchanged [see Eq.(5)]. In this section, we discuss how to use $\boldsymbol{L}^{*}$ and $\boldsymbol{M}^{*}$ to dynamically update cross-media correlation matching results in relevance feedback.

An intuitive description of our algorithm is: if the query example is an image, a smoothing factor is initially assigned to positive and negative audio examples as ranking score; then they spread their scores to audio neighbors in an arithmetic progression. When the query example is an audio clip, we implement similar steps in relevance feedback to refine returned images. Formally, we define smoothing factor $R(i, j)$ and updated cross-media correlation matrix $\boldsymbol{A}=(a_{ij})$ as:

$$\boldsymbol{A}: a_{ij} = \lambda \cdot Crodis(i, j) + (1-\lambda) \cdot R(i, j), \quad (14)$$

where λ is a constant parameter in (0, 1). It can be seen that updated cross-media correlation is influenced by two parts: general distance $Crodis(i, j)$ and smoothing factor $R(i, j)$. The former reflects cross-media topology in SSCPM subspace, while the latter is calculated according to optimized image cluster results and audio cluster results in Section 4.

Let r be an image query example, P denote the set of positive audio marked by the user in a round of relevance feedback, and N denote the set of negative audio. The pseudo-code to calculate $R(i, j)$ is shown below:

Pseudo-code 1: dynamic correlation update

Input: $Crodis(i, j)$, feedback examples P, N, optimized audio similarity matrix $\boldsymbol{M}^*$.
Output: matrix $\boldsymbol{A}$ and smoothing factor $R(i, j)$.

```
Initialize R(i, j)=0;
Value a constant −τ1 (τ1>0);
For each positive audio p_i∈P do {
    R(r, p_i)=−τ1;
    Find p_i's k_i-nearest audio neighbors T: {t_1, …, t_{k_i}}
        according to matrix M*;
    Rank T in ascending order by the distances to p_i;
    d=τ1/k_i;
    For each t_j∈T do R(r, t_j)=−τ1+j·d;
        A(r, t_j)=λ·Crodis(r, t_j)+(1−λ)·R(r, t_j);
}
Value a constant τ2 (τ2>0);
For each negative audio n_i∈N do {
    R(r, n_i)=τ2, τ2>0;
    Find n_i's k_i-nearest audio neighbors H: {h_1, …, h_{k_i}}
        according to matrix M*;
    Rank H in ascending order by the distances to n_i;
    d=τ2/k_i;
    For each h_j∈H do R(r, h_j)=τ2−j·d;
        A(r, h_j)=λ·Crodis(r, h_j)+(1−λ)·R(r, h_j);
}
```

The topology of SSCPM subspace is updated with the relevance feedback process, which has a long-term influence on future cross-media retrieval. However, there are common and personal relevance feedbacks when different users interact with the system (Tan *et al*., 2006). So we employ the above long-term strategy in training stage, which provides common cross-media correlation feedback. For short-term effect on current retrieval session after training, we refine the correlation values, but not memorize them.

EXPERIMENTAL RESULTS

Dataset

We collect an image-audio dataset consisting of 15 semantic categories, such as dog, car, bird, explosion, tiger, goal, piano, zither, etc. In each semantic category there are 100 images and 50 audio segments. Most of them are collected from the Corel image galleries and the Internet. Some other audio clips are extracted from movies. A retrieved media object is considered correct if it belongs to the same category of the query example.

The extracted visual features include Color Histogram (in HSV space), Color Coherence Vector (CCV), and Tamura Texture. The combined visual feature vector is 500-dimensional. Auditory features we use are made up of Centroid, Rolof, Spectral Flux, and RMS (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 s, and employ Fuzzy Clustering algorithm (Zhao *et al*., 2002) on auditory features for dimension reduction to get isomorphic feature vectors.

SSCPM correlation preserving testing

SSCPM is based on canonical correlation between visual features and auditory features, while traditional feature analysis methods only cope with content features of single modality. Thus we compare the mapping results of our SSCPM method with the dimensionality reduction method of principal components analysis (PCA). We do not compare it with other feature analysis methods because PCA has been shown to be useful for feature transformation and

selection by finding the uncorrelated components of maximum variance.

