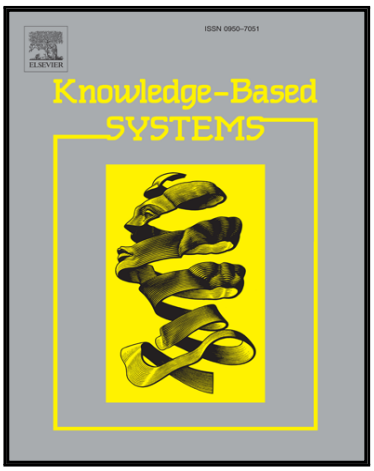


Background Music Recommendation Based on Latent Factors and Moods

Chien-Liang Liu, Ying-Chuan Chen

PII: S0950-7051(18)30348-4
DOI: 10.1016/j.knosys.2018.07.001
Reference: KNOSYS 4375

To appear in: *Knowledge-Based Systems*

Received date: 21 July 2017
Revised date: 29 June 2018
Accepted date: 3 July 2018

Please cite this article as: Chien-Liang Liu, Ying-Chuan Chen, Background Music Recommendation Based on Latent Factors and Moods, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.07.001

Highlights

- Devise a background music recommendation algorithm based on latent factor model.
- Design several experiments to assess the performance and the effectiveness of the proposed algorithm.
- The proposed model is promising in accuracy and quantitative research.

Background Music Recommendation Based on Latent Factors and Moods

Chien-Liang Liu[a,1,*], Ying-Chuan Chen[b]

[a]*Department of Industrial Engineering and Management, National Chiao Tung University, 1001 University Road, Hsinchu 300, Taiwan (R.O.C.)*
[b]*Department of Computer Science, National Chiao Tung University,1001 University Road, Hsinchu 300, Taiwan (R.O.C.)*

Abstract

Many mobile devices are equipped with video shooting function, and users tend to use these mobile devices to produce user generated content (UGC), and share with friends or the public owing to the popularity of social media. To make the video to be attractive, embedding appropriate background music into the video is a popular way to enrich user experience, but it is a time-consuming and labor-intensive task to find music that fits the video. This work proposes to use latent factors to recommend a list of music songs for a given video, in which the recommendation is based on the proposed score function, which involves the weighted average of the latent factors for the video and music. Moreover, we use pairwise ranking to design the objective function, and use stochastic gradient descent to optimize the proposed objective function. In the experiments, we specify two hypotheses and design several experiments to assess the performance and the effectiveness of the proposed algorithm from different aspects, including accuracy, quantitative research, and qualitative research. The experimental results indicate that the proposed model is promising in accuracy and quantitative research. Furthermore, this work provides detailed analysis to investigate the fitness of the background music that recommended by the system through interviewing the subjects.

Keywords: background music recommendation system, moods, latent

*I am corresponding author
Email addresses: clliu@mail.nctu.edu.tw (Chien-Liang Liu), edward.chenyc@gmail.com (Ying-Chuan Chen)

factor model, recommender systems, collaborative filtering, multimodal information retrieval

1. Introduction

Multimedia retrieval has attracted much of attention over decades (Rui et al., 1999; Smeulders et al., 2000; Paulin et al., 2017; Cheng et al., 2016; Skowron et al., 2017) due to the huge collections of multimedia data available at hand. The multimedia data is presented in different modalities such as videos, images and sounds, so many information systems about multimedia require to deal with multiple modalities simultaneously. For example, many search engines allow users to search for images with text queries, and the image search involves two modalities, namely, image and text. As compared with traditional information retrieval, multimodal information retrieval (MMIR) is about searching for information in any modality on web with combinations of two or more retrieval models (Bokhari and Hasan, 2013).

With the improvement of the technology, more and more mobile devices are equipped with video shooting function, and users tend to use these mobile devices to produce user generated content (UGC). Moreover, many people tend to share their videos with friends or the public owing to the popularity of social media, so mixing videos with appropriate music files as background music is a popular way to enrich user experience. However, selecting appropriate background music to supplement the content is a cumbersome process. For the users who would like to embed background music in the videos, they have to look for the music manually. Nevertheless, this is a time-consuming and labor-intensive task, explaining why this work focuses on constructing a system to help users find the appropriate background music easier and faster.

Traditional information retrieval system could directly use text query to retrieve relevant documents, since both query and document belong to the same type and could be represented as semantic feature vectors, giving a base for the system to calculate the similarity between query and documents. In contrast, the retrieval problem presented in this work involves two data types, namely, video and music. The videos comprise image frames, each of which is made of pixel; while the music is composed of sounds with a fundamental frequency and overtones. Moreover, another challenging task is that pixels and frequencies fail to reveal semantic information as provided by texts.

This work proposes to use machine learning to build the model, so we have to transform the data into feature representations. Key frames are

extracted to represent the videos, and features for background music should be extracted as well. To associate the features coming from multimodal retrieval, this work proposes to use latent factors in the proposed model owing to the success of latent factors in many application domains. The latent factor models have been widely used in machine learning, which relate observable data to a set of latent variables. The goal is to infer latent factors from a collection of observable data, and these latent factors often help practitioner find insights in data. For example, topic modeling (Blei et al., 2003) discovers the topics that occur in a collection of documents, and topic modeling has been a cornerstone of information retrieval.

In the proposed model, each feature of videos and music is characterized by a latent factor, and each video or music is a linear combination of latent factors of associated features. The key idea is to map videos and music onto a latent feature space, and then calculate the similarities on the new feature space. The previous research studies have shown that using latent factor models could benefit from good scalability while keeping high predictive performance (Rendle, 2010a). Based on latent factors and the characteristics of the problem, we formulate the retrieval problem as an optimization problem, and propose to use stochastic gradient descent (SGD) to optimize the proposed objective function.

Training data is required to train a machine learning model. To label data, one requires to associate background music with videos, but this is a challenging task since both audios and videos are highly related to personal interests and emotional state. Audio and visual signals are the two major indicators of human emotion (Guan et al., 2009), so we use emotion as the surrogate to associate videos and music to obtain positive combinations and negative combinations, all of which are used as training examples. Once the labeling process is completed, we use the training data to train the proposed model. In the testing phase, the proposed model could generate a music list for a specific video.

In performance assessment, accuracy is a commonly used metric, but the fitness between background music and videos involves subjective factors. Consequently, this work considers not only accuracy, but also user experience by inviting subjects to participate the experiments to perform quantitative research and qualitative research. Meanwhile, the subjects have to complete a questionnaire after watching videos embedded with recommended background music, and interviewed by us to understand the criteria of their judgements for further analysis.

The main contributions of this work are listed as follows. First, we propose to use latent factor model to deal with multimedia retrieval problem, in which a list of appropriate music songs is recommended for a given video. In the proposed model, we use pairwise ranking to design the objective function, and use stochastic gradient descent to optimize the proposed objective function owing to the consideration of training time. Second, we conduct experiments to compare with several classical methods, and the experimental results indicate that the proposed method outperforms the other alternatives in accuracy. In the experiments, labeled examples are required for model training, but defining whether a music song is relevant to a specific video is influenced by subjective factors, such as personal interest and emotional state. This work proposes to use emotion as a surrogate to label the data, and we believe that other research studies that are related to emotion could benefit from this approach to label data samples. Third, to further understand subjects' responses, we conduct quantitative and qualitative experiments. The proposed method shows competitive performance on accuracy evaluation, and significantly fits users' preferences. Furthermore, we interview the subjects to investigate how and why they select the songs they like or dislike to perform the qualitative research. This work conducts a thorough and comprehensive analysis of the proposed method to understand the reasons that user may consider when selecting background music.

The rest of this paper is organized as follows. Section 2 presents related surveys and techniques. Section 3 introduces the proposed algorithm. Section 4 describes the architecture of the proposed system. Section 5 shows the experimental results. We further discuss the experimental results in Section 6, and the conclusions and future work are presented in Section 7.

2. Related Work

This work proposes to use latent factors to recommend background music to a given video, so we introduce multimedia retrieval and latent factor model in the following sections.

2.1. Multimedia Retrieval

The last decade has witnessed a massive explosion of multimedia content on the web, and multimedia retrieval is always an important research topic. In multimedia retrieval, using text-based representations to retrieve multimedia is still the most popular means, since users are used to using text

queries to retrieve relevant content in search engines. For example, multimedia search provided by search engines such as Google or Bing allows users to issue text query, and the search engines return matched images or videos. Consequently, many researchers have devoted to the design of models that can account for multiple content modalities. For example, Rasiwasia et al. (2010) used the hidden topic model learned with latent Dirichlet allocation (LDA) (Blei et al., 2003) to represent text component, and used bags of visual features to represent images. Once the feature representations for the two components are completed, the correlations between the two components are learned with canonical correlation analysis (CCA) (Thompson, 2005). Collaborative filtering (CF) suffers from cold-start problem, as new items do not have feedback information, and new users do not have historical behavior information. Oramas et al. (2017a) proposed to use deep network approach to tackle this problem. In the network architecture, the artist embeddings are learned from biographies and track embeddings are learned from audio signals, and then both feature embeddings are combined in a multimodal network to predict song recommendations of cold-start artists.

