2013

Query-Document-Dependent Fusion: A Case Study of Multimodal Music Retrieval

Zhonghua LI

Bingjun ZHANG

Yi YU

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Ye WANG

Citation

LI, Zhonghua; ZHANG, Bingjun; YU, Yi; SHEN, Jialie; and WANG, Ye. Query-Document-Dependent Fusion: A Case Study of Multimodal Music Retrieval. (2013). *IEEE Transactions on Multimedia*. 15, (8), 1830-1842. Research Collection School Of Information Systems.
Available at: http://ink.library.smu.edu.sg/sis_research/1822

Query-Document-Dependent Fusion: a Case Study of Multimodal Music Retrieval [1]

Zhonghua Li[1]*, Bingjun Zhang[1], Yi Yu[1], Jialie Shen[2], Ye Wang[1]

[1]School of Computing, National University of Singapore

[2]School of Information Systems, Singapore Management University

[1]{lizhongh,bingjun,yuy,wangye}@comp.nus.edu.sg, [2]jlshen@smu.edu.sg

Abstract

In recent years, multimodal fusion has emerged as a promising technology for effective multimedia retrieval. Developing the optimal fusion strategy for different modality (e.g. content, metadata) has been the subject of intensive research. Given a query, existing methods derive a unified fusion strategy for all documents with the underlying assumption that the relative significance of a modality remains the same across all documents. However, this assumption is often invalid. We thus propose a general multimodal fusion framework, query-document-dependent fusion (QDDF), which derives the optimal fusion strategy for each query-document pair via intelligent content analysis of both queries and documents. By investigating multimodal fusion strategies adaptive to both queries and documents, we demonstrate that existing multimodal fusion approaches are special cases of QDDF and propose two QDDF approaches to derive fusion strategies. The dual-phase QDDF explicitly derives and fuses query- and document-dependent weights, and the regression-based QDDF determines the fusion weight for a query-document pair via a regression model derived from training data. To evaluate the proposed approaches, comprehensive experiments have been conducted using a multimedia data set with around 17K full songs and over 236K social queries. Results indicate that the regression-based QDDF is superior in handling single-dimension queries. In comparison, the dual-phase QDDF outperforms existing approaches for most query types. We found that document-dependent weights are instrumental in enhancing multimedia fusion performance. In addition, efficiency analysis demonstrates the scalability of QDDF over large data sets.

Index Terms

query-document-dependent fusion, multimodal, information retrieval

I. Introduction

With the dramatic growth of multimedia information on the Internet, information retrieval has emerged as a promising technology for large-scale multimedia exploration and data management. Since media documents (e.g., images, songs) generally contain information or cues from various modalities (e.g., metadata, content), multimodal fusion, which combines multiple complementary modalities, has become an effective approach to boost information retrieval performance. For example, in a web retrieval system, web pages are retrieved by comparing both body texts and links to other web pages [1]. A similar approach is also used in many multimedia search systems, where textual metadata (e.g., titles, tags, descriptions) and content features (e.g., motion intensity, texture, timbre) are combined to rank or rerank videos, images and songs [2], [3], [4], [5]. Recently, researchers have also conducted benchmarking activities (e.g., MusiCLEF [6]) to promote multimedia access and retrieval using multimodal methods.

The earliest multimodal fusion approach is query-independent fusion (QIF), which uses the same fusion strategy for all incoming queries (e.g., [7], [8]). However, since different queries have different information needs, using the same fusion strategy generally results in low retrieval accuracy. To address this limitation, query-dependent fusion (QDF) was developed. The most straightforward QDF applies different fusion strategies to different classes of queries (e.g., [2], [9]). Recent research has sought to do so for each individual query (e.g., [10]). However, fusion strategies in both QIF and QDF take only query dependence into consideration. Once a fusion strategy is determined for a given query, all documents associated with this query use the same fusion strategy to combine different modalities.

Both QIF and QDF implicitly assume that the relative importance of a modality is the same across all documents associated with a query. However, this is often not the case. As shown in Figure 1, given a query "Male alternative", two songs (song 1: "Karma Police", song 2: "Pearly") by Radiohead are retrieved and ranked at different positions in the ranking lists by two retrieval experts (e.g., text matching and content similarity measure as detailed in Section V-B). Looking further into the modalities of each song, we find that the descriptive power of the same modality vary significantly between these two songs.

For example, song 1 has as many as 61 tags (guitar, male vocalist, alternative rock, etc.), while song 2 has only two tags ("alternative" and "favorite"), which provide much less information. Therefore, other modalities, such as content, would be more descriptive to song 2 (e.g., the relative importance between content and text is 0.9:0.1) than to song 1 (e.g., 0.4:0.6). If we use the uniform fusion weight (w) derived by QIF or the query-dependent weight (w_q) derived by QDF, both songs, along with all other songs in the ranking list, would fuse different modalities using the same weight. Therefore, QIF and QDF fail to accurately reflect the contributions of different modalities to a song's relevance with the query, resulting in sub-optimal performance. Although in previous studies, such as Max, Min, CombMNZ [7], and WSUM/WMNZ [11], a document's ranking scores or occurrence frequency are used to adjust fusion strategies, the performances of these approaches rely on that of each retrieval system and fail to address whether document content can contribute to the final fusion strategy.

Hence, we investigate whether and how fusion strategy should adapt to both queries and documents. Our preliminary work [12] proposed a document-dependent fusion (DDF) method and validated the efficacy of applying document information in fusion strategy derivation. This paper goes beyond our previous study in the following ways:

- We propose a general multimodal fusion framework, query-document-dependent fusion (QDDF), which derives the optimal fusion strategy for each query-document pair by analyzing the content of both queries and documents.
- Based on how fusion strategies adapt to queries and documents, we re-categorize existing multimodal fusion approaches and demonstrate they are special cases of QDDF.
- We formalize a dual-phase approach for QDDF as well as a regression-based QDDF to directly

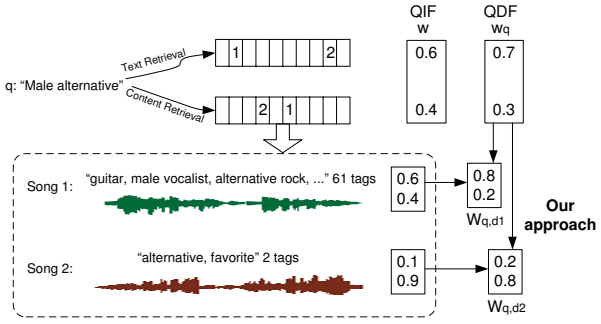


Fig. 1. A retrieval and fusion example.

learn the optimal fusion strategy for each query-document pair.

- A comprehensive evaluation is conducted to compare the retrieval performances of both dual-phase QDDF and regression-based QDDF with existing approaches. Experimental results demonstrate the effectiveness and efficiency of the proposed approaches.

The rest of the paper is organized as follows: Section II reviews different multimodal fusion approaches. Section III introduces the QDDF framework and examines its relationship with other approaches. Section IV introduces both the dual-phase and regression-based QDDF methods. Section V presents our experimental configurations. Experimental results are provided and analyzed in Section VI. Section VII concludes our work and proposes possible directions for future study.

II. Related Work

A. Multimodal Fusion Overview

Multimodal fusion is generally performed at two levels: feature level and decision level [13], [14]. Feature-level fusion combines features extracted from different modalities before classification or ranking. For example, combining audio and text (e.g., lyrics, tags, web documents) features has been widely used in music genre classification (e.g., [15], [16]), mood classification (e.g., [17], [18]), and retrieval (e.g., [19], [20]). Examples also exist beyond music domain, such as video analysis [14]. Most of these works focus on exploring features of different modalities to enhance retrieval performance. By fusing features together, multiple modalities generally assume equal importance. Decision-level fusion relaxes the constraints on features and combines the output of different classifiers or ranking algorithms with the flexibility of prioritizing different modalities using different weights. For instance, Laurier et al. [17], Hu and Downie [18] combined audio- and lyrics-based systems using different combination weights and evaluated the performance in classifying music into different moods. Mayer and Rauber [21] adopted estimated accuracy as combination weights to combine audio- and lyrics-based classification results in music genre classification. Decision-level fusion has also been widely used in meta search (e.g., [7], [11]), video retrieval (e.g., [2]), etc.

At each fusion level, different fusion methods have been proposed and evaluated. For example, Kittler et al. [22] provided a theoretical introduction to a number of rule-based fusion methods, such as linear weighted fusion (sum or product), majority voting, AND/OR, etc. Linear weighted fusion combines

information from different modalities in a linear fashion by assigning appropriately normalized weights to different modalities. This method has become an effective tool in many multimedia applications, such as video classification (e.g., [2]), image retrieval (e.g., [23]), music classification (e.g., [18]), etc. A number of classification-based methods also fuse different modalities non-linearly. Examples include the super-kernel method in video concept detection (e.g., [24], [25]) and the kernel-based partial order embedding method in artist similarity measure (e.g., [26]). A comprehensive survey on multimodal fusion can be found in [13]. In the following section, we will discuss linear weighted fusion methods at decision levels in more detail.

B. Linear Multimodal Fusion Approaches

Previous studies on linear multimodal fusion can be categorized into two families: query-independent fusion (QIF) and query-dependent fusion (QDF).

1) QIF: In the early development of multimodal fusion, QIF derived a unified fusion strategy and applied it for all incoming queries and all related documents. The simplest fusion strategy is to weight each modality equally, such as the CombSUM [7] and Average [8] methods in meta search. Other methods such as Max, Min, and CombMNZ [7], all of which modify the equally weighted strategy using document ranking scores or occurrence frequency, are also frequently used in meta search. In addition, fusion strategies can be identified over validation sets. For example, Bartell et al. [27], [28] derived a fusion strategy by maximizing the average performance over a broad range of queries in a validation set. Mayer and Rauber [21] adopted the estimated accuracy on training data set as the fusion strategy to combine audio and lyric modalities in music genre classification. By using the same fusion strategy for every query, QIF assumes that different modalities make fixed contributions regardless of query types and content. QIF remains simple and effective as long as queries are all limited to the same topic; otherwise, it becomes inadequate.

