KNOWLEDGE DISCOVERY OVER COMMUNITY-SHARING MEDIA: FROM SIGNAL TO INTELLIGENCE

Winston Hsu [1], *Tao Mei* [2], *Rong Yan* [3]

[1] National Taiwan University, Taipei, Taiwan
[2] Microsoft Research Asia, Beijing, P. R. China
[3] IBM T. J. Watson Research Center, New York, USA

ABSTRACT

The explosive growth of photos/videos and the advent of media-sharing services have drastically increased the volume of user-contributed multimedia resources, which bring profound social impacts to the society and pose new challenges for the design of efficient search, mining, and visualization methods for manipulation. Besides plain visual or audio signals, such large-scale media are augmented with rich context such as user-provided tags, geo-locations, time, device metadata, and so on, benefiting a wide variety of potential applications such as annotation, automatic training data acquisition, contextual advertising, and visualization. We review the research advances for enabling such applications and present a brief outlook on open issues and major opportunities.

Index Terms— social media, multimedia annotation, multimedia search, multimedia advertising, visualization, machine learning, survey

1. INTRODUCTION

The prevalence of capture devices and growing practice of media sharing in community-contributed multimedia sites like Flickr [1] and YouTube [2] have brought profound social impact to the society and posed various new challenges for manipulating the rich contents.

Aside from the visual signals in social media, there are rich textual and visual cues [3], device metadata, and user interactions for context-aware social and organizing purposes [4]. The textual cues come form user-provided tags, descriptions for the media, and so on; the viewers might leave some comments or ratings for the social media, bookmark as favorites, or even mark "notes" (visual annotation) surrounding certain regions in the media. The capture devices can also provide geo-location, time, camera settings (e.g., shutter speed, focal length, flash, etc.), which reflect the capturing environment for the scene.

For these billion-scale social media, efficient methods for search and visualization have become emerging needs. Besides, due to the rich social interaction [4] and user-contributed contents, the social media sites have attracted large volumes of visitors (e.g., 100 million videos being watched everyday [5]). Therefore, multimedia advertising has become a key topic for social media monetization. Moreover, such rich media and user interactions have inspired researchers to leverage the enormous scale of (noisy) annotations to be an emerging resource for automatically learning media semantics. In this paper, we discuss the topics above and review the advances discovering the knowledge in social media. We also present a short list of important open issues.

2. SOCIAL MEDIA ANNOTATION

The most popular way nowadays to manage online multimedia content is manual annotation, or generally called, "tagging." The popularity of manual annotation partially stems from its high annotation quality for self-organization/retrieval purpose, and its social bookmarking functionality in online communities [4]. It allows users to annotate images with a chosen set of "tags" from an uncontrolled vocabulary. This type of approaches can be implemented in a variety of ways with respect to interface designs and user incentives. For example, Flickr encourages users to create free-text tags for each uploaded image. ESP Game [6] motivates users to annotate photos with freely chosen keywords in a gaming environment. However, it is very time-consuming for users to type and verify new keywords. For example, it takes nearly 15 seconds to obtain one tag in one ESP Game.

In light of the growing contents from social media, developing automatic annotation algorithms has drawn much more attentions from the research community. For example, the ALIPR system [7] used advanced statistical learning techniques to provide fully automatic and real-time annotation for user-uploaded digital pictures. Kennedy *et al.* considered using image search results to improve the annotation quality [8]. Moreover, many automatic annotation systems have considered leveraging external information sources to improve annotation performance, e.g., speech recognition, external semantic networks and location information from GPS [9]. For example, a location-tag-vision-based approach has been proposed to retrieve images of geography-related landmarks and features from the Flickr data set [10]. A recent work has attempted to characterize the time-evolving patterns of group photo streams [11]. Specifically, both image content and context information are leveraged in a joint matrix factorization framework for theme discovery and tag prediction.

While fully automatic annotation still achieved limited success [12], Internet-based annotation which is characterized by collecting crowd-sourcing knowledge, as well as combining human and computer for active tagging (i.e., ontology-free annotation), is a promising direction for annotation. For example, a recent work presents an active tagging approach to combine the power of human and computer for recommending tags to images [13]. The research on social tagging has proceeded in another dimension which aims to differentiate the tags with various degrees of relevance [14] [15]. The tags with different relevance can benefit visual search performance and in turn improve the relevance in any social media application.

3. MULTIMEDIA SEARCH

Current social media search approaches are mostly restricted to text-based solutions which process keyword queries against the tags or descriptions that are provided by users via some lightweight annotation tools. The associated tags may contain abundant information, yet their qualities are not uniformly guaranteed. Tags are therefore often inaccurate, wrong or ambiguous [10]. In particular, due to the complex motivations behind tag usage [4], tags do not necessarily describe the content of the image [8].

To remedy the limitation in (noisy or missing) tags, the authors propose ContextSeer, which formulates the keyword search as a ranking problem and fuses rich context cues (e.g., time, geo-tags, user-contributed tags, visual features, etc.) of shared consumer photos to improve the search result and even to recommend relevant tags and canonical images [3]. In [16], the authors exploit visual annotations (i.e., "notes" in Flickr) to enhance keyword-based photo search. The notes generally highlight a certain region (of interest) in the photo and associate a tag with the region—providing consistent visual and textual coherence.

A promising aspect for user-contributed photos and videos is their small-world phenomenon [5]. They are shown contextually correlated and can be embedded in a graph weighted by visual and context cues automatically. Such (context) graphs have shown effective for keyword-based photo [17] or video [18] [19] search by efficient random-walk-like methods.

4. MULTIMEDIA ADVERTISING

While research on advertising has been predominantly studied in the text domain since the end of 1990s, there has been an emerging trend arising from multimedia advertising. This trend particularity evolves with the unprecedented online delivery of social media, as well as the fast and growing online advertising market in recent years. In general, there are three key problems in an advertising system: (1) *relevance*—which ads are to be selected from an ad database according to a given image or video, (2) *position*—where the selected ads are to be embedded, and (3) *displaying*—how the selected ads are to be displayed at the detected positions. An effective advertising system aims at maximizing the relevance between the media and ads while minimizing the ad intrusiveness by taking the above three problems in to account.

Conventional advertising treats image and video advertising as general text advertising by displaying textually relevant ads based on the contents of Web page. Compared with text, social media like images and videos have unique advantages which consequently make them become more effective information carriers for advertising [20]: they are more attractive and salient than plain text, thus they can grab users' attention instantly; they carry more information that can be comprehended intuitively.

Recently, researchers have invented intelligent context-aware multimedia advertising technologies that can take advantages of the visual form of information representation. This new generation of multimedia advertising selects the ads contextually relevant to the media contents rather than the general Web pages, as well as seamlessly embeds the ads within rather than around the media. For example, the ads in [21] [22] are matched against the media according to the multimodal relevance which consists of textual and conceptual relevance, as well as visual similarity, while the ads in [23] are selected by searching the exact local patches. By leveraging vision techniques, a set of appropriate ad insertion positions can be detected within video streams (i.e., spatio-temporal positions on the frames) [21] [24] or images (i.e., non-salient areas) [22]. For example, VideoSense detects ad positions on the timeline based on the different combinations of content dissimilarity and attractiveness [21], while the virtual content insertion system finds the low attentive regions in the high attentive video shots as ad insertion positions by visual saliency analysis [23]. While fully automatic detection of non-salient areas within images for advertising is very challenging, ImageSense finds the image corners which are with the lowest saliency for ad embedding [22]. The next generation of multimedia advertising is envisioned to be game-like and more impressive [20].

5. SOCIAL MEDIA VISUALIZATION

The large amount of social media contents available on the Internet is typically unstructured. An effective visualization of such large-scale media collections can allow efficient indexing, browsing, and even effective world exploration and social interaction. It has proved effective to list tags of interest for a given location and visualize representative media associated to these tags.

The *World Explorer* system analyzes the tags associated with six million geo-referenced Flickr photos, and exposes the "representative tags" for each map region and zoom level by a multi-level text clustering process [25]. When a user points the mouse over a tag, photos associated with that tag and from that area are visualized. Kennedy *et al.* further studied how to select canonical views from these photos to represent a landmark [26]. It suggests that visual and geographic features can to be used to learn the visual models for selecting representative photos, on the basis of the photos shared by many individuals.

Rather than visualize social media by selecting representative tags and views, the efforts from vision community have predominantly focused on enabling users to navigate in a virtual tour which is created from a large photo collection. The *Photo Tourism* presents a 3D interface for interactively browsing and exploring large collections of unstructured photos of a scene [27]. It first estimates the pose of camera (location, orientation, and field-of-view) based on keypoint detection and matching, registers the photos in a global geometry coordinate, and then provides smooth transitions between photos by morphing techniques. As a result, it is able to present several modes for navigating through a scene: (1) free-flight navigation, (2) moving between related views, (3) object-based navigation, and so on. Another work in [28] organizes the photos in themes and constructs a virtual 3D space for photo navigation based on the visual similarities between the photos. The users are free to move from one image to the next using intuitive 3D controls. In response to user controls, the system retrieves the most similar images from several million images, displays them in correct geometric and photometric alignment with respect to the current photo.

For user-contributed videos, Kennedy proposed a system for synchronizing and organizing YouTube videos for live music events [29]. The work aggregates videos of the same venue and improves the representation of the event content, including identifying key moments of interest and descriptive text for important segments of the show.

6. TRAINING DATA CROWDSOURCING

The sheer volume of these resources poses not only opportunities, e.g., a simple non-parametric recognition approach supported by the 80 million tiny images collected from the Internet [30], but also challenges for the existing learning algorithms [31], most of which cannot scale well to this volume straightforwardly, e.g., Support Vector

Machines (SVMs), the state-of-the-art concept models. The common solutions for improving the scalability of the learning algorithms includes down-sampling the training collections, adopting advanced feature indexing / hashing methods, and resorting to distributed learning algorithms.

The crowd-sourcing nature of social annotations also brings another challenge when learning media semantics, i.e., the inaccurateness and incompleteness of the crowd annotations. It stems from several factors including user subjectivity, video- and album- level annotation, and lack of controls on annotation quality. This setback can greatly affect the learning performance if "noisy" tags are directly applied without being cleansed. In order to address this issue, the most straightforward approach is to manually clean up noisy images from the training collection. While manual expert cleanup is feasible for hundreds of thousands of images, the entire process can become very time- consuming and cannot easily go beyond the current limitation.

In contrast, a more scalable approach is to automatically identify the correct association between labels and images without any human intervention. To extract finer-granularity information from YouTube's video-level annotation, Ulges *et al.* presented a nonparametric probabilistic method to model the relevant keyframes to a given concept in presence of irrelevant content [32]. The frame-level annotation is iteratively determined by an Expectation-Maximization (EM) algorithm. Similarly, Fergus *et al.* developed the TSI-pLSA model, which extends the probabilistic Latent Semantic Analysis (pLSA) to handle large proportion of irrelevant images from Google Image retrieval results [33]. Empirical results show that these methods can successfully learn concepts out of even highly noisy labels, and can be used to re-rank the original retrieval outputs. However, the average performance of learning with automatically refined annotations is still noticeably lower than that of manually filtering counterpart.

Domain diversity is the third challenge for learning social annotation, because socially uploaded media data can originate from an unlimited number of sources, and there is little domain knowledge from which sources the users upload. Previous studies have demonstrated that, if the training and testing data come from non-identical sources, semantic learning performance will be dramatically degraded [20]. In view of these observations, several cross-domain learning methods have been developed, e.g., "adaptive" [34] and "cross-domain" SVMs [35]. They provide efficient and effective methods to adapt existing semantic models (i.e., SVMs) to new domains. Setz *et al.* shows social tagged training images can help to improve video search on broadcast videos, particularly after manual disambiguation [36].

7. OPEN ISSUES

The exciting developments in leveraging (implicit) knowledge in social media for these applications above are actually accompanied with many challenging open issues. We list some of the important ones below.

7.1. Context-aware fusion and context graph construction

The rich contextual cues mined from user-generated photos, videos, tags, or metadata have been shown effective for search, visualization, annotation, and event detection, for example, personalized recommendation based on user profile and interest, as well as concept or event detection by mining use behavior from geo-tags. However, it is still unknown how to effectively and automatically fuse these diverse (and sometimes missing) context cues for different applications. Meanwhile, the social-media entities (e.g., videos, photos, tags, users, and so on) are contextually correlated and able to boost the applications mentioned above by constructing the context graphs. However, beyond the experimental small-scale benchmarks, for the enormous billion-scale entities, it is still unknown how to build the graph efficiently, and meanwhile, how to utilize and align these diverse graphs for further manipulations.

7.2. Contextual and impressive multimedia advertising

How can we make multimedia advertising more impressive so that users are more willing to interact with the advertisements? This problem might be partially tackled from the following perspectives. (1) Design advertising in a game form to make users participate the game and get some incentives simultaneously. (2) Leverage techniques in computer graphics to render the advertisements in a more visually appealing way. (3) Rather than directly push the products or services around the media, we can automatically provide rich and related valuable information along with the advertisement. In this way, the user experience can be enriched by connecting more useful information with the ad contents (e.g., weather, discount information, traffic, education, traditional TV commercials, etc.) while consuming the advertisements.

7.3. Annotating social media

How can we annotate social media in a more robust and efficient way? We propose some promising directions to exploit. (1) Develop more efficient manual annotation interface and algorithms in a social environment. (2) Leverage additional metadata such as GPS and time to improve annotation accuracy. (3) Automatically suggest relevant tags to users in the manual tagging process. (4) Discovery user interest through media content and context for improving tag recommendation [37].

7.4. Leverage social annotation as training data

Online social annotation has proved a valuable resource for learning media semantics. However, there are several open issues worth further pursuing in this direction. (1) Scale the statistical learning algorithms to deal with web-scale training data; (2) Manage and mitigate the inaccurateness and incompleteness of crowd annotations. (3) Analyze and transfer training data from one specific domain to others (e.g., from YouTube videos to Flickr photos).

8. REFERENCES

[1] Flickr, http://www.flickr.com/.

[2] YouTube, http://www.youtube.com/.

[3] Y. Yang and P.-T. Wu, "ContextSeer: context search and recommendation at query time for shared consumer photos," in *Proceedings of ACM Multimedia*, 2008, pp. 199–208.

[4] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proceeding of SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 971–980.

[5] X. Cheng, C. Dale, and Jiangchuan Liu, "Statistics and social network of youtube videos," in *Proceedings of International Workshop on Quality of Service*, 2008, pp. 229–268.

[6] L. Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceeding of SIGCHI Conference on Human Factors in Computing Systems*, 2004, pp. 319–326.

[7] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," in *Proceedings of ACM Multimedia*, 2006, pp. 911–920.

[8] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label? predicting the performance of search-based automatic image classifiers," in *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, 2006.

[9] J. Kustanowitz and B. Shneiderman, "Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives," *Technical report, Univ. of Maryland*, 2004.

[10] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections.," in *Proceedings of ACM Multimedia*, 2007.

[11] Y.-R. Lin, H. Sundaram, M. D. Choudhury, and A. Kelliher, "Temporal Patterns in Social Media Streams: Theme Discovery and Evolution Using Joint Analysis of Content and Context," in *Proceedings of IEEE International Conference on Multimedia & Expo*, Cancun, Mexico, June 2009.

[12] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang, and M. Y. Chen, "Extreme video retrieval: Joint maximization of human and computer performance," in *Proceedings of the ACM International Conference on Multimedia*, 2006.

[13] K. Yang, M. Wang, and H.-J. Zhang, "Active tagging for image indexing," in *Proceedings of International Workshop on Internet Multimedia Search and Mining, in conjunction with ICME*, 2009.

[14] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang, "Tag ranking," in *Proceedings of International World Wide Web Conference*, 2009.

[15] E. Moxley, J. Kleban, J. Xu, and B. S. Manjunath, "Not all tags are created equal: Learning Flickr tag semantics for global annotation," in *Proceedings of IEEE International Conference on Multimedia & Expo*, Cancun, Mexico, June 2009.

[16] X. Olivares, M. Ciaramita, and R. Zwol, "Boosting image retrieval through aggregating search results based on visual annotations," in *Proceedings of ACM Multimedia*, 2008, pp. 189–198.

[17] Y. Jing and S. Baluja, "PageRank for product image search," in *Proceedings of International World Wide Web Conference*, Beijing, China, 2008.

[18] S. Baluja, R. Seth, D. Sivakumar, and et al., "Video suggestion and discovery for youtube, taking random walks through the view graph," in *Proceedings of International World Wide Web Conference*, 2008.

[19] W. Hsu and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proceedings of ACM Multimedia*, Augsburg, Germany, 2007, pp. 971–980.

[20] X.-S. Hua, T. Mei, and S. Li, "When multimedia advertising meets the new internet era," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, 2008, pp. 1–5.

[21] T. Mei, X.-S. Hua, L. Yang, and S. Li, "VideoSense: Towards effective online video advertising," in *Proceedings of ACM Multimedia*, 2007, pp. 1075–1084.

[22] T. Mei, X.-S. Hua, and S. Li, "Contextual in-image advertising," in *Proceedings of ACM Multimedia*, 2008, pp. 439–448.

[23] W.-S. Liao, K.-T. Chen, and W. H. Hsu, "AdImage: video advertising by image matching and ad scheduling optimization," in *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*, 2008, pp. 767–768.

[24] H. Liu, S. Jiang, Q. Huang, and C. Xu, "A generic virtual content insertion system based on visual attention analysis," in *Proceeding of the ACM International Conference on Multimedia*, 2008, pp. 379–388.

[25] S. Ahern, M. Naaman, R. Nair, and J. Yang, "World explorer: Visualizing aggregate data from. unstructured text in geo-referenced collections," in *Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries*, 2007.

[26] L. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proceedings of International World Wide Web Conference*, Beijing, China, 2008, pp. 297–306.

[27] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collection in 3d," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 835–846, 2006.

[28] J. Sivic, B. Kaneva, A. Torralba, S. Avidan, and W. Freeman, "Creating and exploring a large photorealistic virtual space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, June 2008.

[29] L. Kennedy and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," in *Proceedings of International World Wide Web Conference*, 2009.

[30] A. Torralba, "80 million tiny images: A large data set for non-parametric object and scene recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[31] L. Xie and R. Yan, *Extracting Semantics from Multimedia Content: Challenges and Solutions*, Book Chapter of Multimedia Content Analysis: Theory and Applications, 2008.

[32] A. Ulges, C. Schulze, D. Keysers, and T. Breuel, "Identifying relevant frames in weakly labeled videos for training concept detectors," in *Proceedings of ACM International Conference on Image and Video Retrieval*, 2008, pp. 9–16.

[33] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proceedings of International Conference on Computer Vision*, 2005, pp. 1816–1823.

[34] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proceedings of ACM Multimedia*, 2007, pp. 188–197.

[35] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui, "Cross-domain learning methods for high-level visual concept classification," in *Proceedings of IEEE International Conference on Image Processing*, 2008.

[36] A. Setz and C. G. M. Snoek, "Can social tagged images aid concept-based video search?," in *Proceedings of IEEE International Conference on Multimedia & Expo*, Cancun, Mexico, June 2009.

[37] J. Yu, D. Joshi and J. Luo, "Connecting People in Photo-Sharing Sites by Photo Content and User Annotations," in *Proceedings of IEEE International Conference on Multimedia & Expo*, Cancun, Mexico, June 2009.