# Multimodal Analysis of User-Generated Content in Support of Social Media Applications

Rajiv Ratn Shah
School of Computing, National University of Singapore, Singapore
rajiv@comp.nus.edu.sg

## ABSTRACT

The number of user-generated multimedia content (UGC) online has increased rapidly in recent years due to the ubiquitous availability of smartphones, cameras, and affordable network infrastructures. Thus, it attracts companies to provide diverse multimedia-related services such as preference-aware multimedia recommendations, multimedia-based e–learning, and event summarization from a large collection of multimedia content. However, a real-world UGC is complex and extracting semantics from only multimedia content is difficult because suitable concepts may be exhibited in different representations. Modern devices capture contextual information in conjunction with a multimedia content, which greatly facilitates in the semantics understanding of the multimedia content. Thus, it is beneficial to analyse UGC from multiple modalities such as multimedia content and contextual information (*e.g.*, spatial and temporal information). This doctoral research studies the multimodal analysis of UGC in support of above-mentioned social media problems. We present our proposed approaches, results, and works in progress on these problems.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.9 [**Image Processing and Computer Vision**]: Applications

## Keywords

Multimodal analysis; UGC; social media applications

## 1. INTRODUCTION AND MOTIVATION

Due to advancements in technologies, capturing UGC such as user-generated images (UGIs) and user-generated videos (UGVs) anytime and anywhere, and then instantly sharing them on social media platforms has become a very popular activity. Therefore, it necessities social media companies to understand the semantics and sentics of UGC in order to provide diverse multimedia-related services. However, it is very challenging due to the following reasons: (i) the difficulty in capturing the semantics of UGC, (ii) the existence of noise in available metadata, (iii) the difficulty in handling big size datasets, (iv) the difficulty in learning user preferences, and (v) the insufficient accessibility and searchability of video content. Since multimodal information augments knowledge bases by inferring semantics and sentics from unstructured multimedia content and contextual information, we leverage it in our proposed approaches to the following three social media problems. First, we answer the automatic soundtrack recommendation task for UGVs [16, 17]. Next, we work on the task of automatic lecture video segmentation [14, 15]. Finally, we address the automatic event summarization task from a large collection of UGIs [13].

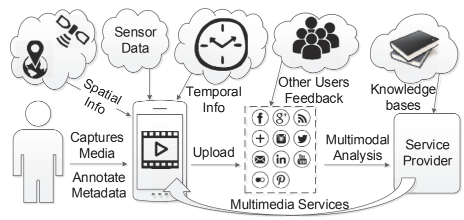


**Figure 1: Overview of our approach.**

Sound is a very important aspect that contributes greatly to the appeal of a UGV when it is being viewed. However, many outdoor UGVs lack a certain appeal because their soundtracks consist mostly of ambient background noise (*e.g.*, environmental sounds such as cars passing by). Thus, it entails to replace the background noise of the UGV with a matching soundtrack that matches with scenes, locations, and users' preferences by exploiting both content and contextual information. However, generating soundtracks for the UGV is not easy in mobile environment due to the following reasons: (i) traditionally it is tedious and time-consuming for a user to add a custom soundtrack to the UGV and (ii) an important aspect is that a good soundtrack should match and enhance the overall mood of the UGV. We present a fast and effective heuristic ranking approach based on heterogeneous late fusion by jointly considering three aspects: venue categories, visual scenes, and the listening history of the user. First, we predict scene moods from a real-world video dataset that was collected from the user's daily outdoor activities. Second, we perform heuristic rankings to fuse the predicted confidence score of multiple models. Finally, we customize the video soundtrack recommendation functionality to make it compatible with mobile devices. Further, we consider other areas such as eduction where UGVs have a significant impact on society.


*ICMR'16, June 06-09, 2016, New York, NY, USA*

DOI: http://dx.doi.org/10.1145/2911996.2912032

Due to rapid growth in digital lecture videos online, a multimedia-based e–learning system such as MIT OpenCourseWare[1] has become an important learning environment which uses electronic educational technologies as a platform for teaching and learning activities. Since a specific topic of interest is often discussed in only a few minutes of a long lecture video recording, it is often relatively easy to find relevant videos in an archive but difficult to find proper positions within the videos. Websites such as VideoLectures.NET enable students to access different topics within lecture videos. However they determine the topics based on the manual annotation of segment boundaries, which is a very time consuming, subjective, error prone, and costly process. Thus, it necessities a video segmentation system that can automatically determine segment boundaries accurately from a lecture video despite its quality is not sufficiently high. We argue that the presence of contextual information in conjunction with UGVs assists in an effective temporal segmentation. Thus, we present the ATLAS and TRACE systems that leverage multimodal information and late fusion techniques in our solutions. Specifically, we combine confidence scores produced by models constructed from visual, transcriptional, and Wikipedia features in the late fusion. Additionally, we consider areas where UGIs are exploited to provide multimedia-related services.