2) QDF: QDF approaches seek to derive different fusion strategies for different queries. A basic approach, class-based QDF, derives different fusion strategies for a number of query classes either defined manually [2], [9], [29], or discovered automatically [30], [31], [32], [33]. Using language processing techniques, an incoming query is then classified into one of these classes or represented as a mixture of multiple classes [32]. Class-based QDF has been widely used and evaluated in video retrieval [2], [9],

[30], [32], [33], [34], image retrieval [23], [29], and web retrieval [1] systems. However, its efficacy is still constrained by the small number of query classes and the limited accuracy of query classification.

TABLE I
CLASSIFICATION OF MULTIMODAL FUSION APPROACHES BASED ON WHETHER FUSION WEIGHTS DEPEND ON QUERIES OR DOCUMENTS.

Fusion Type	Query	Document	Methods	Example Works
QDIF	$\times$	$\times$	Equal weight method Validation set-based method	CombSUM [7], Average [8], [27], [28], [21]
QDF	$\surd$	$\times$	Class-based method Difficulty prediction method Regression-based method	[2], [9], [29], [30], [31], [32], [35], [36], [37], [10]
DDF	$\times$	$\surd$	Document content-based method Rank score/occurrence-based method	[12], Max, Min, CombMNZ [7], [11]
QDDF	$\surd$	$\surd$	Query and document-based method	This work

To address this weakness, Xie et al. [35] proposed the dynamic formation of query class. Each query is linearly represented by its *K* nearest neighbors. However, the high computational cost for searching nearest neighbors restricts its application in larger data sets. Yom-Tov et al. [36] proposed another approach that, for a given query, predicts its "difficulty" by statistically measuring performances of different retrieval systems and weights them accordingly. Although this approach is effective, how to apply query difficulty predictions to query adaptation awaits further investigation. Recently, QDF was modeled as a mapping from each query to its fusion weight [10]. A regression model was trained to directly predict the fusion strategy for each query, avoiding query matching and lowering computational complexity.

3) Summary: QIF and QDF only consider fusion strategy dependence on queries. Therefore, they prioritize only the documents among which modalities have fixed importances and are thus sub-optimal for other documents. Since the descriptive abilities of different modalities may vary considerably across documents, our previous work [12] addressed this problem and confirmed the efficacy of incorporating document dependence in fusion weight derivation. In this work, we introduce a general framework, query-document-dependent fusion (QDDF), which takes the dependence of both queries and documents into consideration when deriving the optimal fusion strategies.

III. QDDF Framework and Its Relation to Other Approaches

A. QDDF Framework Overview

We now present the QDDF framework. Some frequent notations used in this and belows sections are provided in Table II. Suppose a multimodal retrieval system contains n unimodal retrieval experts to search relevant documents on different modalities. Given a query q_i from the query set $Q = \{q_1, q_2, ..., q_i, ...\}$, each retrieval expert returns a document list ranked by their relevances to the query. Let

TABLE II
TABLE OF NOTATIONS

Symbol	Explanation
$A_{\cdot,j,k}$	Descriptive ability of document $d_{\cdot,j}$ on the k-th modality
$d_{i,j}$	The j-th unique document returned for query q_i
$D_{i,k}$	Document set returned by the k-th retrieval expert for query q_i
D_i	A set of unique documents returned for query q_i
D	Document set
$f(\mathbf{V}_{i,j})$	Regression function on $\mathbf{V}_{i,j}$
$g_{i,j}$	Ground truth relevance of document $d_{i,j}$ to query q_i
l	Loss function
m'	Number of samples chosen per iteration in Pegasos solver
M	Keyword set in the music social query space
M_k	A subset of M on the k-th music dimension
n	Number of retrieval experts
q_i	A query
Q	Query set
$R_{\cdot,j,k}$	Relative score ratio for document $d_{\cdot,j}$ on the k-th modality
$\mathbf{S}_{i,j}$	Ranking score vector of document $d_{i,j}$
$\widehat{S}_{i,j}$	Fusion score of document $d_{i,j}$
$\mathbb{T}, \mathbb{C}$	Textual modality, content modality
$\mathbf{V}_{i,j}$	Feature vector of query-document pair $(q_i, d_{i,j})$
$\mathbf{W}_{i,j}$	Fusion weight for query-document pair $(q_i, d_{i,j})$
$\widehat{\mathbf{W}}_{i,j}$	The optimal fusion weight for query-document pair $(q_i, d_{i,j})$
$\mathbf{W}_{i,\cdot}$	Query-dependent weight
$\mathbf{W}_{\cdot,j}$	Document-dependent weight
$\mathbf{W}_{\cdot,\cdot}$	Query-document-independent weight
β	Balance factor between query weight and document weight
η	Learning rate in Pegasos solver

$D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,j}, ...\}$ denote the set of unique documents associated with q_i. For each document $d_{i,j} \in D_i$, a ranking score vector $\mathbf{S}_{i,j} = [S_{i,j,1}, S_{i,j,2}, ..., S_{i,j,n}]^T$ can be computed according to how $d_{i,j}$ is relevant to q_i on different modalities. To rank the documents, a fusion score $\widehat{S}_{i,j}$ is calculated for each

document $d_{i,j} \in D_i$ as:

$$\widehat{S}_{i,j} = \sum_{k=1}^{n} S_{i,j,k} W_{i,j,k} = \mathbf{W}_{i,j}^T \cdot \mathbf{S}_{i,j} \ , \tag{1}$$

where $\mathbf{W}_{i,j} = [W_{i,j,1}, ..., W_{i,j,n}]^T$ ($0 \leq W_{i,j,k} \leq 1$ and $\sum_k W_{i,j,k} = 1$) is the fusion weight (or fusion strategy) that is applied to document $d_{i,j}$ given query q_i. Unlike existing approaches, $\mathbf{W}_{i,j}$ can be theoretically different for every document and related query. Because fusion weights play a key role in ranking the documents, deriving the optimal fusion weight $\widehat{\mathbf{W}}_{i,j}$ is a critical step in a multimodal retrieval system.

B. Multimodal Fusion Approaches

Depending on whether fusion weights depend on queries or documents, multimodal fusion approaches can be grouped into four classes: query-document-independent fusion (QDIF), query-dependent fusion (QDF), document-dependent fusion (DDF), and query-document-dependent fusion (QDDF). Based on this categorization, existing works can be summarized in Table I. As shown in this table, for some QIF approaches reviewed in Section II, such as CombSUM [7] and Average [8], fusion strategy of different retrieval experts is invariable to both queries and documents. Therefore, these approaches are categorized into QDIF. An example of DDF can be found in [12] and its key idea is to combine document-dependent weights with equal query weights. Other methods, such as Max, Min, CombMNZ [7], and [11], whose fusion strategies rely on document ranking scores or occurrence frequency exclusively can also be categorized into DDF.

As a general multimodal fusion framework, QDDF can be simplified to the other three types of fusion approaches by relaxing the dependence on queries, documents, or both (Fig. 2). To examine their relationships, we formalize them as different approaches to optimize fusion weights by minimizing different loss functions between the ideal retrieval results and the predicted results.

1) QDDF: QDDF derives the optimal fusion weight for each query-document pair. Given query q_i and its associated documents D_i, the optimal fusion weight $\widehat{\mathbf{W}}_{i,j}$ for $d_{i,j} \in D_i$ can be derived as:

$$\widehat{\mathbf{W}}_{i,j} = arg \min_{\mathbf{W}_{i,j}} l(q_i, d_{i,j}, g_{i,j}, \mathbf{W}_{i,j}) \ , \tag{2}$$

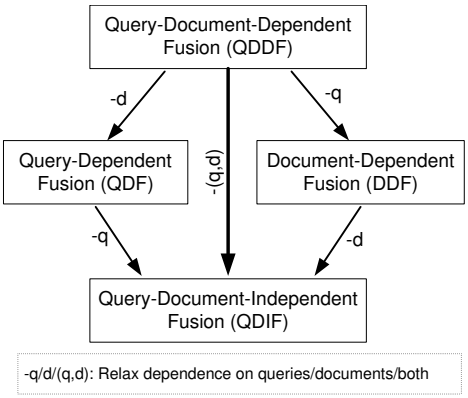


Fig. 2. Transitions between different multimodal fusion approaches.

where $l(q_i, d_{i,j}, g_{i,j}, \mathbf{W}_{i,j})$ denotes the loss function defined at document level, and $g_{i,j}$ is the ground truth relevance of $d_{i,j}$ to q_i. The fusion weight $\widehat{\mathbf{W}}_{i,j}$ of QDDF enables every document to fuse its modalities in the optimal way.

2) QDF: By relaxing the dependence on each document, QDF approaches derive the optimal fusion weight for each query. All documents associated with this query use the same fusion strategy to combine different modalities. We can formalize this procedure as:

$$\widehat{\mathbf{W}}_{i,j} = \mathbf{W}_{i,\cdot} = arg \min_{\mathbf{W}_{i,j}} \sum_{j=1}^{|D_i|} l(q_i, d_{i,j}, g_{i,j}, \mathbf{W}_{i,j}) \ , \tag{3}$$

where $\mathbf{W}_{i,\cdot}$ denotes a fusion weight invariable to documents, as $\mathbf{W}_{i,j} = \mathbf{W}_{i,j'}$ for any two unique documents $d_{i,j}$ and $d_{i,j'}$ ($1 \leq j, j' \leq |D_i|$ and $j \neq j'$).

