Video Background Music Recommendation Based on Multi-level Fusion Features

Xin Zhao
Zhengzhou University
Xiaobing Li (lxbmusic@188.com)
Central Conservatory of Music
Yun Tie
Zhengzhou University
Lin Qi
Zhengzhou University

Research Article

Keywords: cross-modal recommendation, music recommendation, deep learning, convolutional neural network

Posted Date: June 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3037240/v1

Additional Declarations: No competing interests reported.

Video Background Music Recommendation Based on Multi-level Fusion Features

Xin Zhao[1], Xiaobing Li[2*], Yun Tie[1], Lin Qi[1]

[1]School of Electrical and Information Engineering, Zhengzhou University, Street, Zhengzhou, 450001, Henan, China.
[2*]Department of Music Artificial Intelligence, Central Conservatory of Music, Street, Beijing, 100091, China.

*Corresponding author(s). E-mail(s): lxbmusic@188.com;
Contributing authors: zhaoxin0227@gs.zzu.edu.cn; ieytie@zzu.edu.cn; ielqi@zzu.edu.cn;

Abstract

People resonate more with music when exposed to visual information, and music enhances their perception of video content. Cross-modal recommendation techniques can be used to suggest appropriate background music for a given video. However, there is not a simple correspondence between the different modal data. Therefore, to explore the association between the two modalities of video and music, we propose MFF-VBMR, a video background music recommendation model based on multi-level fusion features. The model uses the cross-modal information of static, dynamic and emotional content of video and music to realize the task of matching and recommending suitable background music for a given video. We propose a feature normalized convolutional similarity algorithm network FNC, which takes into account the pairwise similarity of visual and acoustic regions without losing region details. Experimental results show that the proposed model outperforms other existing models in terms of performance and achieves satisfactory results for video background music recommendation.

Keywords: cross-modal recommendation, music recommendation, deep learning, convolutional neural network

1 Introduction

With the rapid development of the Internet and self-media, various types of multimedia information have proliferated and flooded people's lives like a spring. Video and music are two widely-watched media on the Internet, and people's perceptions of them are highly correlated. The stimulation of visual information resonates with people when they listen to music, which echoes the video in visual perception and adds color to the scene.

There are related studiesCowen et al (2020)-Sievers et al (2013) suggest that human perception of music may have evolved from an ancient skill, namely the ability to interpret emotions from movements. It has been found that when people pair an emotion with a melody or a video animation, they choose combinations that have the same temporal and spatial characteristics, such as the same tempo, rhythm and smoothness, for the pairing. Surprisingly, this holds true for people from completely different cultural backgrounds. In other words, although different cultures do not understand music in exactly the same way, music is able to retain similar subjective experiences and emotional arousal across culturesCowen et al (2020) . People's response to audio-visual stimuli not only exists in the cognitive level, but also is an involuntary and natural state.

Therefore, the two modalities of video and music are not simply in correspondence, but there are many correlations that make people unconsciously react in a similar way. At the same time, short video platforms are becoming increasingly popular in people's lives, and finding a suitable soundtrack is a concern when sharing videos on the web in order to increase the completeness of the video. Therefore, the topic of how to obtain the deep connection between video and music and how to better match and integrate video and music has great theoretical research value and social application value, which has attracted the attention of many researchers from different angles.

Video and music are both widely used media, but the connections between them have not been well explored. Currently, some studies have focused on analyzing the underlying semantic features of both video and audio modalitiesKuo et al (2013)-Lin and Shan (2017), but only on low-level features of video and music. These shallow features are not closely enough connected to capture the key information of video and music, resulting in a loose connection of multimodal features in the feature space and poor retrieval results in cross-modal retrieval recommendation tasks. Another part of the study focused on the emotional connection between video and music by constructing a cross-modal neural network modelSharma et al (2021)-Tsai et al (2022)to build a sentiment communal embedding space to bridge the heterogeneity gap between different data modalities. However, from a practical application point of view, the results of recommendations are more or less unsatisfactory due to the different understanding of video and music by each individual. Meanwhile, existing cross-modal recommendation methods mostly use the construction of a common subspaceSurís et al (2018)-Jin et al (2020)that rely on the mathematical relationship between the two feature vectors to capture similarity, which tends to lead to too homogeneous multimodal information and poor model results.

Therefore, in order to explore the association between the two modalities of video and music, as well as the prevalence of people's perception of multimodal information,

this paper proposes a method for video background music recommendation based on multi-level fusion features, which utilizes cross-modal information based on static, dynamic and emotional content corresponding to the different levels generated by video and music to achieve the task of matching and recommending appropriate background music for a given video. Firstly, in order to obtain the deep correspondence between video and music data, we make a comprehensive summary of the features of video and audio, systematically and scientifically analyze the feature information of video at the level of key frame, shot and scene, and the feature data of music at the level of audio spectrum, note and melody. Secondly, this paper chooses HIMV-200K benchmark datasetHong et al (2018), Pop music videos dataset Lin and Yang (2021) and self-built dataset for experiments, and uses extended convolutional neural network and acoustic feature processing tools respectively to extract features from videos and music. Finally, we fed the fused multimodal features of video and music into our proposed feature-normalized convolutional similarity algorithm network, which considers the pairwise similarity of visual and acoustic regions without losing regional details.

Specifically, the contributions of this paper are divided into three areas.

1) Aiming at the task of recommending suitable background music for video, we propose a video background music recommendation model MFF-VBMR based on multi-level fusion features. The proposed model is able to synchronously process pre-processed multimodal feature data and find key pairs of multimodal information by exploiting the correlation framework of video and music.

2) In terms of feature selection, in order to explore the correlation between the two modes of video and music, we summarized the similarities and differences between the features of video and music, and extracted the features comprehensively. Instead of analyzing the content semantically or emotionally, the features are extracted from static, dynamic and emotional aspects of video and music respectively, which better represent their deeper content and are complemented by contextual information, making the cross-modal information more closely linked.

