Unsupervised Latent Aspect Discovery for Diverse Event Summarization*

Wen-Yu Lee, Yin-Hsi Kuo, Peng-Ju Hsieh, Wen-Feng Cheng, Ting-Hsuan Chao,
Hui-Lan Hsieh, Chieh-En Tsai, Hsiao-Ching Chang, Jia-Shin Lan, Winston Hsu
National Taiwan University, Taipei, Taiwan

ABSTRACT

Recently, the fast growth of social media communities and mobile devices encourages more people to share their media data online than ever before. Analyzing data and summarizing data into useful information have become increasingly popular and important for modern society. Given a set of event keywords and a dataset, this paper performs event summarization, aiming to discover and summarize what people may concern for each event from the given dataset. More specifically, this paper extracts latent sub-events with diverse and representative attributes for each given event. This paper proposes effective methods on detecting events with (1) human attribute discovery, such as human pose and clothes, (2) scene analysis, (3) image aspect discovery, and (4) temporal and semantic analysis, to provide people different perspectives for the events they are interested in. For practical implementation, this paper studied and conducted experiments on YFCC100M, which is a dataset with 100 million of photos and videos, provided by Yahoo!. Finally, a comprehensive and complete system is created accordingly to support diverse event summarization.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

Keywords

Event summarization, multimodal, visualization

1. INTRODUCTION

As the growth of social media, more and more people like to share their daily life and activities through social websites. Over the decades, thousands of photos and videos per minute have been uploaded to social photography sharing websites. Contributed by different people, the photos and videos on the cloud do offer rich information. As a result, detecting events of interest from large media collection has been receiving increasing attention for various media applications, such as earthquake detection by Twitter tweets [8].

*The work was supported in part by MOST 103-2622-E-002-034, MediaTek, and Intel-NTU Connected Context Computing Center.

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2809935.

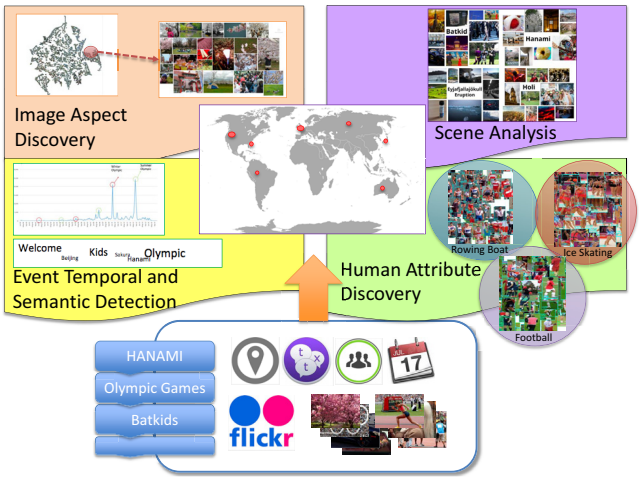


Figure 1: A system framework of extracting various perspectives for a large-scale dataset from the Flickr website.

Mostly, event detection attempts to answer the questions, "What are popular/happened?", "What they are doing?", and "What can be found?". As a global solution, event summarization gathers results of event detection and then combines useful information so as to provide summarized results to people. Event summarization automatically uncovers the structure of a large collection of photos and videos in terms of detecting and identifying events, followed by summarizing them succinctly.

This work focuses on event summarization. We consider a dataset of Flickr, with 100 million photos and videos. We extract (1) image contents, (2) locations, (3) time information, and (4) texts (e.g., title and user tags), and expected to unify them for summarization. Compared to previous works, our work not only detects special events, but also summarizes sub-events for each event. We then develop an effective system [1], see Figure 1, that discovers multiple diverse attributes and contents from the given dataset. Meanwhile, we present new methods to handle and leverage a massive amount of user-contributed data, including more than 100 million photos and videos. To sum up, the contributions of the paper include (1) developing new methods to unify and summarize different complementary information for data analytics, (2) discovering the latent behaviors, such as human poses and clothes (i.e., human attributes), (3) applying image aspect analysis and scene analysis to recognize and visualize the view of the given events which would be like, (4) performing unsupervised mining on the image contents of random top-

[1] `http://bit.ly/1DwJnTN`

Figure 2: Given an event, such as Olympics, we observed that we can automatically discover sub-events by analyzing human factors such as clothing.

ics to describe the given events, (5) showing robust detection and learning methods for massive and noisy user-contributed data, and (6) providing diverse contents for the given events.

2. MINING SPECIAL EVENT FROM FLICKR

2.1 Scene Diversity Analysis

Scene recognition can uncover the distribution of places representation among images in a specific event. It is important for users to quickly grasp the concept (e.g., outdoor, indoor) of an event's attribute. Deep convolutional neural networks (CNNs) show excellent results [4, 12] on scene recognition and they can transfer the generic representation from public scene datasets to our task (i.e., by the fine-tuning approach). Hence, for learning scene concepts, we train a deep convolutional neural network using Caffe [3] with the Places [12] pre-trained model, which we denote by Places-CNN. The Places dataset contains about 2.5 million scene-centric images from 205 scene categories. Finally, we use the activations of the last fully-connected layer as our scene feature.