3) DDF: Similarly, DDF ignores the dependence of individual queries and learn a fusion weight which relies only on the document. This document-dependent weight can be derived by summarily analyzing the performances of all training queries:

$$\widehat{\mathbf{W}}_{i,j} = \mathbf{W}_{\cdot,j} = arg \min_{\mathbf{W}_{i,j}} \sum_{i=1}^{|Q|} l(q_i, d_{i,j}, g_{i,j}, \mathbf{W}_{i,j}) \ , \tag{4}$$

where $\mathbf{W}_{\cdot,j}$ denotes the fusion weight invariable to queries.

4) QDIF: QDIF ignores the dependence on both queries and documents. A constant fusion weight can be derived and applied to all documents given every query.

$$\widehat{\mathbf{W}}_{i,j} = \mathbf{W}_{\cdot,\cdot} = arg \min_{\mathbf{W}_{i,j}} \sum_{i=1}^{|Q|} \sum_{j=1}^{|D_i|} l(q_i, d_{i,j}, g_{i,j}, \mathbf{W}_{i,j}) \ . \tag{5}$$

The weight invariability to both queries and documents is denoted by $\mathbf{W}_{\cdot,\cdot}$.

As a general framework, QDDF achieves the finest mapping from query/document to fusion weight among these four approaches. When applying these approaches, fusion weights can be derived in different ways to adapt to the difficulties of collecting ground truth relevance. In this paper, we mainly discuss how QDDF learns fusion weights.

IV. Fusion Weight Learning by QDDF

Fig. 2 points to two possible approaches to derive fusion strategies of QDDF. The first is to combine query-dependent weights with document-dependent weights. Therefore, we propose a dual-phase QDDF approach by first deriving these weights separately and then combining them together. The second approach is to directly derive fusion strategies by considering queries and documents simultaneously. We thus extend the regression-based QDF approach [10] to QDDF by modeling fusion weight derivation as a regression from query-document pairs to their fusion weights.

A. Dual-phase Fusion Weight Learning

As shown in Fig. 3, dual-phase QDDF consists of three main components: query-dependent weight learning, document-dependent weight learning, and weight fusion. Since derivations of query-dependent weights are well-known, we will only provide a step-by-step derivation of document-dependent weights and explain its fusion with query-dependent weights.

1) Descriptive Ability and Music Social Query Space: Textual metadata and content are basic modalities among most multimedia documents, and in this study we equally consider both for each music dimension. Therefore, among the total n modalities, there are $n/2$ textual modalities and $n/2$ content modalities. We define the descriptive ability for each modality as the extent to which it satisfies users' information needs and derive document-dependent weights by assigning more weights to modalities

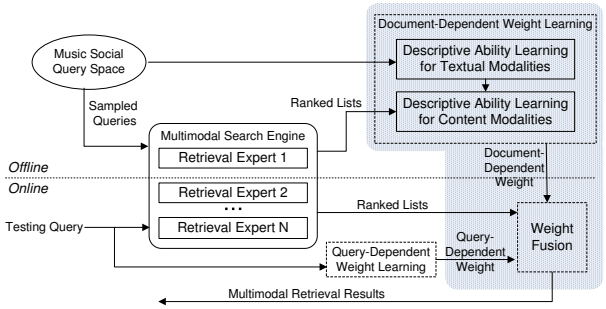


Fig. 3. Dual-phase fusion weight learning diagram

with greater descriptive abilities. To quantify descriptive ability, we construct an online folksonomy-based music social query space [10] to simulate users' information needs in music retrieval. The data information of this space in provided in Table III.

TABLE III
DATA INFORMATION OF THE MUSIC SOCIAL QUERY SPACE.

Dimensions	No. of Tags	Music Styles
Genre	244	Classical, Country, Electronic, HipHop, Jazz, Metal, Pop, Rock
Mood	286	Angry, Joy, Pleasure, Sad
Vocalness	13	Female, Male, Mixed, Nonvocal
Instrument	454	Brass, Percussion, Strings, Woodwinds

2) Derive Descriptive Ability for Textual Modalities: Descriptive abilities for textual modalities are calculated by matching document metadata (e.g., tags) with the music social query space. Let the indicator function $[w \in d_{\cdot,j}] = 1$ if a word w in the music social query space M appears in a document $d_{\cdot,j}$[2]; otherwise let it equal to 0. The descriptive ability (A) of the k-th ($k = 1, 2, \ldots, n/2$) textual modality of $d_{\cdot,j}$ is then defined as:

$$A_{\cdot,j,\mathbb{T}_k} = \frac{\sum_{w \in M_k} [w \in d_{\cdot,j}]}{\sum_{w \in M} [w \in d_{\cdot,j}]}, \tag{6}$$

where $\mathbb{T}$ denotes the textual modality and $M_k \subset M$ denotes a set of words in the k-th music dimension.

3) Derive Descriptive Ability for Content Modalities: Because content modalities are not directly comparable with the music social query space, we propose a relative-score-based method to calculate the descriptive abilities for content modalities. We first generate a query set Q from the query space and then retrieve relevant documents for each query using both text and audio retrieval experts (see Section V-B

[2] $\cdot$ denotes the independence to queries.

for more details). For each unique document $d_{\cdot,j}$, we record its ranking score $S_{i,j,k}$ in the k-th ranked list for every q_i and sum it for all the queries. The average score for the k-th ($k = 1, 2, \ldots, n$) modality of document $d_{\cdot,j}$ is:

$$\bar{S}_{\cdot,j,k} = \sum_{i=1}^{|Q|} S_{i,j,k} / \sum_{i=1}^{|Q|} [d_{\cdot,j} \in D_{i,k}] \ . \tag{7}$$

Given q_i, if the document set $D_{i,k}$ returned by the k-th retrieval expert contains document $d_{\cdot,j}$, $[d_{\cdot,j} \in D_{i,k}]$ is set to 1, otherwise 0.

The average score of a modality approximates its importance to a document. For each music dimension, the relative score between textual and content modalities can be used to derive their relative importance. Thus, the descriptive ability for the k-th ($k = 1, 2, \ldots, n/2$) content modality of document $d_{\cdot,j}$ is computed as:

$$A_{\cdot,j,\mathbb{C}_k} = R_{\cdot,j,k} A_{\cdot,j,\mathbb{T}_k} \ , \tag{8}$$

where $\mathbb{C}$ denotes the content modality and $R_{\cdot,j,k}$ the content-to-text ratio based on the average score of the content modality and that of the textual modality. When an average score is zero, possibly as a result of unbalanced query samples or mismatch between queries and documents on a modality, $R_{\cdot,j,k}$ is set to 1.

4) Document-Dependent Weight Learning: Given the descriptive abilities of all the modalities of a document, document-dependent weights are derived by linearly assigning more weights to modalities with greater descriptive abilities. Assume $W_{\cdot,j,\mathbb{T}_k}$ and $W_{\cdot,j,\mathbb{C}_k}$ are the weights for the k-th ($k = 1, 2, \ldots, n/2$) textual and content modalities of document $d_{\cdot,j}$, the document weight vector is represented as $\mathbf{W}_{\cdot,j} = [W_{\cdot,j,\mathbb{T}_1}, ..., W_{\cdot,j,\mathbb{T}_{n/2}}, W_{\cdot,j,\mathbb{C}_1}, ..., W_{\cdot,j,\mathbb{C}_{n/2}}]^T$, where $W_{\cdot,j,\mathbb{T}_k} = \alpha A_{\cdot,j,\mathbb{T}_k}$, $W_{\cdot,j,\mathbb{C}_k} = \alpha A_{\cdot,j,\mathbb{C}_k}$, and α is a normalization factor which can be calculated by solving:

$$\sum_{k=1}^{n/2} (W_{\cdot,j,\mathbb{T}_k} + W_{\cdot,j,\mathbb{C}_k}) = 1 \ . \tag{9}$$

5) Weight Fusion: The document-dependent weight $\mathbf{W}_{\cdot,j}$ and the query-dependent weight $\mathbf{W}_{i,\cdot}$ can be fused in various ways, such as the multiplication method [12]. In this paper, the fusion weight $\widehat{\mathbf{W}}_{i,j} = [W_{i,j,1}, ..., W_{i,j,n}]^T$ is calculated by combining them linearly.

$$W_{i,j,k} = \beta W_{i,\cdot,k} + (1 - \beta) W_{\cdot,j,k} \ , \tag{10}$$

where $k = 1, 2, \ldots, n$ and β is a parameter to balance the contributions of query and document weights.

By incorporating document-dependent weights, the relative significance of different modalities, such as tags and audio of the example songs in Section I, can be used to fine tune query-dependent weights and lead to better multimodal fusion results.

B. Regression-based Fusion Weight Learning

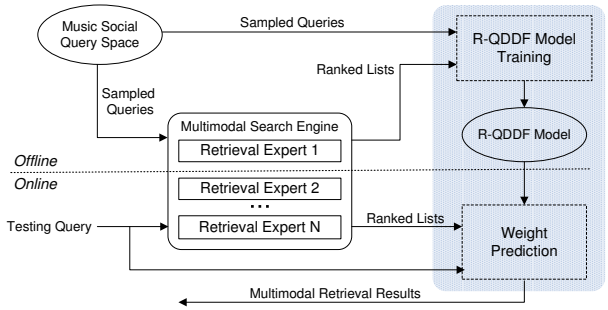


Fig. 4. Regression-based fusion weight learning diagram

1) Model Definition: Fusion weight derivation of QDDF can be modeled as a regression function from query-document pairs to their fusion strategies (Fig. 4). A textual query is represented as a document vector [38] over a vocabulary of words in the music social query space, with 1 and 0 denoting the presence and absence of a word in the query. For each document, a *tf-idf* weight vector is constructed to represent its metadata information. Its content information is represented as a fuzzy music semantic vector (FMSV), which is a probability vector obtained using a multi-class support vector machine (SVM) to classify multiple audio features [20]. A feature vector $\mathbf{V}_{i,j}$ combining query and document features is then constructed for the query-document pair $(q_i, d_{i,j})$, and the optimal fusion strategy of $(q_i, d_{i,j})$ is denoted as $\widehat{\mathbf{W}}_{i,j} = [W_{i,j,1}, W_{i,j,2}, ..., W_{i,j,n}]^T$. Fusion weight derivation can thus be modeled as a multiple regression $\widehat{\mathbf{W}}_{i,j} = f(\mathbf{V}_{i,j})$, which can be further decomposed into n single regression models that correspond to n retrieval experts $W_{i,j,k} = f_k(\mathbf{V}_{i,j})$. Finally, the weights predicted from single regression models are normalized and combined as the final fusion strategy.