3) In the retrieval task, in order to make cross-modal information matching recommendations more accurate, we abandon the traditional common subspace network and improve the convolutional neural network algorithm. We propose a feature-normalized convolutional similarity algorithm, FNC. To make each feature vector contribute equally to the similarity calculation, the extracted feature vectors are L2 normalized; and a self-attentiveness mechanism is introduced to weight the captured video and music feature vectors according to them. We feed the feature matrix into a convolutional neural network and use a mean-max filter in the last layer, which can suppress possible spurious similarity results.

2 Related work

People can get information from images, sounds, text and so on. In other words, our world is a multimodal world. When a research problem or dataset contains more than one modality, it can be handled by multimodal techniques. Multimodal techniques can be applied in a variety of fields. For example, one of the earliest applications of multimodal research is audiovisual speech recognition (AVSR)Bourlard and Dupont

(1996), which uses visual information to improve the accuracy of speech recognition. Multimodal technology has also played an important role in such fields as emotion recognitionValstar et al (2013), image descriptionHodosh et al (2015), VQAAntol et al (2015) and traffic event detectionChen et al (2021).

Recommender systems are a tool to help users quickly discover useful information and have been widely used in recent years in the study of cross-modal recommendations. Related work on cross-modal recommendation can be divided into: approaches based on probabilistic graphical models, approaches based on matrix analysis, etc. Probabilistic graph models have been widely used because of their good scalability and theoretical foundation. RoyRoy et al (2012) proposed the OSLDA model for topic modelling of Twitter streams and recommending videos based on the relevance of the tags corresponding to the videos to these topics. The transmedia LDATan et al (2014) model assumes that users and all media types are associated with the common topic space, so the association between users and transmedia objects can be obtained by comparing their distribution on the common topic. Kumar et al.Kumar et al (2014) used WordNet ontologies to compare semantic correlations between words in different domains. The clustered semantic dictionaries were then modelled using the PLDA model to obtain the potential semantic space shared by multiple domains. In order to model different media types on two different platforms, Min et al.Min et al (2015) proposed a cross-platform multimodal topic pattern model. Unlike the other probabilistic graph models mentioned above, this model modelled the relationship between the two aspects simultaneously. Chen et al.Li et al (2013) implemented a matrix decomposition-based system to provide a variety of recommendation results, including item recommendations, friend recommendations and group recommendations. For each type of recommendation, the system uses knowledge from the other two aspects as a secondary knowledge source to improve system performance. In addition, other work has implemented cross-modal recommendations. Chen et al.Chen et al (2013) and Zhang et al.Jia et al (2014) use tensor decomposition to find the potential space between different modalities. Hoxha et al.Hoxha (2014) modelled the semantic content and user browsing behavior between multi-modal objects and used support vector machines to learn the correlation between recommended objects and decide whether to recommend them. Wang et al.Wang et al (2015) used two CNNs (convolutional neural networks) to learn potential feature representations for both text and image modalities. A one-to-many learning framework was then used to learn the relationship between the latent features of the two modalities. HeitmannHeitmann et al (2012) proposed a semantic interest graph to model user preferences for multiple modalities to enable cross-modal recommendations.

On the study of background music recommendation, Kuo et al.Kuo et al (2005) proposed a framework to discover associations between sentiment and music features for music recommendation. They investigated the extraction of music features and suggested the use of affinity graphs to discover associations. Yu et al.Yi et al (2012) used location information from UGV to map geotags to emotion tags by investigating categories on the website and then comparing them to music emotions, but did not consider the visual content of the videos. Shan et al.Shah et al (2014) improved the system by using modelled scene emotions. A sequence of visual and geographic features

was trained to predict scene mood and compare it with the collected music mood. Wang et al.Wang et al (2012) proposed an audiovisual emotion Gaussian modelling algorithm to learn the relationship between video, music and emotion. The emotion distribution of video and music was measured by KL scatter to measure similarity. Lin et al.Lin et al (2015) proposed an EMV framework that uses an emotional time course model to learn the relationship between temporal phase sequences of music or video, and uses a Hidden Markov Model and an expectation maximization algorithm to predict sequences and valence-evoking emotional quadrants, and then compares temporal phase sequences and emotional quadrants by string matching.

Researchers have also made many improvements in the study of algorithms for making matching recommendations. Cristani et al.Cristani et al (2010) proposed a recommendation strategy when driving a car. Given a video of the driving scene, and a matching audio is selected. They used Pearson's correlation coefficient to calculate the association between audio and video features. Kuo et al.Kuo et al (2013) proposed a background music recommendation system that learns the relationship between music and video by using multimodal latent semantic analysis and calculates the alignment of the recommended music and the given video. Yet, the depth of individual features used in the above methods is not sufficient, and the mapping relations of the cosubspace network are somehow different. To address the limitations of single feature and similarity algorithms, we use different levels of multimodal fusion information and construct similarity learning networks to calculate the similarity between video and music. The model takes into account the pairwise similarity between visual and acoustic regions.

3 Method model

To achieve the task of video background music recommendation, this paper proposes a video background music recommendation model (MFF-VBMR) based on multi-level fused features. First, we perform feature extraction based on the deep-level feature correspondence between video and music data. Second, we fuse multimodal features; finally, we feed the fused multimodal features of video and music, into our proposed feature normalized convolutional similarity algorithm network to obtain the optimal solution for music recommendation, and the model framework is shown in Figure 1.

3.1 Multimodal feature analysis

When people watch a video with music playing in the background, they relate to the video and the music, and this feeling of similarity is present not only in the video, but also in the music. In order to match the recommended soundtrack to the video, we explored the cross-modal alignment between video and music at three levels of analysis: static, dynamic and emotional, as shown in Table 1.