Figs.3a and 3b show the scatter plots of the images that are projected to a two-dimensional subspace identified by the first two principal components and the first two SSCPM components. Circles correspond to the category of "bird" and triangles correspond to the other 4 categories (dog, explosion, tiger, zither).

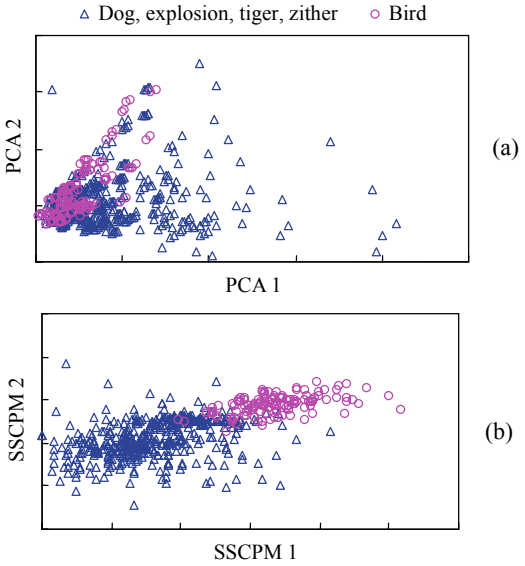

Fig.3 Scatter plots of the images. (a) Principal components analysis (PCA); (b) Semi-supervised correlation preserving mapping (SSCPM)

Compared with PCA in Fig.3a, SSCPM in Fig.3b can better separate data from different semantic categories. It can be concluded that our SSCPM algorithm preserves cross-media correlation during mapping process so as to learn latent semantic information.

Cross-media retrieval results

Fig.4 shows the statistical results of image-audio cross-media retrieval performance for overall evaluation. Fig.4a is the mean result of retrieving images by examples of audio. When the number of returned results is 35, the number of correct results is 24.15 at the second round of relevance feedback, while originally it is 12.95.

Fig.4b shows the mean result of retrieving audio by image examples. When the number of returned results is 40, the number of correct results is 22.51 at the second round of relevance feedback.

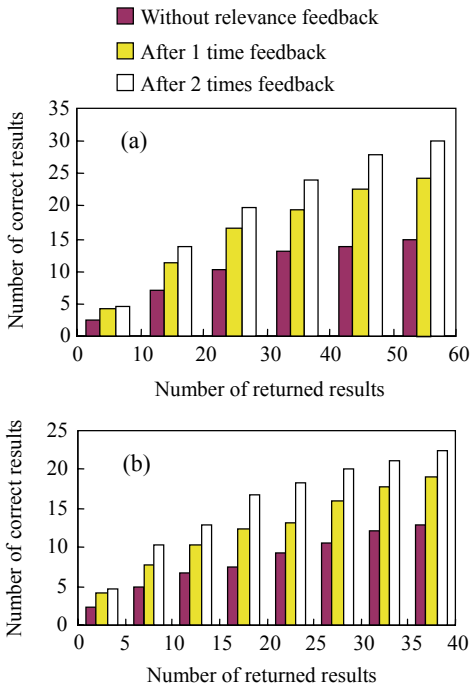

Fig.4 (a) Retrieval images by audio example; (b) Retrieval audio by image example

This observation confirms that: (1) image cluster and audio cluster both have good quality after subspace optimization, which is basic for relevance feedback algorithm to take effect; (2) updated cross-media correlation becomes more and more consistent with human perceptions as relevance feedback is incorporated. In the above experiments the "mean results" means: (1) a query is formulated by randomly selecting a sample media object from the dataset; (2) we generate 10 random image queries and 10 random audio queries for each category, and conduct 2 rounds of relevance feedback for dynamic correlation update.

Fig.5 shows an example of cross-media retrieval results. By submitting a 4.8-s audio clip of "bird", we get 13 images of "bird" in the top 15 returned image results. The number under each image in Fig.5 is the correlation value which represents how close this image is related to the query audio clip in semantics.