Support Vector Regression (SVR) is one of the most powerful and widely adopted regression methods. We assume that each single regression model here fits the SVR model, $y = \langle \mathbf{w}, \mathbf{x} \rangle + b$ [39]. Let $\mathbf{x} = \mathbf{V}_{i,j}$ and $y = W_{i,j,k}$. Given a training dataset (see Section V-C for more details) with m samples, i.e., $\mathcal{U} = \{(x_i, y_i) \mid i = 1, 2, \ldots, m\}$, a single regression model is solved by minimizing the following

objective function:

$$f(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{(\mathbf{x},y)\in\mathcal{U}} \ell(\mathbf{w};(\mathbf{x},y)) \ , \tag{11}$$

where λ is a regularization factor and

$$\ell(\mathbf{w};(\mathbf{x},y)) = max\{0, |y - \langle \mathbf{w}, \mathbf{x}\rangle| - \epsilon\} \tag{12}$$

is the ϵ-insensitive empirical loss function.

2) Regression Solver using Pegasos: To solve Eq. 11, we employ the regression Pegasos [10], which performs t iterations and randomly selects m' samples to compute the sub-gradient at each iteration. Initially, $\mathbf{w}_1$ is set to a zero vector. At each iteration i, using the set $\mathcal{U}_i' \subseteq \mathcal{U}$ with m' chosen samples, we obtain the following function to approximate Eq. 11.

$$f(\mathbf{w};\mathcal{U}_i') = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m'}\sum_{(\mathbf{x},y)\in\mathcal{U}_i'} \ell(\mathbf{w};(\mathbf{x},y)) \ . \tag{13}$$

Next, $\mathbf{w}_i$ is updated in two steps. First, the weight vector $\mathbf{w}_{i+\frac{1}{2}}$ is formulated as:

$$\mathbf{w}_{i+\frac{1}{2}} = \mathbf{w}_i - \eta_i \bigtriangledown_i \ , \tag{14}$$

where $\eta_i = 1/(\lambda i)$ is the learning rate,

$$\bigtriangledown_i = \lambda \mathbf{w}_i - \frac{1}{|\mathcal{U}_i'|}\sum_{(\mathbf{x},y)\in\mathcal{U}_i^+} sign(y - \langle \mathbf{w}_i, \mathbf{x}\rangle)\mathbf{x} \tag{15}$$

is the sub-gradient of $f(\mathbf{w};\mathcal{U}_i')$ at $\mathbf{w}_i$, and $\mathcal{U}_i^+$ is a sample set for which $\mathbf{w}$ has a non-zero loss. Then $\mathbf{w}_{i+\frac{1}{2}}$ is projected onto the set $\{\mathbf{w} \mid \|\mathbf{w}\| \leq \frac{1}{\sqrt{\lambda}}\}$ which contains the optimal $\mathbf{w}$. Consequently,

$$\mathbf{w}_{i+1} = \min\{1, 1/(\sqrt{\lambda}\|\mathbf{w}_{i+\frac{1}{2}}\|)\}\mathbf{w}_{i+\frac{1}{2}} \ . \tag{16}$$

The weight after a predefined number of iterations is output as the final weight.

With the trained regression models, the fusion weight for a query-document pair can be directly predicted. This approach jointly considers the dependence of queries and documents, The relative significance of different modalities does not need to be explicitly studied. Therefore, it can be better generalized to different multimedia documents in various retrieval systems.

V. Experimental Configuration

A. Data Collection

Since most existing benchmarking data sets lack ground truth annotations on multiple music dimensions, we adopt the music data set which was previously annotated and used in [10] to evaluate our proposed approaches. This data set consists of two main components:

1) Music Collection: It contains 17174 music tracks along with their metadata, including titles, descriptions, keywords, comments and tags. While tags were crawled from Last.fm, audio tracks and the other metadata were crawled from YouTube using their open APIs. Four music dimensions were studied, including genre, mood, instrument and vocalness. Socially tagged ground truth annotations involving four music dimensions in 20 music styles (Table III) were collected and crossed checked by amateur musicians with reference to Last.fm.

2) Query Collection: To assess retrieval performance using various numbers of modalities, queries related to different music dimensions were generated by sampling tags from the music social query space. The probability of each tag being sampled grows linearly with its popularity. Since four music dimensions are involved, each generated query contains at most four tags from different dimensions. This also guarantees that tags with conflicting meanings, e.g., "happy sad", will not appear concurrently in a query. According to the number of related music dimensions, the generated queries range from low complexity (with tags from only one dimension) to high complexity (with tags from multiple dimensions) to simulate different complexities of users' information needs. A total of 236973 queries are generated, examples including "violin", "classical piano", "jazz saxophone male", "happy classical violin female", etc. Table IV presents the overall categorization and quantity information of these queries.

TABLE IV
SOCIAL QUERY DISTRIBUTION OVER DIFFERENT MUSIC DIMENSION COMBINATIONS. G, M, V, I REPRESENT GENRE, MOOD, VOCALNESS AND INSTRUMENT DIMENSIONS. QUERIES ARE GROUPED ACCORDING TO THEIR RELATED MUSIC DIMENSIONS. NO. INDICATES THE NUMBER OF QUERIES IN EACH GROUP.

1-dimension		2-dimension		3-dimension		4-dimension	
Query	No.	Query	No.	Query	No.	Query	No.
G	158	GM	13165	GMV	19209	GMVI	106642
M	173	GV	2546	GMI	21395		
V	13	GI	13373	GVI	19305		
I	309	MV	2951	MVI	19658		
		MI	13783				
		VI	4293				

B. Multimodal Retrieval Experts

We consider both textual and audio modalities for each music dimension, and a unimodal retrieval expert is constructed on each modality. Every incoming query (e.g., "male alternative song") is parsed by comparing with the music social query space, and query keywords (e.g., "male" and "alternative") are then fed to the corresponding text and audio retrieval experts.

1) Text Retrieval Experts: A text retrieval expert retrieves relevant music items by matching a query keyword with the metadata (e.g., tags) of these music items. We adopted the vector space model [40] to represent music metadata, by tokening metadata, eliminating stop words, and stemming using Porter's algorithm [41]. Each music item is then represented as a *tf-idf* weight vector. Given a query keyword, the OKAPI BM-25 method [42] is applied to rank different music items.

2) Audio Retrieval Experts: An audio retrieval expert retrieves music items that match a query keyword in the audio dimension. For each audio track, multiple audio features (e.g., timbral, spectral, rhythmic features) are first extracted [20]. A multi-class SVM is then used to classify these features into different music styles within different music dimensions. The activation probabilities of a track being classified into different music styles form a fuzzy music semantic vector (FMSV). For every query keyword, a query FMSV is generated by assigning the value of 1 to the matched music style and 0 to the others. For example, the query keyword "male" will result in a FMSV with only the music style "Male" in the "Vocalness" dimension being 1 and others 0. Euclidean distance is calculated between the query FMSV and the FMSVs of all the audio tracks. A ranked list is constructed by sorting tracks in an ascending order of their Euclidean distances. The audio analysis was implemented based on Marsyas [43].

3) Result Fusion and Multi-Valued Relevance: Given multiple ranked lists, the ranking score $S_{i,j,k}$ of music item $d_{i,j}$ in the k-th ranked list is calculated as $S_{i,j,k} = 1 - Rank_k(d_{i,j})/N$, where N, the number of music items studied in each ranked list, is set to 100 in our experiment. When a fusion strategy is available, ranking scores from different retrieval experts are linearly combined (Eq. 1) to rank music items into a final result list.

The performance of a final ranked list is measured by examining the relevance of every music item with the query. Instead of using binary relevance (1 for relevant, 0 for irrelevant), we use multi-valued relevance to address the partial match between a music item and a query. This relevance is defined as the number of matched dimensions over the total number of dimensions required by the query.

C. Evaluation Methodology

To evaluate the effectiveness of the proposed framework, different types of multimodal fusion approaches are implemented and compared. Average precision (AP) is used to measure the retrieval performance of a ranked list and mean average precision (MAP) the performance over all testing queries. Moreover, efficiency (measured by runtime) of different QDDF approaches is also analyzed.

1) Comparison Approaches: We compared QDDF with the other three types of fusion approaches, including one QDIF approach, one DDF approach, and four QDF approaches.

QDIF was implemented by applying the same fusion strategy for all training queries and their associated music items. Grid search was then used to find the optimal fusion strategy which achieves the highest retrieval accuracy. *DDF* was implemented by integrating document-dependent weights with the above QDIF approach [12].

For QDF, we implemented: a class-based QDF approach based on single-class query matching (*QDF-Single*) [30]; another class-based QDF approach based on mixture-of-classes query matching (*QDF-Mixture*) [32]; a dynamic-class-based QDF approach which matches a query using *K* nearest neighbors (*QDF-KNN*) [35]; and a regression-based QDF approach which predicts a fusion strategy for each query using a regression model (*QDF-Reg*) [10].