Most mobile devices are equipped with cameras, and users have used to using mobile devices to make user-generated videos (UGVs) and share with friends. To make UGVs more attractive, Shah et al. (2014) proposed a personalized video soundtrack recommendation system to replace the ambient background noise of a UGV with a soundtrack. The key idea is to predict video scene moods based on the combination of geo-categories and video visual features, in which geo-categories reflects the environmental atmosphere associated with moods and a color histogram is used as a visual feature to represent the mood in video content. Then, the concatenated geo-categories and visual features are the inputs of SVM^{hmm} (Joachims et al., 2009), and the output is to predict a sequence of scene moods. The predicted mood tags are used by music recommendation sub-system to find the music tracks relevant to a UGV from a database. Their system focuses on extracting moods from geographic context and video content, so it is different from the proposed method. However, their work inspires us to use emotion as the surrogate to label data. Many mobile devices are embedded with global positioning system (GPS), explaining why many researchers developed context-aware recommendation systems (Liu et al., 2010; Cheng and Shen, 2016). For example, Liu et al. (2010) considered to use location and emotion to propose an interactive music system. Park et al. (2007) developed a location-based recommendation system, in which the recommendation is based on location,

time, user mood and contextual information. Cheng and Shen (2016) also focused on location-aware music recommendation system, which identifies suitable songs for various types of popular venues. They proposed to use topic modeling (Blei et al., 2003) technique to extract features of songs, and devised an algorithm called location-aware topic model (LTM) to capture the connections between the venues and the music. Benzi et al. (2016) formulated a song recommender system as a matrix completion problem, and proposed to use non-negative matrix (NMF) (Lee and Seung, 1999) technique to tackle this problem, in which they used a weighted Kullback-Leibler (KL) divergence as distance metric and considered the outside information given by the songs and playlists graphs as regularization terms. Although recommender systems are popular in many application domains, recommendation techniques have been rarely used in the business-to-business (B2B) environment, since items or user profiles often present complicated structures in B2B applications. For example, the product categories are always presented as tree structures, and online users preferences are often vague and fuzzy. Wu et al. (2015) proposed a novel method for modeling fuzzy tree-structured user preferences and developed a Web-based B2B recommender system. Besides the applications mentioned above, Lu et al. (2015) have conducted a thorough and comprehensive analysis of recommender system applications, including up-to-date application developments of recommender systems, and eight main application domains of recommender systems.

Many research studies have focused on the recommendation of background music over the last decade (Guan et al., 2017; Lee et al., 2017; Åman and Liikkanen, 2017; Cheng and Shen, 2016; Yoshida and Hayashi, 2016), since background music could make the content more impressive. For example, Yoshida and Hayashi (2016) developed a system called OtoPittan, which recommends background music for video based on the valence and arousal model. They considered to use 20 audio features along with 30 visual features to represent a feature vector, and used principal component analysis (PCA) to reduce the dimension of the feature vector. The recommendation is based on impression, which is represented as a pair of the valence level and the arousal level, and the prediction of impression is based on multiple regression model. Lin et al. (2014) proposed a semantic-based home video background music recommender system, which uses texts to represent semantics of the videos and music. In video aspect, they used pre-defined objects (Li et al., 2010) to recognize the corresponding objects in the input video. As a result, the name of each object is viewed as the text semantics

of the input video. On the other hand, they used lyrics to represent the semantics of the songs. Once videos and music are represented as text representations with tf-idf (Salton and McGill, 1986), the similarity between videos and music could be obtained. Moreover, the post-process stage used motion-direction histogram (MDH) and pitch tempo pattern (PTP) to process the motion of videos and tempo of candidate music, respectively. By calculating the similarity between these two features with approximate string matching (ASM) (Yeh and Cheng, 2008), they re-ranked the candidate list to get the final recommendation. The system proposed by Lin et al. (2014) relies on pattern recognition technique to identify the objects presented in the video frames, such that the subsequent matching procedure could be achieved in terms of texts. It is apparent that their system is different from the proposed work. Some researchers (Lin and Shan, 2017; Kuo et al., 2013) proposed to use correlation-based approach to discover the correlation between audio and video, in which the correlation is obtained by multiple-type latent semantic analysis (MLSA) (Wang et al., 2006) and cross-modal factor analysis (CFA). In recommendation, side information about items has proven to boost the performances of collaborative-filtering techniques (Ning and Karypis, 2012). Oramas et al. (2017b) proposed to use knowledge graph to enrich the data, in which the semantic entities are extracted from item textual descriptions and linked to external knowledge graphs such as WordNet (Miller, 1995) and DBpedia (Auer et al., 2007) for gathering additional knowledge.

2.2. Latent Factor Model

Collaborative filtering (CF) recommender systems attracted much of attention in the past decade (Zhang and Min, 2016; Hernando et al., 2016; Pirasteh et al., 2015). Matrix factorization (Koren et al., 2009) is one of the most popular collaborative filtering methods, in which the users and items are both mapped onto a latent space, and calculate the similarities with inner products. The matrix factorization is one of the main algorithms used by the winner of Netflix competition, explaining why many research studies have used latent factors to devise algorithms over the last decade (Liu and Wu, 2016b; Rendle et al., 2009).

Factorization machines (FM) (Rendle, 2010b; Rendle et al., 2011) combines the advantages of support vector machines (SVM) with factorization models, and it has become a state-of-the-art algorithm in recommender systems. Chen et al. (2012) proposed a framework called feature-based matrix

factorization to incorporate various features into the model. The SVDFeature is an extension of FM. Liu and Wu (2016a) combined latent factors and ℓ_2 norm to formulate the recommendation problem as a k-nearest-neighbor (kNN) problem, and proposed a novel recommendation algorithm, which uses locality-sensitive hashing (LSH) (Indyk and Motwani, 1998) to reduce search time complexity. Rendle et al. (2009) proposed a pairwise ranking algorithm called Bayesian Personalized Ranking (BPR) to use learning-to-rank technique to consider the ordering of the recommendation items. Deep learning has become an active research topic in machine learning, and Van den Oord et al. (2013) proposed to use convolutional neural networks (CNNs) to predict latent factors from music audio when latent factors cannot be obtained from usage data. It is apparent that the purpose of this work is different from our approach, since the purpose of this work is to predict latent factors for a given song from the corresponding audio signal, while we propose to use latent factor to associate music and videos. Moreover, the deep learning approaches always require enormous training examples for model training, and that may be a problem for some application domains, since data labeling is a time-consuming and labor-intensive task.

3. Background Music Recommender System

This section describes the problem and the proposed algorithm in this work.

3.1. Problem Specification

Let $\mathbf{V} = \{v^{(1)}, \ldots, v^{(|\mathbf{V}|)}\}$ be the set of videos, and $\mathbf{M} = \{m^{(1)}, \ldots, m^{(|\mathbf{M}|)}\}$ the set of music. Given a video $v \in \mathbf{V}$, the goal is to find appropriate music songs from $\mathbf{M}$ that are relevant to v. We can transform this problem into a machine learning problem, and the goal is to learn a scoring function $f : \mathbf{V} \times \mathbf{M} \to \mathcal{Y}$, in which $\mathcal{Y} \in \mathbb{R}$. The output score is the relevance score of a given music song with respect to a given video, and a high score indicates high relevance.

Given N training examples, each of which can be represented as a tuple $\{(v, m, y)\}$, where video $v \in \mathbf{V}$, music $m \in \mathbf{M}$, and $y \in \mathcal{Y}$. This work focuses on binary classification problem, namely $y \in \{0, 1\}$, in which $y = 0$ denotes negative class and $y = 1$ is positive class. Moreover, the system should provide a list of matching music for a specific video.

Once the data labeling process is completed, one can obtain the training set $\mathbf{X} = \{\mathbf{x} = (v, m, y_{vm})\}$, in which y_{vm} is the relevance between music m and video v, $y_{vm} = 1$ means positive relevance, and $y_{vm} = 0$ means negative relevance.

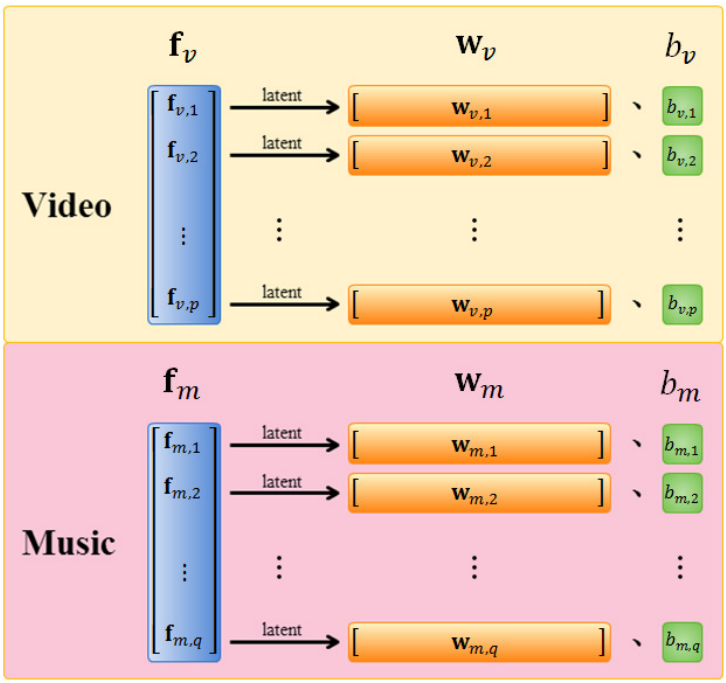


Figure 1: An illustration of features and their corresponding latent vectors

3.2. Scoring Function

This work proposes to use latent factor model to find the relevance between videos and music. The basic assumption behind the proposed method is that one can project both videos and music onto a latent emotional space, and calculate the relevance scores on that space.

