ADVISOR – Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings

Rajiv Ratn Shah, Yi Yu, Roger Zimmermann
School of Computing, National University of Singapore
Singapore 117417
{rajiv,yuy,rogerz}@comp.nus.edu.sg

ABSTRACT

Capturing videos anytime and anywhere, and then instantly sharing them online, has become a very popular activity. However, many outdoor user-generated videos (UGVs) lack a certain appeal because their soundtracks consist mostly of ambient background noise. Aimed at making UGVs more attractive, we introduce *ADVISOR*, a personalized video soundtrack recommendation system. We propose a fast and effective heuristic ranking approach based on heterogeneous late fusion by jointly considering three aspects: venue categories, visual scene, and user listening history. Specifically, we combine confidence scores, produced by SVM^{hmm} models constructed from geographic, visual, and audio features, to obtain different types of video characteristics. Our contributions are threefold. First, we predict scene moods from a real-world video dataset that was collected from users' daily outdoor activities. Second, we perform heuristic rankings to fuse the predicted confidence scores of multiple models, and third we customize the video soundtrack recommendation functionality to make it compatible with mobile devices. A series of extensive experiments confirm that our approach performs well and recommends appealing soundtracks for UGVs to enhance the viewing experience.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor fusion*

Keywords

Video soundtrack generation; geographic category; user preference; scene mood prediction; music retrieval

1. INTRODUCTION

In the era of ubiquitous availability of mobile devices with wireless connectivity, user-generated videos (UGV) have become popular since they can be easily acquired using most modern smartphones or tablets and are instantly available for sharing on social media web sites (*e.g.*, YouTube, Vimeo, Dailymotion). In addition, people enjoy listening to music online. Thus, various user-generated data of online activities (*e.g.*, sharing videos, listening to music) can be rich sources containing users' preferences. It is very interesting to extract activity-related data with a user-centric point of view. Exploiting such data may be very beneficial to individual users, especially for preference-aware multimedia recommendations [35]. We consider location (*e.g.*, GPS information) and online listening histories as user-centric preference-aware activities. We categorize user activity logs from different data sources, correlate them with user preferences by using semantic concepts, *i.e.*, moods, and leverage them to complement recommendations for personal multimedia events. To enhance the appeal of a UGV for viewing and sharing, we have designed ADVISOR, which replaces the ambient background noise of a UGV with a soundtrack that matches both the video scenes and a user's preferences. A generated music video (the UGV with the recommended soundtrack) enhances the video viewing experience because it not only provides the visual experience but simultaneously renders music that matches the captured scenes and locations. ADVISOR can be used in many applications such as to recommend music for a slideshow of sensor-rich Flickr images or for outdoor UGV live streaming, *etc.*

In terms of the target environment, this work mainly studies soundtrack recommendations for outdoor UGVs in places where different geo-categories such as beach, temple, *etc*, would be relevant. Thus, ADVISOR may not work well for indoor scenes (*e.g.*, parties). The reader may imagine the following scenario: a mom brings her son outdoors where she records a video of the little boy playing on a beach and swimming in the sea. Subsequently they would like to add music of their own style to this video to make it more appealing. Since video and audio have quite different low level features they are linked via a high level semantic concept – moods – in this work. As shown in Figure 1, the ADVISOR system consists of two parts: an offline training and an online processing component. Offline a training dataset with geo-tagged videos is used to train SVM^{hmm} models that map videos to mood tags. The online processing is further divided into two modules: a smartphone application and a server backend system. The smartphone application allows users to capture sensor-annotated videos[1]. Geographic contextual information (*i.e.*, geo-categories such as *Theme*

MM '14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654919.

[1]We use the terms sensor-annotated videos and UGVs interchangeably in this study to refer to the same outdoor videos acquired by our custom Android application.

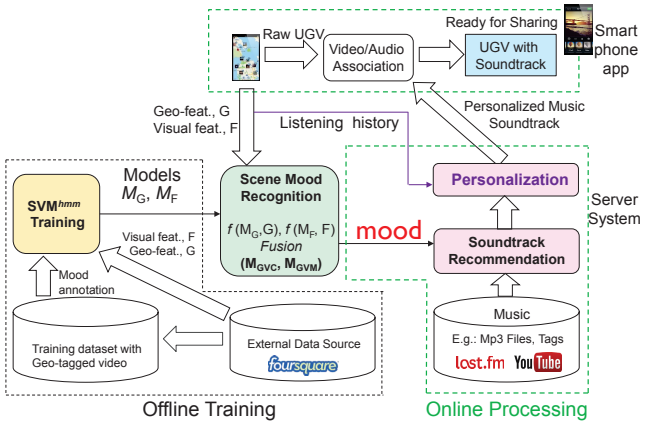

Figure 1: System overview of soundtrack recommendations for UGVs with ADVISOR.

Park, *Lake*, *Plaza* and others derived from Foursquare[2]) captured by geo-sensors (GPS and compass), can serve as an important dimension to represent valuable semantic information of multimedia data while video frame content is often used in scene understanding. Hence, scene moods are embodied in both the geographic context and the video content. The sensor data streams are mapped to a geo feature G, and a visual feature F is calculated from the video content. With the trained models, G and F are mapped to mood tags. Then, songs matching these mood tags are recommended. Among them, the songs matching a user's listening history are considered as user preference-aware.

In the ADVISOR system, first, we classify the 20 most frequent mood tags of Last.fm[3] into four mood clusters (see Table 1) and then use these mood tags and mood clusters to generate ground truths for the collected music and video datasets. Next, in order to effectively exploit multimodal (geo, visual and audio) features, we propose late fusion methods to predict moods for a UGV. We construct two offline learning models (see M_{GVM} and M_{GVC} in Figure 1) which predict moods for the UGV based on the late fusion of geo and visual features. Furthermore, we also construct an offline learning model (Figure 2, M_{Eval}) based on the late fusion of visual and concatenated audio features (MFCC, mel-spectrum and pitch [33]) to learn from the experience of experts who create professional *soundtracks* in Hollywood movies. We leverage this experience in the automatic selection of a matching soundtrack for the UGV using M_{Eval} (see Figure 2). We deploy these models (M_{GVM}, M_{GVC} and M_{Eval}) in the backend system. To generate the music soundtrack for the UGV, the Android application first uploads its recorded sensor data and selected key-frames to the backend system. Next, the backend system computes geo and visual features for the UGV and forwards these features to M_{GVM} and M_{GVC} to predict scene *mood tags* and *mood clusters*, respectively, for the UGV. Moreover, we also construct a novel heuristic method to retrieve a list of songs from an offline music database based on the predicted scene moods of the UGV. The soundtrack recommendation component of the backend system re-ranks a list of songs retrieved by the heuristic method based on user preferences and recommends them for the UGV (see Figure 5). Next, the backend system determines the most appropriate song from the recommended list by comparing the characteristics of a composition of a selected song and the UGV with a *soundtrack* dataset of Hollywood movies of all movie genres using the learning model M_{Eval}. Finally, the Android application generates a music video using that song as a soundtrack for the UGV. The remaining parts of this paper are organized as follows. In Section 2, we review the related literature, and we describe the ADVISOR system in Section 3. The experiments and results are presented in Section 4. Finally, we conclude the paper with a summary in Section 5.

