Multimodal-based Multimedia Analysis, Retrieval, and Services in Support of Social Media Applications

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

ABSTRACT

The rapid growth in the amount of user-generated content (UGCs) online necessitates for social media companies to automatically extract knowledge structures (concepts) from user-generated images (UGIs) and user-generated videos (UGVs) to provide diverse multimedia-related services. For instance, recommending preference-aware multimedia content, the understanding of semantics and sentics from UGCs, and automatically computing tag relevance for UGIs are benefited from knowledge structures extracted from multiple modalities. Since contextual information captured by modern devices in conjunction with a media item greatly helps in its understanding, we leverage both multimedia content and contextual information (*e.g.*, spatial and temporal metadata) to address above-mentioned social media problems in our doctoral research. We present our approaches, results, and works in progress on these problems.

Keywords

Multimodal analysis; UGCs; Social media applications

1. INTRODUCTION AND MOTIVATION

Due to advancements in technologies, capturing UGCs anytime and anywhere, and then instantly sharing them on social media platforms has become a very popular activity. This activity attracts social media companies to provide diverse multimedia-related services leveraging multimedia content. However, providing such services are very challenging because it is difficult to capture semantics and sentics from a large collection of real-world UGIs and UGVs. Since multimodal information augments knowledge bases by inferring semantics and sentics from unstructured multimedia content and contextual information [23], it is beneficial in several significant multimedia-related applications such as tag ranking for social media photos [28], automatic lecture video segmentations [26, 27], adaptive news video uploading [24], SMS-based FAQ retrieval systems [32, 33]. We leverage multimodal information in first answering automatic soundtrack recommendation for user-generated videos (UGVs) [30, 31]. Next, we deal with the semantics understanding of UGIs and automatically generate a multimedia summary for a given event in real-time [25]. Finally, in our works in progress, we focus on sentics understanding and computing tag relevance for UGCs [29].

Since many outdoor UGVs consist of ambient background noise, it necessitates replacing the noise with matching soundtracks that matches with scenes, locations, and users' preferences. However, manually generating soundtracks for a UGV is tedious and time-consuming. Thus, we present a fast and effective heuristic ranking approach based on heterogeneous late fusion by jointly considering three aspects: venue categories, visual scenes, and listening histories of users. First, we predict scene moods for the UGV. Next, we perform heuristic rankings to fuse the predicted confidence scores of multiple models. Finally, we customize the video soundtrack recommendation functionality to make it compatible with mobile devices. Additionally, we consider areas where UGIs are exploited to provide services.

We present the EventBuilder[1] system that deals with the semantics understanding of user-generated images (UGIs) aggregated in social media sharing platforms and automatically generates a multimedia summary for a given event. It has three novel characteristics: (i) leveraging Wikipedia as event background knowledge to obtain additional contextual information about the input event during event detection, (ii) visualizing an interesting event on a Google Maps in real-time with a diverse set of social media activities, and (iii) solving an optimization problem to produce text summaries for the event. Subsequently, we present the EventSensor system that aims to address the sentics understanding of UGCs and produces a multimedia summary for a given mood. It extracts concepts and mood tags from the visual content and textual metadata of UGCs and exploits them in supporting several significant multimedia-related services such as a musical multimedia summary. Finally, we present a tag ranking system which computes tag relevance for UGIs based on voting from neighbors derived leveraging the three proposed high-level features. It is very helpful in the analysis, search, and retrieval of UGCs on social media.

In this doctoral research, we aim to exploit the multimodal information of UGCs in the support of the above-mentioned problems. Figure 1 shows an overview of our approach that leverage contextual information (*e.g.*, temporal, spatial, and sensor data) in conjunction with captured multimedia content in our solutions. Moreover, we also exploit knowledge bases in the semantics and sentics understanding of UGCs.