Let $\mathbf{f}_v \in \mathbb{R}^p$ denote the feature vector for video v, and $\mathbf{f}_m \in \mathbb{R}^q$ denote the feature vector for music m. Each feature in $\mathbf{f}_v$ and $\mathbf{f}_m$ can be mapped to a latent vector $\mathbf{w} \in \mathbb{R}^k$ and a bias term $b \in \mathbb{R}$, where k is the number of latent factors. Figure 1 illustrates the mappings of features as described above.

Based on the latent factor concept, this work proposes a scoring function as listed in Equation (1).

$$f(v, m) = \mathbf{w}_v^{\mathbf{T}} \mathbf{w}_m + b_v + b_m \tag{1}$$

, where

$$\begin{aligned}
\mathbf{w}_v &= \frac{1}{\sum_i \mathbf{f}_{v,i}} \sum_i \mathbf{f}_{v,i} \mathbf{w}_{v,i} \\
\mathbf{w}_m &= \frac{1}{\sum_i \mathbf{f}_{m,i}} \sum_i \mathbf{f}_{m,i} \mathbf{w}_{m,i} \\
b_v &= \frac{1}{\sum_i \mathbf{f}_{v,i}} \sum_i \mathbf{f}_{v,i} b_{v,i} \\
b_m &= \frac{1}{\sum_i \mathbf{f}_{m,i}} \sum_i \mathbf{f}_{m,i} b_{m,i}.
\end{aligned}$$

As presented in the above equations, we use a vector $\mathbf{w}_v$ to denote a video v, in which $\mathbf{w}_v$ is the weighted average of all $\mathbf{w}_{v,i}$, $i = 1, \ldots, q$. The same setting is applied to music m and the two bias terms, namely b_v and b_m. Note that the parameters in the proposed scoring function are denoted as $\mathbf{\Theta} = \{\mathbf{w}_v, \mathbf{w}_m, b_v, b_m\}$.

As presented in Equation (1), the similarity or relevance between video v and music m is mainly determined by the inner product. When the outcome of inner product is larger, it means that the two latent vectors are closer to each other. In that case, it is more likely that v and m have positive correlation and vice versa. To achieve the goal, this work uses learning-to-rank approach to design loss function.

3.3. Loss Function

We introduce a symbol $\succ$ to define the rank of data, i.e., assume that $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$, where $\mathbf{X} = \{\mathbf{x} = (v, m, y_{vm})\}$, if the rank of $\mathbf{x}_i$ is higher than $\mathbf{x}_j$, it will be denoted by $\mathbf{x}_i \succ \mathbf{x}_j$. Normally, there will be a function $f \in \mathbf{F}$ which determines the rank of data. If we apply the above method to our scoring function, the goal is to make $\forall \mathbf{x}_i \succ \mathbf{x}_j : f(\mathbf{x}_i|\Theta) > f(\mathbf{x}_j|\Theta)$. To achieve this goal, we have to learn a proper latent factor parameter set Θ.

This work uses pairwise hinge loss function (Joachims, 2002) presented in Equation (2) as the loss function, in which we denote the positive data with

+ and negative data with −, that is, $f(+)$ means $f(v^{(+)}, m^{(+)})$ and $f(-)$ is $f(v^{(-)}, m^{(-)})$. The goal is to let the scores of positive examples be larger than those of negative examples.

$$\ell(\Theta) = \max(0, 1 + f(-) - f(+)) \tag{2}$$

Equation (2) can be rewritten as Equation (3).

$$\ell(\Theta) = \begin{cases} f(-) - f(+) + 1 & \text{, if } 1 > f(+) - f(-) \\ 0 & \text{, otherwise} \end{cases} \tag{3}$$

, where 0 is non-differentiable. We only focus on the differentiable part, which is the case $1 > f(+) - f(-)$, since $\ell(\Theta) = 0$ means the model is able to rank the input pair correctly, and model needs not be updated. Thus, the original problem is formulated as a classification problem of pairwise instances as used by Herbrich et al. (1999).

When the ranking of a pair is violated, model should be adjusted to rank this pair correctly. Equation (4) shows the loss function for the available pairs, in which λ_Θ is a ℓ_2 regularization term to prevent the model from over-fitting.

$$\mathcal{L}(\Theta) = \sum_{((+),(-)) \in \mathbf{X}^2} \ell(\Theta) + \frac{\lambda_\Theta}{2} \|\Theta\|_2 \tag{4}$$

3.4. Learning Algorithm

Although hinge loss is a convex function, but it is non-differentiable at zero. This work only focuses on differentiable part of hinge loss, so one can use gradient descent algorithm to optimize it. However, using gradient descent to update model parameters once requires to loop over all training data to obtain the sum of gradients. It is apparent that the update is computationally expensive when one is confronted with large-scale data sets.

This work uses stochastic gradient descent (SGD) algorithm (Bottou, 2010) to optimize the proposed objective function, since each update requires only a single training example. Moreover, the SGD algorithm does not need

to remember which examples were visited during the previous iterations, it can process examples on the fly in a deployed system.

With SGD optimization algorithm, we randomly select a pair from the training data, and update the model parameter using the update rule listed in Equation (5), in which $\theta \in \mathbb{R}^k$ is any parameter in the set Θ, and η is the learning rate. The partial derivatives of the loss function with respect to all parameters are listed in Equation (6).

$$\theta = \theta - \eta \left(\frac{\partial}{\partial \theta} \ell(\Theta) + \lambda_\Theta \theta \right) \tag{5}$$

$$\begin{aligned}
\frac{\partial \ell(\Theta)}{\partial \mathbf{w}_{v,j}} &= \frac{\mathbf{f}_{v,j}^{(-)} \mathbf{w}_m^{(-)}}{\sum_i \mathbf{f}_{v,i}^{(-)}} - \frac{\mathbf{f}_{v,j}^{(+)} \mathbf{w}_m^{(+)}}{\sum_i \mathbf{f}_{v,i}^{(+)}} \\
\frac{\partial \ell(\Theta)}{\partial \mathbf{w}_{m,j}} &= \frac{\mathbf{f}_{m,j}^{(-)} \mathbf{w}_v^{(-)}}{\sum_i \mathbf{f}_{m,i}^{(-)}} - \frac{\mathbf{f}_{m,j}^{(+)} \mathbf{w}_v^{(+)}}{\sum_i \mathbf{f}_{m,i}^{(+)}} \\
\frac{\partial \ell(\Theta)}{\partial b_{v,j}} &= \frac{\mathbf{f}_{v,j}^{(-)}}{\sum_i \mathbf{f}_{v,i}^{(-)}} - \frac{\mathbf{f}_{v,j}^{(+)}}{\sum_i \mathbf{f}_{v,i}^{(+)}} \\
\frac{\partial \ell(\Theta)}{\partial b_{m,j}} &= \frac{\mathbf{f}_{m,j}^{(-)}}{\sum_i \mathbf{f}_{m,i}^{(-)}} - \frac{\mathbf{f}_{m,j}^{(+)}}{\sum_i \mathbf{f}_{m,i}^{(+)}}
\end{aligned} \tag{6}$$

Algorithm 1 shows the learning algorithm, which is based on the update rules listed in Equation (5). In Algorithm 1, $\mathbf{X}^{(+)}$ and $\mathbf{X}^{(-)}$ are positive and negative training data sets. The *EscapeFlag* in Line 10 of algorithm 1 is for preventing infinite loop when the system fails to select the data that satisfies the condition $f(-) + 1 > f(+)$. Note that Algorithm 1 is an online learning algorithm, and it can update model parameters without re-training the whole model once new data instances are observed.

The time complexity of the scoring function presented in Equation (1) is $O(k)$, since it only involves the inner product of two vectors. Thus, the prediction of the proposed algorithm is very efficient. On the other hand, the training of the proposed algorithm is based on SGD algorithm, which is a commonly used optimization technique on large-scale data sets. The complexity for updating Equation (5) depends on the partial derivative of loss function $\ell(\Theta)$ as listed in Equation (6), which involves the summation

of the scoring function as listed in Equation (1). The time complexity for updating Equation (5) is $O(k(p+q))$. Practically, the number of latent factors, and the lengths of feature vector for video and music are all small numbers. Thus, the time complexity of the updating is also very efficient.

Algorithm 1: Latent Factor Model Algorithm

Input:
Positive examples $\mathbf{X}^{(+)}$, negative examples $\mathbf{X}^{(-)}$, model parameters Θ, and learning rate η
Output:
latent factors Θ

foreach *latent factor* θ *in* Θ **do**
  Initialize $\theta \sim N(0, \sigma^2)$
end
repeat
  repeat
    Randomly pick $\mathbf{x}^{(+)}$ from $\mathbf{X}^{(+)}$
    Randomly pick $\mathbf{x}^{(-)}$ from $\mathbf{X}^{(-)}$
    $f(+) = \mathbf{w}_v^{(+)T}\mathbf{w}_m{}^{(+)} + b_v{}^{(+)} + b_m{}^{(+)}$
    $f(-) = \mathbf{w}_v^{(-)T}\mathbf{w}_m{}^{(-)} + b_v{}^{(-)} + b_m{}^{(-)}$
    $EscapeFlag = \text{RANDINT}(0, 9)$
  until $f(-) + 1 > f(+)$ *or* $EscapeFlag == 0$;
  if $f(-) + 1 > f(+)$ **then**
    Make a gradient step to minimize:
    $f(-) + 1 - f(+)$
  end
until *convergence*;

4. System Architecture

The proposed system in this paper can be divided into two main parts as shown in Figure 2. The left-hand side is the core of the system, which includes two main databases, namely video and music. The purpose of this part is to use available videos and music to learn a recommendation model, so several steps are required to achieve the goal.