Selection and effect of parameters

As described in Section 4, parameters α, β are the weights of image similarity matrix and audio similarity matrix, respectively. In our experiments α, β are selected as empirical values: α=0.2, β=0.8. α is

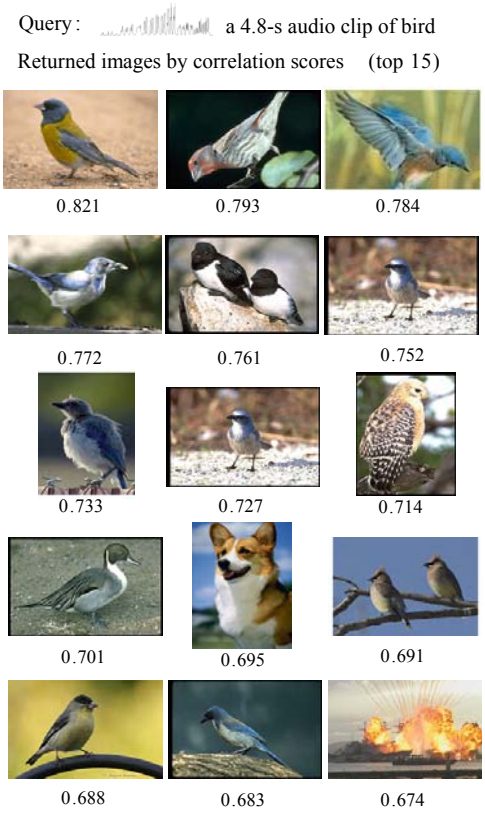

Fig.5 Returned images by submitting a bird audio

smaller than β because in SSCPM subspace audio dataset has a better cluster quality than image dataset does. The decay factor γ ensures that the propagated correlations are weaker than the original correlations. We select an empirical value γ=0.75.

Another important parameter is the dimensionality of SSCPM subspace, which is an open problem in feature analysis and dimension reduction. In general, it has to be large enough to keep most of the semantic correlation structures and small enough to remove some noise. We evaluate the effect of SSCPM dimensionality on the retrieval performance. Fig.6 shows the mean cross-media retrieval precision vs. the dimensionality of SSCPM subspace. In our experiments the precision value is defined in the same way as that used in content-based image retrieval (Wang *et al.*, 2004). And the precision listed in y-axis in Fig.6 is the mean result of query-image-by-audio and query-audio-by-image when the number of returned media objects is 25.

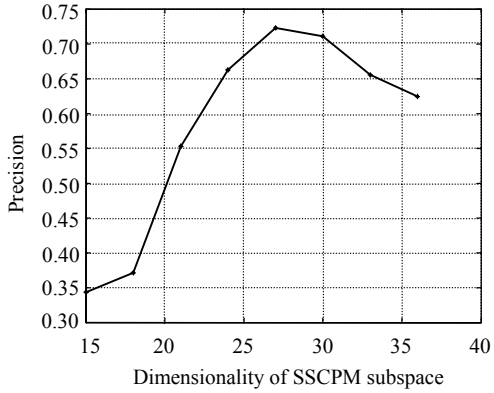

Fig.6 Mean precision under different dimensions

It can be seen that the retrieval performance highly depends on the dimensionalities of SSCPM subspace, and works best in our datasets SSCPM subspace whose dimensionality is between 26 and 30.

CONCLUSION

In this paper, we developed a discriminative learning approach to explore the underlying correlations between heterogeneous content features of image and audio data for cross-media retrieval. The approach provides a solution for the cross-media retrieval, which judges the semantic correlation between media objects of different modalities from low-level features. The encouraging results of experiments on image-audio dataset demonstrate that this is an effective approach for cross-media retrieval.

Since this approach is based on content feature vectors, it is applicable to other multi-modal analysis and correlation learning, which are frequently observed in recent research issues, such as talking face detection by audio signals and multi-modal retrieval of web pages, etc. The main limitation is that the size of image-audio database is comparatively small. Future work includes further study on large-scale image-audio retrieval and unsupervised subspace mapping.

