Bridging Music and Image via Cross-Modal Ranking Analysis

Xixuan Wu, Yu Qiao, *Senior Member, IEEE*, Xiaogang Wang, *Member, IEEE*, and Xiaoou Tang, *Fellow, IEEE*

***Abstract*—Human perceptions of music and image are closely related to each other, since both can inspire similar human sensations, such as emotion, motion, and power. This paper aims to explore whether and how music and image can be automatically matched by machines. The main contributions are three aspects. First, we construct a benchmark dataset composed of more than 45 000 music-image pairs. Human labelers are recruited to annotate whether these pairs are well-matched or not. The results show that they generally agree with each other on the matching degree of music-image pairs. Secondly, we investigate suitable semantic representations of music and image for this cross-modal matching task. In particular, we adopt lyrics as a middle-media to connect music and image, and design a set of lyric-based attributes for image representation. Thirdly, we propose *cross-modal ranking analysis* (CMRA) to learn the semantic similarity between music and image with ranking labeling information. CMRA aims to find the optimal embedding spaces for both music and image in the sense of maximizing the ordinal margin between music–image pairs. The proposed method is able to learn the non-linear relationship between music and image, and to integrate heterogeneous ranking data from different modalities into a unified space. Experimental results demonstrate that the proposed method outperforms state-of-the-art cross-modal methods in the music–image matching task, and achieves a consistency rate of 91.5% with human labelers.**

***Index Terms*—Cross-modal, feature embedding, lyric-based image attribute, music-image matching, ordinal regression.**

I. Introduction

THE amount of multimedia content people can access increase exponentially along with the rapid evolvement of Internet techniques and the popularization of digital audiovisual devices. Content receivers thus require efficient and effective tools to perceive and utilize huge multimedia contents. Music and image are two types of widely used media. However, the connection between them is not well explored so for. Effective matching techniques between music and image have various applications in cross-modal retrieval, music exploration [1], [2], and automatic music video (MV) generation [3], [4]. For example, music only may be tedious, but appears with image or video clips will bring more acousticvisual enjoyment. So that, modern shows in movie and TV always mix with music and images. Creating music videos by adding personal music and images has become popular applications.[1] However, the lack of professional domain skills will limit the choices of music and images for amateur users. Querying music via track name, artist, genre, or other musical attributes are the main functions music providers have. Users want to click once to deliver rich information to search for their wanted songs, such as taking a photo by mobile devices. Representing music with its album cover has inspired a lot of works [1], [2]. But customizing the cover for every single music still remains an problem since an album always contain more than one song. Music generation for photo show and video have been studied in [3], [4], where emotion and contextual sensor information are utilized to help connecting music and video. Given the variety of user needs, in this work, we concentrate on the matching of music and image, one of the multimedia cross-modal matching tasks.

Many psychology and cognition studies [5]–[8] indicate that brain information processing of vision and audio can be integrated together. Juslin and Västfjäll [5] argue that visual imagery is a mechanism through which music can call out emotion and "emotions experienced are the result of a close interaction between the music and the images." For example, people may react with positive emotions and imagine a beautiful nature scene at the same time, when listening to melodic movement as "upward." Osborne [6], Quittner and Glueckauf [7] suggest that music is effective to stimulate visual imagery. For persons with synaesthesia [8], auditory and visionary sensory can be connected automatically. In one type of synaesthesia, hearing sounds is in response to visual motion and flicker. Meyer [9] concludes that "it seems probable that ... image processes play a role of great importance in the musical affective experiences of many listeners."

To our best knowledge, there are few studies aiming to directly connect music (not audio) and image. This paper aims to deeply

Manuscript received January 11, 2015; revised August 17, 2015; accepted April 07, 2016. Date of publication April 21, 2016; date of current version June 15, 2016. This work was supported in part by SenseTime Group Limited, in part by the General Research Fund sponsored by the Research Grants Council of the Hong Kong SAR under Grant CUHK 416713, in part by the Sino-Dutch Joint Scientific Thematic Research Program under Grant JSTP172644KYSB20150019, in part by the Guangdong Innovative Research Program under Grant 2015B010129013 and Grant 2014B050505017, and in part by the Shenzhen Research Program under Grant KQCX2015033117354153. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan.

X. Wu and X. Tang are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: xixuan1124@gmail.com; xtang@ie.cuhk.edu.hk).

Y. Qiao is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China (e-mail: yu.qiao@siat.ac.cn).

X. Wang is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: xgwang@ee.cuhk.edu.hk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org provided by the authors. This includes descriptions of the image labeling tools, music and image features, detailed feature calculating processes, and music-image matching applications. This material is 2.64 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2557722

[1][Online]. Available: http://www.magisto.com/

explore whether and how music and image can be matched based on their contents. One of the challenges in this task is the lack of ground truth of well-matched music-image pairs. Extraction of music-image pairs is much harder than acquisition of image-text from accessible sources: image-text pairs can be easily obtained from online webpages, but grabbing music-image pairs from movie or TV shows is more difficult. In this paper, we mainly utilize music videos as the main source to extract music-image pairs for two reasons. Firstly, there exist large number of music videos with rich and diversified contents which can be used for analysis. Secondly, music videos are already well-organized, where audio and visual parts are matched by professional artists.

Since music usually has a long temporal duration with contents changing with time, we limit the scope of this paper to match music segments and single images. But this still remains challenging. First, music-image matching is a subjective task for which different people have different criteria. Even for human, it is difficult to quantize these subjective matching degrees. In this paper, we circumvent this problem by leveraging the side-information, i.e., ranking order of matching degree for music-image pairs. Previous studies [10]–[12] also show that it is more reliable for users to give relative comparisons, which can be regarded as a type of side-information, than quantitative scores for subjective tasks.

Different from previous methods for other cross-modal tasks such as image-text and text-audio matching [13]–[17], music and image are much more complex than text. One key problem in our task is to construct effective representation for music and image for cross-modal matching. There exist extensive studies on semantic analysis of music [18]–[21] and image [22]–[26]. Attribute-based representations prove effective in classification and retrieval for music and image. But they do not share the same categories of attributes: music is usually classified using album, artist, and genre, while image is generally recognized through color, texture, and object. Furthermore, many image attributes studied in previous work are not related to music, which may prevent them from using in matching music directly.

The main contributions of this paper are as follows. Firstly, we construct a music-image dataset consisted of about $45\,000$ pairs, and conduct human annotation on it to explore whether human agree with each other on sensing the cross-modal similarity between music and image. These annotation results indicate that human have consensus on evaluating degree of matching between music and image. Secondly, we investigate how to construct effective music and image representations for this cross-modal task. It is found that high-level abstraction of music and image can benefit music-image matching. Specially, we develop a set of lyric-based attributes for image representation which prove effective in our experiments. Finally, we propose a general kernel-based ranking framework to model and estimate the music-image matching degree. We formulate the music-image matching problem in their joint embedding space to maximize the ordinal margins of cross-modal pairs. Experimental results demonstrate that our framework can clearly improve the matching performance compared with other state of the art methods.

The rest of this paper is organized as follows. Section II briefly introduces the previous work related to our task. Section III presents human evaluation results for music-image pairs extracted from music video. Several semantic representations for music and image are discussed in Section IV. Section V presents the details of our cross-modal matching method. The experimental settings and results are shown and discussed in Section VI. This work is summarized in Section VII.

II. Related Work

A. Multi-Modal Learning

Many research efforts have been devoted into how to fuse multimodal information in multimedia analysis. This can be done by feature fusing [27] [28] and classification output fusing [29] [30]. Besides image-text fusion, audiovisual information fusion is another widely explored task. In [31], Dupont and Luettin integrate audio and visual information for speech recognition. Dunker *et al.* in [32] introduce a common mood space for music and image, and make use of audio-visual features to classify mood. An overall review about multi-modal interaction is given in [33]. Multiple kernel learning is also explored in multimodal analysis [11] to learn an unified data representation for heterogeneous data. However, different from these multi-modal researches which integrate or fuse information from individual modality, cross-modal tasks aim to analyze the intrinsic relationship or similarity between two modalities. Moreover, multimodal queries, composed of different modalities of features, are required for most of these multi-modal approaches.

B. Cross-Modal Analysis

Cross-modal researches receive much more interests in recent years partly due to wide applications in cross-modal multimedia analysis. One popular research topic is to jointly model image and text in multimedia documents. In [16], [17], Rasiwasia *et al.* use high-level semantic vectors to represent images, and learn a common subspace to maximize the correlation between images and texts by canonical correlation analysis (CCA) [34] and Kernel CCA [35]. Positive and negative correlations among different modalities provide also feasible cues for estimating the cross-modal similarities. The Wikipedia concepts and their negative links is utilized in [36] to improve the cross-modal image-text retrieval performance. Maximum covariance unfolding (MCU) [37] utilizes relationships among neighborhood in single modalities to reduce the cross-modal data dimensions. Jia *et al.* [38] encode the correlation between image and text by sharing the topics learned with Markov random field of topic models.