For QDDF, we implemented both the dual-phase QDDF (*D-QDDF*) and the regression-based QDDF (*R-QDDF*). D-QDDF was implemented by integrating document-dependent weights with each of the four QDF approaches.

Since our focus was multimodal fusion using unimodal retrieval systems, we did not consider approaches that are used mainly in meta search.

2) Training and Evaluation: To train and test QDF approaches, ground truth weights need to be obtained for every query in the data set (Section V-A2). For each query, grid search was applied to find the fusion weight which produces the highest AP. This weight (also termed oracle combination weight) is used as the ground truth weight. However, grid search is too computationally complex for R-QDDF as it requires ground truth weight for every query-document pair. Therefore, in R-QDDF, the training weight for query-document pair $(q_i, d_{i,j})$ was derived by solving a constrained linear least square problem based

on comparing the predicted relevance with its ground truth relevance $g_{i,j}$:

$$l(q_i, d_{i,j}) = \frac{1}{2}(\mathbf{W}_{i,j}^T \cdot \mathbf{S}_{i,j} - g_{i,j})^2 \ . \tag{17}$$

Having obtained the ground truth weights of all queries, we randomly sampled 200K queries as the training data set. The remaining 36973 queries were used for both validation and testing.

Since many multimodal fusion approaches involve training, training data set availability or the effort of constructing such training data sets is a critical concern in real life multimodal retrieval systems. Therefore, it is worthwhile to study the performance when different amounts of training data are available. In our experiments, we randomly selected different numbers (0.1K, 0.4K, 1.6K, 6.4K, 25.6K, 102.4K, and 200K) of queries from the 200K query bank to form different training data sets. The performances when different approaches were trained using these training data sets were examined. For R-QDDF, the training samples are the query-document pairs generated by combining every query and its associated documents. A random sampling (with a sampling rate of $1/100$) was performed among the query-document pairs of every query in the training data sets. However, for clearer comparison, we used the same notations for training sizes when presenting the results of R-QDDF with other methods.

Evaluation was first performed on validation data sets to determine the optimal parameter settings, which were then used for testing. For robust parameter tunning, the 36973 samples were equally divided into four subsets, and each subset took turns being used as the validation data set, while the others were used for testing. MAP was computed over all tested samples and then averaged across different folds for presentation. Since three trials of random sampling were performed for training R-QDDF models, the presented R-QDDF results for both validation and testing are the averages across different trials. The average runtime of both D-QDDF and R-QDDF were also collected. All the experiments were conducted on a DELL Optiplex 755 PC, with a dual-core CPU (Intel Core 2 E6550 @ 2.33GHZ) and 4GB memory.

VI. Results and Analysis

A. Dual-phase QDDF (D-QDDF)

Since D-QDDF was implemented by integrating document-dependent weights with QDF approaches, its performance may be affected by two factors: (1) the different QDF approaches to derive query-dependent weights and (2) the different methods used to fuse them. We thus compared the performance

of D-QDDF based on four QDF approaches using both linear fusion (D-QDDF-LNR, in Section IV-A5) and multiplication fusion methods (D-QDDF-MUL) [12].

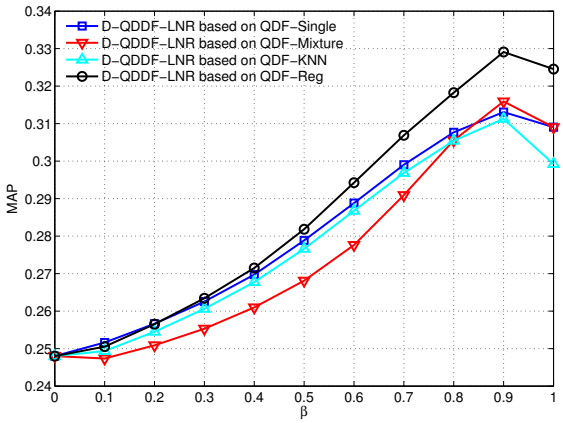
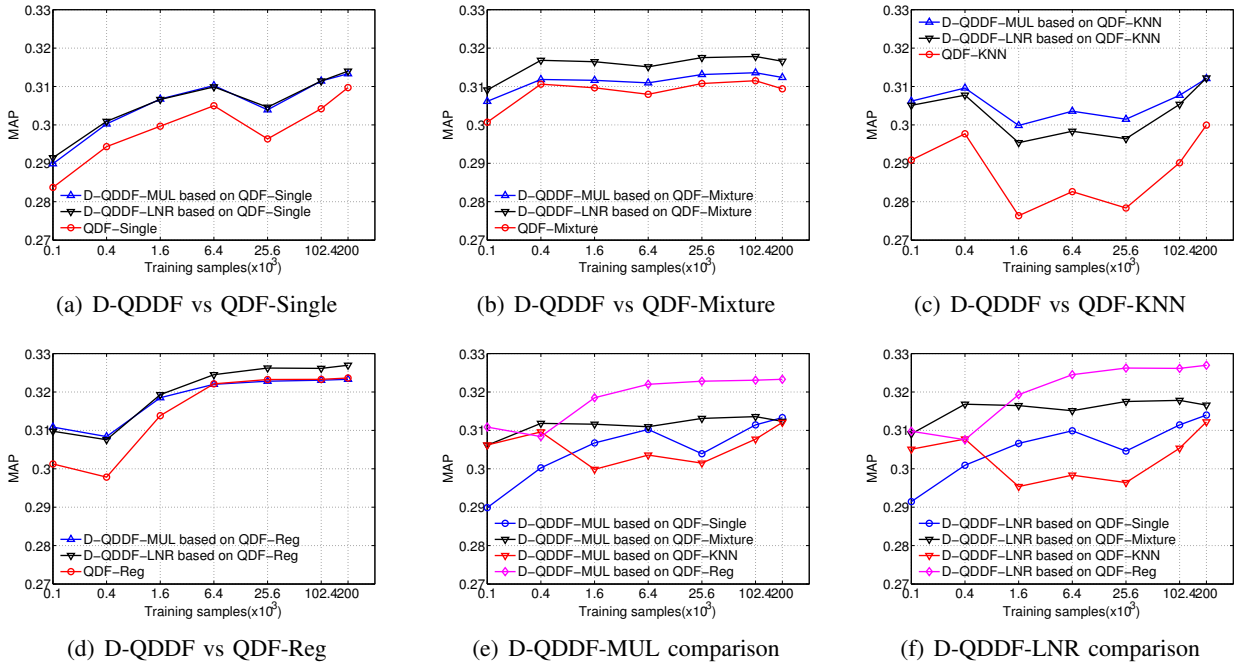


Fig. 5. Retrieval accuracy (MAP) comparison of D-QDDF-LNR approaches (trained with the data set of 200K) using different β.

1) On Effects of Parameter Tuning: We first implemented four QDF approaches and evaluated each one using different parameter settings on the validation sets. For class-based QDF approaches, different numbers $(10i, i = 1, 2, \ldots, 6)$ of query classes were generated based on oracle combination weights. Different numbers $(5i, i = 1, 2, \ldots, 5)$ of nearest neighbors were tested for QDF-KNN. For QDF-Reg, $m' = 1, 2, \ldots, 32$ training samples were tested to compute the sub-gradient at each iteration. The parameters yielding the best performance on validation sets were adopted: QDF-Single adopted one out of 40 query classes, QDF-Mixture used top 10 out of 50 query classes, QDF-KNN adopted 5 nearest neighbors, and m' was set to 5 in QDF-Reg.

We then fused said QDF approaches with document-dependent weights to form their respective D-QDDF approaches. We tested different β (Eq. 10) values $(\beta = 0, 0.1, \ldots, 1)$ to examine the importance of query- and document-dependent weights in the fusion process. As shown in Fig. 5, the performances of all D-QDDF-LNR approaches (using the training data set of 200K) follow a similar trend as β varies and reach the optimum at $\beta = 0.9$. The same optimum was observed using training data sets of different sizes. This importance factor reflects the primary role of query-dependent weights in the retrieval process but also illustrates the necessity for document-dependent weights in improving fusion strategy. In the following experiments, β is set to 0.9.



(a) D-QDDF vs QDF-Single (b) D-QDDF vs QDF-Mixture (c) D-QDDF vs QDF-KNN

(d) D-QDDF vs QDF-Reg (e) D-QDDF-MUL comparison (f) D-QDDF-LNR comparison

Fig. 6. Retrieval accuracy (MAP) comparison between D-QDDF and QDF approaches. (a) $\sim$ (d) show that by integrating document dependence how much D-QDDF approaches improve upon the four QDF approaches. (e) and (f) compare the performance of D-QDDF-MUL and D-QDDF-LNR approaches based on different QDF approaches.

2) D-QDDF Evaluation: We tested the performances of D-QDDF-LNR and D-QDDF-MUL with training data sets of different sizes. As shown in Fig. 6 (a) $\sim$ (d), by integrating document-dependent weights, the performances of QDF-Single, QDF-Mixture, and QDF-KNN are improved by 2.2%, 1.9%, and 5.5%, respectively, using D-QDDF-LNR, and by 2.2%, 1.2%, and 6.2%, respectively, using D-QDDF-MUL. D-QDDF-LNR outperforms QDF-Reg when using training data sets of all sizes with an average improvement of 1.7%. On the other hand, D-QDDF-MUL improves upon QDF-Reg by as much as 3.7% when using the training data set of 0.4K and performs as well as QDF-Reg when the size of training data set reaches 6.4K. However, as shown in our later experiment (Table V), when further comparing their performances with respect to query types, QDDF approaches can outperform QDF-Reg for most types of queries.