We present the EventBuilder system to efficiently obtain the multimedia summary of an event from the large collection of UGIs aggregated in social media platforms. It has three novel characteristics: (i) leveraging Wikipedia as event background knowledge to obtain additional contextual information about the event during event detection, (ii) visualizing the event on Google Map in real-time with a diverse set of social media activities, and (iii) solving an optimization problem to produce text summaries for the event.

In this doctoral research we aim to exploit both the content and contextual information of UGC in the support of the three problems discussed above. Figure 1 shows an overview of our approach. Moreover, we exploit knowledge bases for a better semantics and sentics understanding of UGC. In our work in progress, we mainly focus on sentics determination from the multimedia content leveraging the fusion of multimodal information.

## 2. STATE OF THE ART

The area of music recommendation for UGC is largely unexplored. Earlier works [8, 21] add soundtracks to the slideshow of UGIs. There exist a few approaches [6,20,22] to recognize emotions from videos but the field of video soundtrack recommendation for UGVs [4,24] is largely unexplored. Multi-feature late fusion techniques have been advocated in various applications such as video event detection and object recognition [23]. Moreover, Lu *et al.* [10] used heuristic approaches for querying desired songs from a music database by humming a tune. Such earlier works inspired us to build the ADVISOR system [16] by performing heterogeneous late fusion to recognize moods and retrieve a ranked list of songs using a heuristic approach for UGVs.

Due to the high cost and rapidly growing sizes of databases, it is not feasible to manually determine segment boundaries of lecture videos. Earlier approaches [7, 9, 14] attempted to segment videos automatically by exploiting visual, audio, and linguistic features. However, these approaches fail when the quality of a lecture video is not sufficiently high. This motivates us to leverage knowledge bases such as Wikipedia and other contextual information in conjunction with the content of the lecture video in our multimodal approach for temporal segmentation [15].

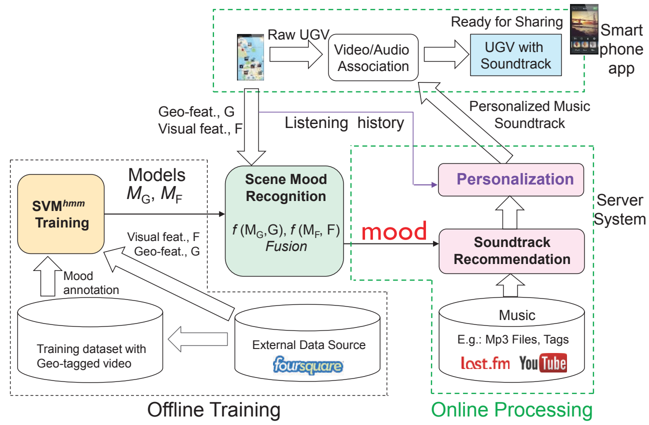


**Figure 2: System overview of soundtrack recommendations for UGVs with ADVISOR [16].**

Significant works [3, 11] have been done in the area of event modeling, detection, understanding, and summarization from multimedia over the past few years. Fabro *et al.* [5] presented an algorithm for the summarization of real-life events based on community-contributed multimedia content using photos from Flickr and videos from YouTube. In our novel event summarization system [13], we leverage multimodal information of UGIs and knowledge bases such as Wikipedia in event detection and summarization.

In our earlier work [12] we have shown that multimodal information is useful in uploading videos over adaptive middleboxes to news servers in weak network infrastructures. Moreover, we presented a system [18,19] for SMS-based FAQ retrieval by performing a match between SMS queries and FAQ database. In the future, we want to extend this concept to build an SMS-based news retrieval system leveraging information from multiple modalities.

## 3. PROPOSED APPROACH

As discussed in the previous sections, there are several multimdia-related problems which emerged after rapid growth in UGC online. Some problems are not addressed yet, and rest need efficient solutions after more contextual information is available. To the best of our knowledge, our work (**OBJ-1**) is the first attempt to recommend soundtracks for outdoor UGVs that correlates preference-aware activities from different behavioral signals of individual users, *e.g.*, online listening activities and physical activities. Furthermore, our work (**OBJ-2**) is the first attempt to compute segment boundaries from lecture videos leveraging knowledge bases such as Wikipedia. Finally, we present a novel event summarization and visualization system (**OBJ-3**) for the large collection of UGIs. The following is more details about our approaches for above objectives.