Cross-modal acousticvisual researches are limited to some particular types of audio and video. Kidron *et al.* [39] localize visual events associated with sound sources based on CCA. The typical spatial sparsity of audio-visual events is exploited to remove the inherent ill-posedness. In [40], Alameda-Pineda *et al.* explore the geometric and physical properties of an audio-visual sensor. The sounds such as clothe chafing and footsteps emitted by people are used for human identification under the view of camera. CCA also benefits cross-modal and multi-modal retrieval for video and audio [41], in which the queries can be

single-modal (image) or multi-modal (combination of image, audio, and location). Perhaps the closest work to our problem is [14], proposed by Zhang *et al.*, which investigates the cross-modal relationship between an animal's image and its corresponding sounds.

This study focuses on the matching between music and image, a particular cross-modal matching task. Various acoustic and visual features can be applied to this task, but improper ones may limit the matching performance. Therefore, one of our goals is to identify the affective representations of music and image in music-image matching task. Moreover, we take account of ranking information available in annotation to model the non-linear relationship among heterogeneous feature spaces. Overall, this work aims to yield a thorough study on modeling the cross-modal music-image relationship by utilizing effective semantic representations of music and image, and ranking orders of labeled pairs.

III. Dataset and Annotation

As we mentioned previously, one main challenge in music-image matching is the lack of data with annotations. In this section, we describe the details of the construction of our music-image dataset and the human annotation process.

A. Dataset

Remind that we choose music video, not movie or TV shows, as our data source for two main reasons. Firstly, the visual part of MV is created elaborately by conductors and singers, who are professional artist and have great understanding about the song. Based on this observation, we can assume that the images in MV are generally well-matched with the song/music. Secondly, music video is much easier to collect in contrast with film and documentary. For films we need to detect when the background music starts and ends, and cut it out properly, which cannot be finished automatically by computer. But the whole process of extracting music-image pairs from music video could be done by automatic algorithms.

We searched about $1,500$ MVs with various genres and styles from Internet. MVs usually last for several minutes, in which both music and video parts contain variations. A still image cannot be well-matched to every segment of the song. To simplify the problem, we conduct segmentation for each MV first. The audio (music) part is extracted and divided into several small music segments by dynamic texture model [42]. It is found through experiments that a short music segment is able to deliver enough information for matching image. For each music segment, we extract its corresponding video segment with the same temporal duration, and select one or two key frames from this video segment. An example of this process is shown in Fig. 1. During this process, we can obtain about 20 pairs of music segment and image in average for each MV. In the remainder of this paper, "music" is used to represent the "audio segment" extracted from music videos without special notification. In total, we collect a dataset of $47\,888$ music-image pairs.

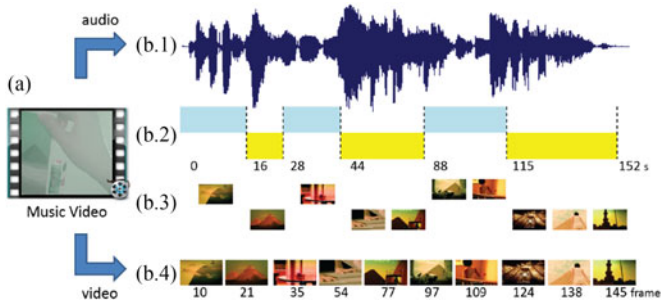


Fig. 1. Example of constructing music-image pairs. The audio part (b.1) and frames (b.3) are first extracted from a music video (a). Then we divide the audio part by its rhythm information (b.2). Finally, music-image pairs, whose music segments and frames are matched by time point, are collected (b.4).

B. Human Annotation

Although our music-image pairs are extracted from music videos, which are elaborately created by experts (singers and conductors), there still remains a question: whether people agree with these experts on the matching degree of these music-image pairs. It is noted that there do not exist strict and clear criterion for estimating the matching degree between music and image. Different persons may have different opinions. However, we argue that human own consensus on evaluating most of the music-image pairs.[2] We recruit human labelers to evaluate the matching degree of the music-image pairs. Our objectives are two folds. Firstly, we examine whether people have consensus on this correlation, i.e., whether people agree with each other on the matching degree of music-image pairs. Secondly, the music-image pairs labeled by human yield training and testing data for us to develop automatic similarity estimation algorithms.

In practice, we asked six human labelers to compare the matching degree between music and images. Three of them have art background, while the others do not. An online system is developed for labeling. The details are given in supplementary material. The system presents one music segment and two images each time. Labelers are required to select the image that matches the music better. Among the two presenting images, one comes from the music-image pairs obtained from music video, another is randomly chosen from the whole images in our dataset. All music videos are new to the labelers. The labeling music list is randomly generated for every labeler in order to avoid the effect of contextual pairs. Each labeler annotated the same numbers of music-image pairs independently without discussions. Their opinions are recorded by back-end database for further analysis. (Features of the music-image dataset and labeling results are available in the project page.[3] So is supplementary material.)

[2]This is in essence similar to how human perceive the aesthetics of photos. Although people have personal preference, they generally agree with each other on some photos, which meet the requirement of aesthetics.

[3][Online]. Available: http://mmlab.siat.ac.cn/musicimage_matching/index.html

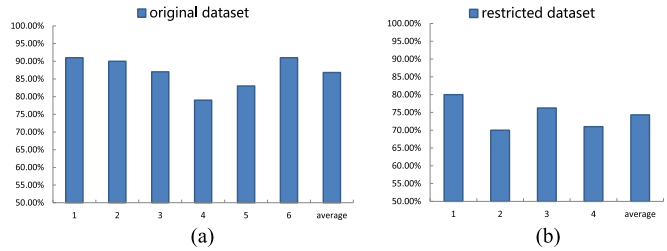


Fig. 2. (a) Accuracy of each labelers on our dataset. X axis represents the labeler ID. The first three labelers have art background, while the others not. (b) Accuracy of each labelers on our restricted dataset. X axis denotes the labeler ID. The first one labeler is a musical student, and the others are not.

TABLE I
CONSISTENCY OF LABELER'S CHOICES

$K/6$	N	Total pair #	Consistency rates
$6/6$	$17,257$	$23,922$	72.1%
$5/6$	$21,340$	$23,922$	89.2%
$4/6$	$22,827$	$23,922$	95.4%

C. Analysis and Annotation Results

Totally, we have 22 632 music segments with 47 888 music-image pairs. Some music segments are associated with more than one frame, since the visual content of its video part vary a lot. About half of them (23 922 pairs) are labeled by human labelers. Firstly, we examine the annotations of each labelers, to investigate whether labelers' backgrounds affect their decisions. The accuracy is evaluated by for what percentage of pairs a labeler select image from original MV. The experimental results are depicted in Fig. 2(a), which show that all labelers with or without art background, prefer the music-image pairs extracted from music video to the random ones. The labelers choose MV music-image pairs rather than random ones with an average percentage of 86.78%. To further study the effect of labelers' background knowledge for music-image labeling, we conduct a more strict experiment in which the source of two matched images is limited. They are both chosen from western songs and singers of the same gender, which implies that people cannot identify image through background knowledge. We ask 4 students to do the experiment. One is a musical student, and the others are not. The accuracy results are shown in Fig. 2(b), which indicate that background information only has a small effect on labeling music-image pairs.

The consensus among the labelers on the preference to music-image pairs is also explored. We count the number of pairs which are agreed by 4, 5, and 6 labelers, and calculate the consistency rates. The results are summarized in Table I, where K is the required minimum number of labelers who reach agreement on a pair, and N is the number of pairs from MVs which are preferred by at least K labelers. It is proved that all the labelers highly agree on the majority of pairs from MVs. Near 90% pairs are labeled consistently by at least five labelers.

We conclude several observations from the above results. First, all labelers exhibit a strong ability to perceive the relationship between music and image extracted from music video, whether they have art background or not. The labelers prefer to choose the MV pairs with much higher probability than the random pairs, even without any extra song or lyric information. Secondly, all the labelers show consensus on the preference to MV music-image pairs. We count the pairs which are chosen correctly by $4, 5, 6$ labelers, and calculate the consistency rates (Table I). The experimental results show that people highly agree on the majority of pairs. Near 90% pairs are labeled correctly and consistently by at least five labelers. Moreover, a many-to-many matching relationship may exist in music-image matching. In some cases, both candidate images are chosen as a matching for the given music segment. Apparently, a music segment can be matched to more than one image, and vice versa.

IV. SEMANTIC REPRESENTATIONS OF MUSIC AND IMAGE

It is essential to effectively represent music and image in cross-modal music-image matching. In this section, we discuss how to extract suitable semantic representations for both modalities.