3.1.1 Video features

In general, videos can be represented in a hierarchy: frames, shots and scenes. A shot is a video segment consisting of consecutive actions, and a scene consists of one or

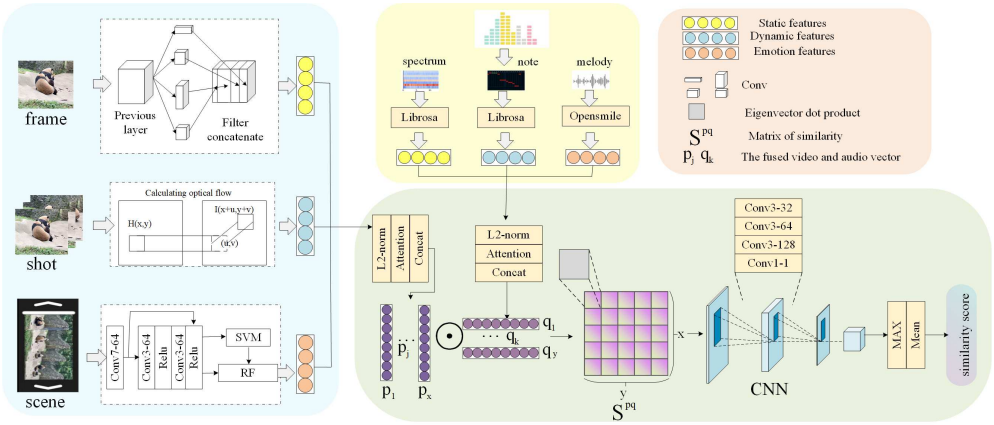

Fig. 1 MFF-VBMR model

more shots forming a semantic unit. The key frames of a video can be characterized by color, texture and light, and shots can be represented by trajectories of motion, while these factors can also have a strong impact on emotion, which is the most important factor in representing a video scene. In the present work, we extract visual features from the categories described above.

Color features Color symbolism varies from culture to culture, but in general, warm tones help to intensify visual perceptions such as warmth, excitement and intensity, while cool tones help to highlight effects such as serenity, depth and solitude. The descriptors of color characteristics in this paper include color energy and saturation ratioWang and Cheong (2006). The color energy is calculated based on color contrast and angular distance to blue and red respectively, and the saturation ratio is based on the proportion of low-saturation pixels.

Texture features Textures add depth to key frames and are an important element of human visual perception. In this paper we extract one of the most widely used texture features, the grey scale co-occurrence matrix (GLCM), whose descriptors are shown in Table 1.

Light features The descriptors of light features adopted in this paper are median light value and shadow ratioWang and Cheong (2006). The median light value is the median value of brightness, and the shadow ratio is the proportion of the shaded area measured in the frame.

Motion features Movement is a highly expressive element that triggers an emotional response from the audience. We used the camera-level optical flow featureSimonyan and Zisserman (2014) as the motion feature of the entire video.

3.1.2 Music Features

The mood of the viewer is greatly influenced by sound effects and music. Music can create a specific atmospheric tone (including temporal and spatial characteristics) for

Table 1 Multimodal feature classification

modality	species	type	feature
video	static	color	color energy
			saturation ratio
		texture	Angular second moment
			Contrast ratio
			correlation
			Difference of phase
			entropy
			Property of homogeneity
			Mean gray scale
			Variance of gray scale
		light	Median brightness value
			Ratio of shadows
	dynamic	motion	Optical flow
			Excitement of vision
	emotional	emotion	Characteristics of emotion
music	static	timbral texture	MFCC
			Spectral center of mass
			Attenuation of spectrum
			Flux of spectrum
	dynamic	rhythm	Beat histogram
		high level	Nature of dance
			Duration of time
			energy
			key
			loudness
			model
			rhythm
			Characteristic of time
	emotional	emotion	Characteristics of emotion

parts of a video or the whole, thus deepening the visual effect and enhancing the impact of the picture. The musical features used in this paper are as follows.

Timbral Texture features In terms of musicality, five features widely used in audio classification and speech recognition are used, namely zero-crossing rate, spectral roll-off, spectral centroids, spectral flux and Mayer spectral coefficients.

Rhythm features The rhythm descriptor of music is the beat histogram proposed inTzanetakis (2001), which is established by the autocorrelation function of the signal.

High-Level features High-level descriptors of music include dance ability, duration, energy, key, loudness, pattern, rhythm and time features, and this paper uses the rhythmic features and high-level features of music as its dynamic features.

Emotion features Emotional features are shared by video and music, and people can feel the impact of emotions visually and audibly. Video and music are two different types of media that show a strong connection and relevance to each other. Music not only enhances our emotional response to video and images, but also improves our understanding of visual effects, while video and images not only enrich our emotional response to music but also convey and express the atmosphere of the song. In this paper, emotional features are extracted based on video understanding and music understanding respectively.

3.2 Feature extraction

3.2.1 Video feature extraction

Static features In this paper, static features are extracted from key frames of video using Inception network, which can reduce the number of optimization parameters in deep learning networks, greatly reduce the computational effort and optimize the computational speed, and has been widely used for feature extraction in recent years. In this paper, the pre-processed video data is fed into the Inception network, and the output is a video feature vector for each frame. The model will decode the video, and the decoded video data will be fed into the Inception network, followed by the ReLU activation function for computation. The final output feature vector for each video frame is quantified by principal component analysis and calculated to output a 1024-dimensional video frame-level feature vector.

Dynamic features This paper uses optical flow features as dynamic features of video, which are useful tools for analyzing video motion. Generally speaking, optical flow $f_t(x, y) \in \mathbb{R}^{H\times W\times 2}$ is the measurement of two consecutive frames I_t , the $I_{t+1} \in \mathbb{R}^{H\times W\times 3}$the displacement of a single pixel between them. Similar to distance and velocity, we define the optical flow amplitude F_t as the average of the absolute optical flow to measure the amplitude of motion in frame t.

$$F_t = \frac{\sum_{x,y} f_t(x, y)}{HW} \tag{1}$$

The video is extracted with T key frames and the motion relationship between the key frames is to be calculated. The motion saliency at frame t is calculated as the average positive change in optical flow in all directions between two consecutive frames. We then obtain a series of visual beats by selecting the frame with the maximum local motion saliencyDavis and Agrawala (2018) When the key frame has a sudden visible change, the saliency will have a larger value. The corresponding optical flow feature vector is finally obtained.