**OBJ-1.** Figure 2 shows the architecture of our proposed music video generation system, called ADVISOR [16]. It consists of two parts: an offline training and an online processing component. The online processing is further divided into two modules: a smartphone app and a server backend system. This app allows users to capture sensor-annotated videos. Geographic contextual informa-

[1]www.ocw.mit.edu/

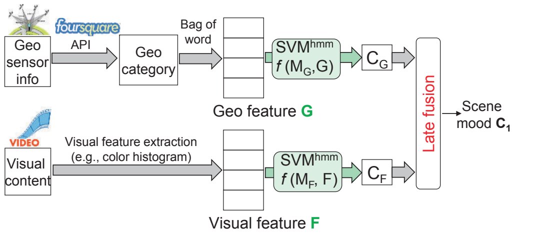


**Figure 3: Mood recognition from UGVs [16].**

tion (*i.e.*, geo-categories such as *Park* and *Lake* derived from Foursquare[2]) for a UGV serves as an important dimension to represent valuable semantic information while its video frame content is often used in scene understanding. During offline processing, models $M_G$ and $M_F$ are trained by exploiting geo- and visual features ($G$ and $F$) to predict geo- and visual aware mood tags ($C_G$ and $C_F$), respectively, using a $SVM^{hmm}$ technique from a dataset with geo-tagged videos. Next $C_G$ and $C_F$ are fused, and mood tags with high likelihoods are regarded as scene moods $C_1$ of the UGV (see Figure 3). Then, songs matching the scene moods are recommended. Among them, the songs matching a user's listening history are considered as user preference-aware songs.

We propose a heuristic music retrieval method to recommend a list of songs for input scene moods. We calculate the total score of each song based on the likelihood of predicted mood tags for a UGV and then retrieve a ranked list of soundtracks. Further, our system extracts audio features including MFCC and pitch from a user's frequently listened audio tracks. We re-rank the retrieved list by correlating it with the computed audio features, and then recommending a list of user preference-aware songs. Next, the soundtrack selection component automatically chooses the most appropriately matching song from this list and attaches it as the soundtrack to the UGV (see Figure 4). We leverage the soundtrack of Hollywood movies to select an appropriate UGV soundtrack since such music is generated by professionals and ensures a good harmony with the movie content. We learn from the experience of such experts using their *professional soundtracks* of Hollywood movies through a $SVM^{hmm}$ learning model. We construct this model 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. The soundtrack selection process consists of two components. First a music video generation model that maps visual features $F$ and audio features $A$ of the UGV with a soundtrack $S_t$ to mood tags $C_2$ based on the late fusion of $F$ and $A$. Second, a soundtrack selection component that attaches $S_t$ to the UGV if $C_2$ is similar to $C_1$ (mood tags predicted based on geo- and visual features).

**OBJ-2.** The proposed temporal segmentation system [15] has the following main contributions: (i) the extraction of a novel linguistic-based Wikipedia feature which is useful in finding segment boundaries of low quality videos, (ii) a $SVM^{hmm}$ model which learns temporal transition cues from visual features, and (iii) the investigation of the late fusion of video segmentations derived from state-of-the-art methods. Figure 5 shows the architecture of segment boundaries detection from SRT using the proposed linguistic-based method. It leverages Wikipedia texts of subjects which lecture videos belong. We assume that the subject (say, *Artificial Intelligence*) of a lecture video is known. We use the Wikipedia API to find related Wikipedia articles and texts of different topics. Say, $b_w$ is the block of texts corresponding to a Wikipedia topic. We determine POS tags for Wikipedia texts and SRT of the lecture video. Next, we find a block $b_s$ of 120 words from SRT which matches closely with the Wikipedia block $b_w$ for each topic in Wikipedia texts. Specifically, first we create linguistic feature vectors based on noun phrases in the entire Wikipedia texts. Next, we compute cosine similarities between a Wikipedia block $b_w$ and all SRT blocks $b_s$. In this way determine the closest matching SRT block corresponding to each Wikipedia block, hence segment boundaries for the lecture video. Next, we compute temporal transitions derived from other modalities such as visual content and SRT using state-of-the-arts, and investigate the effect of their fusion with boundaries determined leveraging Wikipedia. We fuse segment boundaries by replacing two transitions less than ten seconds apart by their average transition times and keeping rest transitions as the final temporal transitions for the lecture video [14].