A. Music Semantic Representation

We construct music semantic representation from two types of low-level music features. The first is composed of the Echo Nest Song (ENS) features [43], including basic features like tempo and mode, as well as some modified features like mean section length. Another is Mel-Frequency Cepstral Coefficient (MFCC) vector, which a widely used audio feature due to its good performance in many speech and music processing tasks [44]. In practice, following [44] we extend MFCC feature vector by adding MFCC-Delta and MFCC-Delta Delta vectors. The total dimension of our MFCC feature is 39 for a frame. By sliding window along the audio file with half overlapping, we obtain a long sequence of MFCC vectors (about 5200 frames per minute). Following [21], we randomly sample $10\,000$ frames to reduce the computation time for each music segment.

Based on the two low-level feature types mentioned above, a semantic representation of music is further extracted. Each dimension of the semantic features corresponds to a descriptive word. We firstly select a pre-defined word list $\mathrm{W} = \{\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_H\}$ to annotate songs. Each word w_i represents a music semantic concept such as "happy," "warm," "school," "classic," and "guitar." For each song in our database, we obtain its associated words by collecting such semantic concepts from common music sites, such as Google Music and Yahoo music. Finally, we adopt 101 semantic words, which are categorized into nine types: Emotion, Instrument, Rhythm, Usage, Style, Genre, Region, Time, and Special Class. Based on ENS and MFCC feature, we apply hierarchal GMM [21] to calculate a posterior probability distribution over our musically-relevant tag list as the music semantic features. Details are available in supplementary material. Finally, we use these posterior probabilities to construct the semantic representation for music $\mathcal{M}$: $x_{\mathcal{M}} = [p(\mathrm{w}_1|\mathcal{M}), p(\mathrm{w}_2|\mathcal{M}), ..., p(\mathrm{w}_H|\mathcal{M})]$. Each $p(w_i|\mathcal{M})$ denotes the posterior probability of word w_i. $x_{\mathcal{M}}$ can

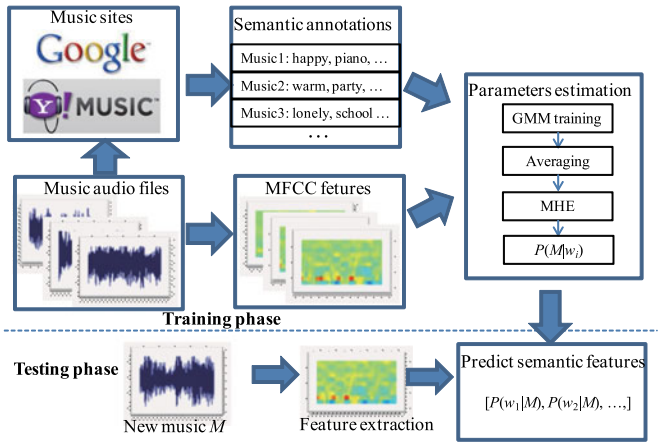


Fig. 3. Diagram of extracting semantic features from music.

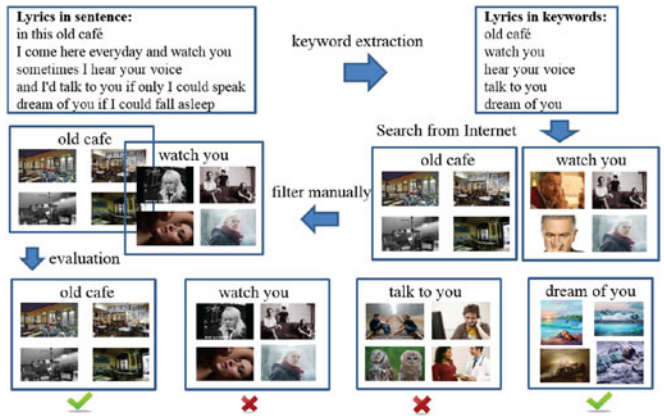


Fig. 4. Process of semantic abstraction based on lyrics.

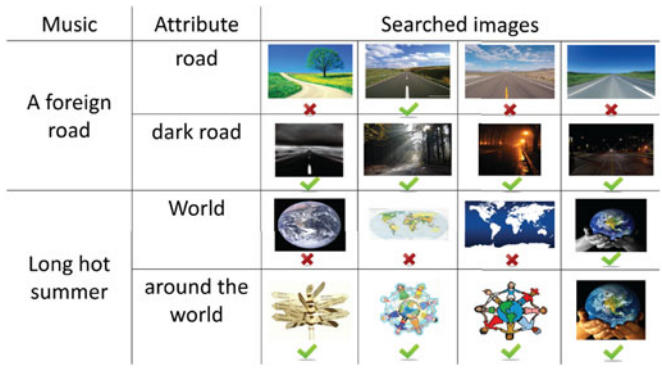


Fig. 5. Comparison of searched images between single-word attributes and lyric-based attributes in music-image matching task. As can be seen, single word attributes may be too general to describe images for cross-media matching, such as "world."

be introduced as an attribute vector. The process of extracting music semantic features is depicted in Fig. 3.

B. Image Semantic Representation

Like music, several low-level features are used to estimate image semantic representation. These features can be categorized into three categories: 1) color features, including attention guided color signature [45] and color spatialet [46], 2) texture and appearance features, which contain Daubechies wavelet [46] and gist features [47], and 3) shape features, including Histogram of Gradient (HoG) [48] and Multi-layer rotation invariant edge orientation histogram (MLR-EOH) [49]. Details about these features could be found in supplementary material.

Attribute-based image representation is widely used due to its ability to describe high-level semantic information. Various image attributes have been proposed to describe objects [50], scenes [24], faces [51], and actions [52]. However, we argue that for cross-modal music-image matching, such attributes are not always effective, since many of them fails to contain useful cues for matching music. In order to select appropriate image attributes, we hypothesize that emotional or functional image attributes should be more suitable for cross-modal music-image matching. The expected matching images should be able to describe certain aspects of music. To test our hypothesis, we use lyric as a source for constructing musical descriptive image attributes, since it is usually highly correlated with the audio part of music compared with scene attributes [24], which is expected to be the most likely type of image attributes for describing music content compared with other object, face, and action attributes. The whole attribute extraction process is shown in Fig. 4. It includes two phases: deciding image attributes, and training statistical predictable models for them.

1) Attribute Vocabulary: In practice, for constructing musical descriptive attributes, we collect 300 songs as our data source to extract lyrics. Most lyric lines are sentences, which make them too long to be attributes. Stanford Parser [53] is utilized to extract adjectives, nouns, and phrases from all the lyric sentences, and four labelers are then asked to refine these phrases. For each lyric sentence, we only reserve one attribute. Phrases, emotional adjectives, and scenery words are preferred more as lyric attributes for music, and useless verbs and pronouns are removed. For example, as shown in Fig. 5, we try to identify images that are matched to two songs. One song is sad with low-tempo, whose lyrics contain "dark road," and another is warm with lyrics "around the world." We can see that images searched by "dark road" show a clear emotional sign for matching this low-tempo and sad music, while images searched by "road" are ambiguous in emotion expression. The images searched by "world" are very concrete, which prevent themselves from matching the warm song. So the emotional or functional words or phrases are mostly extracted from this process. The detailed guidance for attribute extraction is described in supplementary material.

Having extracted the attributes, we use them as keywords to search online images. The searched images will be used as training examples to learn attribute predictors. However, current image search engines such as Google and Bing may return noisy and unrelated images. It is also possible that the searched images are not suitable for the corresponding music. So we start with 4500 phrases (attributes), for each of which we retrieve 120 images using Google Image. Three labelers are asked to remove the unqualified images. A tool is developed for this task, details of which are available in supplementary material. We find that about 50 percent of images and 30 percent of attributes

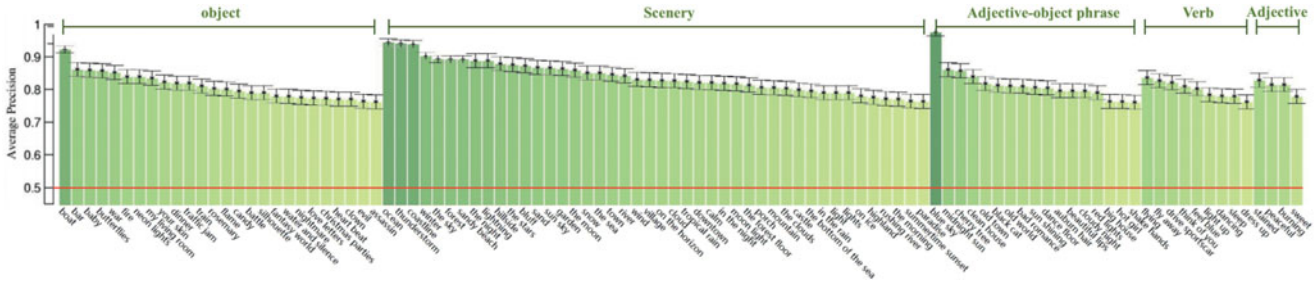


Fig. 6. Average precision (AP) for attributes in our lyric-based attribute database. The AP of chance selection is marked by the red line.

are removed during this process. We also conduct music-image matching experiments on the extracted attributes without manually cleaning. The performance is poorer compared with the filtered ones.