Emotion features This paper uses a video shot boundary detection algorithmWei et al (2021) that divides the video into multiple shots. A representative frame is then randomly selected in each shot for sentiment saliency estimation, which not only saves time but also avoids the appearance of redundant frames. In this paper, the difference in color histogram between frames is used to detect video shot boundaries.

The use of the histogram method is effective in avoiding differences caused by the motion of objects in the footage, thus improving robustness. Typically, the inter-frame variance in a shot is stable over a small range, and when a shot shift occurs, the inter-frame variance is significantly larger than the mean. Therefore, frames with inter-frame differences greater than the average should be identified as lens boundaries.

Deep visual features are widely used in feature extraction for sentiment recognition. In this paper, a representative depth model is chosen to extract deep visual features from video frames: ResNet-101He et al (2015). ResNet-101 consists of a convolutional layer, 33 building blocks and a fully connected layer. The output of the average pooling layer is used as a depth feature, which is recorded as an object feature. By extracting key frames, each video can be represented as a key frame of length X. The frame-level features extracted using the ResNet-101 model are represented as R, as follows.

$$\begin{aligned} R &= (R_1, R_2, ...R_X)^T \\ &= \begin{bmatrix} R_1[1] & \cdots & R_1[k] & \cdots & R_1[2048] \\ \vdots & & \vdots & & \vdots \\ R_i[1] & \cdots & R_i[k] & \cdots & R_i[2048] \\ \vdots & & \vdots & & \vdots \\ R_N[1] & \cdots & R_N[k] & \cdots & R_N[2048] \end{bmatrix}. \end{aligned} \tag{2}$$

Model SVM and RF are traditional models for recognizing video emotionsLy et al (2019)-Samadiani et al (2019) . SVMCortes and Vapnik (1995) is a discriminative method for learning boundaries between classes. It has been widely used in image sentiment classification due to its good generalization ability. We use an RBF kernel suitable for high-dimensional features. RFHo (1995) is an important integrated learning method based on Bagging, which consists of multiple basic learners. Randomness provides RF with powerful preventive overfitting properties, so it is often used to construct predictive models for classification and regression problems. Before using the traditional model, a Max pool is used to convert the frame-level features to video-level features and then perform feature normalization to obtain the corresponding video sentiment feature vector.

3.2.2 Music feature extraction

Static, dynamic features This paper uses the audio feature extraction tool liborsaMcFee et al (2015) to extract static and dynamic features from music. For the spectral analysis of the music, the fast Fourier transform and discrete wavelet transform are first applied to the windowed signal in each local frame. Based on the results of the amplitude spectrum, features including spectral shape centers, spectral bandwidth, and spectral roll-off are calculated. In order to extract more meaningful features, the Mel-scale spectrogram and Mel frequency cepstrum coefficients are calculated for each frame. To capture the change in timbre over time, incremental MFCC features are used. The number of time-domain over-zero points is also extracted in order to detect the amount of noise in the audio signal. Finally this paper calculates

the mean and variance of each frame-level feature and the maximum top K-order statistic, and concatenates all the static-level features. Also, liborsa can extract the tempo and beat of the music, estimate the tempo and tune speed, and the beats per minute are transformed into a matrix to calculate the beat histogram, which allows the extraction of the high-level feature vectors in Table 1.

Emotion features This paper uses openSMILEEyben et al (2010) to extract the emotional features of music. OpenSMILE is a command-line tool that extracts audio features by configuring a config file. It is mainly used for speech recognition, emotion computing and music information acquisition. OpenSMILE provides a variety of standard feature sets for emotion recognition, in this paper we use "emobase2010" with some adjustments to the normalization of duration and location features. This feature set contains a significantly enhanced set of low-level descriptors (LLDs), as well as a list of functions that are more finely selected than in 'emobase'. It is recommended that this feature set be used as a reference for comparing new emotion recognition feature sets and methods, as it represents the most current state-of-the-art in emotion and language recognition.

3.3 FNC Network

3.3.1 Similarity calculation

This paper presents the feature normalized convolutional similarity algorithm FNC, a network consisting of two parts, the processing of representative video music features and the calculation of similarity between video and music pairs. In this paper, the extracted features are normalized and weighted based on an attention mechanism. To estimate the similarity between video and music, the similarity matrices of the video and music with similarity are fed into a CNN network that can perform similarity learning to obtain the similarity of the video-music pairs, and then the chamfer distance (CD)Zhang et al (2020) is used to calculate the final score. The chamfer similarity (CS) is the similarity counterpart of the chamfer distance. Consider two sets of items x and y, with a total number of items N and M respectively, and their similarity matrices $S \in \mathbb{R}^{N \times M}$, CS is calculated as the average similarity of the most similar items in set y for each item in set x.

In order to make each feature vector contribute equally to the similarity calculation and to consider all feature vectors equally, this paper introduces L2 normalization for the extracted feature vectors. However, there are some problems. In the video region, different key frames or scenes produce different effects. Similarly, in music data, the impact of a clip with sound should be different from that of a clip without sound. Therefore, this paper invokes the attention mechanism to weight the feature vectors of video and music.