[1] eventbuilder.geovid.org

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2971471

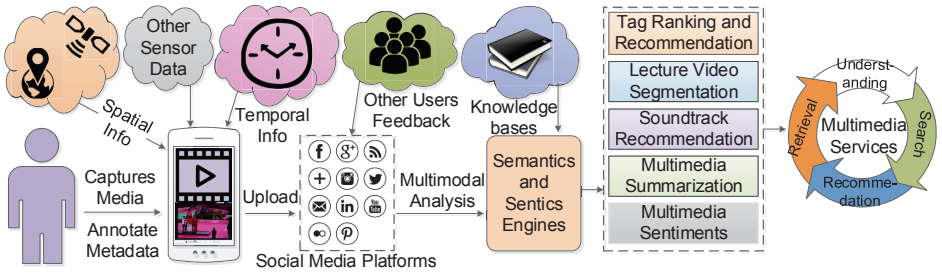

Figure 1: Overview of our approaches in this doctoral research studies.

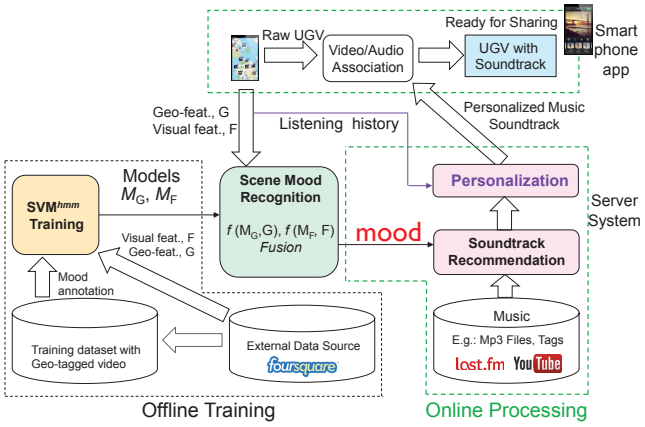

Figure 2: System overview of soundtrack recommendations for UGVs with ADVISOR [30].

2. STATE OF THE ART

Earlier works [11, 34, 37] recognize emotions from videos but the field of soundtrack recommendation for UGVs [9, 39] is largely unexplored. Our ADVISOR system [30] to recommend a matching soundtrack to a UGV is inspired by prior works [7, 16, 38]. It performs heterogeneous late fusion to recognize moods and retrieve a ranked list of songs using a heuristic approach for sensor-annotated videos (UGVs). Moreover, significant prior works [8, 10, 17, 18, 40] in the area of event modeling, understanding, detection, and summarization from multimedia, inspired us to leverage multimodal information and knowledge bases in an effective event detection and summarization from UGIs [25]. Earlier works [13, 35] added soundtracks to the slideshow of photos. Moreover, notable contributions [3, 4, 6, 19, 21, 22] in the area of sentiment analysis motivated us to exploit multimodal information in the sentics understanding of UGIs. Furthermore, regarding computing tag relevance for photos, Liu *et al.* [15] and Li *et al.* [14] proposed methods based on random walk and neighbor voting, respectively. However, they computed neighbors using costly low-level visual features.

3. PROPOSED APPROACH

In this doctoral research we focus on the following three objectives. In our first objective (**OBJ-1**), our goal is 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, in our second objective (**OBJ-2**), we address the problem of semantics and sentics understanding of UGCs. Finally, in our third objective (**OBJ-3**), we focus on computing the tag relevance of UGIs.

OBJ-1. Figure 2 shows the architecture of our proposed music video generation system, called ADVISOR [30]. 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 information (*i.e.*, geo-categories such as *Park* and *Lake* derived from Foursquare [2]) for a UGV serves as an important dimension to represent valuable semantics 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. 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 proposed 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 3). We leverage soundtracks 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 experiences of such experts using their *professional soundtracks* of Hollywood movies through a SVM^{hmm} learning model M. 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 (predicted based on G and F).

OBJ-2. Figure 4 shows the architecture of EventBuilder which detects events 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 ξ, λ, γ, μ, and ρ are similarity functions for the given

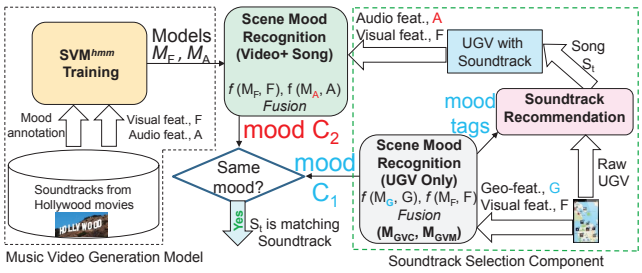

Figure 3: Soundtrack selection process in [30].

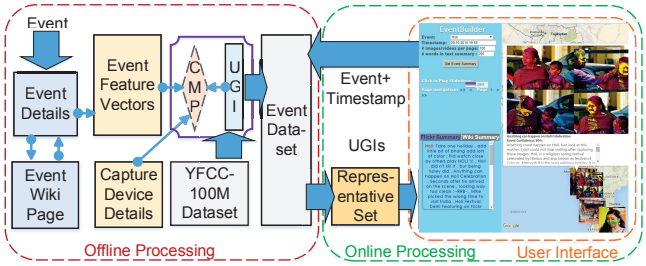

Figure 4: System framework of EventBuilder [25].

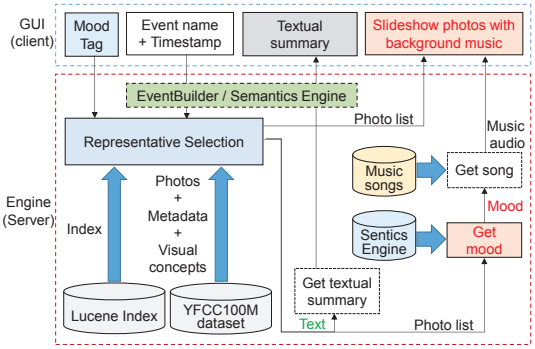

Figure 5: System framework of EventSensor [29].

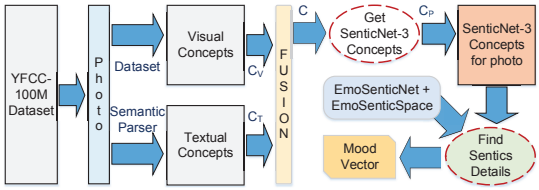

Figure 6: System framework of sentics engine [29].

p and e with respect to event name, temporal information, spatial information, keywords, and camera model, respectively, as described in [25]. EventBuilder also produces text summaries from UGIs and the Wikipedia article 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 *concepts*.

Figure 5 depicts the architecture of the EventSensor system consisting of two components: (i) a client which accepts a user's inputs such as a mood tag, an event name, and a timestamp, and (ii) a backend server which contains semantics and sentics engines. EventSensor leverages the semantics engine (EventBuilder) to obtain the representative set of UGIs R for a given event and timestamp. Subsequently, it uses its sentics engine to generate a mood-based event summarization. It attaches soundtracks to the slideshow of UGIs in R. The soundtracks are selected corresponding to the most frequent mood tags of the UGIs derived from the sentics engine in EventSensor.

Figure 6 shows the system framework of the sentics engine which leverages multimodal information to perform sentiments analysis. Specifically, we exploit concepts (knowledge structures) from the visual content and textual metadata of UGCs. We extract visual concepts for each multimedia item and compute concepts from the textual metadata of multimedia content using the semantic parser API [20]. Next we fuse the extracted visual and textual concepts. After determining the fused concepts C for the multimedia content, we compute the corresponding SenticNet-3 [5] concepts C_P since they bridge the conceptual and affective gap and contain sentics information. Finally, we compute a six-dimensional mood vector M_P of p by combining mood vectors of all concepts in C_P using an arithmetic mean. Experimental results indicate that the arithmetic mean of different mood vectors for concepts performs better than their geometric and harmonic means. Semantics and sentics information computed in earlier steps are very useful in providing different multimedia-related services to users. For instance we provide multimedia summaries from UGIs aggregated on social media such as Flickr. Once the affective information of UGCs is known, it can be used to provide different services related with affect. For instance, we can query Last.fm to retrieve songs for the determined mood tags and enable users to obtain a musical multimedia summary.

OBJ-3. Figure 7 shows the system framework of our tag ranking system based on neighbor voting. We propose three novel high-level features based on geo, visual, and textual content to compute neighbors of UGIs. Initial results indicate that the proposed features complement each others in computing tag relevance. We compute the weights for different modalities based on their recall scores, *i.e.*, the proportion of a seed UGI's tag covered by different modalities, and perform their late fusion to compute tag relevance.

We leverage the Foursquare API to map the GPS location of a UGI to geo concepts (*e.g.*, cafe, hotel, and office). Next, we treat each geo concept as a word and exploit the bag-of-words model [12] on a set of 1194 different geo concepts to create feature vectors. Similarly, we construct 1732-dimensional feature vectors corresponding to visual concepts (*e.g.*, nature, building, and art) present in the YFCC100M dataset [36]. Finally, to construct feature vectors from the textual metadata, we extract semantics concepts from the title, description, and tags of UGIs using the semantic parser provided by Poria *et al.* [20]. Next, we leverage SenticNet-3 knowledge base to construct an unified vector space, which is a publicly available resource for concept-level sentiment analysis [5] and consists of 30,000 common and commonsense concepts such as *food* and *accomplish_goal*. With the bag-of-words model, we construct 13727-dimensional feature vectors from textual metadata. After computing feature vectors for different modalities, we compute their k nearest neighbors using the cosine similarity metric.

After computing neighbors for a seed UGI p, we compute the relevance score of the seed UGI's tag as follows:

$$s(t,p) = \sum_{i=1}^{m} w_i \ (v_i(t,p) - prior(t,k)) \qquad (1)$$

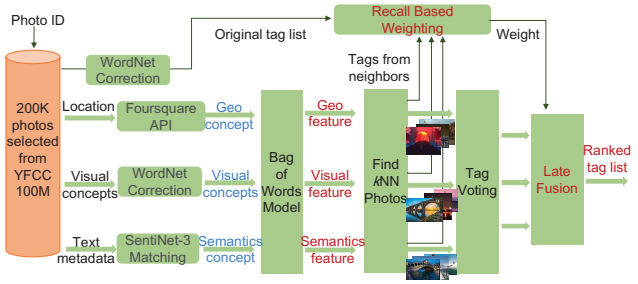

Figure 7: Architecture of our tag ranking system [28].

where m is the number of modalities, w_i is weight for different modalities such that $\sum_{i=1}^{m} w_i = 1$, $s(t,p)$ is the tag t's final relevance score, and $v_i(t,p)$ is the vote count from the k nearest neighbors derived for the ith modality. $prior(t,k)$ indicates the prior frequency of the tag t.

$$prior(t,k) = k\frac{M_t}{N} \quad (2)$$

where N and M_t are a total number UGIs and the number of UGIs tagged with t, respectively, in experimental dataset. For fast processing, we perform Apache Lucene [1] indexing of tags and UGIs. Finally, we rank the tags $t_1, t_2, ..., t_n$ based on their relevance score as follows:

$$rank(s(t_1,p), s(t_2,p), ..., s(t_n,p)). \quad (3)$$

4. RESULTS

To evaluate **OBJ-1** we used 402 soundtracks D_1 from Hollywood movies, 1213 sensor-annotated videos D_2, 729 songs D_3 from ISMIR, and 20 most frequent mood tags from Last.fm. To investigate scene moods prediction accuracy for UGVs, we randomly divided D_2 into training and testing datasets. After 10-fold cross validation, our experiments confirm that the model based on late fusion of geo- and visual features outperforms the models when they are trained with geo- and visual features alone. Moreover, to investigate the accuracy of soundtrack selection process, we randomly divided D_1 into a training and a testing dataset with a 80:20 ratio, and performed 5-fold cross validation experiments to calculate the scene moods prediction accuracy of M for UGVs in the test dataset. We predict the scene mood C_2 of a UGV from D_2 with a recommended soundtrack S_t using M, and compared with C_1. If both C_1 and C_2 are similar then we treat S_t as matching soundtrack since the prediction accuracy of C_1 is high. In this way we achieved the accuracy of 70.0% for 80 UGVs from D_2, which is comparable to the mood prediction accuracy of 68.8% for 80 soundtrack videos from D_1 (see [30] for detailed results).

We evaluate our event detection system [25] in **OBJ-2**, on the YFCC100M dataset [36] with 100 million photos and videos. We performed an extensive user study on results derived from baseline and EventBuilder. A photo consists the name of a given event in its metadata is considered as a result from the baseline. We randomly selected four photos each for the seven events used in EventBuilder [25] and asked 63 users to evaluatef the results by selecting photos which are relevant to the input events. We accepted 52 responses which fulfilled the annotation consistency criteria. Moreover, we also created ground truths for the results leveraging their content and contextual information. We compared responses of users with ground truths based on two metrics (i) precision, recall, and F measure, and (ii) cosine similarity. Figure 8 confirms that EventBuilder outperforms the proposed baseline in event detection.

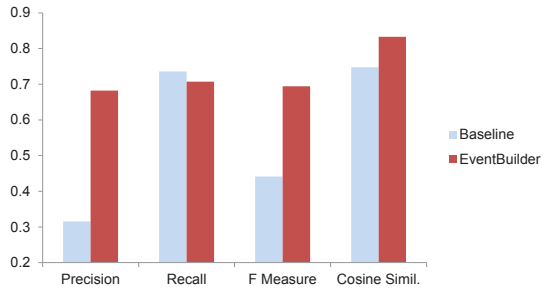

Figure 8: Results for event detection from 52 users, where x- and y-axis indicate evaluation metrics and scores between 0 to 1, respectively [29].

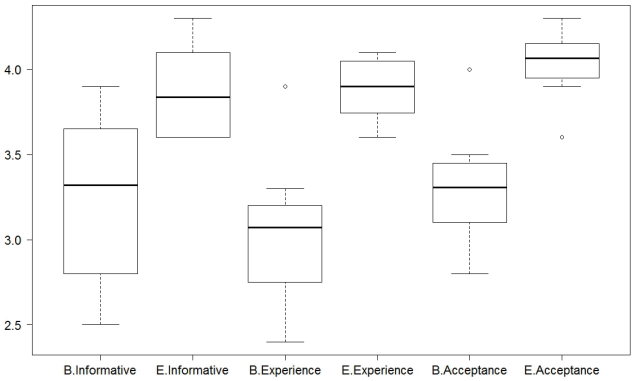

Figure 9: Boxplot for the informative, experience, and acceptance ratings of text summaries, where prefix *B* and *E* in x-axis indicate baseline and EventBuilder, respectively. In y-axis ratings range from 1 (low satisfaction) to 5 (high satisfaction) [29].

Next, we asked ten evaluators to assess produced text summaries by providing scores from 1 to 5, with a higher score indicating better satisfaction, based on the following three perspectives [25]: (i) informativeness, (ii) experience, and (iii) acceptance. We asked users to rate both the Flickr summary (baseline) which is derived from descriptions of UGIs and the Wikipedia summary which is derived from Wikipedia articles of events. Experimental results in Figure 9 indicate that users generally think that the Wikipedia summary is more informative than the baseline and can help them to obtain a quick overview of the events. However, the Flickr summary is also very helpful since they give an overview about what users thinks about the events.

5. CONCLUSIONS

This doctoral research studied the following three problems: (i) soundtracks recommendation for outdoor UGVs, (ii) semantics and sentics understanding of UGIs, and (iii) computing tag relevance for UGIs. We perform multimodal analysis and exploit knowledge bases to solve above problems. Experimental results confirm that they augment semantics and sentics understanding from multimedia content.

ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Apache Lucene. https://lucene.apache.org/core/, April 2016. Java API: Last Accessed April, 2016.

[2] FourSquare. https://developer.foursquare.com/, March 2016. API: Last Accessed March, 2016.

[3] B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain. Concept-level Sentiment Analysis with Dependency-based Semantic Parsing: A Novel Approach. *In Springer Cognitive Computation*, pages 1–13, 2015.

[4] 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.

[5] 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*, pages 1515–1521, 2014.

[6] E. Cambria, S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi. The CLSA Model: A Novel Framework for Concept-Level Sentiment Analysis. In *Springer Computational Linguistics and Intelligent Text Processing*, pages 3–22. 2015.

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

[8] 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.

[9] 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.

[10] M. Del Fabro, A. Sobe, and L. Böszörmenyi. Summarization of Real-life Events based on Community-contributed Content. In *MMEDIA*, pages 119–126, 2012.

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

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

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

[14] X. Li, C. G. Snoek, and M. Worring. Learning Social Tag Relevance by Neighbor Voting. *In IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.

[15] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag Ranking. In *ACM WWW*, pages 351–360, 2009.

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

[17] 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.

[18] P. Pakray, S. Pal, S. Poria, S. Bandyopadhyay, and A. Gelbukh. JU_CSE_TAC: Textual Entailment Eecognition System at TAC RTE-6. In *System Report, TAC RTE Notebook*, 2010.

[19] S. Poria, B. Agarwal, A. Gelbukh, A. Hussain, and N. Howard. Dependency-based Semantic Parsing for Concept-level Text Analysis. In *Springer Computational Linguistics and Intelligent Text Processing*, pages 113–127. 2014.

[20] S. Poria, E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain. Sentiment Data Flow Analysis by Means of Dynamic Linguistic Patterns. *In IEEE Computational Intelligence Magazine*, 10(4):26–36, 2015.

[21] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang. Towards an Intelligent Framework for Multimodal Affective Data Analysis. *Elsevier Neural Networks*, 63:104–116, 2015.

[22] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang. Sentic Patterns: Dependency-based Rules for Concept-level Sentiment Analysis. *In Elsevier Knowledge-Based Systems*, 69:45–63, 2014.

[23] R. R. Shah. Multimodal Analysis of User-Generated Content in Support of Social Media Applications. In *ACM ICMR*, pages 423–426, 2016.

[24] 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, 2016.

[25] 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.

[26] 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.

[27] 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*, pages 217–220, 2015.

[28] R. R. Shah, Y. Yu, S. Tang, S. Satoh, A. Verma, and R. Zimmermann. Concept-Level Multimodal Ranking of Flickr Photo Tags via Recall Based Weighting. In *MMCommons Workshop at ACM MM*, 2016.

[29] R. R. Shah, Y. Yu, A. Verma, S. Tang, A. D. Shaikh, and R. Zimmermann. Leveraging Multimodal Information for Event Summarization and Concept-level Sentiment Analysis. *In Elsevier KBS*, pages 1–8, 2016.

[30] 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.

[31] 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.

[32] A. D. Shaikh, M. Jain, M. Rawat, R. R. Shah, and M. Kumar. Improving Accuracy of SMS Based FAQ Retrieval System. In *Springer FIRE*, pages 142–156. 2013.

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

[34] 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.

[35] A. Stupar and S. Michel. Picasso: Automated Soundtrack Suggestion for Multi-modal Data. In *ACM CIKM*, pages 2589–2592, 2011.

[36] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The New Data in Multimedia Research. *In Communications of the ACM*, 59(2):64–73, 2016.

[37] H. L. Wang and L. F. Cheong. Affective Understanding in Film. *In IEEE TCSVT*, 16(6):689–704, 2006.

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

[39] 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.

[40] M. Zaharieva, M. Zeppelzauer, and C. Breiteneder. Automated Social Event Detection in Large Photo Collections. In *ACM ICMR*, pages 167–174, 2013.