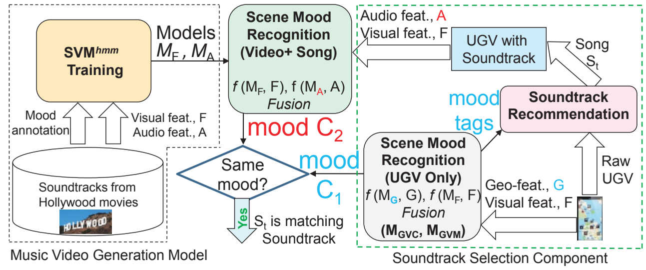


**Figure 4: Soundtrack selection process in [16].**

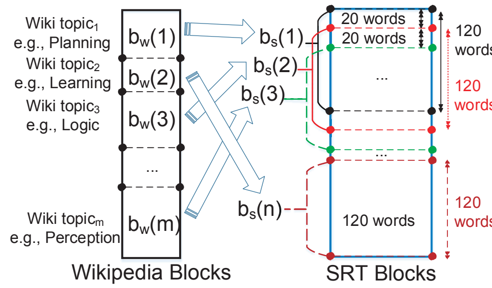


**Figure 5: Architecture for segment boundary detection using SRT and Wikipedia [15].**

**OBJ-3.** Figure 6 shows the architecture of EventBuilder. It detects events from UGIs by computing the relevance score $u(p, e)$ of a UGI $p$ for a given event $e$. It is computed by combining confidence scores from different modalities as follows: $u(p, e) = w_1\ \xi + w_2\ \lambda + w_3\ \gamma + w_4\ \mu + w_5\ \rho$, where $w_{i\,i=1}^{\,5}$ are weights for different modalities such that $\sum_{i=1}^{5} w_i = 1$, and $\xi$, $\lambda$, $\gamma$, $\mu$, and $\rho$ are similarity functions for the given $p$ and $e$ with respect to event name, temporal information, spatial information, keywords, and camera model, respectively, as described in EventBuilder [13]. EventBuilder also produces text summaries from UGIs and Wikipedia articles of $e$. First, it determines important *concepts* (*e.g.*, *kid-play-holi* for the event named *holi*) from available texts. Next, it solves an optimization problem by selecting the minimal number of sentences which cover the maximal number of important *concepts* from matrix constructed by the available texts and the extracted concepts. Finally, EventBuilder presents an interactive visualization of the event on Google Maps from a representative set of UGIs for $e$.

[2]www.foursquare.com

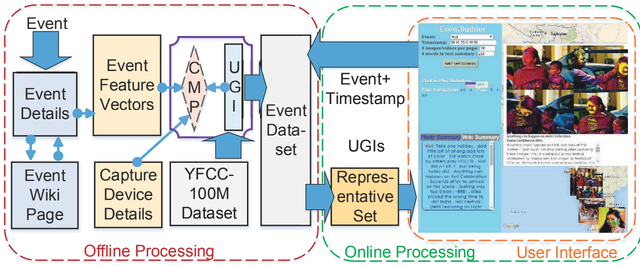


**Figure 6: System framework of EventBuilder [13].**

**Work in progress.** Sentiments are very useful in personalized search, retrieval, and recommendation systems [1]. We determine sentiments for a photo by leveraging concepts determined from its visual content and contextual information. Further we use SenticNet-3 knowledge base [2] which bridges the conceptual and affective gap between word-level natural language data and the concept-level sentiments conveyed by them. We perform sentiment analysis to determine mood tags associated with photos, and subsequently provide sentics based multimedia services to users.

# 4. RESULTS

To evaluate **OBJ-1** we utilized 402 soundtracks from Hollywood movies, 1213 sensor-annotated videos, 729 songs from ISMIR, and 20 most frequent mood tags from Last.fm [16,17]. Next, we used 133 lecture videos with several metadata from the VideoLectures.NET and NPTEL[3] to evaluate **OBJ-2** [14, 15]. Finally, we evaluate **OBJ-3** on the YFCC100M dataset, a collection of 100 million photos and videos from Flickr [13].