2) Estimation and Evaluation of Attribute: Once we have obtained a set of attributes and their associated images, we test the quality of these attributes by their associated images. Specifically, we have about 110 images for each attribute, of which we use 80 images for training and the other 30 for testing. Negative training samples are selected from images of other attributes. We extract a set of low-level features from each image, including colorSig, colorCell, HOG, MLR-EOH, wavelet, and GIST. With these features as input, we train a one-vs-rest SVM classifier for each attribute. Both the training and testing set are equally divided into positive and negative samples. We conduct five-fold cross-validation for each attribute, and calculate average precision as a measure of the reliability of SVM predictor.

Fig. 6 shows the SVM prediction results of 100 attributes, which are randomly selected from those whose average precision higher than 75 percent. These attributes can be roughly categorized into five categories: object, scenery, adjective-object phrase, verb and verb-object phrase, and adjective. As Fig. 6 shows, it is easier to recognize object and scenery attributes than others. Moreover, compared with traditional object and scenery attributes, lyric attributes are more easy to imagine their emotional signs, like "war," "candy," and "in the rain." Finally, we adopt 800 attributes with the highest average precisions to construct a semantic representation for image. Following [17], a sigmoidal transformation of the multi-class SVM scores are taken as their posterior class probabilities, which form the image semantic feature vectors.

V. Cross-Modal Matching Model

In this section, we propose a cross-modal ranking framework which connect cross-modal data samples with extra ranking information into a shared space, in which the cross-modal correlation can be maximized and estimated. Large amount of pairwise ranking annotations obtained from labeling process are exploited in the training phase of our framework. In the next, we first briefly introduce linear transformation (CCA), then develop ranking CCA for utilizing pairwise ranking information, and finally generalize it to nonlinear case due to the complexity of music-image matching. The music and image space is denoted as as $\mathcal{X}$ and $\mathcal{Y}$, respectively. The whole framework is shown in Fig. 7, which aims to estimate the similarity function $\mathcal{S}(x, y)$ between a music segment x and a single image y.

A. Canonical Correlation Analysis

Canonical Correlation Analysis [34] has been widely used for cross-modal analysis. Here we briefly introduce it for completeness. We denote the embedding functions $g_{\mathcal{X}} : \mathcal{X} \rightarrow \mathcal{Z}$ and $g_{\mathcal{Y}} : \mathcal{Y} \rightarrow \mathcal{Z}$ for music and image feature spaces respectively, where $\mathcal{Z}$ represents the shared space of music and image

$$g_X(x) = Ax \tag{1}$$

$$g_Y(y) = By \tag{2}$$

where A and B are transformation matrices for music feature x and image feature y. Let $X = [x_1; x_2; \ldots; x_n]$ and $Y = [y_1; y_2; \ldots; y_n]$, where x_i and y_i represent the feature vectors of music and image in the ith training pair respectively, and n is the number of training pairs. CCA aims to find vectors a, b such that the correlation of data projections $\langle a^\top X, b^\top Y \rangle$ is maximized. With transformation matrices $A = [a_1; a_2; \ldots; a_J]$ and $B = [b_1; b_2; \ldots; b_J]$. Mathematically the objective function is

$$\max_{A,B} \sum_{j=1}^{J} \sum_{i=1}^{n} a_j x_i b_j y_i. \tag{3}$$

Then, by obtaining the optimal A and B (the optimization details for CCA and kernel CCA are shown in supplementary material), the similarity function of CCA is defined as

$$\mathcal{S}_{\mathrm{CCA}}(x, y) = \langle Ax, By \rangle. \tag{4}$$

B. Marginal Ranking CCA by Quadratic Programming

Then we try to add the pairwise ranking information yielded from manual annotations III-B, which is that one pair has a higher similarity score than another. The key idea is to make the two embedding functions g_X and g_Y map the heterogeneous features to a shared space, and also take advantage of the pairwise ranking data among the training samples. To do so, we introduce a transformation matrix M into the embedding functions g_X and g_Y as

$$g_X(x) = MAx \tag{5}$$

$$g_Y(y) = MBy. \tag{6}$$

where A and B are obtained by CCA (3). Please note that pairwise ranking information is not encoded in M. Here M is

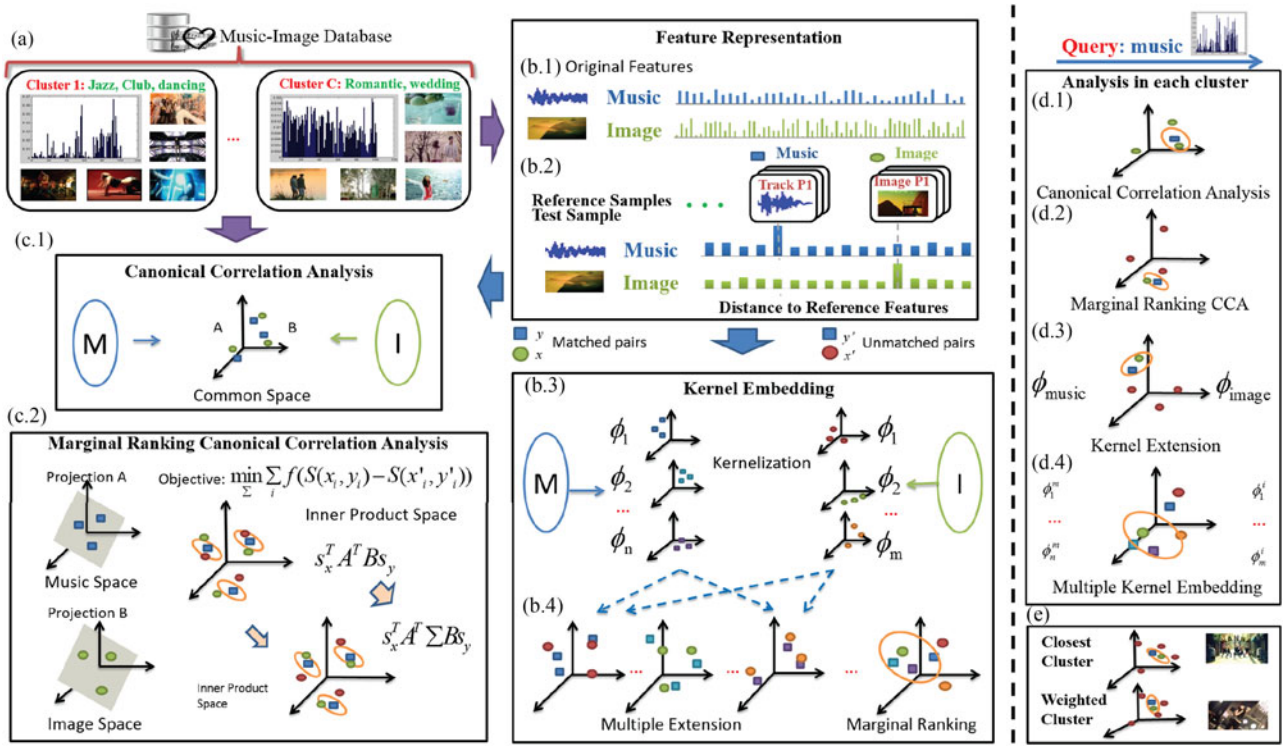


Fig. 7. Framework of our cross-modal ranking analysis. (a) Music-image pairs are clustered by music representation. Methods (c.1)-c–4 are trained in each cluster. (b) The original feature and similarity-based feature (Section VI-A) are extracted for each music and image. (c.1) The illustration for CCA. The blue square and green ellipse represent music and image object, respectively (similarly hereinafter). (c.2) The illustration for MR-CCA. The red ellipse represents the ill-matched image for the music in a given music-image pair (similarly hereinafter). (c.3) c–4. The illustration for KR-CCA and MKR-CCA. The music and image space are represented by several kernel spaces, which are jointly combined together. d. Similarity calculation for a new music-image pair using CCA, MR-CCA, KR-CCA, and MKR-CCA in a single cluster. (e) Combine the similarity scores in each cluster as the final similarity for a new pair.

regarded as a key matrix in estimating pairwise similarity. Our goal is to find appropriate M which satisfies as many pairwise ranking constraints as possible. Then the inner product matrices of the embedding samples are characterized as

$$I_x = x^\top A^\top M^\top M A x \tag{7}$$

$$I_y = y^\top B^\top M^\top M B y. \tag{8}$$

The similarity function for marginal ranking CCA is then defined as

$$\mathcal{S}_{R-\mathrm{CCA}}(x,y) = x^\top A^\top \Sigma B y \tag{9}$$

where $\Sigma = M^\top M$.