The first step is to extract features and mood tags from music and videos. It is apparent that music and videos are two different data formats, in which music is composed of sounds with a fundamental frequency and overtones, while videos are composed of a considerable number of frames. As a result, different approaches are required to extract features from videos and music as presented in the following sections.

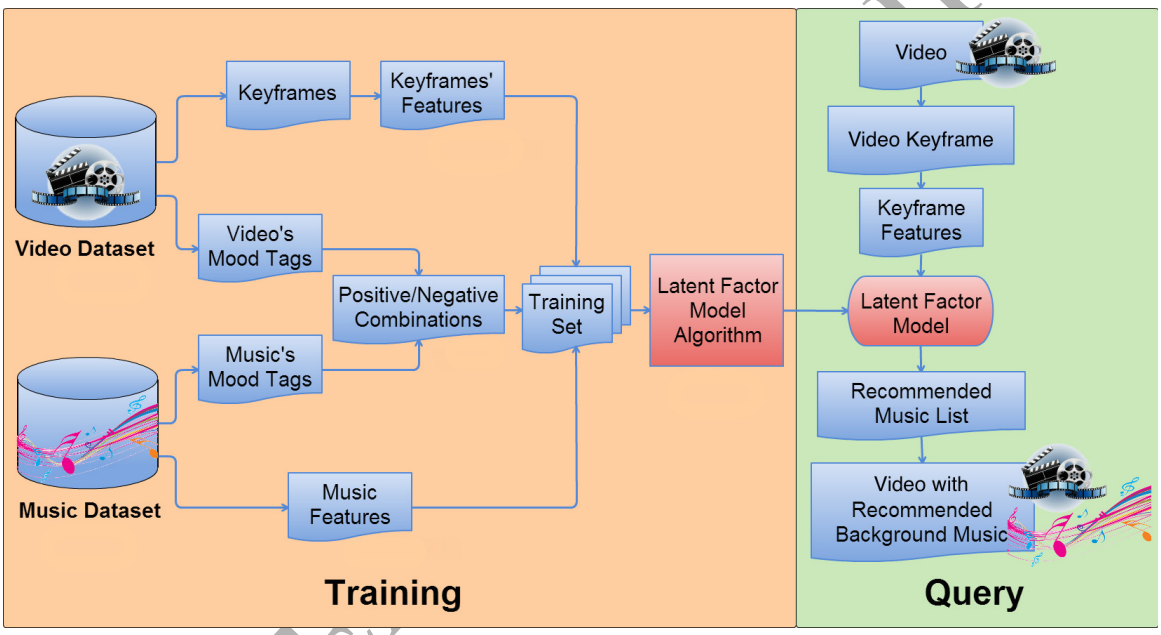


Figure 2: Flowchart of proposed background music recommender system

The second step is to generate training examples from the databases, so that we can use the obtained training examples to train a recommendation model. This work relies on Thayers model of mood (Thayer, 1990) along with the mood tags provided by video and music databases to label data. The detailed labeling process is described in the following section. Once feature extraction and the generation of training data are completed, one can use the proposed latent factor model algorithm to train a predictive model.

The right-hand side is the flowchart of how users use this system. When the user provides a specific video as input, the system analyzes this video and extracts important features from the video. Then, the recommendation engine generates a list of recommended background music. Finally, the user can select one of them or discard the list.

4.1. Video Features

To obtain features from videos, we identify the key frames of videos by using FFmpeg [1], since these key frames could be viewed as the representatives of the videos. Then we use image processing methods to extract features of the key frames to represent these key frames.

Image features can be generally divided into text features and visual features. Typical text features are keywords or abstract, and most visual features belong to three main types, including color, texture, and shape. This work only uses visual features, which are briefly introduced below.

- BilVideo-7
  MPEG-7 is a multimedia content description standard, and BilVideo-7 is an MPEG-7-compatible video indexing and retrieval system which was established by Bastan et al. (2010). The MPEG-7 develops an MPEG-7 profile to represent the videos by decomposing them into shots, key frames, still regions and moving regions. This work uses MPEG-7 to obtain descriptors, including color structure descriptor (CSD), color layout descriptor (CLD) (Ohm et al., 2001), scalable color descriptor (SCD), homogeneous texture descriptor (HTD), and edge histogram descriptor (EHD) (Wu et al., 2001). Note that CSD, SCD, and CLD are color descriptors, while HTD and EHD are texture descriptors.

- Scale-invariant feature transform (SIFT)
  The SIFT (Lowe, 2004) is an algorithm to detect and describe local features in images, and it transforms an image into a large collection of feature vectors, each of which is invariant to image translation, scaling, and rotation. It has been widely used on computer vision applications, including, but not limited to, object recognition, gesture recognition, and video tracking.

- Bag of features
  Bag of features, or bag of visual words, is originated from natural language processing technique, namely, bag of words model. In this model, an image is represented by frequencies of its features.

[1] FFmpeg: https://ffmpeg.org

Once the above feature extraction process is completed, we further concatenate extracted features to generate an image feature vector to represent the original video.

4.2. Music Features

The raw data of an audio file is signal information. In general, audio features comprise statistical features, temporal features, timbre features, rhythm features, and high-level features. The music data set and features used in this work are from MusiClef 2012 data set, which provides block-level features (Seyerlehner et al., 2010) and PS09 features (Pohle et al., 2009). The former includes several temporal features and the latter involves timbre and rhythm features.

This work uses temporal, timbre, and rhythm features, all of which are available in the music data set. For each music, these features are concatenated to generate a music feature vector.

4.3. Data Labeling

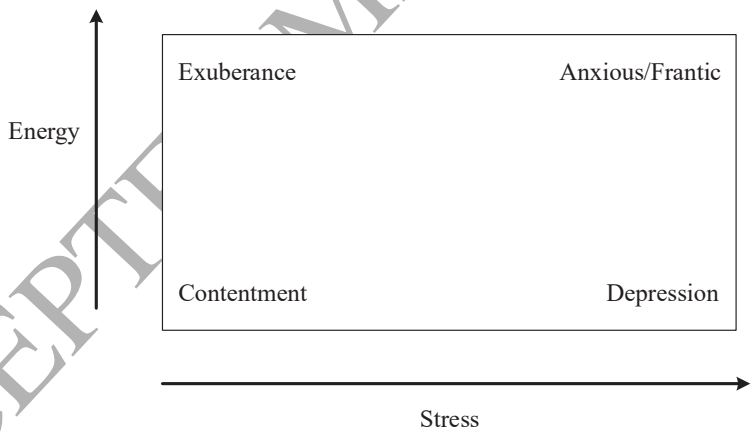


Figure 3: Thayer's model of mood

It is apparent that data labeling in this work is a challenging task, since defining whether a music song is relevant to a specific video is influenced by subjective factors, such as personal interest and emotional state. To let the data labeling task be objective, this work uses the video and music data sets

that comprise user tags as the data sources in the experiments. These tags are provided by users, so they can provide semantic information of the videos or music songs.

Moreover, this work relies on the Thayers model of mood (Thayer, 1990) to label data. Thayer's model of mood was proposed by Thayer (1990), who categorized emotions by two axes, stress and energy, as shown in Figure 3. The higher the stress is, the more negative the emotion is. The higher the energy is, the more energetic the emotion is.

Based on these two axes, emotions can be categorized into four quadrants, high energy/high stress, high energy/low stress, low energy/low stress, and low energy/high stress. This work maps the mood tags of music and video to four different quadrants as presented in Figure 3. For example, when people are confronted with high stress and low energy, they may suffer from depressed mood or loss of interest in activities they once enjoyed. In contrast, exuberance mood would appear when people are in the situation of low stress and high energy.

Once the mapping procedure is completed, all music songs and videos have their own mapping quadrants. The data labeling is based on the mapping, and those music songs and videos mapping to the same quadrant are viewed as positive examples, and the others are regarded as negative examples.

5. Experiments

This work focuses on two hypotheses as listed below, and designs experiments to verify whether the hypotheses are true.

H_1: The accuracy of recommendation can be effectively improved by using emotions and latent factors.
H_2: The background music which is recommended based on emotions and latent factors is more suitable than the methods without using emotions and latent factors.

5.1. Data sets

This work focuses on recommending background music that is relevant to a given video. Thus, the data sources should involve video and music datasets, and the two datasets should include tag information for labeling purpose. In the experiments, this work uses two data sets, in which UGV is

a video data set and MusiClef 2012 is a music data set. The introductions for the two data sets are presented below.