The following attention mechanism is constructed in this paper: for the feature vectors of video and music, respectively $V_{i,j}$ and $M_{i,k}$, where $i \in [1, 2, 3]$, and $j \in [1, X], k \in [1, Y]$. We introduce the contextual unit vector u and use it to measure the importance of each region vector. To do this, we use the context vector u to calculate the importance of each $V_{i,j}$ and $M_{i,k}$ dot product between the region vectors to obtain a weight score a_{ij} that b_{ik} . Since all vectors are unit vectors after the normalization

process, the a_{ij} , the b_{ik} will remain between [-1, 1]. To make the direction of the region vectors consistent, we normalize the weight fraction a_{ij} ,b_{ik} divided by 2 and added 0.5 to control within [0, 1].

$$a_{ij} = \sum_{i=1}^{3} \sum_{j=1}^{x} V_{i,j} softmax(s(V_{i1,j1}, V_{i2,j2})) \tag{3}$$

$$b_{ik} = \sum_{i=1}^{3} \sum_{k=1}^{Y} M_{i,k} softmax(s(M_{i1,k1}, M_{i2,k2})) \tag{4}$$

Where $s(x, y)$ is the dot product model, $s(x, y) = x^T y$.

Then, we connect the processed video and music feature vectors into a group according to each feature level, and obtain the video feature vector group p_j and music feature vector group q_k. The dot product calculation is carried out to obtain the feature matrix $S^{pq} \in \mathbb{R}^{X \times Y}$.

$$S^{pq} = p_j \odot q_k \tag{5}$$

The generated feature matrix is then S^{pq} into a four-layer convolutional network that has the ability to capture segment-level temporal patterns of video and music similarities. Due to the fact that the convolution results in CNN networks can be considered as the inner product of vectors composed of convolution kernels and convolution regions, and the inner product represents the similarity between two vectors, the model can learn similar patterns in CNN subnets by manipulating the similarity matrix between feature vectors. These similarity matrices containing all paired feature vectors are fed to a CNN for training a video music level similarity model. We chose three commonly used 3×3 convolutional kernels, requiring significantly fewer parameters than a 7×7 convolutional kernel, which definitely reduces the complexity of the model, speeds up training and preserves as much detailed data as possible. We replaced the fully-connected layer with a 1×1 convolution kernel, which does not destroy the spatial structure of the feature matrix image and is no longer subject to the requirement of a fully-connected layer with fixed inputs. The maximum pooling layer we chose a 2×2 filter with a 2×2 stride, and the convolutional layer stride was set to 1×1. The convolutional layer setup is shown in Figure 2.

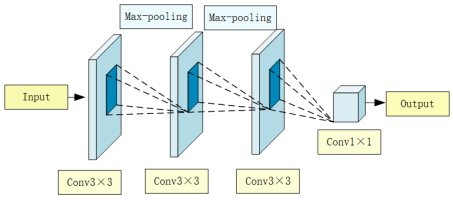

Fig. 2 Structure of convolutional network for similarity calculation

In order to calculate the final similarity result of video music pairs, the output value of the convolutional network will pass through the Htanh activation function,

and the chamfered similarity CS is actually a mean-maximum filter processing on the output value, so that the similarity score F can be obtained, as shown in the formula:

$$F = Score = \frac{1}{X_{output}} \sum_{j=1}^{X_{output}} \max_{k \in [1, Y_{output}]} Htanh(S^{pq}(j,k)) \tag{6}$$

3.3.2 Loss function

F, as the target music similarity score of the video, should be higher for relevant music and lower for irrelevant music. Based on this objective, we used the most reasonable triadic loss function Schroff et al (2015) that allows the network to match higher scores to positive video music pairs and vice versa to match lower scores to negative video music pairs. We use the input video data as the baseline (Anchor), music with similarity as Positive music, and music that differs too much from the video as Negative music. The loss function has the following equation, where α is the margin parameter.

$$L_1 = \max\{F(A,P) - F(A,N) + \alpha, 0\} \tag{7}$$

In addition, we introduce a similarity regularization loss function, which provides significant performance improvement. The range of Hard tanh activation function is [-1,1], and the mechanism of loss function can drive the network to generate output matrix within this range. The sum of the values of all output similar matrices outside this range is the regularization loss.

$$L_2 = \sum_{j=1}^{X_{output}} \sum_{k=1}^{Y_{output}} \max\{S^{pq}(j,k) - 1, 0\} + \min\{S^{pq}(j,k) + 1, 0\} \tag{8}$$

The final loss function we define as:

$$L = L_1 + \beta L_2 \tag{9}$$

Where β is the regularization hyperparameter that adjusts the contribution of the similarity regularization to the total loss.

4 Experiment

4.1 Datasets

HIMV-200KHong et al (2018) , a dataset consisting of 200,500 video-music pairs. These video-music pairs were obtained from YouTube-8M, a large-scale tagged video dataset consisting of millions of YouTube video IDs and associated tags. Throughout

the videos associated with thousands of entities, all videos tagged with "music video" were downloaded and then divided into video and audio components using FFmpeg. A total of 205,000 video-music pairs were obtained, of which 200K were used for training, 4K for validation and 1K for testing.

Pop music videos datasetLin and Yang (2021) , pop music videos have a large number of camera angles, shots and movements that help to learn the relationship between the various videos and the corresponding music. 1280 music videos and corresponding music were collected from YouTube channels and Warner Music. Segmenting each video music pair, one can obtain 5120 samples, and we divide this dataset into: 3600 training video music pairs, 1320 validation video music pairs and 200 test video music pairs.

The self-built dataset, in order to demonstrate the practical applicability of the recommendation task, is collected from Tik Tok and PMEmoZhang et al (2018) respectively, downloading the required video and music data. The music data contains 1794 songs with sentiment annotations and a corresponding collection of videos with similar category labels. Of these, 1600 data were used for training, 150 for validation and 44 for testing.

4.2 Parameter setting and evaluation criteria

Training the above architecture requires the organization of the dataset used for video music triad training. Therefore, we extract video music pairs with relevant background music content to be used as anchor positive pairs during training. We extract positive pairs from both the pop music video dataset and the Tik Tok video dataset by selecting video music pairs whose distance between the video and music feature vectors is less than a certain value. We then create video triples based on the positive pairs by selecting videos that serve as hard negative examples.In other words, we select the feature space where the Euclidean distance is less than the distance between the anchored positive pairs plus the edge value d of all anchored negative pairs, i.e. $D(A, N) < D(A, P) + d$, where $D(.,.)$ denotes the Euclidean distance between any pair of video music. d value is set to 0.15 based on experience. To train the data, we can only supply the network with one video music triad at a time due to GPU memory limitations. We used Adam optimization with the learning rate set to 1×10^{-5} .

