Title: Leveraging Multimodal Information in Semantics and Sentics Analysis of User-Generated Content

Speaker: Mr Rajiv Ratn Shah

Date/Time: 5 January 2017, Thursday, 10:00 AM to 11:30 AM

Venue: Executive Classroom, COM2-04-02

Supervisor: Dr Roger Zimmermann, Associate Professor, School of Computing

Abstract:

The amount of user-generated multimedia content (UGC) has increased rapidly in recent years due to the ubiquitous availability of smartphones, digital cameras, and affordable network infrastructures. To benefit users and social media companies from an automatic semantics and sentics understanding of UGC, this thesis focuses on developing effective algorithms for several significant multimedia analytics problems. Sentic are common affective patterns associated with natural language concepts exploited for tasks such as emotion recognition from text/speech or sentiment analysis. Knowledge structures derived from UGC are beneficial in an efficient multimedia search, retrieval, and recommendation. However, real-world UGC is complex, and extracting the semantics and sentics from only multimedia content is very difficult because suitable concepts may be exhibited in different representations. Moreover, due to the increasing popularity of social media sites and advancements in technology, it is possible now to collect significant amount of important contextual information (e.g., spatial, temporal, and preference information). Thus, it necessitates analyzing the information of UGC from multiple modalities to facilitate different social media applications. Specifically, applications related to multimedia summarization, tag ranking and recommendation, preference-aware multimedia recommendation, and multimedia-based e--learning, are built by exploiting the multimedia content (e.g., visual content) and associated contextual information (e.g., geo-, temporal, and other sensory data). However, it is very challenging to address the above-mentioned problems efficiently due to the following reasons: (i) difficulty in capturing the semantics of UGC, (ii) the existence of noisy metadata, (iii) difficulty in handling big datasets, (iv) difficulty in learning user preferences, and (v) the insufficient accessibility and searchability of video content.

Exploiting information from multiple sources helps in addressing aforementioned challenges and facilitating different social media applications. Therefore, in this thesis, we leverage information from multiple modalities and fuse the derived knowledge structures to provide
effective solutions for several significant multimedia analytics problems. Our research focuses on the semantics and sentics understanding of UGC leveraging both content and contextual information. First, for a better semantics understanding of an event from a large collection of user-generated images (UGIs), we present the EventBuilder system. It enables people to automatically generate a summary for the event in real-time by visualizing different social media such as Wikipedia and Flickr. In particular, we exploit Wikipedia as the event background knowledge to obtain more contextual information about the event. This information is very useful in effective event detection. Next, we solve an optimization problem to produce text summaries for the event. Subsequently, we present the EventSensor system that aims to address sentics understanding and produces a multimedia summary for a given mood. It extracts concepts and mood tags from the visual content and textual metadata of UGIs and exploits them in sentics-based multimedia summary. EventSensor supports sentics-based event summarization by leveraging EventBuilder as its semantics engine component. Moreover, we focus on computing tag relevance for UGIs. Specifically, we leverage personal and social contexts of UGIs and follow a neighbor voting scheme to predict and rank tags. Furthermore, we focus on semantics and sentics understanding from user-generated videos (UGVs).

Since many outdoor UGVs lack a certain appeal because their soundtracks consist mostly of ambient background noise, we solve the problem of making UGVs more attractive by recommending a matching soundtrack for a UGV by exploiting content and contextual information. In particular, 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.

Furthermore, we address the problem of knowledge structure extraction from educational UGVs to facilitate e-learning. Specifically, we solve the problem of topic-wise segmentation for lecture videos. To extract the structural knowledge of a multi-topic lecture video and thus make it easily accessible, it is very desirable to divide each video into shorter clips by performing an automatic topic-wise video segmentation. However, the accessibility and searchability of most lecture video content are still insufficient due to the unscripted and spontaneous speech of speakers. We present the ATLAS and TRACE systems to automatically perform the temporal segmentation of lecture videos. In our studies, we construct models from visual, transcript, and Wikipedia features to perform such topic-wise segmentations of lecture videos. Moreover, we investigate the late fusion of video segmentation results derived from state-of-the-art methods by exploiting the multimodal information of lecture videos.