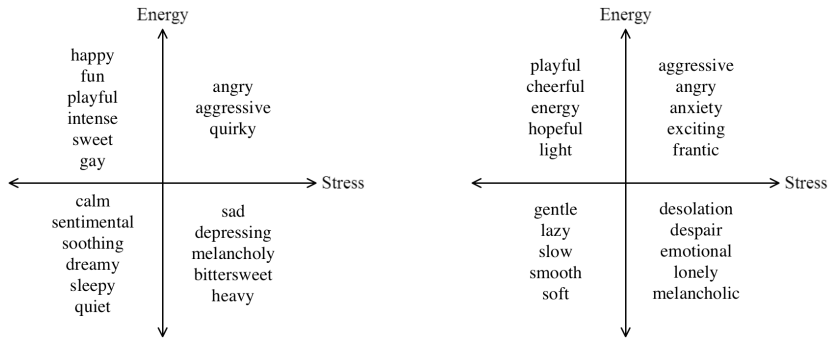
- UGV
  This data set is generated by Shah et al. (2014), and it comprises 1,265 user generated videos (UGV) that were handheld shooting with Android smart phones. Besides the videos, information from smart phone sensors are also included in the data set, such as GPS, gyro, and position. This work only focuses on the videos in the experiments. Moreover, the mood tags of videos are also available in this data set (Shah et al., 2014), in which the tags were obtained from crowd-sourcing. Some mood tags and their corresponding quadrants are presented in Figure 4a.

- MusiClef 2012
  MusiClef 2012 data set was first used in "MusiClef Multimodal Music Tagging Task at MediaEval 2012" contest, and it comprises 1,355 songs originated by 218 different musicians, singers, or groups. This work uses the data set pre-processed by Schedl et al. (2013) in the experiments. This data set comprises various information related to songs and singers, but we only uses music features and mood tags.

  Some mood tags and their corresponding quadrants are presented in Figure 4b. Note that the data set involves 64 different mood tags, and we only list five of them in each quadrant owing to the limit of space. The number of tags for each song is different, and the tags for each song may fall in different quadrants. For each song, we use the quadrant which most mood tags falling in as the quadrant of the song. Finally, for the sake of accuracy, we remove those songs whose quadrants can not be determined. The number of songs in the experiments is 1,035.

  The MusiClef 2012 data set does not comprise audio files, but they are required in the experiments. Therefore, we use the artists and song names to collect corresponding preview files from 7digital [2] website, in which the lengths for most of them are 30 seconds to 1 minute.

[2] 7digital https://www.7digital.com/



(a) Mood tags of videos (b) Mood tags of songs

Figure 4: Mood tags and their corresponding quadrants based on Thayer's model of mood

5.2. Evaluation Metric

The recommendation of songs is highly related to subject's interest or emotion state, so this work evaluates the proposed method using two different metrics. The first metric is area under the curve (AUC), which focuses on performance evaluation of the model and the correlation between videos and songs. The second one involves quantitative research and qualitative research, in which we invite subjects to participate our control experiment and scenario experiment.

- Area under the curve (AUC)
  A receiver operating characteristics (ROC) curve is a plot used to illustrate the performance of a binary classier system as its discrimination threshold is varied. In a ROC curve the true positive rate is plotted in function of the false positive rate for different cut-off points. The area under this ROC curve is AUC, which is a value between 0 and 1. More importantly, the AUC of a classifier is equivalent to the probability that the classier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). Many approaches have been proposed to calculate AUC, and we use the trapezoidal rule here. Let $\mathbf{P}$ be the positive set and $\mathbf{N}$ be the negative set, the AUC can be represented as:

$$AUC = \frac{1}{|\mathbf{N}||\mathbf{P}|} \sum_{j \in \mathbf{N}} \sum_{i \in \mathbf{P}} \mathbf{I}(f(\mathbf{x}_i) > f(\mathbf{x}_j))$$

- Quantitative research
  Quantitative research is systematic and empirical investigation of observable phenomena via statistical, mathematical or computational techniques. Many methods have been used in the evaluation, and this work uses Likert, which was introduced by Likert (1932), to evaluate our system, since it is a psychometric scale broadly involved in survey research that employs questionnaires.

  The Likert scale is usually presented in questionnaires to collect responses from the subjects, in which responses are scored along a range. For each question, the subjects specify level of agreement or disagreement on a symmetric agree-disagree scale. Thus, the range captures the intensity of their feelings for a given question.

  In normal questionnaires, several sub-scales are included, and each of which has several questions. The questionnaires designed for this paper will be introduced in the following sections.

- Qualitative research
  Qualitative research is a broad methodological approach that encompasses many research methods. Compared with quantitative research, no statistics or mathematics are presented to support the experimental results. In contrast, qualitative methods examine why and how of decision making, and smaller but focused samples are more often used than large samples to gather an in-depth understanding of human behavior and the reasons that govern such behavior. In the experiments, we investigate the fitness of the background music that recommended by the system through interviewing the subjects. More detailed experimental design is introduced in the following section.

5.3. Experimental Results

Based on the hypotheses mentioned above, this work designs two experiments. The first one is effectiveness evaluation with AUC which corresponds to H_1, and this experiment does not involve subjects. The second experiment is designed for H_2, and subjects should participate in the experiments. We further introduce control experiment and scenario experiment in the second experiment so that we can perform deeper analysis through quantitative research and qualitative research.

5.3.1. Performance Evaluation

This experiment is designed to verify hypothesis H_1, and we apply hold-out method and randomly sample 80% of the data as training data and the remaining as testing data. Both sets comprise half of positive combinations and half of negative combinations. The value ranges for all features are different, so we apply normalization process to all of the features, so that the values in all dimensions are between zero and one. Finally, we use dimensionality reduction technique introduced by Kamarainen et al. (2001) to speed up the experiments.

The experiments use six comparison methods to compare with the proposed method, including random, "SGD SVM" (Zhang, 2004), "pegasos SVM" (Shalev-Shwartz et al., 2011), "perceptron with margins" (Krauth and Mézard, 1987), "ROMMA" (Li and Long, 2002), and "logistic regression with pegasos projection (LRPP)". Most of them are from sofia-ml (Sculley, 2010) package. The sofia-ml package provides the implementation of these fast incremental algorithms, which can be used for training models for classification, regression, ranking, or combined regression and ranking. Additionally, these comparison methods are classical algorithms in machine learning, so this work uses them to compare with the proposed latent factor model. The goal is to verify whether the recommendation accuracy can be effectively improved by using emotions and latent factors.

In sofia-ml package, we set the loop type option to be *roc* to optimize the ROC curve. Besides the methods provided by sofia-ml package, we also use a random approach to recommend songs, which acts as the baseline.

5.3.2. Control experiment

Control experiment is the quantitative part of the second experiment which is designed to verify hypothesis H_2. The flow of this experiment is presented in the upper and lower left part of Figure 5. First, we randomly select a video from the video data set, and then use different methods to recommend background music to the selected video, in which the methods include random method, the proposed method, and a comparison method. Each method provides five songs as the recommendation, so we can obtain 15 background songs for a video.

In control experiment, we ask subjects to watch the original video without background music first. For each method, we randomly select a song from the five recommended songs. Then, let subjects watch the video along with three songs recommended by different methods in random order. Once the

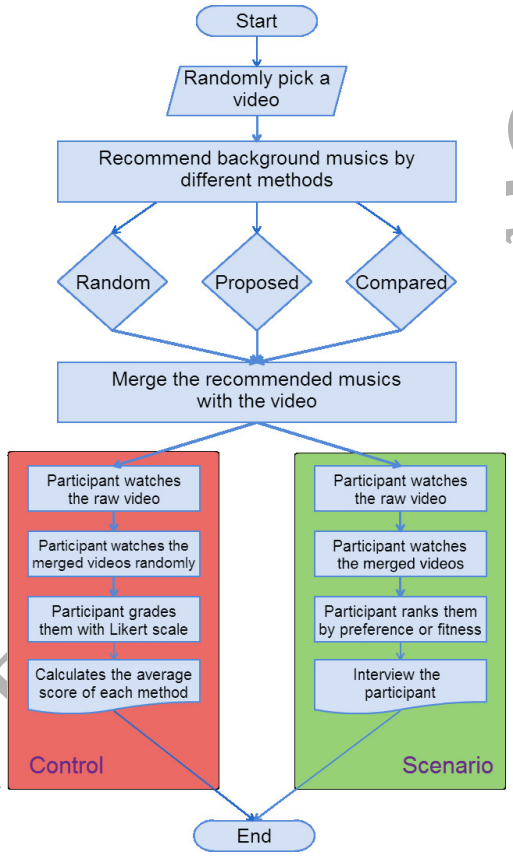


Figure 5: The flowchart of control experiment and scenario experiment

subjects finish watching a video, they answer the questionnaire of Likert scale as mentioned above. Repeat this step three times to complete one trial.

The purpose of this experiment is to investigate whether the proposed method recommends more suitable background music than other methods. Therefore we eliminate all the factors other than the recommending methods, and present the video and combinations of songs in random order. Moreover the information about song recommended by which method is unavailable to the subjects, and the goal is to make the experiment as objective as possible.

5.4. Scenario experiment

Scenario experiment is the other part of the second experiment for verifying hypothesis H_2, which includes both quantitative and qualitative research. The flow of this experiment is presented in Figure 5. Note that the upper part is the same as control experiment. After acquiring 15 songs from three different methods, we let subjects watch all of them in random order, then ask them to select the top three songs they like and another top three songs they dislike. The favorite song that is ranked as the first place is marked as three points, the second place as two points, the third place as one point, and no point to the songs they dislike. Finally, we can compare the performance of each method based on the scores. Furthermore, we interview the subjects to investigate how and why they select the songs they like or dislike to perform the qualitative research. By eliminating external influences, we want to verify the proposed method and also investigate the reasons that user may consider when choosing background music.