For each period, T=1000 triples were selected for each pool. The model was trained for 100 periods, i.e. 200K iterations, and the best network was selected based on the mean accuracy (mAP) on the validation set. Other parameters were set to α=0.5, β=0.1 and W=64, and the weights of the feature extraction CNN and the whitening layer were kept constant.

We refer to other methods, and finally choose Recall rate, Precision and Mean Average Precision as the evaluation criteria.

1. Recall Rate is the proportion of correct predictions that are positive to all actual positive predictions.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Table 2 Experimental results of the model on HIMV-200K dataset(%)

Modality	R@1	R@10	R@25	mAP
Expected Value	0.1	1	2.5	
SF	0.11	1.11	2.68	54.3
DF	0	1.08	2.59	53.5
EF	0.12	1.17	2.94	55.9
Ours(-L2-AT)	0.11	1.22	3.25	58.2
Ours(+L2+AT)	**0.15**	**1.33**	**3.72**	**42.2**

Table 3 Experimental results of the model on the dataset of Pop music videos(%)

Modality	R@1	R@10	R@25	mAP
Expected Value	0.5	5	12.5	
SF	0.53	5.38	12.69	53.1
DF	0.59	5.84	13.94	55.7
EF	0.52	5.26	14.77	54.6
Ours(-L2-AT)	0.61	5.97	15.23	57.3
Ours(+L2+AT)	**0.66**	**6.24**	**18.27**	**64.8**

We applied the percentage recall at top K (Recall@K) metric, which is widely used for cross-modal searches. For a given value of K, Recall@K represents the number of relevant top-K ranked items divided by the total number of relevant items. From this we can obtain a baseline data across different datasets as a way of judging the performance of the model.

2. Precision, also known as the accuracy rate, is the proportion of positive correct forecasts to all positive forecasts.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Accuracy represents the degree of predictive accuracy in positive sample results, while precision represents the overall predictive accuracy, including both positive and negative samples.

3. Mean Average Precision

$$AP = \frac{1}{R} \sum Precision(rank) = L_1 + \beta L_2 \tag{12}$$

$$MAP = \frac{1}{C} \sum AP \tag{13}$$

4.3 Experiment

4.3.1 Ablation experiments

In order to evaluate the performance of MFF-VBMR, a video background music recommendation model based on multi-layered fusion features proposed in this paper, and to analyze the impact of each layer of features on the recommendation effect

Table 4 Experimental results of the model on self-built dataset(%)

Modality	R@1	R@10	R@25	mAP
Expected Value	2.27	22.73	56.82	
SF	2.35	24.05	62.39	54.2
DF	2.31	23.89	60.54	53.8
EF	2.47	25.95	65.71	58.6
Ours(-L2-AT)	2.82	28.66	68.51	59.5
Ours(+L2+AT)	**3.22**	**32.77**	**74.56**	**63.9**

of the model, we conducted ablation experiments on the HIMV-200K dataset, Pop music videos dataset and self-built data respectively. The experiments were divided into three groups as follows.

1. In the feature extraction stage, we extract only the static features, dynamic features and emotion features of the video and music respectively. In other words, we want to analyze the image features of the video key frames and the spectral features of the music, the rhythm and tempo of the video and the music, and the emotional state of the video and the emotion respectively. On the basis of these single features obtained, they are fed into the FNC network, the similarity algorithm in this paper, and the performance of the model is judged on the basis of recall and average precision mean.

2. This set of experiments uses the method proposed in this paper for music recommendation based on multi-level fusion features. The complete set of three global features is fed into the FNC network and compared with the above experiments to judge the performance of the model and to evaluate the model in this paper.

3. Based on the second set of experiments, we no longer apply the feature normalization and attention mechanisms to compare differences in model performance.

The results of the experiments are shown in Tables 2-4, where SF, DF and EF are using only static features, for dynamic features and emotional features respectively, and AT is the attention network.The expected values in the table refer to the theoretical value of R@K under this test dataset and are used to evaluate whether the method achieves a passing score.

It is evident from the experimental results that, regardless of the dataset on which the experiments were conducted, the recall and mean accuracy of the model using multi-layered fusion features for recommendation were higher than those using only single-layered features, and our proposed model performed well on the background music recommendation task. This is because although single-layer features make connections between the complex relationships between visual and auditory elements, only local object features of video and music are captured, when other global contextual information is particularly important and needs to complement the single features.

It is also experimentally demonstrated that the performance of the model improves somewhat when the method in this paper is not subjected to L2 normalization and attention mechanism, which is due to the fact that the multi-level features make the connection between video and music closer, but the performance of the model is further improved when L2 normalization and attention mechanism are added. This is because

Table 5 Comparison results of different models on HIMV-200K dataset(%)

Method	R@1	R@10	R@25	mAP
Expected Value	0.1	1	2.5	
DCCA	0.11	1.11	2.74	59.8
TBVMN	0.11	1.15	2.89	55.6
CBVMR	0.12	1.23	2.78	57.3
AVCA	0.14	1.25	3.09	58.4
EMVGAN	0.13	1.19	2.84	59.1
Ours	**0.15**	**1.33**	**3.72**	**62.2**

this feature processing method is equivalent to imposing a hard constraint on the two branching features of video and music, resulting in an increase in the recognition accuracy of the model.

The videos and songs in the Pop music videos dataset are highly correlated, and in particular the dance moves and musical rhythms within the videos snap together well, so the model performs slightly better than the HIMV-200K dataset and the self-built dataset in this respect when using dynamic features alone for their music recommendations.