The performances of all approaches follow an overall ascending trend as the size of training data set grows. For example, having more training samples consistently yields better regression models for both QDF-Reg and the D-QDDF approaches based on QDF-Reg. The performances of QDF-Single, QDF-Mixture, and QDF-KNN were not always enhanced with larger data sets as their fusion weights are determined by a class or several classes of queries. By introducing document-dependent weights, both

D-QDDF-MUL and D-QDDF-LNR exhibit similar performance fluctuations with the corresponding QDF approaches over different training data sets. This can be explained by the dual-phase learning procedure. Document-dependent weights, which capture the importance of different modalities of a document, can be considered as an intrinsic characteristic of a document and is independent of how query-dependent weights are derived. Therefore, integrating document-dependent weight does not affect the variances in performance caused by different query training data sets.

Fig. 6 (e) and (f) compare the retrieval performances of D-QDDF-MUL and D-QDDF-LNR, respectively, based on different QDF approaches. We observe that D-QDDF approaches based on QDF-Reg achieve the best performances in both cases and are thus used in our later experiments.

B. Regression-based QDDF (R-QDDF)

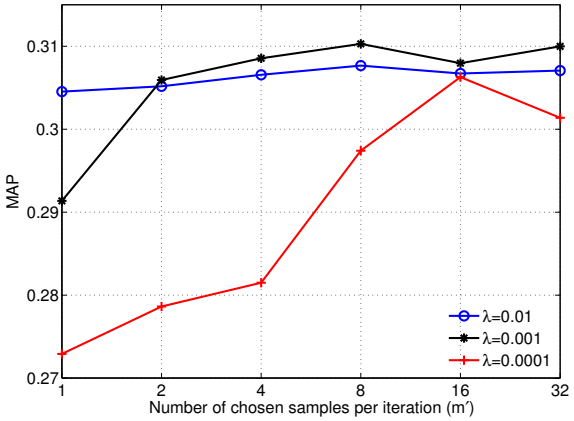


Fig. 7. Retrieval accuracy (MAP) of R-QDDF (trained with the data set of 200K) using different regularization factors (λ) and numbers of chosen samples per iteration (m').

1) On Effects of Parameter Tuning: We examined R-QDDF's performance using different regularization factors ($\lambda = 0.0001, 0.001, 0.01$) and with different numbers of samples ($m' = 2^i$, and $i = 0, 1, \ldots, 5$) in each iteration the validation data sets. Fig. 7 compares the performance of R-QDDF (using training data set of 200K) under different parameter settings and indicates that it is influenced by both factors. As m' increases, the performance shows a roughly growing trend. The smoothness of this growth can be influenced by different regularization factors. For example, after m' reaches 4, R-QDDF performs relatively smoothly for $\lambda = 0.01$ and $\lambda = 0.001$, but changes dramatically for $\lambda = 0.0001$. Grid search determined that the optimal values of λ and m' to be 0.001 and 8, respectively, for all sizes of training

TABLE V

RETRIEVAL ACCURACY (MAP) OF QDDF, QDF, DDF, AND QDIF WITH RESPECT TO QUERY TYPES WHEN THE SIZE OF TRAINING DATA IS 200K. G, M, V, I REPRESENT THE FOUR MUSIC DIMENSIONS: GENRE, MOOD, VOCALNESS, AND INSTRUMENT. ALL INDICATES ALL QUERY TYPES. BOLD VALUE REPRESENTS THE HIGHEST MAP AMONG ALL METHODS GIVEN A QUERY TYPE.

Method	All	G	M	V	I	GM	GV	GI	MV	MI	VI	GMV	GMI	GVI	MVI	GMVI
D-QDDF-LNR	**.327**	.639	.486	.562	.848	**.326**	.349	.438	**.398**	**.447**	.474	**.284**	**.304**	.326	**.347**	**.258**
D-QDDF-MUL	.323	.647	.480	.559	.852	.325	**.353**	.440	.388	.447	**.481**	.273	.298	**.328**	.342	.252
R-QDDF	.310	**.663**	.487	**.605**	**.913**	.294	.306	.427	.378	.427	.467	.258	.280	.310	.329	.246
QDF-Reg	.324	.647	.472	.551	.843	.317	.343	**.441**	.386	.440	.471	.280	.301	.324	.343	.254
DDF	.315	.643	**.488**	.569	.847	.314	.329	.435	.390	.437	.470	.263	.288	.318	.335	.245
QDIF	.311	.653	.473	.555	.844	.283	.297	.433	.374	.430	.469	.257	.287	.323	.331	.247

TABLE VI

IMPROVEMENT t-TEST OF QDDF APPROACHES (D-QDDF-LNR, D-QDDF-MUL, R-QDDF) OVER QDF-REG, DDF, AND QDIF WITH RESPECT TO QUERY TYPES WHEN THE SIZE OF TRAINING DATA IS 200K. G, M, V, I REPRESENT FOUR MUSIC DIMENSIONS: GENRE, MOOD, VOCALNESS, AND INSTRUMENT. ALL INDICATES ALL QUERY TYPES. + OR - INDICATE QDDF APPROACHES IMPROVE OR DETERIORATE THE PERFORMANCES OF THE OTHER APPROACHES. $<$ REPRESENTS $p < 0.05$, WHILE $>$ MEANS $p > 0.05$.

Method	Compared to	All	G	M	V	I	GM	GV	GI	MV	MI	VI	GMV	GMI	GVI	MVI	GMVI
D-QDDF-LNR	QDF-Reg	+<	-<	+<	+<	+<	+<	+<	->	+<	+<	+>	+<	+>	+>	+<	+<
	DDF	+<	->	->	-<	+>	+<	+<	+>	+<	+<	+<	+<	+<	+<	+<	+<
	QDIF	+<	-<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+>	+<	+<
D-QDDF-MUL	QDF-Reg	->	+>	+<	+<	+<	+<	+<	->	+>	+<	+<	-<	->	+<	->	->
	DDF	+<	+<	-<	-<	+<	+<	+<	+<	->	+<	+<	+<	+<	+<	+<	+<
	QDIF	+<	-<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<	+<
R-QDDF	QDF-Reg	-<	+<	+<	+<	+<	-<	-<	-<	-<	-<	->	-<	-<	-<	-<	-<
	DDF	-<	+<	->	+<	+<	-<	-<	-<	-<	-<	->	-<	-<	-<	-<	+>
	QDIF	->	+<	+<	+<	+<	+<	+<	-<	+<	->	->	+>	-<	->	->	->

data sets. The overall performance variances (using the training data set of 200K) are 0.262×10^{-6} and 0.026×10^{-6} across different validation folds and trials of random sampling, respectively.

2) R-QDDF Evaluation: We then tested the performances of R-QDDF using different sizes of training data sets and queries with different complexities.

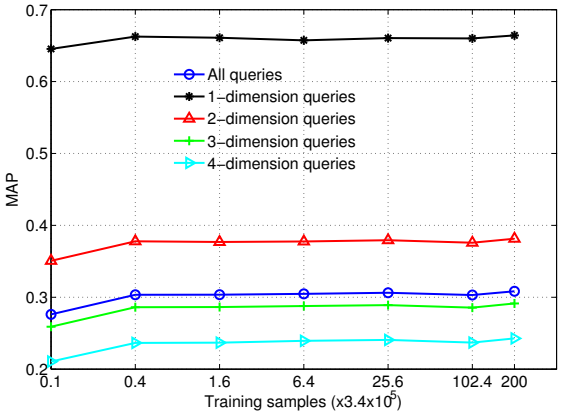


Fig. 8. Comparison of retrieval accuracy (MAP) given different query complexities when R-QDDF is trained with data sets of different sizes.

As shown in Fig. 8, the overall retrieval accuracy reveals a smoothly ascending trend as the training

data size grows, which is consistent with the regression-based learning principles. Given queries with different complexities, the retrieval accuracy shows a similar growing trend across different training data sets. Moreover, as the query complexity increases, the performance of R-QDDF decreases. In addition to R-QDDF, we observe the same phenomenon for all tested approaches (Table V). This is mainly because of the requirement to match more dimensions and the increasing sparsity of annotated ground truth relevance for complex queries. Since the training weights of R-QDDF were generated by comparing fusion scores with their ground truth relevance (Eq. 17), sparse and incomplete ground truth annotations result in sub-optimal training weights. R-QDDF based on different trials of random sampling shows similar performance. When using the training data set of 200K, the overall performance variances are 0.210×10^{-6} and 0.124×10^{-6} across different folds and trials, respectively.

C. QDDF vs. Other Approaches

Table V compares the retrieval accuracy of the proposed QDDF approaches along with the three existing methods (QDF, DDF and QDIF) given various query types. The data shows the performance of the approaches under the best parameter settings as determined by the evaluation (Section VI-A and Section VI-B). Using t-test, we assess whether each QDDF approach significantly improves (indicated by $+<$) upon QDF-Reg, DDF, and QDIF (Table VI).

We first observe that integrating query dependence or document dependence can both improve the retrieval performance. Among these three existing methods, QDF-Reg and DDF both outperform QDIF, with QDF-Reg achieving the best overall performance. While QDIF produces better higher MAPs for single dimension queries than QDF-Reg, the latter outperforms QDIF for all types of multi-dimension queries and does so significantly in all but two cases. DDF significantly improves QDIF for 11 out of 15 query types without bias towards the query types being tested, since document dependence is independent of query types. The comparison between QDF-Reg and DDF highlights the significance of document dependence in the fusion process especially for simple queries, for which DDF outperforms QDIF for all but one query type while QDF-Reg is outperformed in all cases.

When considering both query dependence and document dependence, we can observe that most of the best performance is achieved by QDDF approaches. D-QDDF-LNR and D-QDDF-MUL can significantly (Table VI) improve QDF-Reg, the best baseline method, by an average of 1.7% and 1.0%, respectively,

over all training data sets. The improvements of D-QDDF-LNR over DDF and QDIF can reach 3.8% and 5.1%, respectively. Overall, D-QDDF-LNR achieves the best performance among all tested approaches across both simple queries and complex queries. Regardless of the fusion strategy used, D-QDDF can achieve a finer mapping from query/document to fusion weight and thus yield fusion scores that better mirror the actual document-query relevance, resulting in superior performance.