6. Discussion

Table 1 presents the results of performance evaluation. The AUC of random approach is about 0.5, which conforms to the intuition. The experimental results indicate that the proposed method outperforms the other alternatives, indicating that the proposed latent factor model could achieve good performance in recommendation problems. The hypothesis H_1 is verified by this evaluation result.

6.1. Quantitative research

- Control experiment
  We use questionnaires to conduct quantitative research in control experiment. Two sub-scales are involved in the questionnaires. One of

Table 1: Performance evaluation

Method	AUC
Random	0.5094
SGD SVM	0.6557
Pegasos SVM	0.6557
Perceptron with Margins	0.6397
ROMMA	0.6773
Logistic Regression with Pegasos Projection (LRPP)	0.7001
Latent Factor Model	**0.7674**

them is "fitness" sub-scale, and an example question from this sub-scale is "Do you think this background music fits this video?". The purpose of this sub-scale is to verify whether the recommended music is appropriate from music aspect. The other one is "engagement" sub-scale, and an example question is "Do you feel more engaged at this video after embedding background music?". The purpose is to verify the system effectiveness from video aspect.

In this experiment, we randomly select 16 videos and conduct 80 trials by 80 different subjects, in which each subject is responsible for one trial. Note that we only use random and LRPP as comparison methods in control experiment, since LRPP achieves the best performance among the comparison methods, while random method could be the baseline. The experimental results are presented in Figure 6. The experimental results indicate that the proposed method gets higher scores than comparison methods. We further carry out T-test to test the significance as listed in Table 2, in which we compare the proposed method with random and LRPP methods, respectively.

Table 2: The results of T-test on the total scores of control experiment

	Proposed/Random	Proposed/LRPP
scores	0.0009 ***	0.04293 *

Note that a *p-value* could help the practitioners determine the significance of experimental results when performing a hypothesis test in statistics. Thus, the result for each T-test is presented with *p-value*. In Table 2 and the following T-test tables, the score labeled with "*"

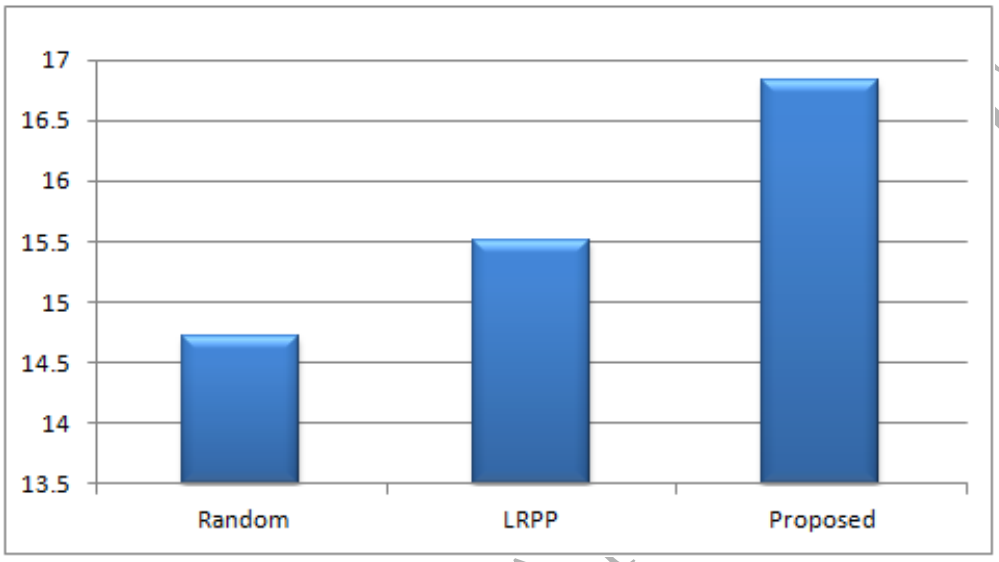


Figure 6: Total scores of control experiment

stands for *p-value* < 0.05, which means the result is "significant", the score labeled with "**" stands for *p-value* < 0.01, which means the result is "very significant", and the score labeled with "***" stands for *p-value* < 0.001, which means the result is "extremely significant".

The experimental results indicate that the proposed method significantly outperforms random and LRPP methods. To perform deeper analysis, we separate the scores by sub-scales as mentioned above, and the results are presented in Figure 7.

Figure 7 shows that the proposed method outperforms other methods on both sub-scales. Then, we conduct T-test on the two sub-scales and the results are listed in Table 3.

Table 3: The T-test results for scores of control experiment on sub-scales

	Proposed/Random	Proposed/LRPP
Sub-scale 1	0.0005 ***	0.1046
Sub-scale 2	0.0094 **	0.0185 *

Table 3 shows that the proposed method significantly outperforms ran-

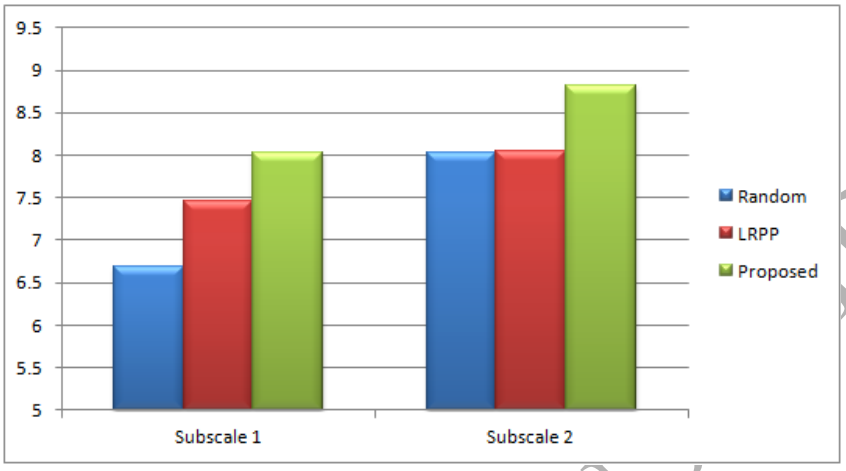


Figure 7: The scores of control experiment on sub-scale

dom method on both sub-scales, but only significantly outperforms LRPP method on sub-scale 2, namely "engagement" sub-scale. To know the reasons why the proposed method fails to achieve significant result on sub-scale 1 when compared with LRPP, we further separate the scores by questions and obtain the following results presented in Figure 8.

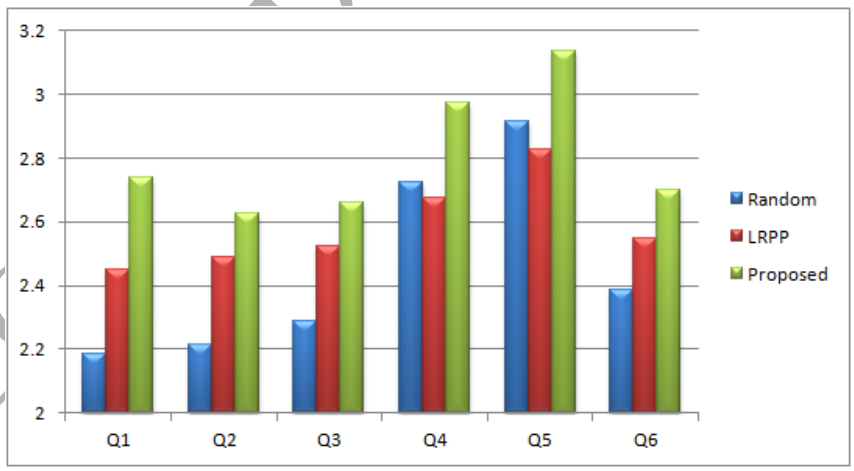


Figure 8: The scores of control experiment on questions

Table 4 shows the questionnaires used in the experiments, in which Q1

Table 4: Questionnaires used in the experiments

ID	Question
Q1	Do you think this background music 'fits' this video?
Q2	Do you think that the 'emotion' of this background music and the 'emotion' of this video meet?
Q3	Do you think that the 'scenario' of this background music and the scenario of this video meet?
Q4	Are you more willing to watch the video after embedding background music?
Q5	Are you more impressed with the video after embedding background music?
Q6	Do you feel more engaged at this video after embedding background music?

to Q3 belong to "fitness" sub-scale, and Q4 to Q6 belong to "engagement" sub-scale. As shown in Figure 8, the proposed method outperforms the other methods for all of the questions. The results of T-test for this part are presented in Table 5.

Table 5: The T-test results for scores of control experiment on questions

	Proposed/Random	Proposed/LRPP
Q1	6E-05 ***	0.033 *
Q2	0.0019 **	0.1886
Q3	0.0067 **	0.2018
Q4	0.0246 *	0.0138 *
Q5	0.0445 *	0.0083 **
Q6	0.0213 *	0.1619

As shown in Table 5, the T-test results between the proposed method and random method conform to the previous result, namely the proposed method significantly outperforms random method. In contrast, the T-test results for the proposed method and LRPP indicate that performance difference is significant for Q1, Q4, and Q5, but insignificant for Q2, Q3 and Q6. We further explore the reasons behind the results.