4.3.2 Comparative experiments

In the comparison experiments, this paper selects some mainstream cross-modal recommendation modeling approaches to compare their effectiveness with the proposed MFF-VBMR model. TBVMNJin et al (2020) and CBVMRHong et al (2018) both have a two-branch structure, which enables video music retrieval function by constructing a common subspace, while CBVMR also enables music-to-video inverse retrieval. EMVGAN Tsai et al (2022) achieves the task of music recommendation by constructing a sentiment public embedding space to bridge the heterogeneity gap between different data modalities. AVCALi et al (2019) extracts global and local features from visual and audio signals, and then constructs a unified framework consisting of global and local embedding networks for sentiment video content analysis. The DCCA Andrew et al (2013) model is often used for the task of cross-modal graphical retrieval, using this model also performs relatively well on the task of video background music recommendation. In this paper, comparative experiments are conducted on the HIMV-200K dataset, Pop music videos dataset and self-built dataset respectively.

Tables 5-7 respectively shows the recall rates and average accuracy averages of different cross-modal recommendation methods in the three data sets. From the experimental results, it can be seen that the values of Recall@K all exceed the desired data, indicating that all these methods achieve the desired goals and are good at the task of cross-modal retrieval and recommendation. Compared with other traditional methods, the performance of the model designed in this paper is more superior.

The PR curves shown in Figures 3-5 visually show the performance of the six methods on the three datasets, from which the performance differences between the models can be clearly and intuitively seen. The model has achieved good results on the three datasets.

Table 6 Comparison results of different models on Pop music videos dataset(%)

Method	R@1	R@10	R@25	mAP
Expected Value	0.5	5	13.5	
DCCA	0.56	5.54	15.92	55.8
TBVMN	0.57	5.77	14.46	59.6
CBVMR	0.61	5.6	16.79	57.4
AVCA	0.62	5.96	17.33	59.3
EMVGAN	0.63	5.86	16.2	60.2
Ours	**0.66**	**6.24**	**18.27**	**64.8**

Table 7 Comparison results of different models on self-built dataset(%)

Method	R@1	R@10	R@25	mAP
Expected Value	2.27	22.73	56.82	
DCCA	2.46	25.32	65.71	57.5
TBVMN	2.5	27.88	67.89	56.9
CBVMR	2.67	28.34	69.15	57.8
AVCA	2.74	29.52	70.09	58.3
EMVGAN	2.81	30.07	69.83	60.4
Ours	**3.22**	**32.77**	**74.56**	**63.9**

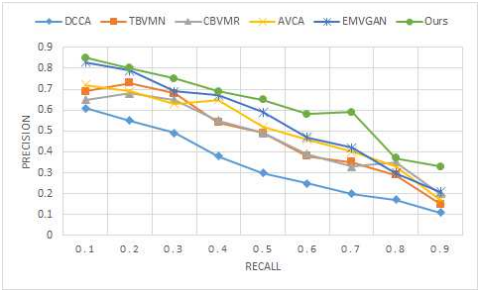

Fig. 3 Model comparison results on HIMV-200K dataset

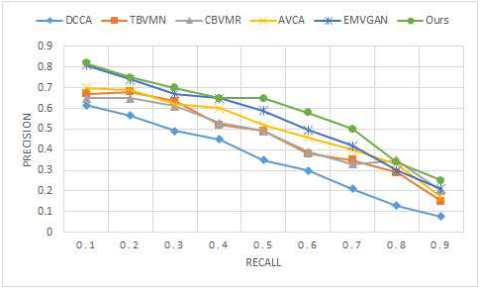

Fig. 4 Model comparison results on Pop music videos dataset

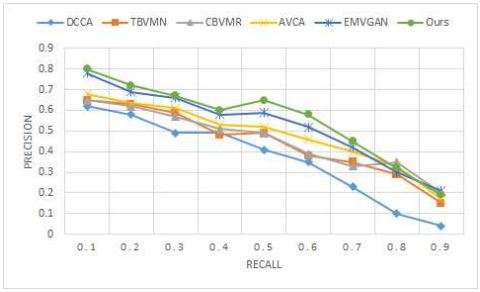

Fig. 5 Model comparison results on self-built dataset

5 Conclusion

In this paper, we propose a video background music recommendation method based on multi-level fusion features, and design a convolutional similarity calculation network. This approach exploits the multimodal information of video and music to achieve the task of matching appropriate background music recommendations for a given video. Experimental results show that the proposed model improves the recommendation performance and achieves higher accuracy. This method has a certain reference value for the future cross-mode recommendation application. In the future, our study will further investigate other more cross-modal recommendation scenarios and consider improving the algorithm to improve computational efficiency based on large amounts of data information.

Acknowledgments. Not Applicable.

Author contributions. Xin Zhao wrote the main manuscript text, Xiaobing Li, Yun Tie, and Lin Qi contributed to Figure 1 and the experimental section. All authors reviewed the manuscript.

Funding. Not Applicable.

Data Availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical Approval. Not Applicable.

Competing interests. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Andrew G, Arora R, Bilmes JA, et al (2013) Deep canonical correlation analysis. In: International Conference on Machine Learning

Antol S, Agrawal A, Lu J, et al (2015) Vqa: Visual question answering. International Journal of Computer Vision 123(1):4–31

Bourlard H, Dupont S (1996) A mew asr approach based on independent processing and recombination of partial frequency bands. Proceeding of Fourth International Conference on Spoken Language Processing ICSLP '96 1:426–429 vol.1

Chen Q, Wang W, Huang K, et al (2021) Multi-modal generative adversarial networks for traffic event detection in smart cities. Expert Systems with Applications p 114939

Chen W, Hsu W, Lee ML (2013) Making recommendations from multiple domains. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining

Cortes C, Vapnik VN (1995) Support-vector networks. Machine Learning 20:273–297

Cowen AS, Fang X, Sauter D, et al (2020) What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. Proceedings of the National Academy of Sciences 117(4):1924–1934