References

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.I., 2003. Matching words and pictures. *J. Machine Learning Research*, **3**(6):1107-1135. [doi:10.1162/153244303322533214]

Chen, Z., Liu, W.Y., Zhang, F., Li, M.J., Zhang, H.J., 2001. Web mining for web image retrieval. *J. Amer. Soc. Inf. Sci. & Tech.*, **52**(10):831-839. [doi:10.1002/asi.1132.abs]

Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A., 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. Proc. 7th European Conf. on Computer Vision, p.97-112.

Hotelling, H., 1936. Relations between two sets of variables. *Biometrika*, **28**:321-377. [doi:10.1093/biomet/28.3-4.321]

Jeon, J., Lavrenko, V., Manmatha, R., 2003. Automatic Image Annotation and Retrieval using Cross-media Relevance Models. Proc. Int. ACM Conf. on Research and Development in Information Retrieval, p.119-126.

Lu, T.C., Chang, C.C., 2007. Color image retrieval technique based on color features and image bitmap. *Int. J. Inf. Processing and Management*, **43**(2):461-472. [doi:10.1016/j.ipm.2006.07.014]

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, **264**:746-748. [doi:10.1038/264746a0]

Slaney, M., Covell, M., 2000. FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. Proc. Neural Information Processing Systems, p.814-820.

Snoek, C., Worring, M., Smeulders, A.W.M., 2005. Early versus Late Fusion in Semantic Video Analysis. Proc. ACM Multimedia, p.399-402.

Tan, B., Shen, X.H., Zhai, C.X., 2006. Mining Long-term Search History to Improve Search Accuracy. Proc. Int. Conf. on Knowledge Discovery and Data Mining, p.718-723.

Wang, X.J., Ma, W.Y., Xue, G.R., Li, X., 2004. Multi-model Similarity Propagation and its Applications for Web Image Retrieval. Proc. ACM Multimedia, p.944-951.

Wu, F., Zhang, H., Zhuang, Y.T., 2006. Learning Semantic Correlation for Cross-media Retrieval. Proc. Int. Conf. on Image Processing, p.1465-1468. [doi:10.1109/ICIP.2006.312707]

Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S., 2003. Distance Metric Learning with Application to Clustering with Side-information. Proc. Neural Information Processing Systems, **15**:505-512.

Yang, J., Hauptmann, A., 2004. Naming Every Individual in News Video Monologues. Proc. ACM Multimedia, p.580-587.

Ye, J.P., Li, Q., 2005. A two-stage linear discriminant analysis via QR-decomposition. *IEEE Trans. on Pattern Anal. Machine Intell.*, **27**(6):929-941. [doi:10.1109/TPAMI.2005.110]

Zhang, C., Chen, T., 2002. An active learning framework for content-based information retrieval. *IEEE Trans. on Multimedia*, **4**(2):260-268. [doi:10.1109/TMM.2002.1017738]

Zhang, H., Zhuang, Y.T., Wu, F., 2007. Cross-modal Correlation Learning for Clustering on Image-Audio Dataset. Proc. ACM Multimedia, p.273-276.

Zhang, Z.Y., Liu, Z.C., Adler, D., Cohen, M.F., Hanson, E., Shan, Y., 2004. Robust and rapid generation of animated faces from video images: a model-based modeling approach. *Int. J. Computer Vision*, **58**(2):93-119. [doi:10.1023/B:VISI.0000015915.50080.85]

Zhao, R., Grosky, W.I., 2002. Negotiating the semantic gap: from feature maps to semantic landscapes. *Pattern Recognition*, **35**(3):593-600. [doi:10.1016/S0031-3203(01)00062-0]

Zhao, X.Y., Zhuang, Y.T., Wu, F., 2002. Audio Clip Retrieval with Fast Relevance Feedback based on Constrained Fuzzy Clustering and Stored Index Table. Proc. 3rd Pacific-Rim Conf. on Multimedia, p.237-244.