The first three questions belong to "fitness" sub-scale, but Q1 leads to significant difference while Q2 and Q3 do not. Thus, we analyze the reasons behind the phenomenon. As listed in Table 5, Q1 is about the fitness issue, while Q2 and Q3 are associated with emotion and scenario.

Emotion and scenario are highly related to the contexts of subjects, such as their emotional sates, interests or environments. Thus, given a song that is suitable for the video, the subjects may not agree with that the emotion and scenario of songs and videos meet with each other.

The last three questions belong to "engagement" sub-scale, but Q4 and Q5 lead to significant difference while Q6 does not. The contents of the videos are mainly daily life scenery such as markets, parks, or beaches, rather than the movie clips which have plot, so it is hard for the subjects to feel engaged at the videos even with background music as presented in Q6. Even so, the subjects are more willing to watch the video and feeling more impressed with the video as presented in Q4 and Q5. These results verify that the proposed method could recommend the background music that fits users' preferences.

- Scenario experiment
  The quantitative research in scenario experiments is based on the scores, which are determined by the rankings given by the subjects. We randomly select 20 videos, each of which has 15 clips with different background music and recommended by three different methods as described above. The 20 subjects are asked to rate the 15 clips derived from the same video, and the experimental results are presented in Figure 9.

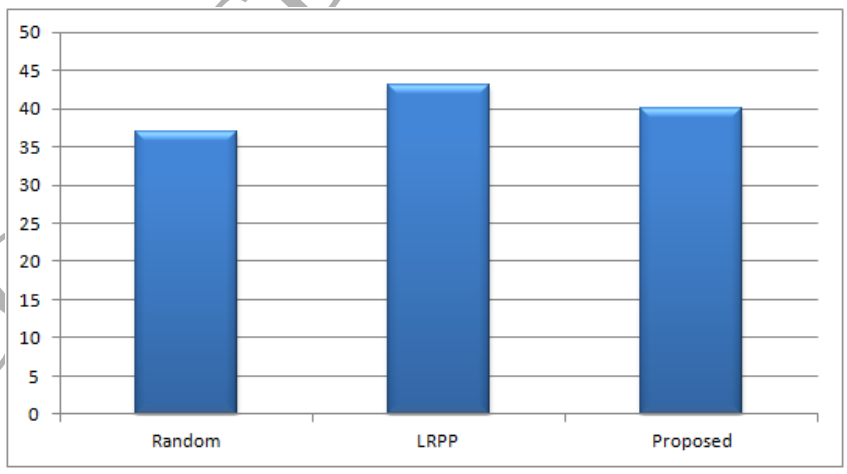


Figure 9: Total scores of scenario experiment

The experimental results indicate that the differences between these three methods are insignificant, and that is against the hypothesis H_2. Thus, we interviewed the subjects in this experiment, and use qualitative research to analyze the reasons leading to the results.

6.2. Qualitative research

In the scenario experiment, after subjects finish ranking the clips, we interviewed the subjects about how they rank the clips and the reasons why they select the clips that they like or dislike. Each subject is given an identifier, such as P8, in the following description.

Among the 20 subjects, most of them, accounting for 60% of the subjects, stated that they rank the clips depending on whether the rhythm of music fits the content of the video: *"The content of the video is playing basketball, and that is more appropriate for rhythmic music"* (P8). As mentioned above, we use the mood tags to determine whether the relation between music and video is positive or negative, while most of the subjects use the rhythm of music and the content of videos as their criteria. We believe that this is the reason why the differences presented in Figure 9 are insignificant.

As for the remaining of the subjects, three subjects stated that their criteria were based on mood: *"Both of the video and the music are lively, so they are more appropriate"* (P15). Four subjects said that they ranked the clips *"depending on the motion of the camera and the tempo of music, and if do not apparently fit each other, it is a poor clip"* (P18).

Besides, several subjects watch the clips very carefully: *"There is fountain in the video, and the top 3 music songs are more synchronous with the fountain"* (P13). This situation is out of our expectation when we were designing the experiments. Since the videos were taken with smartphone, another subject mentioned that *"The video content is too meaningless to focus, and it is hard to choose the music."* (P11). We believe that this is another reason that makes the differences to be insignificant as shown in Figure 9.

6.3. Comparison with other Latent Factor Model

The proposed framework relies on the latent factor model to recommend a list of music for the given video, so the other latent factor models that have the ability of recommending items could be used in the proposed framework. This work conducts experiments on a benchmark dataset, namely movie-

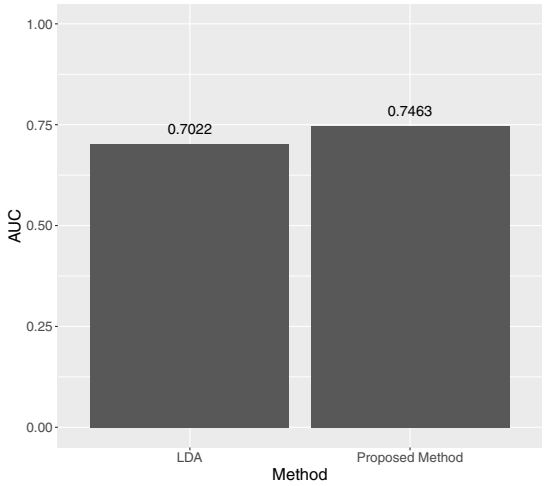


Figure 10: Comparison of LDA and the Proposed Method

lens 1M (ml-1m) [3], and compares the proposed method with latent Dirichlet allocation (LDA) (Blei et al., 2003).

The ml-1m dataset comprises 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. Ratings are made on a 5-star scale, and we use 3-star as the criterion to label data, in which the movies with ratings greater than 3-star are positive examples, while the remaining movies are negative examples. Once the labeling process is completed, the number of positive examples is 575,281 and the number of negative examples is 424,928.

In the experiments, we specify the number of latent factors as 100 and only use id information as the feature. Besides, the experiments use 80% of the data as training set and the remaining as test set, and apply 10-fold cross validation to evaluate the proposed latent factor model and LDA. We use the average AUC of the 10-fold cross validation to present experimental result as shown in Figure 10, in which the standard deviations for LDA and the proposed method are 0.0031 and 0.0010, respectively.

The experimental results indicate that the proposed method outperforms LDA on ml-1m dataset. The proposed method considers learning-to-rank technique in the loss function to consider the ordering of recommendation

[3]MovieLens 1M Dataset: https://grouplens.org/datasets/movielens/1m/

items, while LDA is originally designed for discovering topics from a collection of documents. We conclude that the ranking mechanism is the main reason why the proposed model could work well on movie recommendation problem.

7. Conclusion and Future Work

This work proposes a background music recommender system for user generated videos, in which we use latent factors of features to train the model. We provide detailed deviation of the proposed algorithm, and propose to use SGD to optimize the proposed algorithm owing to the consideration of training time. The goal of the proposed latent algorithm is to infer latent factors from a collection of observable data, and these latent factors often help practitioner find insights in data. We believe that the proposed method could be applied to other music recommendation contexts, since the proposed method belongs to data-driven method. More importantly, the experiments in this work focus on not only technical aspect, but also human aspect. In the technical aspect, we compare with several alternatives to show that latent factor model is comparative. Moreover, we also invite subjects to conduct quantitative research and qualitative research. The findings indicate that it is possible for the proposed method to discover the latent factors behind the recommendation, and we also provide detailed analysis about the results. We propose to use emotion as a surrogate to label the data, and we believe that other research studies that are related to emotion could benefit from this approach to label data samples.

Several applications are possible based on this work. For example, one can apply the proposed model to develop a mobile app, so that the users could select appropriate background music once they finish the video recording. Additionally, the proposed system could be a multimedia retrieval system. Given a video file, the system returns relevant background music. More importantly, the proposed method could be extended to the other modalities, since the latent factors are not limited to videos and music songs. The future work is to use deep learning technique to map video and music features to a latent space to learn discriminative feature representation.

Acknowledgment

This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant no. MOST 106-2221-E-009-100 and MOST 106-2218-E-009-031.

References

Åman, P., Liikkanen, L. A., 2017. Interacting with context factors in music recommendation and discovery. International Journal of Human–Computer Interaction 33 (3), 165–179.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data. The semantic web, 722–735.

Bastan, M., Cam, H., Gudukbay, U., Ulusoy, O., 2010. Bilvideo-7: an mpeg-7-compatible video indexing and retrieval system. IEEE MultiMedia 17 (3), 62–73.

Benzi, K., Kalofolias, V., Bresson, X., Vandergheynst, P., 2016. Song recommendation with non-negative matrix factorization and graph total variation. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. Ieee, pp. 2439–2443.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3 (Jan), 993–1022.

Bokhari, M. U., Hasan, F., 2013. Multimodal information retrieval: Challenges and future trends. International Journal of Computer Applications 74 (14).

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, pp. 177–186.

Chen, T., Zhang, W., Lu, Q., Chen, K., Zheng, Z., Yu, Y., 2012. Svdfeature: A toolkit for feature-based collaborative filtering. Journal of Machine Learning Research 13, 3585–3588.

Cheng, Z., Jialie, S., Hoi, S. C., 2016. On effective personalized music retrieval by exploring online user behaviors. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16. ACM, New York, NY, USA, pp. 125–134.
URL http://doi.acm.org/10.1145/2911451.2911491

Cheng, Z., Shen, J., 2016. On effective location-aware music recommendation. ACM Transactions on Information Systems (TOIS) 34 (2), 13.