[2]www.foursquare.com

[3]Last.fm is a popular music website.

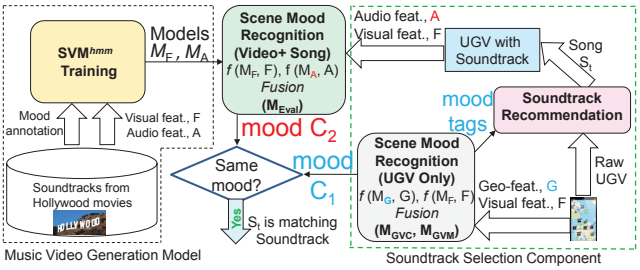

Figure 2: Soundtrack selection process for UGVs in ADVISOR.

Table 1: Four mood clusters.

Mood Cluster	Cluster Type	Mood Tags
$Cluster_1$	High Stress, High Energy	Angry, Quirky, Aggressive
$Cluster_2$	Low Stress, High Energy	Fun, Playful, Happy, Intense, Gay, Sweet
$Cluster_3$	Low Stress, Low Energy	Calm, Sentimental, Quiet, Dreamy, Sleepy, Soothing
$Cluster_4$	High Stress, Low Energy	Bittersweet, Depressing, Heavy, Melancholy, Sad

2. RELATED WORK

Our purpose is to support real-time user preference-aware video soundtrack recommendations via mobile devices. The steps of such a process can be described as follows. (1) A user captures a video on a smartphone. (2) An emotion cognition model predicts video scene moods based on a heterogeneous late fusion of geo and visual features. (3) A list of songs are automatically recommended for the video matching with the user's listening history. (4) The system sets the most appropriate song from the recommended list as the soundtrack for the video by leveraging the experience of professional mood-based associations between music and movie contents. In this section, we briefly provide some recent progress on emotion recognition and music recommendation systems and techniques.

Despite significant efforts that have focused on music recommendation techniques [9, 14, 18, 22] in recent years, little attention has been paid to music recommendation for sets of images or UGVs. Kuo *et al.* [14] investigated the association discovery between emotions and music features of film music and proposed an emotion-based music recommendation system. As of now, the music recommendation area for a set of images has been largely unexplored and consists of only a few state-of-the-art approaches such as an emotion-based impressionism slideshow system from images

of paintings by Li *et al.* [16]. This method extracts features such as the dominant color, the color coherence vector, and the color moment for color and light. It also extracts some statistical measures from the gray level co-occurrence matrix for textures, and computes the primitive length of textures. Furthermore, Wei *et al.* [30] tried to establish an association between color and mood by exploiting the color-related features using an SVM classifier.

There exist a few approaches [6, 27, 29] to recognize emotions from videos but the field of video soundtrack recommendation for UGVs [24, 34] is largely unexplored. Hanjalic *et al.* [6] proposed a computational framework for affective video content representation and modeling based on the dimensional approach to affect. They developed models for arousal and valance time curves using low-level features extracted from video content, which map the affective video content onto a 2D emotion space characterized by arousal and valence. Soleymani *et al.* [27] introduced a Bayesian classification framework for affective video tagging which takes contextual information into account since emotions that are elicited in response to a video scene contain valuable information for multimedia indexing and tagging. Based on this, they proposed an affective indexing and retrieval system which extracts features from different modalities of a movie, such as video, audio, *etc.* To understand the affective content of general Hollywood movies, Wang *et al.* [29] formulated a few effective audiovisual cues to help bridge the affective gap between emotions and low-level features. They introduced a method to extract affective information from multifaceted audio streams and classified every scene of Hollywood domain movies probabilistically into affective categories. They further processed the visual and audio signals separately for each scene to find the audio-visual cues and then concatenated them to form scene vectors which were sent to a SVM to obtain probabilistic membership vectors. Audio cues at the scene level were obtained using the SVM and the visual cues were computed for each scene by using the segmented shots and key-frames.

Cristani *et al.* [4] introduced a music recommendation policy for a video sequence taken by a camera mounted on board a car. They established the association between audio and video features from low-level cross-modal correlations. Yu *et al.* [34] presented a system to automatically generate soundtracks for UGVs based on their concurrently captured contextual sensor information. The proposed system correlates viewable scene information from sensors with geographic contextual tags from OpenStreetMap[4] to investigate the relationship between geo-categories and mood tags. Since the video soundtrack generation system by Yu *et al.* does not consider the visual content of the video or the contextual information other than geo-categories, soundtracks recommended by this system are very subjective. Furthermore, the system used a pre-defined mapping between geo-categories and mood tags, and hence the system is not adaptive in nature. In our recent work [24], we recommend soundtracks for a UGV based on modeling scene moods using a SVM^{hmm} model. In particular, first, the SVM^{hmm} model predicts scene moods based on the sequence of concatenated geo and visual features. Next, a list of matching songs corresponding to the predicted scene moods are retrieved.

Currently, sensor rich media content is receiving increasing attention because sensors provide additional external information such as location from GPS, viewing direction from a compass unit, and so on. Sensor-based media can be useful in applications such as life log recording, location-based queries and recommendations [1]. Kim *et al.* [11] discussed the use of textual information such as web documents, social tags and lyrics to derive an emotion of a music sample. Rahmani *et al.* [19] proposed context-aware movie recommendation techniques based on background information such as users' preferences, movie reviews, actors and directors of movie, *etc.* Chen *et al.* [2] proposed an approach by leveraging a tripartite graph (user, video, query) to recommend personalized videos. Kaminskas *et al.* [9] proposed a location-aware music recommendation system using tags, which recommends songs that suit a place of interest. Park *et al.* [18] proposed a location-based recommendation system based on location, time, the mood of a user and other contextual information in mobile environments. Recently, Schedl *et al.* [22] proposed a few hybrid music recommendation algorithms that integrate information of the music content, the music context, and the user context, to build a music retrieval system. For the ADVISOR system, these earlier works inspired us to mainly focus on sensor-annotated videos that contain additional information provided by sensors and other contextual information such as a user's listening history, music genre information, *etc.*

Multi-feature late fusion techniques have been advocated in various applications such as video event detection and object recognition [32]. Snoek *et al.* [25, 26] performed early and late fusion schemes for semantic video analysis and found that a late fusion scheme performs better. Ghias *et al.* [5] and Lu *et al.* [17] used heuristic approaches for querying desired songs from a music database by humming a tune. These earlier works inspired us to build the ADVISOR system by performing heterogeneous late fusion to recognize moods from videos and retrieve a ranked list of songs using a heuristic approach. To the best of our knowledge, this is the first work that correlates preference-aware activities from different behavioral signals of individual users, *e.g.*, online listening activities and physical activities.

3. SYSTEM DESCRIPTION

To generate a music video for a UGV, ADVISOR first predicts scene moods from the UGV using learning models described next in Section 3.1. The scene moods used in this study are the 20 most frequent mood tags of Last.fm, described in detail in Section 4.1.1. Next, the *soundtrack recommendation component* in the backend system recommends a list of songs, using a heuristic music retrieval method, described in Section 3.2. Finally, the *soundtrack selection component* selects the most appropriate song from the recommended list to generate the music video of the UGV, using a novel method, described in Section 3.3.

3.1 Scene Moods Prediction Models

In our custom Android recording app, a continuous stream of geo-sensor information is captured together with each video using GPS sensors. This sensor information is mapped to geo-categories such as *Concert Hall*, *Racetrack*, and others using the Foursquare API. Then the geo-categories are mapped to a geo-feature G using the bag-of-word model. With the trained SVM^{hmm} model (M_G), mood tags C_G with geo-aware likelihood are generated. Furthermore, a vi-

[4]www.openstreetmap.org

Table 2: SVM^{hmm} models used in this study.

Model	Input-1	Input-2	Output
M_F	F	-	T
M_G	G	-	T
M_A	A	-	C
M_{GVC}	M_G	M_F	C = $f_1(M_G, M_F)$
M_{GVM}	M_G	M_F	T = $f_2(M_G, M_F)$
M_{Eval}	M_A	M_F	C = $f_3(M_A, M_F)$
M_{Cat}	G	F	T = $f_4(G, F)$

G, F, and A represent the geo, visual and audio features, respectively. T and C denote the set of predicted mood tags and mood clusters, respectively. M_{GVC} and M_{GVM} are models constructed by the late fusion of M_G and M_F. M_{Eval} is constructed by the late fusion of M_A and M_F.

sual feature such as a color histogram is calculated from the video content. With the trained SVM^{hmm} model (M_F), mood tags C_F associated with visual-aware likelihood are generated. In the next step, the mood tags associated with location information and video content are combined by late fusion. Finally, mood tags with high likelihoods are regarded as scene moods of this video.

3.1.1 Geo and Visual Features

Based on the geo-information, a UGV is split into multiple segments with timestamps, with each segment representing a video scene. The geo-information (GPS location) for each video segment is mapped to geo-categories using APIs provided by Foursquare. The Foursquare API also provides distances of geo-categories with respect to the queried GPS location, which describe the typical objects near the video scene in the video segments. We treat each geo-tag as a word and exploit the bag-of-words model [12] on a set of 317 different geo-tags in this study. Next, for each video segment, a geo-feature with 317 dimensions is computed from geo-tags with their score used as weights.

A color histogram [13,24] with 64 dimensions is computed from each UGV video frame by dividing each component of RGB into four bins. Next, the UGV is divided into multiple continuously correlated parts (CCP), within each of which color histograms have high correlations. Specifically, starting with an initial frame, each subsequent frame is regarded as part of the same CCP if its correlation with the initial frame is above a pre-selected threshold. Next, a frame with its timestamp, which is most correlated with all the other frames in the same CCP, is regarded as a key-frame. Color histograms of key-frames are treated as visual features.

3.1.2 Scene Moods Classification Model

Wang *et al.* [29] classified emotions for a video using an SVM-based probabilistic inference machine. To arrange scenes depicting *fear*, *happiness* or *sadness*, Kang [10] used visual characteristics and camera motion with hidden Markov models (HMMs) at both the shot and scene levels. In order to effectively exploit multi-modal features, late fusion techniques have been advocated in various applications and semantic video analysis [25,26,32]. These approaches inspired us to use SVM^{hmm} models [8] based on the late fusion of various features of UGVs to learn the relationships between UGVs and scene moods. Table 2 shows the summary of all the SVM^{hmm} learning models used in this study.

To establish the relation between UGVs and their associated scene moods, we train several offline learning models with the *GeoVid dataset* as described later in Section 4.1.2.

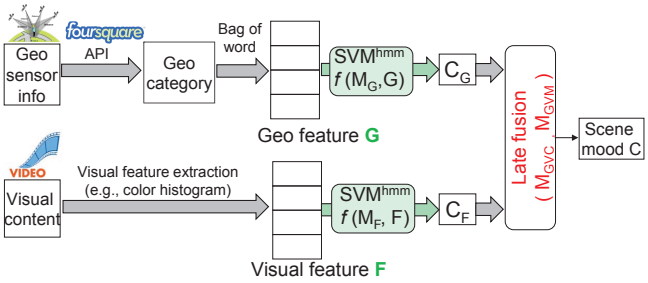

Figure 3: Mood recognition from UGVs using M_{GVC} and M_{GVM} SVM^{hmm} models.

Experimental results in Section 4.2.1 confirm that a model based on late fusion outperforms other models in scene mood prediction. Therefore, we construct two learning models based on the late fusion of geo and visual features and refer to them as *emotion prediction models* in this study. A geo feature computed from geo-categories reflects the environmental atmosphere associated with moods and a color histogram computed from key-frames represents moods in the video content. Next, the sequence of geo-features and the sequence of visual features are synchronized based on their respective timestamps to train *emotion prediction models* using SVM^{hmm} method. Figure 3 shows the process of mood recognition from UGVs based on heterogeneous late fusion of SVM^{hmm} models constructed from geo and visual features. M_{GVC} and M_{GVM} are *emotion prediction models* trained with mood clusters and mood tags, respectively, as ground truths for the training dataset. Hence, M_{GVC} and M_{GVM} predict mood clusters and mood tags, respectively, for a UGV based on heterogeneous late fusion of SVM^{hmm} models constructed from geographic and visual features.

3.1.3 Scene Moods Recognition

UGVs acquired by our Android application are enhanced with geo-information by using sensors such as GPS and compass. When a user requests soundtracks for a UGV then the Android application determines timestamps for multiple video segments of the UGV with each segment representing a video scene based on geo-information of the UGV. Furthermore, the Android application extracts key-frames of the UGV based on timestamps of video segments and uploads them to the backend system along with the geo-information of video segments. The backend system computes geo and visual features of the UGV from the uploaded sensor information and key-frames. The SVM^{hmm} models, M_{GVC}, M_{GVM} and M_{Cat} read the sequence of geo and visual features and recognize moods for the UGV. For example, M_{Cat} is trained with the concatenation of geo and visual features as described in the following sequence (see Figure 4).

$$\langle V, G_1, F_1, m_1 \rangle, \langle V, G_1, F_1, m_2 \rangle, \langle V, G_2, F_1, m_2 \rangle, \ldots$$

In this specific example, in the emotion recognition step, when M_{Cat} is fed with geo features G and visual features F using $f_4(G, F)$, then it automatically predicts a set of scene mood tags $m = \{m_1, m_2, m_2, m_2, m_3\}$ for the UGV.

3.2 Music Retrieval Techniques

We prepared an offline music dataset of candidate songs in all main music genres, with details described later in Section 4.1.3. We refer to this dataset as the *soundtrack dataset*.

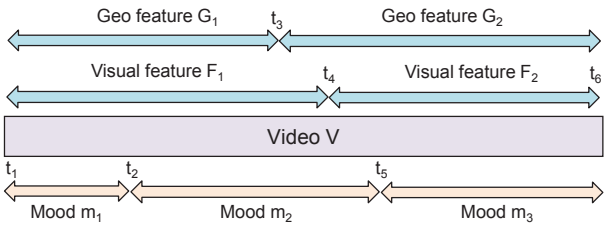

Figure 4: The concatenation model M_{Cat} from [24].

The next step in the ADVISOR system is to find music from the *soundtrack dataset* that matches with both the predicted mood tags and the user preferences. With the given mood tags, the soundtrack retrieval stage returns an initial song list L_1. For this task we propose a novel music retrieval method. Many state-of-the-art methods for music retrieval use heuristic approaches [5, 17]. This inspired us to propose a heuristic method which retrieves a list of songs based on the predicted scene moods by M_{GVM} and M_{GVC}. We take the user's listening history as user preferences, and calculate the correlation between audio features of songs in the initial list L_1 and in the listening history. From the initial list, songs with high correlations are regarded as user specific songs L_2, and recommended to users as video soundtracks.

3.2.1 Heuristic Method for Soundtrack Retrieval

An improvement in mood tag prediction accuracy for a UGV is also an improvement in matching music retrieval because songs in the *soundtrack dataset* are organized in a hash table with mood tags as keys. However, retrieving songs based on only one mood tag suffers from subjectivity because the *mood clusters* prediction accuracy of M_{GVC} is much better than the *mood tags* prediction accuracy of M_{GVM} for an UGV (see Table 5). Since a song may have multiple mood tags, when the *emotion prediction models* predict multiple mood tags, a song may be matched with several tags. Therefore, we calculate the total score of each song to reduce this subjectivity issue and propose a heuristics based on a music retrieval method to rank all the predicted mood tags for the UGV and then normalize them as the likelihood to retrieve the final ranked list L_1 of N songs. Algorithm 1 describes this retrieval process and its *composition operation* $*$ is defined such that it outputs only those most frequent *mood tags* T from the list of mood tags predicted by $f_2(M_G, M_F)$ which belong to the most frequent *mood clusters* predicted by $f_1(M_G, M_F)$. Thus, the *composition operation* $*$ is defined by the following equation:

$$T = f_1(M_G, M_F) * f_2(M_G, M_F)$$

where T, G, F, f_1 and f_2 have the usual meaning, with details described in Table 2.

3.2.2 Post-filtering with User Preferences

A new paradigm shift in music information retrieval (MIR) is currently creating a move from a system-centric perspective towards user-centric approaches. Therefore, addressing user-specific demands in music recommendation is receiving increased attention. User preference-aware music recommendations based on users' preferences observed from their listening history is very common. Music genres of the user's frequently listened to songs are treated as his/her listening preference and later used for the re-ranking of a list of songs

Algorithm 1 Heuristic based song retrieval procedure

procedure HEURISTICSONGSRETRIEVAL(H)
  INPUT: geo and visual features (G, F) of the UGV
  OUTPUT: A ranked list of songs L_1 for the UGV
  $T = f_1(M_G, M_F) * f_2(M_G, M_F)$
  $L = []$ ▷ Initialize with empty list.
  for each mood tag i in T **do**
    $p(i) = likelihood(i)$ ▷ Likelihood of mood tag i.
    $L_t(i) = songList(i)$ ▷ Song list for mood tag i.
    $L = L \cup L_t(i)$ ▷ L has all unique songs.
  end for
  ▷ $isPrsnt$ returns 1 if s is present in $L_t(i)$ else 0.
  ▷ $scr(s, i)$ is the score of song s with mood tag i.
  for each song s in L **do**
    $Score(s) = 0$ ▷ Initialize song score.
    for each mood tag i in T **do**
      $Score(s) += p(i) * scr(s, i) * isPrsnt(s, L_t(i))$
    end for
  end for
  $L_1 = sortSongScore(L)$ ▷ Sort songs.
  Return L_1 ▷ A ranked list of N songs.
end procedure

L_1 recommended by the heuristics method. Our system extracts audio features including MFCC [21] and pitch from audio tracks of the user's frequently listened to songs. These features help in re-ranking the list of recommended songs L_1 by comparing the correlation coefficients of songs matching the genres preferred by the user, and then recommending a list of user preference-aware songs L_2 (see Figure 5). Next, the *soundtrack selection component* automatically chooses the most appropriately matching song from L_2 and attaches it as the soundtrack to the UGV.

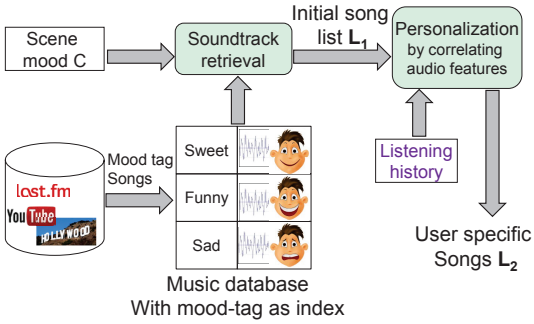

Figure 5: Matching songs with a user's preferences.

3.3 Automatic Music Video Generation Model

Wang *et al.* [29] concatenated audio and visual cues to form scene vectors which were sent to a SVM method to obtain high-level audio cues at the scene level. We propose a novel method to automatically select the most appropriate soundtrack from the list of songs L_2 recommended by our music retrieval system as described in the previous Section 3.2, to generate a *music video* from the UGV.

We use soundtracks of Hollywood movies in our system to select appropriate UGV soundtracks since music in Hollywood movies is designed to be emotional and hence is easier to associate with mood tags. Moreover, music used by Hollywood movies is generated by professionals, which ensures a good harmony with the movie contents. Therefore, we learn from the experience of such experts using their professional *soundtracks* of Hollywood movies through a SVM^{hmm} learning model. We refer to the collection of such soundtracks as

the *evaluation dataset*, with details described later in Section 4.1.4. We construct a *music video generation model* (M_{Eval}) using the training dataset of the *evaluation dataset*, which can predict mood clusters for any music video. We leverage this model to select the most appropriate soundtrack for the UGV. We construct M_{Eval} based on heterogeneous late fusion of SVM^{hmm} models constructed from visual features such as a color histogram and audio features such as MFCC, mel-spectrum and pitch. Similar to our findings with the learning model to predict *scene moods* based on the late fusion of geo and visual features of UGVs, we find that the learning model M_{Eval} based on the late fusion of visual features and concatenated ⟨MFCC, mel-spectrum and pitch⟩ audio features also performs well.

Figure 2 shows the process of soundtrack selection for a UGV. It consists of two components, first, music video generation model (M_{Eval}), and second, a soundtrack selection component. M_{Eval} maps visual features F and audio features A of the UGV with a soundtrack to mood clusters C_2, *i.e.*, $f_3(F, A)$ corresponds to mood clusters C_2 based on the late fusion of F and A. The soundtrack selection component compares moods (C_2 and C_1) of the UGV predicted by M_{Eval} and, M_{GVC} and M_{GVM}.

Algorithm 2 describes the process of the most appropriate soundtrack selection from the list of songs recommended by the heuristic method to generate the music video of the UGV. To automatically select the most appropriate soundtrack, we compute audio features of a selected song and visual features of the UGV and refer to this combination as the *prospective music video*. We compare the characteristics of the *prospective music video* with video songs of the *evaluation dataset* of many famous Hollywood movies. Next we predict mood clusters (C) for the *prospective music video* using M_{Eval}. We treat the predicted mood clusters (C_1) of the UGV by M_{GVC} as ground truth for the UGV, since the mood clusters prediction accuracy of M_{GVC} is very good (see Section 4.2.1). Finally, if the most frequent mood cluster C_2 from C for the *prospective music video* is similar to the ground truth (C_1) of the UGV, then the selected song (S_t) is treated as the soundtrack and the music video of the UGV is generated. If both mood clusters are different then we repeat the same process with the next song in the recommended list L_2. In the worst case, if none of the songs in the recommended list L_2 satisfies the above criteria then we repeat the same process with the second most frequent mood cluster from C, and so on.

Algorithm 2 Music video generation for a UGV

```
1: procedure MusicVideoGeneration(MV)
2:     INPUT: A UGV V by the Android application
3:     OUTPUT: A music video MV for V
4:     m = moodTags(V)            ▷ M_GVM predicts mood tags.
5:     C_1 = moodClusters(V)      ▷ M_GVC predicts clusters.
6:     L_2 = HeuristicSongsRetrieval(m, C_1)
7:     F = visualFeatures(V)      ▷ Compute visual features.
8:     for rank = 1 to numMoodCluster do
9:         for each song S_t in L_2 do
10:            a_1 = calcMFCC(S_t)              ▷ MFCC feat.
11:            a_2 = calcMelSpec(S_t)           ▷ Mel-spec feat.
12:            a_3 = calcPitch(S_t)             ▷ Pitch feat.
13:                          ▷ Concatenate all audio features.
14:            A = concatenate(a_1, a_2, a_3)
15:            C = findMoodCluster(F, A)        ▷ using M_Eval.
16:            C_2 = mostFreqMoodCluster(rank, C)
17:                          ▷ Check for similar mood clusters
18:                          ▷ predicted by M_GVC and M_Eval.
19:            if C_2 == C_1 then
20:                          ▷ Android app generates music video.
21:                MV = generateMusicVideo(S_t, V)
22:                Return MV                    ▷ Music video for V.
23:            end if
24:        end for
25:    end for
26: end procedure
```

4. EXPERIMENTS AND RESULTS

4.1 Dataset and Experimental Settings

The input dataset in our study consists of sensor-annotated videos acquired from a custom Android (or iOS) application running on smartphones. As described in Section 3, we train several learning models to generate a music video from a UGV. To train effective models for the ADVISOR system, it is important to have good ground truths for the training and the testing dataset. However, due to the difference in age, occupation, gender, environment, cultural background and personality, music perception is highly subjective among users. Hence generating ground truths for the evaluation of various music mood classification algorithms are very challenging [11]. Furthermore, there is no standard music dataset with associated mood tags (ground truths) available due to the lack of an authoritative taxonomy of music moods and an associated audio dataset. Therefore, we prepare our own datasets in Sections 4.1.1, 4.1.2, 4.1.3 and 4.1.4 to address the above issues.

4.1.1 Emotion Tag Space

Mood tags are important keywords in digital audio libraries and online music repositories for effective music retrieval. Furthermore, oftentimes, music experts refer to music as the finest *language of emotion*. Therefore it is very important to learn the relationship between music and emotions (mood tags) to build a robust ADVISOR. A number of prior methods [11,15] have described state-of-the-art classifications of mood tags into different emotion classes. The first type of approach is the *categorical approach* which classifies mood tags into emotion clusters such as *happy*, *sad*, *fear*, *anger* and *tender*. Hevner [7] categorized 67 mood tags into eight mood clusters with similar emotions based on musical characteristics such as *pitch*, *mode*, *rhythm*, *tempo*, *melody* and *harmony*. Thayer [28] proposed an energy-stress model, where the mood space is divided into four clusters such as *low energy / low stress*, *high energy / low stress*, *high energy / high stress*, and *low energy / high stress* (see Figure 6). The second type of method is based on the *dimensional approach* to affect, which represents music samples along a two-dimensional emotion space (characterized by arousal and valence) as a set of points.

We consider the *categorical approach* of music mood classification to classify the mood tags used in this work. We extracted the 20 most frequent mood tags of Last.fm from the crawled dataset of 575,149 tracks with 6,814,068 tag annotations in all main music genres by Laurier *et al.* [15]. Last.fm is a music website with more than 30 million users, who have created a site-wide folksonomy of music through end-user tagging. We classified these 20 mood tags into four mood clusters based on mood tag clustering introduced in earlier work [7, 20, 24]. Four mood clusters represent four

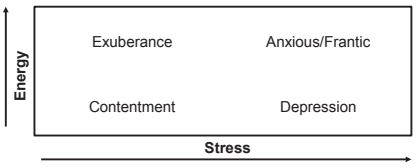

Figure 6: Thayer's [28] model of moods.

quadrants of a 2-dimensional emotion plane with energy and stress characterized as its two dimensions (see Table 1).

However, emotion recognition is a very challenging task due to its cross-disciplinary nature and high subjectivity. Therefore experts have suggested the need for the use of multi-label emotion classification. Since the recommendation of music based on low-level mood tags can be very subjective, many earlier approaches [11, 31] on emotion classification and music recommendation are based on high-level mood clusters. Therefore, in order to calculate the annotator consistency, accuracy and inter-annotator agreement, we compare annotations at four high-level *mood clusters* instead of the 20 low-level *mood tags* in this study. Moreover, we leverage the mood tags and mood clusters together to improve the scene mood prediction accuracy of ADVISOR.

4.1.2 GeoVid Dataset

To create an offline training model for the proposed framework of scene mood prediction of a UGV we utilized 1,213 UGVs which were captured during eight months (4 March 2013 to 8 November 2013) using the GeoVid[5] application. These videos were captured with iPhone 4S and iPad 3 devices. The video resolution of all videos was 720×480 pixels, and their frame rate 24 frames per second. The minimum sampling rate for the location and orientation information was 5 samples per second (*i.e.*, a 200 millisecond sampling rate). In our case, we mainly focus on videos that contain additional information provided by sensors and we refer to these videos as *sensor-annotated videos*. The captured videos cover a diverse range of rich scenes across Singapore and we refer to this video collection as the *GeoVid dataset*.

Since emotion classification is highly subjective and can vary from person to person [11], generating ground truths for the evaluation of the various emotion classifications from video techniques are difficult. It is necessary to use some filtering mechanism to discard bad annotations. In the E6K music dataset for MIREX[6], IMIRSEL assigns each music sample to three different evaluators for mood annotations. They then evaluate the quality of ground truths by the degree of agreement on the music samples. Only those annotations are considered as ground truths where the majority of evaluators selected the same mood cluster. The ground truth of music samples for which all annotators select different mood clusters are resolved by music experts.

For the *GeoVid dataset* we recruited 30 volunteers to annotate emotions (the mood tags listed in Table 1). First, we identified annotators who are consistent with their annotations by introducing redundancy. We repeated one of the videos in the initial sets of the annotation task with ten videos given to each of the evaluators. If any annotated mood tag belonged to a different mood cluster for a repeated video then this annotator's tags were discarded. Annotators passing this criteria were selected for mood annotation of the *GeoVid dataset*. Furthermore, all videos of the *GeoVid dataset* were split into multiple segments with each segment representing a video scene, based on its geo-information and timestamps. For each video segment, we asked three randomly chosen evaluators to annotate one mood tag each after watching the UGV carefully. In order to reduce subjectivity and check the inter-annotator agreement of the three human evaluators for any video, we inspected whether the majority (at least two) of the evaluators chose mood tags that belonged to the same mood cluster. If the majority of evaluators annotated mood tags from the same mood cluster then that particular cluster and its associated mood tags were considered as ground truth for the UGV. Otherwise the decision was resolved by music experts. Due to the subjectivity of music moods, we found that all three evaluators annotated different mood clusters for 298 segments during annotation for the *GeoVid dataset*, hence their ground truths were resolved by music experts (see Table 3).

Table 3: Ground truth annotation statistics with three annotators per video segment.

All Different	Two The Same	All The Same
298	1293	710

4.1.3 Soundtrack Dataset

We prepared an offline music dataset of candidate songs (729 songs altogether) in all main music genres such as *classical*, *electronic*, *jazz*, *metal*, *pop*, *punk*, *rock* and *world* from the ISMIR'04 genre classification dataset[7]. We refer to this dataset as the *soundtrack dataset* and we divided it into 15 *emotion annotation tasks* (EAT). We recruited 30 annotators and assigned each EAT (with 48–50 songs) to two randomly chosen annotators and asked them to annotate one mood tag for each song. Each EAT had two randomly selected repetitive songs to check the annotation consistency of each human evaluator, *i.e.*, if the evaluator-chosen mood tags belonged to the same mood cluster for redundant songs then the evaluator was consistent, otherwise the evaluator's annotations were discarded. Since the same set of EATs was assigned to two different annotators, their inter-annotator agreement is calculated by *Cohen's kappa coefficient* (κ) [3]. This coefficient is considered to be a robust statistical measure of inter-annotator agreement and defined as follows:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where $Pr(a)$ is the relative observed agreement among evaluators, and $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each annotator randomly indicating each category. If $\kappa = 1$ then both annotators for an EAT are in complete agreement while there is no agreement when $\kappa = 0$. According to Schuller *et al.* [23], an agreement level with a κ value of 0.40 and 0.44, respectively, for the music mood assessment with regard to valence and arousal, are considered to be *moderate to good*. Table 4 shows the summary of the mood annotation tasks for the *soundtrack dataset* with a

[5]The GeoVid app and portal at http://www.geovid.org provide recorded videos annotated with location meta-data.

[6]The MIR Evaluation eXchange is an annual evaluation campaign for various MIR algorithms hosted by IMIRSEL (International MIR System Evaluation Lab) at the University of Illinois at Urbana-Champaign.

[7]ismir2004.ismir.net/genre contest/index.htm

Table 4: Summary of the emotion annotation tasks.

Total number of songs	729
Pairs of annotators	15
Common songs per pair	48–50
κ : Maximum	0.67
κ : Minimum	0.29
κ : Mean	0.47
κ : Standard deviation	0.12

mean κ value of 0.47, which is considered to be *moderate to good* in music judgment. For four EATs, annotations were carried out again since evaluators for these EATs failed to fulfill the *annotation consistency* criteria.

For a fair comparison of music excerpts, samples were converted to a uniform format (22,050 Hz, 16 bits, and a mono channel PCM WAV) and normalized to the same volume level. Yang *et al.* [31] suggested to use 25-second music excerpts from around the segment middle to reduce the burden on evaluators. Therefore, we manually selected 25-second music excerpts from near the middle such that the mood was likely to be constant within the excerpt by avoiding drastic changes in musical characteristics. Furthermore, songs were organized in a hash structure with their mood tags as hash keys, so that ADVISOR was able to retrieve the relevant songs from the hash table with the predicted mood tags as keys. We then considered a sequence of the most frequent mood tags T predicted by the *emotion prediction model*, with details described in Section 3.1, for song retrievals.

The *soundtrack dataset* was stored in a database, indexed and used for soundtrack recommendation for UGVs. A song with ID s and k tags is described by a list of tag attributes and scores from $\langle s, tag_1, scr_1\rangle$ to $\langle s, tag_k, scr_k\rangle$, where tag_1 to tag_k are mood tags and scr_1 to scr_k are their corresponding scores. Tag attributes describe the relationship between mood tags and songs, and are organized in a hash table where each bucket is associated with a mood tag. With the aforementioned song s as an example, its k tag attributes are separately stored in k buckets. Since a tag is common to all songs in the same bucket, it is sufficient to only store tuples consisting of song ID and tag score.

4.1.4 Evaluation Dataset

We collected 402 soundtracks from Hollywood movies of all main movie genres such as *action*, *comedy*, *romance*, *war*, *horror* and others. We refer to this video collection as the *evaluation dataset*. We manually selected 1-minute video segments from around the middle for each clip in the *evaluation dataset* such that the emotion was likely to be constant within that segment by avoiding drastic changes in scene and musical characteristics. We ignored segments having dialogues in a scene while selecting 1-minute excerpts. Since the segments in the *evaluation dataset* are professionally produced and their genres, lyrics and context are known, emotions elicited by these segments are easy to determine. Mood clusters (listed in Table 1) were manually annotated for each segment based on its movie genre, lyrics and context and treated as ground truth for the *evaluation dataset*.

4.2 Experimental Results

4.2.1 Scene Moods Prediction Accuracy

To investigate the relationship between *geo* and *visual features* to predict video scene moods for a UGV, we trained four SVM^{hmm} models and compared their accuracy. First, the *Geo model* (M_G) was trained with geo features only, second, the *Visual model* (M_F) was trained with visual features only and third, the *Concatenation model* (M_{Cat}) was trained with the concatenation of both geo and visual features (see Figure 4). Finally, fourth, the *Late fusion models* (M_{GVM}, M_{GVC}) were trained by the late fusion of the first (M_G) and second (M_F) models.

Table 5: Accuracy of emotion prediction models.

SVM^{hmm} Learning Model Accuracy (Performed 10-fold cross validation.)					
Ratio	Model Type	Exp-1 (in %)	Exp-2 (in %)	Exp-3 (in %)	Feature Dimens.
70:30	M_F	18.87	52.62	64.63	64
	M_G	25.56	60.12	74.22	317
	M_{Cat}	24.47	60.79	73.52	381
	M_{GVM}	**37.18**	**76.42**	-	317
	M_{GVC}	-	-	**84.56**	317
80:20	M_F	17.76	51.65	63.93	64
	M_G	24.68	60.83	73.06	317
	M_{Cat}	25.97	61.96	71.97	381
	M_{GVM}	**34.86**	**75.95**	-	317
	M_{GVC}	-	-	**84.08**	317

Exp-1: Model trained at mood *tags* level and predicted moods accuracy checked at mood *tags* level.
Exp-2: Model trained at mood *tags* level and predicted moods accuracy checked at mood *cluster* level.
Exp-3: Model trained at mood *cluster* level and predicted moods accuracy checked at mood *cluster* level.

We randomly divided the *GeoVid dataset* into training and testing datasets with 80:20 and 70:30 ratios. We performed 10-fold cross validation experiments on various learning models as described above to compare their scene mood prediction accuracy for UGVs in the test dataset. We used three experimental settings. First, we trained all models from the training dataset with *mood tags* as ground truth and compared their scene mood prediction accuracy at the *mood tags level* (*i.e.*, whether the predicted *mood tags* and ground truth *mood tags* were the same). Second, we trained all models from the training dataset with *mood tags* as ground truth and compared their scene mood prediction accuracy at the *mood clusters level* (*i.e.*, whether the most frequent mood cluster of predicted *mood tags* and ground truth *mood tags* were the same).

Lastly, we trained all models from the training dataset with *mood clusters* as ground truth and compared their scene mood prediction accuracy at the *mood clusters level* (*i.e.*, whether the predicted *mood clusters* and ground truth *mood clusters* were the same). Our experiments confirm that the model based on late fusion of geo and visual features outperforms the other three models. We noted that the scene mood prediction accuracy at the *mood tag level* does not perform well because the accuracy of the SVM classifier degrades as the number of classes increases. A comparison of the scene mood prediction accuracies for all four models is listed in Table 5.

4.2.2 Soundtrack Selection Accuracy

We randomly divided the *evaluation dataset* into a training and a testing dataset with a 80:20 ratio, and performed 5-fold cross validation experiments to calculate the scene

Table 6: M_{Eval} emotion classification accuracy.

Training M_{Eval}	# Training Videos *Evaluation dataset*	322	Accuracy (in %)
Prediction (Exp-1)	# Test Videos *Evaluation dataset*	80	68.75
Prediction (Exp-2)	# UGVs with soundtrack *GeoVid dataset*	80	70.00
(Performed 5-fold cross validation.)			

mood prediction accuracy of M_{Eval} for UGVs in the test dataset. We performed two experiments. First, we trained M_{Eval} from the training set with *mood clusters* as ground truth and compared their scene mood prediction accuracy at the *mood clusters level* for UGVs in the test dataset of the *evaluation dataset* (*i.e.*, whether the predicted *mood clusters* and ground truth *mood clusters* matched). In the second experiment, we replaced the test dataset of the *evaluation dataset* with the same number of *music videos* generated by our system for randomly selected UGVs from the *GeoVid dataset*. The M_{Eval} maps visual features F and audio features A of a video V to mood clusters C, *i.e.*, $f_3(F, A)$ corresponds to mood clusters C based on the late fusion of F and A (see Figure 2). An input vector (in time order) for M_{Eval} can be represented by the following sequence (see Figure 7): $\langle F_1, A_1\rangle$, $\langle F_1, A_2\rangle$, $\langle F_2, A_2\rangle$, $\langle F_2, A_3\rangle$, ...

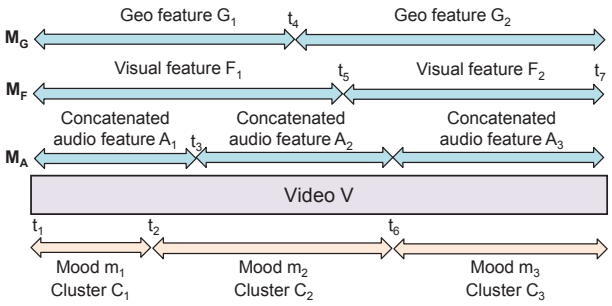

Figure 7: Features to mood tags/clusters mapping.

M_{Eval} reads the above input vector and predicts mood clusters for it. Table 6 shows that the emotions (mood clusters) prediction accuracy (68.75%) of M_{Eval} for *music videos* is comparable to the emotion prediction accuracy at the scene level in movies by state-of-the-art approaches such as introduced by Soleymani *et al.* [27] (63.40%) and Wang *et al.* [29] (74.69%). To check the effectiveness of the ADVISOR system, we generated *music videos* for 80 randomly selected UGVs from the *GeoVid dataset* and predicted their mood clusters by M_{Eval} with 70.0% accuracy, which is again comparable to state-of-the-art algorithms for emotion prediction at the scene level in movies. The experimental results in Table 6 confirm that ADVISOR effectively combines objective scene moods and music to recommend appealing soundtracks for UGVs.

4.3 User Study

Based on the techniques introduced earlier, we implemented the ADVISOR system to generate music videos for UGVs. All UGVs were single-shot clips with sensor metadata, acquired by our Android application designed specifically for recording sensor-annotated videos. We randomly selected five UGVs each for six different sites of Singapore as listed in Table 7, from a set of acquired videos. To judge whether the recommended songs capture the scene moods of videos, we recruited fifteen volunteers to assess the appropriateness and entertainment value of the music videos (UGVs with recommended songs). We asked every user to select one video for each site by choosing the most likely candidate that they themselves would have captured at that site. The predicted scene moods listed in Table 7 are the first three mood tags belonging to the most frequent mood cluster predicted by M_{GVC} for five videos at different sites. A soundtrack for all selected videos was generated using ADVISOR and users were asked to assign a score 1 (worst) to 5 (best) to the generated music videos. Finally, we calculated the average score of music videos for all sites. Table 7 summarizes the ratings and the most appropriate scene moods from a list of predicted mood tags for videos from the aforementioned six sites. The feedback from these volunteers was encouraging, indicating that our technique achieves its goal of automatic music video generation to enhance the video viewing experience.

Table 7: User study feedback from 15 volunteers.

Video Location	Predicted Scene Moods	1	2	3	4	5	Avg. Rating
Cemetery	melancholy, sad, sentimental	0	0	3	4	8	4.3
Clarke Quay	fun, sweet, calm	0	2	5	7	1	3.5
Gardens by the Bay	soothing, fun, calm	0	3	3	9	0	3.4
Marina Bay Sands	fun, playful	0	0	2	6	7	4.3
Siloso Beach	happy, fun, quiet	0	0	1	6	8	4.5
Universal Studios	fun, intense, happy, playful	0	2	5	5	3	3.6
Ratings on a scale from 1 (worst) to 5 (best).							

5. CONCLUSIONS

Our work represents one of the first attempts for user preference-aware video soundtrack generation. We categorize user activity logs from different data sources by using semantic concepts. This way, the correlation of preference-aware activities based on categorization of user-generated heterogeneous data complements video soundtrack recommendations for individual users. The ADVISOR system automatically generates a matching soundtrack for a UGV in four steps. More specifically, first, a learning model based on the late fusion of geo and visual features recognizes scene moods in the UGV. Second, a novel heuristic method recommends a list of songs based on the predicted scene moods. Third, the soundtrack recommendation component re-ranks songs recommended by the heuristics method based on the user's listening history. Finally, our Android application generates a music video from the UGV by automatically selecting the most appropriate song using a learning model based on the late fusion of visual and concatenated audio features. In the future, each one of these steps could be further enhanced. The experimental results and our user study confirm that ADVISOR can effectively combine objective

scene moods and individual music tastes to recommend appealing soundtracks for UGVs.

Acknowledgments

The authors are very grateful to Dr. Suhua Tang and the anonymous reviewers for their insightful and constructive suggestions to improve the quality of this work. The research has been supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office through the Centre of Social Media Innovations for Communities (COSMIC).

6. REFERENCES

[1] K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki. Efficient Retrieval of Life Log based on Context and Content. In *ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, pages 22–31, 2004.

[2] B. Chen, J. Wang, Q. Huang, and T. Mei. Personalized Video Recommendation through Tripartite Graph Propagation. In *ACM International Conference on Multimedia*, pages 1133–1136, 2012.

[3] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[4] M. Cristani, A. Pesarin, C. Drioli, V. Murino, A. Rodà, M. Grapulin, and N. Sebe. Toward an Automatically Generated Soundtrack from Low-level Cross-modal Correlations for Automotive Scenarios. In *ACM Int'l Conference on Multimedia*, pages 551–560, 2010.

[5] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by Humming: Musical Information Retrieval in an Audio Database. In *ACM International Conference on Multimedia*, pages 231–236, 1995.

[6] A. Hanjalic and L.-Q. Xu. Affective Video Content Representation and Modeling. *In IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[7] K. Hevner. Experimental Studies of the Elements of Expression in Music. *The American Journal of Psychology*, pages 246–268, 1936.

[8] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane Training of Structural SVMs. *Machine Learning Journal*, 77(1):27–59, 2009.

[9] M. Kaminskas and F. Ricci. Location-adapted Music Recommendation using Tags. In *User Modeling, Adaption and Personalization*, pages 183–194. Springer, 2011.

[10] H. B. Kang. Affective Content Detection using HMMs. In *ACM International Conference on Multimedia*, pages 259–262, 2003.

[11] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music Emotion Recognition: A State of the Art Review. In *International Society for Music Information Retrieval (ISMIR)*, pages 255–266, 2010.

[12] Y. Ko. A Study of Term Weighting Schemes using Class Information for Text Classification. In *ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 1029–1030, 2012.

[13] O. Kucuktunc, U. Gudukbay, and O. Ulusoy. Fuzzy Color Histogram-based Video Segmentation. *In Computer Vision and Image Understanding*, 114(1):125–134, 2010.

[14] F.-F. Kuo, M.-F. Chiang, M.-K. Shan, and S.-Y. Lee. Emotion-based Music Recommendation by Association Discovery from Film Music. In *ACM International Conference on Multimedia*, pages 507–510, 2005.

[15] C. Laurier, M. Sordo, J. Serrà, and P. Herrera. Music Mood Representations from Social Tags. In *Int'l Society for Music Information Retrieval (ISMIR)*, pages 381–386, 2009.

[16] C. T. Li and M. K. Shan. Emotion-based Impressionism Slideshow with Automatic Music Accompaniment. In *ACM Int'l Conference on Multimedia*, pages 839–842, 2007.

[17] L. Lu, H. You, and H. Zhang. A New Approach to Query by Humming in Music Retrieval. In *IEEE Int'l Conference on Multimedia and Expo (ICME)*, pages 22–25, 2001.

[18] M. H. Park, J. H. Hong, and S. B. Cho. Location-based Recommendation System using Bayesian User's Preference Model in Mobile Devices. In *Ubiquitous Intelligence and Computing*, pages 1130–1139. Springer, 2007.

[19] H. Rahmani, B. Piccart, D. Fierens, and H. Blockeel. Three Complementary Approaches to Context Aware Movie Recommendation. In *ACM Workshop on Context-Aware Movie Recommendation*, pages 57–60, 2010.

[20] J. A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[21] M. Sahidullah and G. Saha. Design, Analysis and Experimental Evaluation of Block based Transformation in MFCC Computation for Speaker Recognition. In *Speech Communication*, volume 54, pages 543–565, 2012.

[22] M. Schedl and D. Schnitzer. Location-Aware Music Artist Recommendation. In *MultiMedia Modeling*, pages 205–213. Springer, 2014.

[23] B. Schuller, C. Hage, D. Schuller, and G. Rigoll. Mister DJ, Cheer Me Up!: Musical and Textual Features for Automatic Mood Classification. *In Journal of New Music Research*, 39(1):13–34, 2010.

[24] R. R. Shah, Y. Yu, and R. Zimmermann. User Preference-Aware Music Video Generation Based on Modeling Scene Moods. In *ACM Int'l Conference on Multimedia Systems (MMSys)*, pages 156–159, 2014.

[25] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus Late Fusion in Semantic Video Analysis. In *ACM Int'l Conference on Multimedia*, pages 399–402, 2005.

[26] C. G. Snoek, M. Worring, J. C. Van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *ACM International Conference on Multimedia*, pages 421–430, 2006.

[27] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun. A Bayesian Framework for Video Affective Representation. In *IEEE Int'l Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009.

[28] R. E. Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, 1989.

[29] H. L. Wang and L. F. Cheong. Affective Understanding in Film. *In IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.

[30] C. Y. Wei, N. Dimitrova, and S.-F. Chang. Color-mood Analysis of Films based on Syntactic and Psychological Models. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 831–834, 2004.

[31] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen. A Regression Approach to Music Emotion Recognition. *In IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, 2008.

[32] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust Late Fusion with Rank Minimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3021–3028, 2012.

[33] Y. Yu, K. Joe, V. Oria, F. Moerchen, J. S. Downie, and L. Chen. Multi-version Music Search using Acoustic Feature Union and Exact Soft Mapping. *In International Journal of Semantic Computing*, 3(02):209–234, 2009.

[34] Y. Yu, Z. Shen, and R. Zimmermann. Automatic Music Soundtrack Generation for Outdoor Videos from Contextual Sensor Information. In *ACM International Conference on Multimedia*, pages 1377–1378, 2012.

[35] R. Zimmermann and Y. Yu. Social Interactions over Geographic-aware Multimedia Systems. In *ACM Int'l Conference on Multimedia*, pages 1115–1116, 2013.