# 5. CONCLUSIONS

This doctoral research addressed the following three problems leveraging multimodal analysis and exploit knowledge bases: (i) the soundtrack recommendation for outdoor UGVs, (ii) the temporal segmentation of lecture videos, and (iii) multimedia event summarization from a large collection of UGIs. Experimental results confirm that our approaches worked well in the semantic understanding from multimedia content. Currently, we are working on the sentics understanding from UGC using multimodal information access.

# ACKNOWLEDGMENTS

This research has been supported in part by Singapore's Ministry of Education (MOE) Academic Research Fund Tier 1, grant number T1 251RES1415.

# 6. REFERENCES

[1] E. Cambria, J. Fu, F. Bisio, and S. Poria. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *AAAI Conference on Artificial Intelligence*, pages 508–514, 2015.

[2] E. Cambria, D. Olsher, and D. Rajagopal. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI Conference on Artificial Intelligence*, 2014.

[3] J. Choi, E. Kim, M. Larson, G. Friedland, and A. Hanjalic. Evento 360: Social event discovery from web-scale multimedia collection. In *ACM MM*, pages 193–196, 2015.

[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 MM*, pages 551–560, 2010.

[5] M. Del Fabro, A. Sobe, and L. Böszörmenyi. Summarization of Real-life Events based on Community-contributed Content. In *MMEDIA*, 2012.

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

[7] A. Haubold and J. R. Kender. Augmented Segmentation and Visualization for Presentation Videos. In *ACM MM*, pages 51–60, 2005.

[8] C. T. Li and M. K. Shan. Emotion-based Impressionism Slideshow with Automatic Music Accompaniment. In *ACM MM*, pages 839–842, 2007.

[9] M. Lin, M. Chau, J. Cao, and J. F. Nunamaker Jr. Automated Video Segmentation for Lecture Videos: A Linguistics-based Approach. *In IJTHI*, 1(2):27–45, 2005.

[10] L. Lu, H. You, and H. Zhang. A New Approach to Query by Humming in Music Retrieval. In *IEEE ICME*, pages 22–25, 2001.

[11] V. Mezaris, A. Scherp, R. Jain, and M. S. Kankanhalli. Real-life Events in Multimedia: Detection, Representation, Retrieval, and Applications. *In Springer MTAP*, 70(1):1–6, 2014.

[12] R. R. Shah, M. Hefeeda, R. Zimmermann, K. Harras, C.-H. Hsu, and Y. Yu. Newsman: Uploading videos over adaptive middleboxes to news servers in weak network infrastructures. In *Springer MMM*, pages 100–113. Springer, 2016.

[13] R. R. Shah, A. D. Shaikh, Y. Yu, W. Geng, R. Zimmermann, and G. Wu. EventBuilder: Real-time Multimedia Event Summarization by Visualizing Social Media. In *ACM MM*, pages 185–188, 2015.

[14] R. R. Shah, Y. Yu, A. D. Shaikh, S. Tang, and R. Zimmermann. ATLAS: Automatic Temporal Segmentation and Annotation of Lecture Videos Based on Modelling Transition Time. In *ACM MM*, pages 209–212, 2014.

[15] R. R. Shah, Y. Yu, A. D. Shaikh, and R. Zimmermann. TRACE: A Linguistic-based Approach for Automatic Lecture Video Segmentation Leveraging Wikipedia Texts. In *IEEE ISM*, 2015.

[16] R. R. Shah, Y. Yu, and R. Zimmermann. ADVISOR: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *ACM MM*, pages 607–616, 2014.

[17] R. R. Shah, Y. Yu, and R. Zimmermann. User Preference-Aware Music Video Generation Based on Modeling Scene Moods. In *ACM MMSys*, pages 156–159, 2014.

[18] A. D. Shaikh, M. Jain, M. Rawat, R. R. Shah, and M. Kumar. Improving Accuracy of SMS Based FAQ Retrieval System. In *Springer Multilingual Information Access in South Asian Languages*, pages 142–156. 2013.

[19] A. D. Shaikh, R. R. Shah, and R. Shaikh. SMS based FAQ Retrieval for Hindi, English and Malayalam. In *ACM FIRE*, page 9, 2013.

[20] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun. A Bayesian Framework for Video Affective Representation. In *IEEE ACII*, pages 1–7, 2009.

[21] A. Stupar and S. Michel. Picasso: automated soundtrack suggestion for multi-modal data. In *ACM CIKM*, pages 2589–2592, 2011.

[22] 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.

[23] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust Late Fusion with Rank Minimization. In *IEEE CVPR*, pages 3021–3028, 2012.

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

[3] www.nptel.ac.in/