Fawcett, T., 2006. An introduction to roc analysis. Pattern recognition letters 27 (8), 861–874.

Guan, C., Fu, Y., Lu, X., Chen, E., Li, X., Xiong, H., 2017. Efficient karaoke song recommendation via multiple kernel learning approximation. Neurocomputing.

Guan, L., Muneesawang, P., Wang, Y., Zhang, R., Tie, Y., Bulzacki, A., Ibrahim, M. T., 2009. Multimedia multimodal methodologies. In: Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. IEEE, pp. 1600–1603.

Herbrich, R., Graepel, T., Obermayer, K., 1999. Large Margin Rank Boundaries for Ordinal Regression. In: Advances in Large Margin Classifiers. The MIT Press, Ch. 7, pp. 115–132.
URL http://www.herbrich.me/papers/nips98_ordinal.pdf

Hernando, A., Bobadilla, J., Ortega, F., 2016. A non negative matrix factorization for collaborative filtering recommender systems based on a bayesian probabilistic model. Knowledge-Based Systems 97, 188–202.

Indyk, P., Motwani, R., 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing. STOC '98. ACM, New York, NY, USA, pp. 604–613.
URL http://doi.acm.org/10.1145/276698.276876

Joachims, T., 2002. Optimizing search engines using clickthrough data. In: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, pp. 133–142.

Joachims, T., Finley, T., Yu, C.-N. J., 2009. Cutting-plane training of structural svms. Machine Learning 77 (1), 27–59.

Kamarainen, J.-K., Kyrki, V., Ilonen, J., Kalviainen, H., 2001. Similarity measures for ordered histograms. In: PROCEEDINGS OF THE SCANDINAVIAN CONFERENCE ON IMAGE ANALYSIS. pp. 699–705.

Koren, Y., Bell, R., Volinsky, C., Aug. 2009. Matrix factorization techniques for recommender systems. Computer 42 (8), 30–37.

Krauth, W., Mézard, M., 1987. Learning algorithms with optimal stability in neural networks. Journal of Physics A: Mathematical and General 20 (11), L745.

Kuo, F.-F., Shan, M.-K., Lee, S.-Y., 2013. Background music recommendation for video based on multimodal latent semantic analysis. In: Multimedia and Expo (ICME), 2013 IEEE International Conference on. IEEE, pp. 1–6.

Lee, D. D., Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.

Lee, W.-P., Chen, C.-T., Huang, J.-Y., Liang, J.-Y., 2017. A smartphone-based activity-aware system for music streaming recommendation. Knowledge-Based Systems.

Li, L.-J., Su, H., Fei-Fei, L., Xing, E. P., 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Advances in neural information processing systems. pp. 1378–1386.

Li, Y., Long, P. M., 2002. The relaxed online maximum margin algorithm. Machine Learning 46 (1-3), 361–387.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of psychology.

Lin, T.-W., Shan, M.-K., 2017. Correlation-based background music recommendation by incorporating temporal sequence of local features. In: Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on. IEEE, pp. 158–164.

Lin, Y.-T., Tsai, T.-H., Hu, M.-C., Cheng, W.-H., Wu, J.-L., 2014. Semantic based background music recommendation for home videos. In: International Conference on Multimedia Modeling. Springer, pp. 283–290.

Liu, C.-L., Wu, X.-W., 2016a. Fast recommendation on latent collaborative relations. Knowledge-Based Systems 109, 25–34.

Liu, C.-L., Wu, X.-W., 2016b. Large-scale recommender system with compact latent factor model. Expert Systems with Applications 64, 467–475.

Liu, H., Hu, J., Rauterberg, M., 2010. Lsm: a new location and emotion aware web-based interactive music system. In: Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference on. IEEE, pp. 253–254.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60 (2), 91–110.

Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G., 2015. Recommender system application developments: a survey. Decision Support Systems 74, 12–32.

Miller, G. A., 1995. Wordnet: a lexical database for english. Communications of the ACM 38 (11), 39–41.

Ning, X., Karypis, G., 2012. Sparse linear methods with side information for top-n recommendations. In: Proceedings of the sixth ACM conference on Recommender systems. ACM, pp. 155–162.

Ohm, J.-R., Cieplinski, L., Kim, H. J., Krishnamachari, S., Manjunath, B., Messing, D. S., Yamada, A., 2001. The mpeg-7 color descriptors. IEEE Transactions on Circuits and Systems for Video Technology.

Oramas, S., Nieto, O., Sordo, M., Serra, X., 2017a. A deep multimodal approach for cold-start music recommendation. In: Proceedings of the 2Nd Workshop on Deep Learning for Recommender Systems. DLRS 2017. ACM, New York, NY, USA, pp. 32–37.
URL http://doi.acm.org/10.1145/3125486.3125492

Oramas, S., Ostuni, V. C., Noia, T. D., Serra, X., Sciascio, E. D., 2017b. Sound and music recommendation with knowledge graphs. ACM Transactions on Intelligent Systems and Technology (TIST) 8 (2), 21.

Park, M.-H., Hong, J.-H., Cho, S.-B., 2007. Location-based recommendation system using bayesian users preference model in mobile devices. Ubiquitous intelligence and computing, 1130–1139.

Paulin, M., Mairal, J., Douze, M., Harchaoui, Z., Perronnin, F., Schmid, C., 2017. Convolutional patch representations for image retrieval: an unsupervised approach. International Journal of Computer Vision 121 (1), 149–168.

Pirasteh, P., Hwang, D., Jung, J. J., 2015. Exploiting matrix factorization to asymmetric user similarities in recommendation systems. Knowledge-Based Systems 83, 51–57.

Pohle, T., Schnitzer, D., Schedl, M., Knees, P., Widmer, G., 2009. On rhythm and general music similarity. In: ISMIR. pp. 525–530.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., Vasconcelos, N., 2010. A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia. MM '10. ACM, New York, NY, USA, pp. 251–260.
URL http://doi.acm.org/10.1145/1873951.1873987

Rendle, S., 2010a. Factorization machines. In: 2010 IEEE International Conference on Data Mining. IEEE, pp. 995–1000.

Rendle, S., 2010b. Factorization machines. In: Proceedings of the 2010 IEEE International Conference on Data Mining. ICDM '10. IEEE Computer Society, Washington, DC, USA, pp. 995–1000.
URL http://dx.doi.org/10.1109/ICDM.2010.127

Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L., 2009. Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. UAI '09. AUAI Press, Arlington, Virginia, United States, pp. 452–461.

Rendle, S., Gantner, Z., Freudenthaler, C., Schmidt-Thieme, L., 2011. Fast context-aware recommendations with factorization machines. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11. ACM, New York, NY, USA, pp. 635–644.
URL http://doi.acm.org/10.1145/2009916.2010002

Rui, Y., Huang, T. S., Chang, S.-F., 1999. Image retrieval: Current techniques, promising directions, and open issues. Journal of visual communication and image representation 10 (1), 39–62.

Salton, G., McGill, M. J., 1986. Introduction to modern information retrieval.

Schedl, M., Orio, N., Liem, C., Peeters, G., 2013. A professionally annotated and enriched multimodal data set on popular music. In: Proceedings of the 4th ACM Multimedia Systems Conference. ACM, pp. 78–83.

Sculley, D., 2010. Combined regression and ranking. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 979–988.

Seyerlehner, K., Widmer, G., Pohle, T., 2010. Fusing block-level features for music similarity estimation. In: Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10). pp. 225–232.

Shah, R. R., Yu, Y., Zimmermann, R., 2014. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, pp. 607–616.

Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A., 2011. Pegasos: Primal estimated sub-gradient solver for svm. Mathematical programming 127 (1), 3–30.

Skowron, M., Lemmerich, F., Ferwerda, B., Schedl, M., 2017. Predicting genre preferences from cultural and socio-economic factors for music retrieval. In: European Conference on Information Retrieval. Springer, pp. 561–567.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., Jain, R., 2000. Content-based image retrieval at the end of the early years. IEEE Transactions on pattern analysis and machine intelligence 22 (12), 1349–1380.

Thayer, R. E., 1990. The biopsychology of mood and arousal. Oxford University Press.

Thompson, B., 2005. Canonical correlation analysis. Encyclopedia of statistics in behavioral science.

Van den Oord, A., Dieleman, S., Schrauwen, B., 2013. Deep content-based music recommendation. In: Advances in neural information processing systems. pp. 2643–2651.

Wang, X., Sun, J.-T., Chen, Z., Zhai, C., 2006. Latent semantic analysis for multiple-type interrelated data objects. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 236–243.

Wu, D., Zhang, G., Lu, J., 2015. A fuzzy preference tree-based recommender system for personalized business-to-business e-services. IEEE Transactions on Fuzzy Systems 23 (1), 29–43.

Wu, P., Ro, Y. M., Won, C. S., Choi, Y., 2001. Texture descriptors in mpeg-7. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 21–28.

Yeh, M.-C., Cheng, K.-T., 2008. A string matching approach for visual retrieval and classification. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. ACM, pp. 52–58.

Yoshida, T., Hayashi, T., 2016. Otopittan: A music recommendation system for making impressive videos. In: Multimedia (ISM), 2016 IEEE International Symposium on. IEEE, pp. 395–396.

Zhang, H.-R., Min, F., 2016. Three-way recommender systems based on random forests. Knowledge-based systems 91, 275–286.

Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the twenty-first international conference on Machine learning. ACM, p. 116.