Side-information is denoted as a set of similarity constraints $\mathcal{C} = \{(x_i, y_j, x_k, y_l)\}$, in which $\mathcal{S}(x_i, y_j) > \mathcal{S}(x_k, y_l)$, our optimization objective function becomes to

$$\min_{A,B,\Sigma} \sum_{\mathcal{C}} h(\mathcal{S}_{R-\mathrm{CCA}}(x_i,y_j) - \mathcal{S}_{R-\mathrm{CCA}}(x_k,y_l)) \tag{10}$$

where h is a hinge penalty function. If $t \le 0$, $h(t) = -t$; otherwise $h(t) = 0$.

We assume that Σ is a diagonal matrix, since the projections obtained by CCA are uncorrelated. Let $W = [w_1, w_2, \ldots, w_J]$ denote the diagonal of Σ. Introduce variables, $z_i^j = a_j x_i b_j y_i$, $z_i = [z_i^1, z_i^2, \ldots, z_i^J]$, $z_i^{j\prime} = a_j x_i' b_j y_i'$, and $z_i' = [z_i^{1\prime}, z_i^{2\prime}, \ldots, z_i^{J\prime}]$. Then (9) can be written into

$$\mathcal{S}_{R-\mathrm{CCA}}(x_i,y_i) = \sum_j w_j a_j x_i b_j y_i = W^\top z_i. \tag{11}$$

Then our objective reduces to optimize W with $\min_W \sum_i f(W^\top z_i - W^\top z_i')$ (10). This is in spirit the same as the optimization of ordinal SVM [54], whose due problem is defined as

$$\min_W \|W\|^2 + \sum \xi_i \tag{12}$$

subject to

$$\xi_i \ge 0,\ W^T z_i - W^T z_i' \ge 1 - \xi_i. \tag{13}$$

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

C. Marginal Ranking CCA by Semidefinite Programming

Following (9), if Σ is not diagonal, we need to adjust the form of the objective function. With the embedding notations, we can formulate the side-information as the constraints of embedding functions g_X and g_Y as

$$\|g_X(x_i) - g_Y(y_j)\|^2 + 1 < \|g_X(x_k) - g_Y(y_l)\|^2 \quad \forall (x_i, y_j, x_k, y_l) \in \mathcal{C}. \tag{14}$$

Please remind that the similarity function $\mathcal{S}_{R-\mathrm{CCA}}(x,y)$ in (9) is just inversely proportional to the distance $\|g_{\mathcal{X}}(x) - g_{\mathcal{Y}}(y)\|^2$. With (5) and (7), the constraint in (14) can be reduced to

$$(Ax_i - By_j)^\top M^\top M(Ax_i - By_j) + 1 < (Ax_k - By_l)^\top M^\top M(Ax_k - By_l). \quad (15)$$

By introducing a slack variable ξ_{ijkl} for each constraint in $\mathcal{C}$, and minimizing the summation of empirical hinge loss over constraint violations $\frac{1}{|\mathcal{C}|}\sum_{\mathcal{C}}\xi_{\mathrm{ijkl}}$. We can get the following optimization problem:

$$\min_{M,\xi\geq 0}\ \mathrm{tr}(M^\top M) + \frac{\gamma}{|\mathcal{C}|}\sum_{\mathcal{C}}\xi_{\mathrm{ijkl}},\ \text{s.t. } (Ax_i - By_j)^\top \times M^\top M(Ax_i - By_j) + 1 \leq (Ax_k - By_l)^\top M^\top \times M(Ax_k - By_l) + \xi_{\mathrm{ijkl}} \quad \forall (x_i, y_j, x_k, y_l) \in \mathcal{C} \quad (16)$$

where γ is a coefficient to control the balance between loss function and regularization. To avoid overfitting, we introduce a regularization term $\mathrm{tr}(M^T M)$. Since M only appears in form of $M^\top M$, we can replace it with semidefinite matrix $H = M^\top M$. Then optimizing (16) becomes a semidefinite programming (SDP) problem where both objectives and constraints are linear in H.

D. Kernel Marginal Ranking CCA

Furthermore, one problem of the above formulation is that the relation between music and image is complex and cannot be well described by linear embedding functions. In this section, we introduce nonlinear kernel analysis during the utilization of side-information. Following [55], we first map music x to reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_X$ via a nonlinear feature mapping $\phi_X(x)$. Similarly, we can map image y to $\mathcal{H}_Y$ via mapping $\phi_Y(y)$. Then we consider linear transformation for $\phi_X(x)$ and $\phi_Y(y)$. The new nonlinear embedding functions are defined as

$$g_X(x) = NU(x),\ g_Y(y) = NV(y) \quad (17)$$

where $U(x)$ and $V(y)$ are linear projections of $\phi_X(x)$ and $\phi_Y(y)$ respectively, and N is an embedding matrix like M for (5). The similarity function of kernel marginal ranking CCA based on the new definition of g_X and g_Y is

$$\mathcal{S}_{\mathrm{KR-CCA}}(x,y) = U(x)^\top N^\top N V(y). \quad (18)$$

Since the dimensionality of the feature space is much larger than the number of samples in the dataset, the linear projection of $\phi_X(x_i)$ and $\phi_Y(y_i)$ must lie in the space spanned by samples $\{\phi_X(x_i)\}$ and $\{\phi_Y(y_i)\}$

$$a = \sum_i \alpha_i \phi_X(x_i),\ b = \sum_i \beta_i \phi_Y(y_i). \quad (19)$$

Therefore, let $\mathbf{X} = [\phi_X(x_1), \phi_X(x_2), \ldots, \phi_X(x_n)]$ and $\mathbf{Y} = [\phi_Y(y_1), \phi_Y(y_2), \ldots, \phi_Y(y_n)]$, where n is the number of samples. The linear transformation of $\mathbf{X}$ and $\mathbf{Y}$ are

$$u(\mathbf{X}) = \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_n\end{bmatrix}\mathbf{X}\mathbf{X}^\top \quad (20)$$

$$v(\mathbf{Y}) = \begin{bmatrix}\beta_1 & \beta_2 & \cdots & \beta_n\end{bmatrix}\mathbf{Y}\mathbf{Y}^\top. \quad (21)$$

Kernel CCA [56] can be used to find the best $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top$ and $\beta = [\beta_1, \beta_2, \ldots, \beta_n]^\top$ to make u and v most correlated. Let $K_x = \mathbf{X}\mathbf{X}^\top$, $K_y = \mathbf{Y}\mathbf{Y}^\top$, the objective of kernel CCA is

$$\max_{\alpha,\beta}\ \alpha^\top K_x K_y \beta \quad (22)$$

$$\text{s.t. } \alpha^\top K_x K_x \alpha = 1, \beta^\top K_y K_y \beta = 1. \quad (23)$$

By obtaining a set of α and β (a and b) after the optimization of kernel CCA, which are named as $\alpha^1, \alpha^2, \ldots, \alpha^m$ and $\beta^1, \beta^2, \ldots, \beta^m$, we then get the new data representation $U(X)$ for X (The same definition as $V(Y)$) as

$$U(X) = \begin{bmatrix} \sum \alpha_i^1 k(x_1, x_i) & \cdots & \sum \alpha_i^1 k(x_n, x_i) \\ \sum \alpha_i^2 k(x_1, x_i) & \cdots & \sum \alpha_i^2 k(x_n, x_i) \\ & \vdots & \\ \sum \alpha_i^m k(x_1, x_i) & \cdots & \sum \alpha_i^m k(x_n, x_i) \end{bmatrix}. \quad (24)$$

With $U(X)$ and $V(Y)$, we can define a distance function with matrix N for music x_i and image y_j as

$$D(i,j) = (U(x_i) - V(y_j))^\top N^\top N (U(x_i) - V(y_j)) = (U_i - V_j)^\top N^\top N (U_i - V_j) \quad (25)$$

where U_i is the ith column of U, V_j is the jth column of V, and N is an embedding matrix to be optimized. The optimization problem is

$$\min_{N,\xi\geq 0}\ \mathrm{tr}(N^\top N) + \frac{\gamma}{|\mathcal{C}|}\sum_{\mathcal{C}}\xi_{\mathrm{ijkl}},\ \text{s.t. } D(i,j) + 1 \leq D(k,l) + \xi_{\mathrm{ijkl}} \quad \forall (x_i, y_j, x_k, y_l) \in \mathcal{C}. \quad (26)$$

The above formulation has a similar form as that of (16) and can also be solved by semidefinite programming.

E. Multiple Kernel Marginal Ranking CCA

In the previous section, we mainly focus on learning the matching relationships between music and image based on single kernel space. However, multimedia data is heterogeneous in nature [11] (described by multiple types of features). It may not be realistic to use a single feature representation to capture the complex structure among data completely. For example, it may be natural to encode genre information for music by spectrum features, and emotion information by tempo features. How to combine the two sources of information into a single kernel space remains a problem. Perhaps the simplest way to use multiple category of features is to concatenate them into a single feature vector. However, this approach ignores the difference among different categories. A better method is to design kernels for different categories and combine them with multiple kernel learning [57]. Assume that domain $\mathcal{X}$ contains n_x^k kernels, and domain $\mathcal{Y}$ contains n_y^k kernels. The simplest combina-

tion is to sum all the kernels $K_x = \sum_{p=1}^{n_x^k} K_x^p, K_y = \sum_{p=1}^{n_y^k} K_y^p$, which cannot distinguish the importance (weight) of different kernels. Another solution is to take a positive-weighted summation $K_x = \sum_{p=1}^{n_x^k} \mu_p K_x^p, K_y = \sum_{p=1}^{n_y^k} \mu_p K_y^p$. However, kernel-based embedding must estimate $K_x K_y$. This multiplication makes it difficult to optimize these weights.

Here, we adopt the joint combination of kernels inspired by [11]. For kernels K_x^s and K_y^t, we can obtain their corresponding $U(s)$ and $V(t)$ by (24). Then the distance between two modalities can be defined as

$$D(i,j) = \sum_{s=1}^{n_x^k} \sum_{t=1}^{n_y^k} (U(s)_i - V(t)_j)^\top N_{s,t}^\top N_{s,t} (U(s)_i - V(t)_j). \tag{27}$$

Then the question is how to optimize $N_{s,t}$. Let $W_{s,t} = N_{s,t}^\top N_{s,t}$, we adopt sub-gradient descent on optimizing $W_{s,t}$. The objective can be rewritten as

$$f_D(W) = \text{tr}(W) + \frac{\gamma}{|\mathcal{C}|} \sum_{\mathcal{C}} h(1 + D(x_i, y_j) - D(x_k, y_l)). \tag{28}$$

We can optimize (28) using an iterative gradient decent algorithm [58]. The gradient of $f_D(W)$ over W can be calculated as

In each iteration, we find out the triples that violate (14), and then use them to calculate the gradient by (29) as shown at the bottom of the page. We restrict $W_{s,t}$ to be diagonal in order to not only simplify the problem to linear programming, but also to yield interpretations of weights of each kernel in the embedding. At each iteration, we update $W_{ii} \to \max(0, W_{ii})$ to ensure positive semidefinite constraint on W.

F. Clustering-Based Similarity Function

Recall that we have 47 888 music-image pairs in our datset. We can apply the methods described in Section V-A–V-E in the whole dataset. However, we are also interested in utilizing the divide-and-conquer technique in our dataset: we cluster music-image pairs according to their inherent structures first, then apply these methods in each cluster of pairs, and finally combine the matching results of each cluster together. In practise, we cluster the music-image pairs by music part since music space has less diversity and exhibits simpler cluster structure. Normalized cut [59] is used to cluster the music-image pairs by their music semantic features. The similarity functions described in Section V-A–V-E are applied to each cluster.

Finally, for a testing pair (x, y), we 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 its nearest cluster. Then

$$\mathcal{S}(x,y) = \mathcal{S}_{c*}(x,y) \tag{30}$$

where $\mathcal{S}(x,y)$ represents the similarity between a single music segment and an image, which can be calculated by CCA (4), R-CCA (9), KR-CCA (18), or MKR-CCA (Section V-E).

(30) is easy to calculate, but it cannot deal well with the samples near the cluster boundary. To overcome this problem, we use softmax function as weights to construct a "soft" similarity function. Let d_c denote the distance between pair (x, y) and the center of cth cluster. The weighted similarity function is defined as

$$\mathcal{S}(x,y) = \sum_{c=1}^{C} \frac{\exp(-d_c/\sigma^2)}{\sum_{j=1}^{C} \exp(-d_j/\sigma^2)} \mathcal{S}_c(x,y) \tag{31}$$

where C is the number of clusters, and σ is a normalization parameter.

VI. Experiment

In this section, we report the experimental evaluation on different feature spaces of music and image, and our proposed cross-modal ranking analysis methods. We start by describing our choice of feature space, evaluation procedure and indicator. This is followed by the comparison of different methods described in Section V-A–V-E, and the parameter settings for our framework.

A. Feature Space

We adopt two types of feature spaces in the experiment:

1) Original Feature Space (OF): It refers to the low-level and high-level features (attribute-based semantic features) which are extracted directly from the music and image, respectively.

2) Similarity-Based Feature Space (SF): A data driven method is used for calculating similarity-based features. We take a set of training pairs (o_x^i, o_y^i) as reference examples. For an instance o_x from $\mathcal{X}$, we calculate its distance to every reference example o_x^i in the training set as $d(o_x, o_x^i)$. The similarity-based feature representation for o_x is denoted as $s_x = [s(o_x, o_x^1), s(o_x, o_x^2), \ldots, s(o_x, o_x^r)]$, where r is the number of the reference examples, and the similarity function s is defined as

$$s(o_x, o_x^i) = \exp\left(-\frac{d(o_x, o_x^i)}{\sigma_X^2}\right). \tag{32}$$

Here σ_X^2 is the variance for examples from $\mathcal{X}$. Similar for the example o_y from $\mathcal{Y}$, we choose the reference example o_y^i in the same pair with o_x^i and calculate the distance $d(o_y, o_y^i)$ to get another similarity feature vector s_y, where the similarity function is $s(o_y, o_y^i) = \exp\left(-\frac{d(o_y, o_y^i)}{\sigma_Y^2}\right)$.

$$\frac{\partial}{\partial W} f_D = I + \frac{\gamma}{|\mathcal{C}|} \sum_{D(i,j) \geq D(k,l)} \text{diag}[(U(s)_i - V(t)_j)(U(s)_i - V(t)_j)^\top - (U(s)_k - V(t)_l)(U(s)_k - V(t)_l)^\top]. \tag{29}$$

TABLE II
COMPARISON BETWEEN FOUR METHODS PROPOSED IN SECTION V: CCA, R-CCA: MARGINAL RANKING CCA, KR-CCA: KERNEL MARGINAL RANKING CCA, AND MKR-CCA: MULTIPLE KERNEL MARGINAL RANKING CCA

Approach	Feature Space	Linear Or Not	Ranking Info	Number of Kernels
CCA (Eq. (4))	OF and SF	linear	Not Used	Single
R-CCA (Eq. (11))	SF	linear	Used	Single
KR-CCA (Eq. (18))	OF	non-linear	Used	Single
MKR-CCA (Eq. (27))	OF	non-linear	Used	Multiple

Compared with OF, SF has several advantages. Firstly, s_x and s_y have the same dimension. Moreover, s_x and s_y are well aligned. The ith element of s_x corresponds to that of s_y, since they represents the distances from the input pair to the same reference pair. It should be noted that SF is essentially incorporated into kernel calculation. So we only consider SF with CCA and R-CCA.

B. Evaluation Procedure and Indicator

The music-image dataset proposed in Section III is used for examining the proposed methods. Recall that the labelers' preference are set as ground truth in the dataset. Therefore, the consistency rates with the human labeling results are used as evaluation criterion. Totally, we have $22\,632$ music segments with $47\,888$ music-image pairs. After filtering the music-image pairs that are not agreed by at least 4 annotators, we randomly select about 40000 music-image pairs as our training samples, and use the rest for testing. Firstly we cluster the training samples into several clusters based on music part, and then apply the cross-modal analysis methods proposed above to each cluster of cross-modal pairs. The final similarity scores are calculated as the one obtained from the closet cluster, or as a weighted summation of the similarity scores of each cluster (Section V-F). We adopt four methods to calculate the consistency rates, whose details are shown in Table II. Please remind that we use marginal ranking CCA solved by quadratic programming to calculate the similarity function for R-CCA in the following experiments.

C. Cross-Modal Matching

1) Comparison Among Feature Representations: In this subsection, we first explore the performance of different music and image feature representations for music-image matching task. For music, we adopt both ENS-based (ENS) and MFCC-based (MFCC) semantic representations (Section IV-A). For image, we explore the effect of low-level features (LowF, Section IV-B), scene attribute-based semantic representations (SA) [24], and our lyric-based semantic representations (LA). CCA and R-CCA are adopted as the embedding method, cluster number K is set as 12, and closest clustering-based similarity function (30) is adopted. The consistency rates are shown in Table III for CCA and Table IV for R-CCA. MFCC-based semantic representations for music clearly beat the ENS-based one by about 7%, and lyric-based attribute improve the consistency rate by about 3%, comparing with the low-level feature.

TABLE III
CONSISTENCY RATE (CR) COMPARISON AMONG DIFFERENT FEATURE REPRESENTATIONS OF MUSIC AND IMAGE FOR CCA

Music	MFCC	MFCC	MFCC	ENS	ENS	ENS
Image	LowF	SA	LA	LowF	SA	LA
CR	77.43%	78.82%	**80.14%**	69.82%	71.83%	73.22%

TABLE IV
CONSISTENCY RATE (CR) COMPARISON AMONG DIFFERENT FEATURE REPRESENTATIONS OF MUSIC AND IMAGE FOR R-CCA

Music	MFCC	MFCC	MFCC	ENS	ENS	ENS
Image	LowF	SA	LA	LowF	SA	LA
CR	79.63%	80.73%	**82.01%**	71.28%	72.41%	74.02%

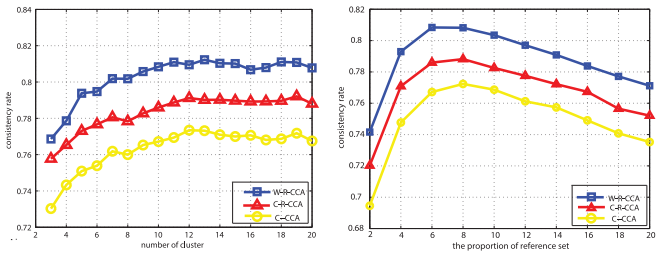
TABLE V
CONSISTENCY RATE COMPARISON BETWEEN CCA, R-CCA, KR-CCA, MKR-CCA, AND MCU [37]

Method	Consistency Rate
CCA	77.43%
R-CCA	79.63%
KR-CCA	86.81%
MKR-CCA	**88.51%**
MCU	85.12%

Scene attributes perform better than low-level features in this case, but fail to compete with lyric-based features. This may be ascribed to three facts: first, the number of scene attributes may not be large enough to cover the images extracted from music videos. Secondly, LA and SA attributes share some common scenery attributes, but LA has more emotional and functional attributes, which make it performs better. Moreover, there exist several scene attributes whose AP scores (Average Precision) are high, but not appropriate for describing images in music videos, such as "no horizon," "man-made," or "asphalt."

2) Comparison Among Cross-Modal Embedding Methods: We then compared our four cross-modal similarity estimation methods, CCA (4), Marginal Ranking CCA [R-CCA, (9)], Kernel Marginal Ranking CCA [KR-CCA, (18)], and Multiple Kernel Ranking CCA (MKR-CCA, Section V-E), with another cross-modal framework maximum covariance unfolding [37]. The MFCC-based semantic representation (Section IV-A) is adopted for music, and the low-level image features (Section IV-B) are combined together to form image representations. In CCA and R-CCA, similarity-based feature space (SF) is adopted. For KR-CCA, we use a Gaussian kernel calculated by the sum of all the low-level image feature distances. For MKR-CCA, we construct a kernel space for each type of image features. Cluster number (Section V-F) K is set as 12, and we adopt the closest clustering-based similarity function (30) to combine the similarity scores calculated in different clusters. The results are shown in Table V.

It can be found that MKR-CCA achieves the best performance among the four methods. The best consistency rate comes to



(a) (b)

Fig. 8. (a) Comparison of consistency rates using C-CCA, C-R-CCA, and W-R-CCA: (a) when changing cluster number of training set, x axis represents the cluster number, (b) when changing the proportion of reference set to training set, x axis represents the inverse of the proportion.

88.5%, which leads to an improvement of consistency rates over CCA and MCU about 11% and 3%. It is clearly shown that utilizing extra ranking labeling results can benefit the music-image matching. The better performance of KR-CCA and MKR-CCA may be expected, as kernelization and utilizing multiple kernels to embed heterogeneous features have been proved effective in cross-modal matching [17] and multi-modal learning [11].

3) Parameter Setting: We first examine the parameters settings for clustering-based similarity functions. Fig. 8(a) exhibits the consistency rates when the cluster number K varies from 3 to 20. It is clearly shown that the precisions of closest CCA, closest ranking CCA, and weighted ranking CCA generally increase with the number of cluster when $K \leq 10$. This is partly because too few clusters cannot model the whole space precisely. When $K > 10$, the precisions increase little. So we set $K = 12$ as default, and it shows that divide-and-conquer technique works in our problem. The weighted clustering-based similarity function performs better than closest one about 2%. This may be simply because that the weighted methods integrate more training information from other clusters.

Then, we change the proportion of the reference set $1/d$ by fixing $K = 10$. Recall that in Section VI-A, we introduce the similarity-based feature space (SF) for non-kernelized methods, like CCA and ranking CCA. We denote the proportion of the reference set to training set as $1/d$. It means that we randomly choose a quarter of training pairs as reference pairs if d is set as 4. In our experiments, d varies from 2 to 20. Fig. 8(b) shows that the best $1/d$ is around $1/6 \sim 1/8$. The reason is that similarity-based features cannot convey enough information when too few reference pairs are used, while too many reference pairs will lead to redundant and noisy representations.

In spite of selecting local references for each cluster, one may suggest to use global reference set which is randomly chosen from the whole training set without clustering for non-kernelized methods. We also make comparison between local reference and global reference with ranking CCA. To be fair for both kinds of reference, we fix 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 VI. All the experiments are run on a machine with 2.80 GHz CPU and 16 GB of RAM. It can be seen

TABLE VI
LOCAL REFERENCE VERSUS GLOBAL REFERENCE

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

TABLE VII
RETRIEVAL PERFORMANCE (MAP SCORES) COMPARISON FOR WIKIPEDIA DATASET BETWEEN RANDOM, SCM [16], MCU [37], AND R-CCA

Method	random	SCM (CCA)	Fast-MCU	R-CCA
text-image	0.118	0.277	0.264	**0.281**
image-text	0.118	0.226	0.198	**0.231**

that local reference has slightly better performance and is much faster, which is more scalable for larger or distributed database. It is because the similarity-based feature representations x, y have lower dimension with local reference than with global reference for the same d.

We also compare original feature space and similarity-based feature space with closest CCA, weighted CCA, closest R-CCA and weighted R-CCA. The proportion of the reference set in similarity-based feature space is fixed as $1/6$. Experiment shows that for all cluster numbers, similarity-based feature space performs much better than original feature space by more than 10%. It proves that it is not appropriate to directly use the original diverse low-level features due to the complexity of this cross-modal problem.

D. Evaluation on Image-Text Dataset

In addition to music-image matching, we also conduct an experiment on Wikipedia dataset, which is used for image-text retrieval [16], with our proposed method KR-CCA and MKR-CCA. Following the evaluation section of [16], we adopt the same text and image features, and use MAP scores for both text-image and image-text retrieval as a measure. Please note that in [37], the text and image are represented as with a LDA model with 20 topics and a bag-of-words representation with 4096 codewords. While in [16] and this work, the topic and codeword number are set as 10 and 128 respectively for fair comparison. We randomly select pairs from different categories as our negative pairs in the training process of KR-CCA and MKR-CCA. For MKR-CCA, we add three more low-level features in addition to the bag-of-words representation for image: attention guided color signature [45], gist features [47], and Histogram of Gradient [48]. The experimental results are shown in Tables VII and VIII. We can see that first, utilizing pair-wise ranking information (R-CCA, KR-CCA) based on CCA and KCCA can slightly benefit cross-modal matching. Secondly, the introduction of more types of features can effectively improves the cross-modal matching performance, which make MKR-CCA get the best MAP scores. One may argue that MKR-CCA utilizes much broader feature set. For SCM and KR-CCA, we further incorporate these image features by combining all the feature kernels.

TABLE VIII
RETRIEVAL PERFORMANCE (MAP SCORES) COMPARISON FOR WIKIPEDIA DATASET BETWEEN SCM [16] (KCCA), OUR KR-CCA AND MKR-CCA WITH MULTIPLE FEATURE TYPES OR NOT

Method	text-image	image-text
SCM (KCCA)	0.324	0.231
SCM (KCCA, multiple)	0.332	0.237
KR-CCA	0.328	0.242
KR-CCA (multiple)	0.334	0.246
MKR-CCA	**0.341**	**0.272**

It can be found in Table VIII that first, SCM and KR-CCA have more discriminative power by utilizing multiple feature types. Secondly, MKR-CCA performs clearly better than direct combination of these kernels due to its ability of generating different weights for different types of kernels. The effects of the kernels may be offset when they are simply combined together.

Overall from the above experiments, we can conclude that first, domain-related abstraction is needed for different cross-modal matching tasks. It has been proved that scene and lyric-based attributes can both benefit music-image matching, while lyric-based attributes perform better due to its descriptive ability for sensing music. Secondly, ranking information generated from annotation process can be utilized to improve the quality of music-image matching. Finally, effective integration of multiple feature spaces also benefit cross-modal matching, which has been proved by MKR-CCA's performance in our music-image and image-text matching experiments. Moreover, by utilizing the ranking labeling information, using weighted clustering-based similarity function, and combining multiple feature representations of music and image by multiple kernel embedding, we obtain a consistency rate about 91.5%, which shows a potential that the music-image matching task can be well solved automatically by machines.

VII. CONCLUSION

New cross-modal matching relationship becomes a big demanding with the daily growth of multimedia interaction and communication. Neuroscience and psychology studies indicate that human perception of music and image are highly correlated. These facts inspire us to investigate cross-modal music-image matching task: whether and how music and image can be matched. We asked six labelers to compare music-image pairs extracted from music video with those randomly generated ones. The results indicate that all labelers prefer the pairs from music videos, and human have consensus on evaluating the matching degree between music and image. Inspired by these observations, we extract image attributes from music lyrics, and present a general cross-modal ranking framework for music-image matching, which can also be extended to other cross-modal applications. Our experiments show that firstly lyric-based image attribute is an effective semantic representation for music-image matching. Secondly, the proposed cross-modal ranking analysis framework helps to improve the performance of cross-modal matching. The best consistency rate by integrating the proposed techniques is 91.5%, which indicates that the relationships between music and image can be automatically modeled, and more music-image applications can be developed based on our proposed techniques. We hope our work can encourage more researches in this new area.

REFERENCES

[1] J. Chao, H. Wang, W. Zhou, W. Zhang, and Y. Yu, "Tunesensor: A semantic-driven music recommendation service for digital photo albums," in *Proc. Int. Semantic Web Conf.*, 2011.

[2] J. Lıbeks and D. Turnbull, "You can judge an artist by an album cover: Using images for music annotation," *IEEE Multimedia Mag.*, vol. 18, no. 4, pp. 30–37, Apr. 2011.

[3] J. Wang, Y. Yang, I. Jhuo, Y. Lin, and H. Wang, "The acousticvisual emotion Guassians model for automatic generation of music video," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 1379–1380.

[4] Y. Yu, Z. Shen, and R. Zimmermann, "Automatic music soundtrack generation for outdoor videos from contextual sensor information," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 1377–1378.

[5] P. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral Brain Sci.*, vol. 31, pp. 559–575, 2008.

[6] J. Osborne, "The mapping of thoughts, emotions, sensations, and images as responses to music," *J. Mental Imagery*, vol. 5, pp. 133–136, 1981.

[7] A. Quittner and R. Glueckauf, "The facilitative effects of music on visual imagery: A multiple measures approach," *J. Mental Imagery*, vol. 7, pp. 105–119, 1983.

[8] M. Saenz and C. Koch, "The sound of change: Visually-induced auditory synesthesia," *Current Biol.*, vol. 18, pp. 650–651, 2008.

[9] L. Meyer, *Emotion and Meaning in Music*. Chicago, IL, USA: Univ. of Chicago Press, 1961.

[10] S. Agarwal *et al.*, "Generalized non-metric multidimensional scaling," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, 2007, pp. 11–18.

[11] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *J. Mach. Learn. Res.*, vol. 12, pp. 491–523, 2011.

[12] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 41–48.

[13] M. Slaney, "Semantic-audio retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, vol. 4, pp. IV-4108–IV-4111.

[14] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 273–276.

[15] Y.-T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 221–229, Feb. 2008.

[16] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[17] J. Costa Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[18] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[19] M. McKinney *et al.*, "Features for audio and music classification," in *Proc. Int. Symp. Music Inf. Retrieval*, 2003, pp. 151–158.

[20] S. Essid, G. Richard, and B. David, "Inferring efficient hierarchical taxonomies for MIR tasks: Application to musical instruments," in *Proc. 6th Int. Conf. Music Inf. Retrieval*, 2005, pp. 324–328.

[21] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 467–476, Feb. 2002.

[22] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 433–440.

[23] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1778–1785.

[24] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2751–2758.
[25] I. Endres, A. Farhadi, D. Hoiem, and D. Forsyth, "The benefits and challenges of collecting richer object annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1–8.
[26] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 951–958.
[27] T. Pham, N. Maillot, J. Lim, and J. Chevallet, "Latent semantic fusion model for image retrieval and annotation," in *Proc. ACM Conf. Inform. Knowledge Manage.*, 2007, pp. 439–444.
[28] T. Westerveld, "Probabilistic multimedia retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 2002, pp. 437–438.
[29] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun, 2009, pp. 1367–1374.
[30] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo, "Combining image captions and visual analysis for image concept classification," in *Proc. Int. Workshop Multimedia Data Mining: Held Conjunction ACM SIGKDD*, 2008, pp. 8–17.
[31] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2002.
[32] P. Dunker, S. Nowak, A. Begau, and C. Lanz, "Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach," in *Proc. 1st ACM Int. Conf. Multimedia Inf.*, 2008, pp. 97–104.
[33] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2002.
[34] H. Hotelling, "Relations between two sets of variates," *Biometrika*, 1936, vol. 28, no. 3/4, pp. 321–377, 1936.
[35] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[36] C. Liu, S. Wu, S. Jiang, and A. Tung, "Cross domain search by exploiting Wikipedia," in *Proc. IEEE 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 546–557.
[37] V. Mahadevan *et al.*, "Maximum covariance unfolding: Manifold learning for bimodal data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 918–926.
[38] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2407–2414.
[39] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 88–95.
[40] X. Alameda-Pineda, V. Khalidov, R. Horaud, and F. Forbes, "Finding audio-visual events in informal social gatherings," in *Proc. Int. Conf. Multimodal Interfaces*, 2011, pp. 247–254.
[41] J. Imura, T. Fujisawa, T. Harada, and Y. Kuniyoshi, "Efficient multi-modal retrieval in conceptual space," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1085–1088.
[42] L. Barrington, A. Chan, and G. Lanckriet, "Modeling music as a dynamic texture," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 602–612, Mar. 2002.
[43] D. Tingle, Y. Kim, and D. Turnbull, "Exploring automatic music annotation with 'acoustically-objective' tags," in *Proc. Int. Conf. Multimedia Inform. Retrieval*, 2010, pp. 55–62.
[44] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Inf. Retrieval*, 2000.
[45] Y. Rubner, L. Guibas, and C. Tomasi, "The earth movers distance, multi-dimensional scaling, and color-based image retrieval," in *Proc. ARPA Image Understanding Workshop*, 1997, pp. 661–668.
[46] X. Wang, K. Liu, and X. Tang, "Query-specific visual semantic spaces for web image re-ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 857–864.
[47] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 1, pp. 273–280.
[48] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.
[49] W. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. Int. Workshop Automat. Face Gesture Recog.*, 1995, pp. 296–301
[50] T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 663–676.
[51] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 365–372.
[52] B. Yao *et al.*, "Human action recognition by learning bases of action attributes and parts," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1331–1338.
[53] D. Klein and C. Manning, "Accurate unlexicalized parsing," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguistics*, 2003, pp. 423–430.
[54] W. Chu and S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, pp. 792–815, 2007.
[55] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 451–458.
[56] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, pp. 2639–2664, 2004.
[57] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, 2006.
[58] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.
[59] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2002.

Xixuan Wu received the B.S. degree from Tsinghua University, Beijing, China, in 2010, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, China, in 2015.

She is currently working with Google, Mountain View, CA, USA. Her research interests include music-image matching, computer vision, and machine learning.

Yu Qiao (S'05–M'06–SM'13) received the Ph.D. degree from the University of Electro-Communications, Tokyo, Japan, in 2006.

He was a JSPS Fellow and a Project Assistant Professor with The University of Tokyo, Tokyo, Japan, from 2007 to 2010. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 90 papers. His research interests include pattern recognition, computer vision, multimedia, image processing, and machine learning.

Prof. Qiao was the recipient of the Lu Jiaxi Young Researcher Award from the Chinese Academy of Sciences in 2012.

Xiaogang Wang (S'03–M'10) received the B.S. degree in electrical engineering and information science from the University of Science and Technology of China, Hefei, China, in 2001, the M.Phil. degree from the Chinese University of Hong Kong, Hong Kong, China, in 2004, and the Ph.D. degree in computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.

He is currently an Assistant Professor with the Department of Electronic Engineering, Chinese University of Hong Kong. His research interests include computer vision and machine learning.

Prof. Wang was the Area Chair of the IEEE International Conference on Computer Vision 2011, European Conference on Computer Vision 2014, and Asian Conference on Computer Vision 2014. He is the Associate Editor of the *Image and Visual Computing Journal* and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He was the recipient of the Outstanding Young Researcher in Automatic Human Behaviour Analysis Award in 2011, the Hong Kong RGC Early Career Award in 2012, and the Young Researcher Award of the Chinese University of Hong Kong.

Xiaoou Tang (S'93–M'96-SM'02–F'09) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1990, the M.S. degree from the University of Rochester, Rochester, NY, USA, in 1991, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1996.

He is a Professor with the Department of Information Engineering and the Associate Dean (Research) of the Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, China. He was the Group Manager of the Visual Computing Group, Microsoft Research Asia, Beijing, China, from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing.

Dr. Tang is a Program Chair of the IEEE International Conference on Computer Vision 2009 and an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision*. He was the recipient of the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition 2009.