Cristani M, Pesarin A, Drioli C, et al (2010) Toward an automatically generated soundtrack from low-level cross-modal correlations for automotive scenarios. In: Proceedings of the 18th International Conference on Multimedea 2010, Firenze, Italy, October 25-29, 2010

Davis A, Agrawala M (2018) Visual rhythm and beat. ACM Transactions on Graphics 37(4CD):1–11

Eyben F, Wllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In: Acm International Conference on Multimedia

He K, Zhang X, Ren S, et al (2015) Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 770–778

Heitmann B, Dabrowski M, Passant A, et al (2012) Personalisation of social web services in the enterprise using spreading activation for multi-source, cross-domain recommendations

Ho TK (1995) Random decision forests. IEEE Computer Society

Hodosh M, Young P, Hockenmaier J (2015) Framing image description as a ranking task: Data, models and evaluation metrics. In: International Conference on Artificial Intelligence, pp 853–899

Hong S, Im W, Yang HS (2018) Cbvmr: Content-based video-music retrieval using soft intra-modal structure constraint. Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval

Hoxha J (2014) Learning relevance of web resources across domains to make recommendations. In: International Conference in Machine Learning and Applications (ICMLA '13)

Jia Z, Zhenming Y, Kai Y (2014) Cross media recommendation in digital library. Springer International Publishing, pp 208–217

Jin C, Zhang T, Liu S, et al (2020) Cross-modal deep learning applications: Audio-visual retrieval. In: ICPR Workshops

Kumar A, Kumar N, Hussain M, et al (2014) Semantic clustering-based cross-domain recommendation. 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) pp 137–141

Kuo FF, Chiang MF, Shan MK, et al (2005) Emotion-based music recommendation by association discovery from film music. In: Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005

Kuo FF, Shan MK, Lee SY (2013) Background music recommendation for video based on multimodal latent semantic analysis. 2013 IEEE International Conference on Multimedia and Expo (ICME) pp 1–6

Li B, Chen Z, Li S, et al (2019) Affective video content analyses by using cross-modal embedding learning features. 2019 IEEE International Conference on Multimedia and Expo (ICME) pp 844–849

Li C, Wei Z, Quan Y (2013) A unified framework for recommending items, groups and friends in social media environment via mutual resource fusion. Expert Systems with Applications 40(8):2889–2903

Lin CT, Yang M (2021) Inversemv: Composing piano scores with a convolutional video-music transformer. ArXiv abs/2112.15320

Lin JC, Wei WL, Wang HM (2015) Emv-matchmaker: Emotional temporal course modeling and matching for automatic music video generation. In: Proceedings of the 23rd ACM International Conference on Multimedia, p 899–902, https://doi.org/10.1145/2733373.2806359, URL https://doi.org/10.1145/2733373.2806359

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

Ly ST, Do NT, Kim S, et al (2019) A novel 2d and 3d multimodal approach for in-the-wild facial expression recognition. Image Vis Comput 92

McFee B, Raffel C, Liang D, et al (2015) librosa: Audio and music signal analysis in python. In: SciPy

Min W, Bao BK, Xu C, et al (2015) Cross-platform multi-modal topic modeling for personalized inter-platform recommendation. IEEE Transactions on Multimedia 17(10):1787–1801

Roy SD, Tao M, Zeng W, et al (2012) Socialtransfer: Cross-domain transfer learning from social streams for media applications. In: Acm International Conference on Multimedia

Samadiani N, Huang G, Luo W, et al (2019) A novel video emotion recognition system in the wild using a random forest classifier. In: International Conference on the Digital Society

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 815–823

Shah RR, Yi Y, Zimmermann R (2014) User preference-aware music video generation based on modeling scene moods. In: Proceedings of the 5th ACM Multimedia Systems Conference

Sharma V, Gaded AS, Chaudhary D, et al (2021) Emotion-based music recommendation system. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) pp 1–5

Sievers B, Polansky L, Casey M, et al (2013) Music and movement share a dynamic structure that supports universal expressions of emotion. Proceedings of the National Academy of Sciences 110(1):70–75

Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems 1

Surís D, Duarte AC, Salvador A, et al (2018) Cross-modal embeddings for video and audio retrieval. In: ECCV Workshops

Tan S, Bu J, Qin X, et al (2014) Cross domain recommendation based on multi-type media fusion. Neurocomputing 127:124–134

Tsai YC, Pan TY, Kao TY, et al (2022) Emvgan: Emotion-aware music-video common representation learning via generative adversarial networks. Proceedings of the 2022 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia

Tzanetakis G (2001) Automatic musical genre classification of audio signals. In: ISMIR 2001, 2nd International Symposium on Music Information Retrieval, Indiana University, Bloomington, Indiana, USA, October 15-17, 2001, Proceedings

Valstar M, Schuller B, Smith K, et al (2013) Avec 2013: the continuous audio/visual emotion and depression recognition challenge. ACM

Wang HL, Cheong LF (2006) Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology 16(6):689–704

Wang J, Cellary W, Wang D, et al (2015) Cross-domain collaborative recommendation by transfer learning of heterogeneous feedbacks 10.1007/978-3-319-26187-4(Chapter 13):177–190

Wang JC, Yang YH, Wang HM, et al (2012) The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In: ACM Multimedia

Wei J, Yang X, Dong Y (2021) User-generated video emotion recognition based on key frames. Multimedia Tools and Applications 80:14343–14361

Yi Y, Shen Z, Zimmermann R (2012) Automatic music soundtrack generation for outdoor videos from contextual sensor information. In: Proceedings of the 20th ACM international conference on Multimedia

Zhang H, Tang Z, Xie Y, et al (2020) A similarity-based burst bubble recognition using weighted normalized cross correlation and chamfer distance. IEEE Transactions on Industrial Informatics 16:4077–4089

Zhang K, Zhang H, Li S, et al (2018) The pmemo dataset for music emotion recognition. Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval