Automatic Music Soundtrack Generation for Outdoor Videos from Contextual Sensor Information

Yi Yu, Zhijie Shen, Roger Zimmermann
School of Computing, National University of Singapore
Singapore 117417
{yuy,z-shen,rogerz}@comp.nus.edu.sg

ABSTRACT

We present a system to automatically generate soundtracks for user-generated outdoor videos (UGV) based on concurrently captured contextual sensor information with mobile apps for the ACM Multimedia 2012 Google challenge: *Automatic Music Video Generation*. Our method addresses the use case of making "a video much more attractive for sharing by adding a matching soundtrack to it." Our system correlates viewable scene information from sensors with geographic contextual tags from OpenStreetMap. The co-occurance of geo-tags and mood tags are investigated from a set of categories of the web site Foursquare.com and a mapping from geo-tags to mood tags is obtained. Finally, a music retrieval component returns music based on matching mood tags. The experimental results show that our system can successfully create soundtracks that are related to the mood and situation of UGVs and therefore enhance the enjoyment of viewers. Our system sends only sensor data to a cloud service and is therefore bandwidth efficient since video data does not need to be transmitted for analysis.

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

Music soundtrack generation, location sensors, mobile videos, geographic tagging, music mood, location-mood categories

1. MOTIVATION AND BACKGROUND

Capturing videos with mobile devices such as smartphones and tablets is now easy and popular and it allows people to share their experiences. However, many user-generated videos (UGV) are recorded without much interesting audio and therefore lack appeal when sharing. This is especially true for outdoor, scenic videos (*e.g.*, captured by tourists on vacation) which may be accompanied mostly by environmental sounds. We propose a system that is designed to enhance UGVs by automatically augmenting them with a music soundtrack that is inspired by the geo-location of where the video was taken. For example a busy city scene that is bustling with people should convey a different atmosphere than a majestic view of a mountain range. Based on investigating social activities in terms of contextual sensor information on Foursquare.com and classifying them into mood tags related to affective atmospheres, we are able to generate a soundtrack with specific moods according to geographic information of the UGVs.

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.

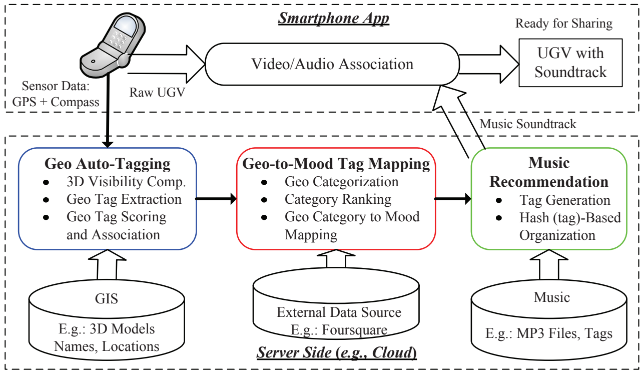

Figure 1: The framework of automatic soundtrack generation for outdoor videos from contextual sensor information.

2. FRAMEWORK DESIGN

Our system is based on several key ideas which together form its contribution. Fig. 1 shows the overall system architecture with its main processing steps as follows:

(i) **Geographic auto-tagging**. UGVs captured on mobile devices are enhanced with geo-information by using sensors such as GPS and compass. Location and orientation meta-data are recorded with the video streams and used to model the coverage areas of the video scenes as spatial objects. We introduced a *viewable scene model* which describes the scenes visible in the video based on the camera's *field of view* (FOV) [1]. Each 3D viewable scene is described by a few parameters (camera position, direction, viewable angles and visibility range). The viewable scenes of a video is modeled as a sequence of FOVs, each having a timestamp t. The annotation process is automated through querying proper data sources with the viewable scene description [3]. A number of objects in the covered region of the *FOVScene* sequence are retrieved from data sources. OpenStreetMap[1], a community-based geo-information system is used as the data source in our prototype system. Next, in each *FOVScene*, a sophisticated geometry compu-

[1] www.openstreetmap.org

Geographic category	Related mood(s)
Arts & Entertainment	quiet, calm
Colleges & Universities	quiet, calm
Food	sweet, happy
Great Outdoors	dreamy
Nightlife Spots	funny, intense, playful
Professional & Other Places	aggressive, heavy
Residences	sweet, sleepy
Shops & Services	happy
Travel & Transport	melancholy, bittersweet, funny

Table 1: Top level Foursquare geographic categories and their corresponding related moods.

tation is conducted on both inputs (*i.e.*, the *FOVScene* and the objects) to determine the visibility of each object. Subsequently the visible objects are retained and their descriptive texts from the data source serve as tags. Then, six relevance criteria are introduced to score the tag importance to the scenes.

(ii) **Geographic-to-mood tag mapping**. We conducted a preliminary study to learn the relationship between locations and moods. Instead of investigating the related moods of each generated geographic tags, we first classify the geographic tags into a number of predefined categories and then determine the related moods on the basis of those geographic categories. In this preliminary study we focus only on the nine top-level geographic categories, listed in Table 1. We leverage the Foursquare API to assign our geographic tags to the categories. Subsequently, the categories of the generated tags of a certain video are also associated with the video. Similar to ranking the geographic tags according to their relevance to a video, we rank the categories as well. The next step is to associate moods with the categories. It is natural that this judgement is somewhat subjective. We recruited several users to determine the relationship through a survey with the results shown in Table 1, where the moods are chosen from the collection suggested by Laurier [2]. After a video's geographic categories have been mapped to mood tags, they can be used as input to the music retrieval engine.

(iii) **Mood-based music recommendation**. The music recommendation sub-system takes mood tags as input and returns corresponding audio clips that form the soundtrack to be associated with the UGVs. First, a music database composed of songs with mood tags is built offline, organized in a hash structure. Typical songs with the chosen mood tags are found from Last.fm[2] and YouTube[2]. Mood tags of each song are sorted based on their relevance to the song. Then, a song with ID s and k tags is described by a list of tag attributes $< s, tag_1, 1 >$, $< s, tag_2, 2 >$, $\cdots$, $< s, tag_k, k >$, where tag_1 to tag_k are mood tags and 1 to k are their ranks. Songs are organized in a hash table, with their tags as hash keys. Given a set of ranked mood tags G inferred from video sensor information, relevant songs are found from the hash table by using each tag in G to locate the buckets. Songs retrieved from the database are further ranked according to the mood preference in G and mood relevance in each song. More specifically, to evaluate the similarity between a song s with a tag set T_s and the input tag set G, we first define the modified mean reciprocal rank (*MMRR*) as $MMRR_s = \frac{1}{|G|}\sum_{k=1}^{|G|}\frac{k}{\max\{k, r_k\}}$, where r_k is the rank of the k^{th} tag g_k in the set T_s. The rank r_k is equal to ∞ if g_k is not found in T_s. In this way, the *MMRR* metric assesses the similarity of two ordered tag sets G and T_s.

Video Name	Geo-tag Category	Recommended Song
Yellow Mountain	Great Outdoors	*Sleeping beauty.*
Marina Bay, shots 1 and 2	Professional & Other Travel & Transport Shops & Services	*I got this feeling.*
Marina Bay, shots 3 and 4	Great Outdoors Travel & Transport Arts & Entertainment Shops & Services	*Look on down from the bridge.*
City of Brig	Great Outdoors Residences	*Settling.*

Table 2: Video shots, the detected geographic categories and recommended songs. The corresponding mood tags are listed in Table 1.

3. EXPERIMENTS

Based on the techniques introduced above, we built a system to generate music soundtracks for outdoor videos and conducted a preliminary experiment where we investigated to find suitable songs for the outdoor video clips from three different sites: one video clip of four shots from the Marina Bay Area of Singapore, one single-shot video clip from the Yellow Mountain in China and another single-shot video clip from the city of Brig, Switzerland. Furthermore, we prepared a collection of candidate songs (98 altogether), which cover a wide range of moods. Table 2 summarizes the detected geographic categories of the three videos and the recommended songs for them. To judge whether the suggested songs capture the moods of the videos, we recruited ten volunteers to assess the appropriateness and entertainment value of our generated video-soundtrack composites by giving them a score from 1 (worst) to 5 (best). The feedback from these volunteers was encouraging (the average scores corresponding to 'Yellow Mountain', 'Marina Bay' and 'City of Brig' equal 4.4, 3.9, 3.8, respectively) indicating that our technique achieves its goal of automatic soundtrack generation for increased enjoyment of UGVs when sharing.

4. CONCLUSIONS

We have designed a system that automatically generates music soundtracks for user-generated videos. Our method leverages contextual sensor information to first obtain relevant geo-tags which are then mapped to corresponding mood tags. A music retrieval component recommends songs based on their mood association. Because of the novel sensor-stream processing, our system is lightweight in its bandwidth use and well suited for mobile app implementation. The results, based on a small user study, are also very encouraging.

Acknowledgment

This 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.

5. REFERENCES

[1] S. Arslan Ay, R. Zimmermann, and S. H. Kim. Viewable Scene Modeling for Geospatial Video Search. In *ACM Multimedia*, pages 309–318, 2008.

[2] C. Laurier. Automatic Classification of Musical Mood by Content-based Analysis. *Tesis Doctorals en Xarxa*, 2011.

[3] Z. Shen, S. Arslan Ay, S. H. Kim, and R. Zimmermann. Automatic Tag Generation and Ranking for Sensor-rich Outdoor Videos. In *ACM Multimedia*, pages 93–102, 2011.

[2] www.last.fm and www.youtube.com