In theory, R-QDDF also can benefit from such finer mapping. However, we note that R-QDDF only significantly outperforms all baseline methods (QDF-Reg, DDF and QDIF) for single-dimension queries, with the average improvements of 5.9%, 4.3%, and 5.4%, respectively. Compared with QDF-Reg, R-QDDF introduces more document information into the regression model but fails to compete with QDF-Reg on multiple-dimension queries. By integrating document dependence, R-QDDF allows documents to contribute to fusion weight derivation but at the same time expands the training space into the document dimension. Therefore, R-QDDF calls for much more comprehensive and thorough annotations for effective training. Meanwhile, because our dataset contains more annotations for single-dimension queries than those for complex queries and because documents have a larger influence in the query-document pairs of simple queries, R-QDDF's performance varies greatly between simple queries and complex queries. In contrast, D-QDDF explicitly fuses query- and document-dependent weights and avoids this training process. It thus outperforms R-QDDF on complex queries and also shows no significant preference among different query types. When comparing the results on simple queries, D-QDDF still fails to compete with R-QDDF as a result of the different fusion process – while R-QDDF fuses queries and documents at the feature level, D-QDDF does not do so until the very end. Therefore, D-QDDF may fail to identify and retrieve some relevant documents into the corresponding ranking lists as it considers less information when measuring the similarity between queries and documents. When dealing with simple queries, fewer modalities are related and the effect may thus become more obvious.

In summary, both R-QDDF and D-QDDF demonstrate the potential of document dependence but require better annotations and more effective fusion methods for effective multimodel fusion.

D. Efficiency Study

1) Efficiency Analysis of D-QDDF: Aside from calculating query-dependent weights, the most computationally intensive components of D-QDDF are: offline document-dependent weight derivation and

online weight fusion.

Let D denote our document collection and M the music social query space. Textual descriptive ability derivation was implemented by first building a sorted index of all the keywords in M then checking the existence of each word in the metadata of a document. Therefore, the computation complexity of this step for all documents is $O(|M|\log|M| + \gamma|D|\log|M|)$, where γ is the average number of words in the metadata of a document. Generally, we can assume $\gamma \ll |M|$, and the complexity becomes $O(|M|\log|M| + |D|\log|M|)$. To learn the relative scores between different modalities, we generated $|Q|$ queries and used n retrieval experts to retrieve relevant documents, resulting in $O(n|Q||D|)$. Weight generation from descriptive abilities takes $O(n|D|)$. Therefore, the total time complexity of calculating document-dependent weights for all documents is

$$O(|M|\log|M| + |D|\log|M|) + O(n|Q||D|) + O(n|D|)\ . \tag{18}$$

Since queries are generated from the music social query space, the following relation exists between $|Q|$ and $|M|$.

$$\begin{aligned} |Q| \leq |Q_T| &\approx \sum_{k=1}^{n/2} \binom{n/2}{k} |M_{av}|^k \times 1^{n/2-k} \\ &= (1+|M_{av}|)^{n/2} \approx (1+|M|/(n/2))^{n/2} \\ &\approx (\frac{2|M|}{n})^{n/2}\ , \end{aligned} \tag{19}$$

where $|Q_T|$ denotes the total number of queries that can be generated from the space, and $|M_{av}| \approx |M|/(n/2)$ represents the average number of keywords in each music dimension. Therefore, Eq. 18 equals to

$$O(n|D|(\frac{2|M|}{n})^{n/2})\ . \tag{20}$$

Since the number of retrieval experts n is generally small, the computational complexity is largely dependent on the size of the document database and the music social query space. In our study, the complete procedure took 550.725s (textual descriptive ability derivation: 30.192s; relative score learning: 519.910s; weight generation: 0.623s).

Given the weights from both queries and documents, the computational complexity of weight fusion

is $O(n)$ for each query-document pair. In our experiment, the average weight fusion time for a query-document pair is 0.020ms and 0.018ms in D-QDDF-MUL and D-QDDF-LNR, respectively.

2) Efficiency Analysis of R-QDDF: The R-QDDF approach involves regression model training and prediction. In our experiment, the training time using different data sets and the average prediction time for each query-document pair can be found in Table VII. Since we performed a further sampling from all the query-document pairs for training, the training time is the average runtime of three random sampling trials.

TABLE VII
TRAINING TIME AND AVERAGE PREDICTION TIME (PER QUERY-DOCUMENT PAIR) OF R-QDDF WHEN USING TRAINING DATA SETS OF DIFFERENT SIZES. NO. DENOTES THE SIZE OF DIFFERENT TRAINING DATA SETS $* = \times 3.4 \times 10^5$.

No.(*)	0.1	0.4	1.6	6.4	25.6	102.4	200
Training(s)	6.360	5.930	7.910	8.330	8.260	10.370	9.310
Prediction(ms)	0.074	0.078	0.079	0.082	0.082	0.083	0.082

As shown in Table VII, there is no significant change in the runtime for both training and prediction as the training data set gets larger. This can be explained by the learning principle of Pegasos. As analyzed in [10], [44], [45], regression Pegasos takes $O(1/(\lambda\epsilon))$ iterations to achieve an ϵ-accurate solution. Therefore, the training complexity is $O(m'|\mathbf{V}|/(\lambda\epsilon))$, where $|\mathbf{V}|$ denotes the size of the feature vector of a query-document pair. Theoretically, the runtime of regression Pegasos is independent of the size of training data set, and it would reveal a reduced training time to achieve a certain generalization performance when the training data set gets larger [45]. Given a trained regression Pegasos model, the prediction is calculated by the inner product of the model weight vectors and the feature vectors of testing query-document pairs, which results in $O(n|\mathbf{V}|)$ for each testing sample. The efficient training process and the linear prediction make R-QDDF scalable to large data sets and practical for real-life multimodal retrieval systems online.

E. Discussion

As the effectiveness study shows, both D-QDDF and R-QDDF are able to improve existing multimodal fusion approaches with respect to query types. In theory, D-QDDF cannot always derive the optimal fusion strategy by explicitly fusing query- and document-dependent weights. However, it can easily leverage existing QDF and DDF approaches and has achieved promising results for complex queries when compared to other approaches. R-QDDF can derive the optimal fusion strategy in theory, but in

reality its efficacy is compromised by the need for comprehensively and thoroughly annotated training samples. Therefore, its performance may vary among queries with different complexities. For example, our results have shown that R-QDDF consistently outperforms other approaches for simple queries.

For efficiency analysis, both approaches take a linear run time for testing. However, R-QDDF is more flexible when adapting to larger databases due to its intrinsic efficiency in both training and prediction. D-QDDF enjoys good efficiency when the document database is static. When documents are frequently updated or new documents join the database, incremental updates, such as those using weight propagation based on document similarity measure between new documents and the database, can be implemented to avoid a thorough update of the database documents. Another possible approach would be to first derive the descriptive abilities for textual modalities of these new documents and then adopt an average content-to-text ratio to calculate document-dependent weights. Complete update of all documents can be done when the number of new (or newly updated) documents reach a certain ratio.

Both D-QDDF and R-QDDF can be easily parallelized to further reduce their computational complexity. For instance, since each retrieval expert retrieves relevant documents on a particular modality, the retrieval processes of multiple experts can be parallelized. In addition, the weight calculation for queries and documents in D-QDDF can also function in parallel.

QDDF generalizes QDF, DDF, and QDIF approaches by considering the dependence of both queries and documents. As a general multimodal fusion framework, it can be applied to not only music retrieval but also other scenarios, such as web, image, video retrieval. As metadata and content are common modalities of different multimedia documents (e.g., image, videos), our implementation methods (i.e., dual-phase QDDF and regression-based QDDF) can be directly applied to these applications. The key step is to construct a social query space that comprehensively covers multiple modalities of the studied media documents. This way, in D-QDDF, the descriptive abilities of multiple modalities can be accurately captured, and R-QDDF model can also be trained with sufficient data. A similar method based on folksonomy data can be adopted to build such social query spaces. However, as the type and number of modalities may vary significantly across different media documents, building such social query spaces using purely folksonomy data may be challenging. For example, most tags of an image are related to the names of objects within the image, while texture information appears as tags far less frequently. Therefore, multiple online resources (such as tags, comments, descriptions) and the semantic relations

between different resources need to be explored. Human efforts may also be required to ensure the correctness and completeness of different modalities.

VII. Conclusions and Future Work

In this paper, we introduce a general multimodal fusion framework QDDF, which derives the optimal fusion strategy for each query-document pair by incorporating the content of both queries and documents. Existing approaches, such as QDF, DDF, and QDIF, are in fact special cases of QDDF when it relaxes the fusion weight dependence on documents, queries or both. We proposed two approaches to derive fusion strategies, D-QDDF and R-QDDF. Experimental results show D-QDDF can outperform the other approaches for most query types and R-QDDF is superior in handling single-dimension queries. Efficiency results also show the scalability of both approaches over large data sets.

Our proposed approaches may be improved in various aspects. For example, better descriptive ability derivation methods and relative score learning methods in D-QDDF may help to accurately capture the importance of different modalities of a document. Fusion methods may also be improved by investigating the importance of queries and documents in different multimodal retrieval systems. For R-QDDF, we would like to explore different feature representations of query-document pairs and construct better quality training data sets to evaluate its performance for complex queries. Since the modalities we investigated are common among most multimedia documents, both D-QDDF and R-QDDF can also be extended to other multimedia documents in different multimodal retrieval applications, such as meta search, and video/image retrieval.

Acknowledgment

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

[1] I. H. Kang and G. Kim, "Query type classification for web document retrieval," in *Proc. of ACM SIGIR*, 2003, pp. 64–71.

[2] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. of ACM Multimedia*, 2004, pp. 548–555.

[3] X. Olivares, M. Ciaramita, and R. van Zwol, "Boosting image retrieval through aggregating search results based on visual annotations." in *Proc. of ACM Multimedia*, 2008, pp. 189–198.

[4] B. Cui, L. Liu, C. Pu, J. Shen, and K.-L. Tan, "Quest: querying music databases by acoustic and textual features," in *Proc. of ACM Multimedia*, 2007, pp. 1055–1064.

[5] J. Shen, J. Shepherd, and A. H. H. Ngu, "Towards effective content-based music retrieval with multiple acoustic feature combination," *IEEE Trans. on Multimedia*, vol. 8, no. 6, pp. 1179–1189, 2006.

[6] N. Orio, D. Rizo, R. Miotto, M. Schedl, N. Montecchio, and O. Lartillot, "Musiclef: a benchmark activity in multimodal music information retrieval," in *Proc. of ISMIR*, 2011, pp. 603–608.

[7] J. A. Shaw and E. A. Fox, "Combination of multiple searches," in *Proc. of TREC-2*, 1994, pp. 243–252.

[8] C. W. Ngo, Y. G. Jiang, X. Y. Wei, F. Wang, W. Zhao, H. K. Tan, and X. Wu, "Experimenting vireo-374: Bags-of-visual-words and visual-based ontology for semantic video indexing and search," in *NIST TRECVID Workshop*, 2007.

[9] T. S. Chua, S. Y. Neo, K. Y. Li, G. Wang, R. Shi, M. Zhao, H. Xu, Q. Tian, S. Gao, and T. L. Nwe, "Trecvid 2004 search and feature extraction task by nus pris," in *NIST TRECVID Workshop*, 2004.

[10] B. Zhang, Q. Xiang, H. Lu, J. Shen, and Y. Wang, "Comprehensive query-dependent fusion using regression-on-folksonomies: a case study of multimodal music search," in *Proc. of ACM Multimedia*, 2009, pp. 213–222.

[11] S. Wu and F. Crestani, "Data fusion with estimated weights," in *Proc. of CIKM*, 2002, pp. 648–651.

[12] Z. Li, B. Zhang, and Y. Wang, "Document dependent fusion in multimodal music retrieval," in *Proc. of ACM Multimedia*, 2011, pp. 1105–1108.

[13] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[14] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. of ACM Multimedia*, 2005, pp. 399–402.

[15] R. Neumayer and A. Rauber, "Integration of text and audio features for genre classification in music information retrieval," in *Proc. of ECIR*, 2007, pp. 724–727.

[16] R. Mayer, R. Neumayer, and A. Rauber, "Combination of audio and lyrics features for genre classification in digital audio collections," in *Proc. of ACM Multimedia*, 2008, pp. 159–168.

[17] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Proc. of ICMLA*, 2008, pp. 688–693.

[18] X. Hu and J. S. Downie, "Improving mood classification in music digital libraries by combining lyrics and audio," in *Proc.of JCDL*, 2010, pp. 159–168.

[19] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Trans. on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

[20] B. Zhang, J. Shen, Q. Xiang, and Y. Wang, "Compositemap: a novel framework for music similarity measure," in *Proc. of ACM SIGIR*, 2009, pp. 403–410.

[21] R. Mayer and A. Rauber, "Music genre classification by ensembles of audio and lyrics features," in *Proc. of ISMIR*, 2011, pp. 675–680.

[22] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," vol. 20, no. 3, pp. 226–239, 1998.

[23] K. Cai, J. Bu, C. Chen, and P. Huang, "Automatic query type classification for web image retrieval," in *Proc. of MUE*, 2007, pp. 1021–1026.

[24] Y. Wu, C.-Y. Lin, E. Y. Chang, and J. R. Smith, "Multimodal information fusion for video concept detection," in *Proc. of IEEE ICIP*, 2004, pp. 2391–2394.

[25] Y. Wu, E. Y. Chang, K. C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. of ACM Multimedia*, 2004, pp. 572–579.

[26] B. McFee and G. Lanckriet, "Heterogeneous embedding for subjective artist similarity," in *Proc. of ISMIR*, 2009, pp. 513–518.

[27] B. T. Bartell, G. W. Cottrell, and R. K. Belew, "Automatic combination of multiple ranked retrieval systems," in *Proc. of ACM SIGIR*, 1994, pp. 173–181.

[28] ——, "Learning to retrieve information," in *Proc. of the Swedish Conference on Connectionism*, 1995, pp. 345–354.

[29] A. B. Benitez, M. Beigi, and S.-F. Chang, "Using relevance feedback in content-based image meta-search," *IEEE Internet Computing*, vol. 2, no. 4, pp. 59–69, 1998.

[30] L. S. Kennedy, A. P. Natsev, and S. F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proc. of ACM Multimedia*, 2005, pp. 882–891.

[31] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, "Learning collection fusion strategies," in *Proc. of ACM SIGIR*, 1995, pp. 172–179.

[32] R. Yan and A. G. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," in *Proc. of ACM SIGIR*, 2006, pp. 324–331.

[33] L. Kennedy, S. F. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 567–588, 2008.

[34] A. Smeaton and P. Over, "TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video," in *Proc. of CIVR*, 2003, pp. 451–456.

[35] L. Xie, A. Natsev, and J. Tesic, "Dynamic multimodal fusion in video search," in *Proc. of IEEE ICME*, 2007, pp. 1499–1502.

[36] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, "Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval," in *Proc. of ACM SIGIR*, 2005, pp. 512–519.

[37] ——, "Meta-search and federation using query difficulty prediction," in *Proc. of ACM SIGIR Query Prediction Workshop*, 2005.

[38] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[39] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[40] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

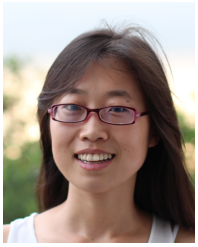
[41] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[42] S. E. Rrobertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," in *NIST Special Publication 500-236:TREC-4*, 1995, pp. 73–96.

[43] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 3, pp. 169–175, 2000.

[44] S. S. Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for svm," in *Proc. of ICML*, 2007, pp. 807–814.

[45] S. S. Shwartz and N. Srebro, "SVM optimization: inverse dependence on training set size," in *Proc. of ICML*, 2008, pp. 928–935.



Zhonghua Li received her B.S. degree in Computer Science from Department of Computer Science, Northwestern Politechnical University, China, in June 2008. She joined the Sound and Music Computing (SMC) Lab at National University of Singapore as a Ph.D. candidate in August 2008. Currently, she is working with her superviser, Dr. Ye Wang, on music information retrieval. Her research interests include music signal processing, information retrieval, multimodal fusion, and music recommendation.

Bingjun Zhang received his Ph.D. degree on music information retrieval under the supervision of Dr. Ye Wang, School of Computing, National University of Singapore in 2010. After that he joined Mozat Pte Ltd. In the first one and a half year with Mozat, he worked as a research engineer focusing on design and development of large scale data storage services. In the next one year with Mozat, he worked as the business head of middle-east market and was in charge of the main revenue stream of the company. From May 2012, he co-founded iCarsclub and since then on he has been a passionate entrepreneur to bring the great benefits of peer to peer car sharing to the Asian society. More details can be found on his LinkedIn profile: http://sg.linkedin.com/in/bingjun/

Yi Yu received the Ph.D. degree in Computer Science in 2009 from Nara Women's University. She worked at different institutions including New Jersey Institute of Technology, University of Milan and Nara Women's University. She currently works at School of Computing, National University of Singapore. Her research interests include social interactions over geo-aware multimedia streams, multimedia/music signal processing, audio classification and tagging, locality sensitive hashing-based music information retrieval, and pest sound classification. She received a best paper award from IEEE ISM 2012.

Jialie Shen is an Assistant Professor in Information Systems, School of Information Systems, Singapore Management University, Singapore. He received his Ph.D. in Computer Science from the University of New South Wales (UNSW), Australia in the area of large-scale media retrieval and database access methods. Jialies main research interests include information retrieval in the textual and multimedia domain, economic-aware media analysis, artificial intelligence (particularly machine perception and its applications on IR and business intelligence) and multimedia systems. His recent work has been published or is forthcoming in leading journals and international conferences including ACM SIGIR, ACM Multimedia, ACM SIGMOD, CVPR, ICDE, WWW, IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT), IEEE Transactions on Multimedia (IEEE TMM), IEEE Transactions on Image Processing (IEEE TIP), ACM Multimedia Systems Journal, ACM Transactions on Internet Technology (ACM TOIT) and ACM Transactions on Information Systems (ACM TOIS). Besides being chair, PC member, reviewer and guest editor for several leading information systems journals and conferences, he is an associate editor of International Journal of Image and Graphics (IJIG) and area editor of Electronic Commerce Research and Applications (ECRA).

Ye Wang is an Associate Professor in the Computer Science Department at the National University of Singapore (NUS) and NUS Graduate School for Integrative Sciences and Engineering (NGS). He received his PhD in Information Technology from Tampere University of Technology, Finland. He established and directed the sound and music computing (SMC) Lab at NUS School of Computing. Before joining NUS he was a member of the technical staff at Nokia Research Center in Tampere, Finland for 9 years. His research interests include sound analysis and music information retrieval (MIR), mobile computing, and cloud computing, and their applications in music edutainment and eHealth, as well as determining their effectiveness via subjective and objective evaluations. Dr. Wang has served as editorial board members of Journal of Multimedia and Journal of New Music Research, as well as on the program committees of international multimedia conferences regularly. He has co-authored the Best Student Paper at ACM Multimedia 2004, and the Best Paper at IEEE International Symposium on Multimedia 2012.