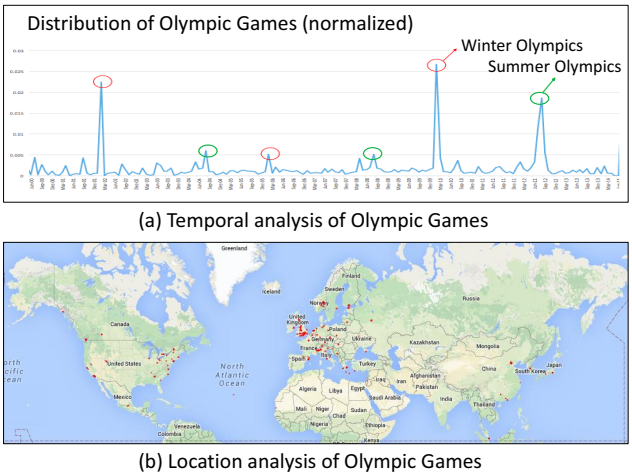
In order to demonstrate the diversity of the places representation, we use KL-divergence

$$D_{KL}(I_A||I_B) = \sum_i P_{I_A}(i) \ln \frac{P_{I_A}(i)}{P_{I_B}(i)},$$

where $P_I(i)$ is the probability of image I to be categorized as scene i. In the beginning, for each event, we draw an image which is closest to centroid of image pool as our first image into candidate pool. Then, we greedily draw the image with maximum KL-divergence against remaining images in image pool into candidate pool until we have enough images. With our method, we can ensure images in candidate pool to have large diversity, allowing users to understand the event with only a few images.

2.2 Automatic Aspect Discovery

Given an event, we attempt to discover various aspects automatically. Inspired by [1] which utilizes latent Dirichlet allocation (LDA) to solve multi-document summarization problem, we propose an automatic aspect discovery for event summarization. LDA, a powerful and unsupervised generative model, is a very promising technique in computer vision tasks such as object recognition [7] and human activity recognition [6]. Meanwhile, with the huge success of deep learning in computer vision [5], we integrate deep CNN features as semantic visual representations with LDA. That is to say, each image now becomes a document consist of a set of semantic visual words. We then apply LDA on images



(a) Temporal analysis of Olympic Games

(b) Location analysis of Olympic Games

Figure 3: Illustration of (a) temporal analysis and (b) location analysis on the given dataset for a pre-selected event, Olympic Games. Based on the results, it is most likely that Olympic Games were held around the time associated to the peaks shown in (a). The Olympic Games are likely to be closely related to the venues shown in (b).

for discovering latent aspects. Moreover, we can further analyze the discovered aspects from those semantic visual representations. We will demonstrate how this important attribute of LDA benefit our event-aspects discovery task in Section 3.4.

2.3 Human Pose and Clothing Analysis

Instead of focusing on text and image analysis, human action and activity are also essential factors to understand an event. As shown in Figure 2, for example; in Olympics, people participate in different games would have specific poses, clothes, etc. Hence, we attempt to investigate and cluster what people are doing during the event based on human pose and clothing analysis. For clothing detection, we apply pose estimator [11] to obtain the positions for arms and legs, and then trim the body box by four limbs. For pose analysis, followed by [2], we apply poselet, a body part detector, to find action patterns. Empirically, we only retain those high confidence results as action patterns. Based on clothing and action pattern detection, similar to prior section, we extract CNN features for obtaining semantic visual representations, and concatenate them for further unsupervised sub-event discovery.

3. VISUALIZATION AND APPLICATIONS

3.1 Data Preprocessing

Before the demonstration of our system, we first introduce our dataset as a preliminary. We used the YFCC100M dataset [9], which was provided by Yahoo! and was generated through data processing and data extraction from the Flickr website. Totally, the dataset contains 100 million of photos and videos. Based on our analysis, for the dataset, the timestamps of most data are between 2004 and 2014, about 48% data have location information, about 32% data have text description, and about 69% data have tags.

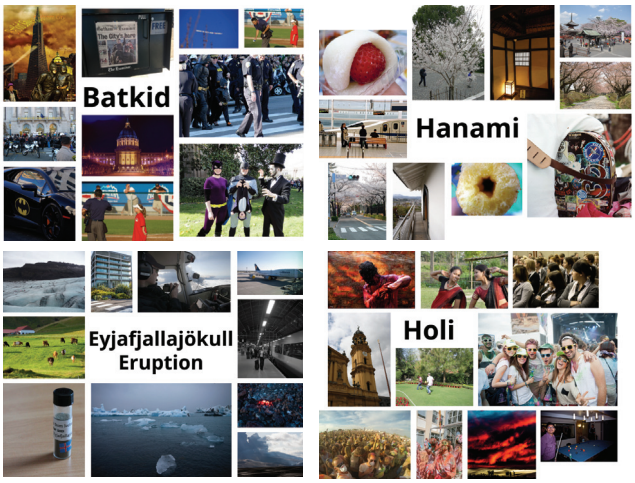


Figure 4: We apply Places-CNN to extract scene features and use it to select a set of images with the highest diversity by KL-divergence. Top-10 images are shown with corresponding event's name. Diverse aspects of place information make users quickly grasp the attributes of an event. For the event 'batkid', scene diversity analysis can reveal that this event took place in *streets*, *stadium*, *city hall*, etc.

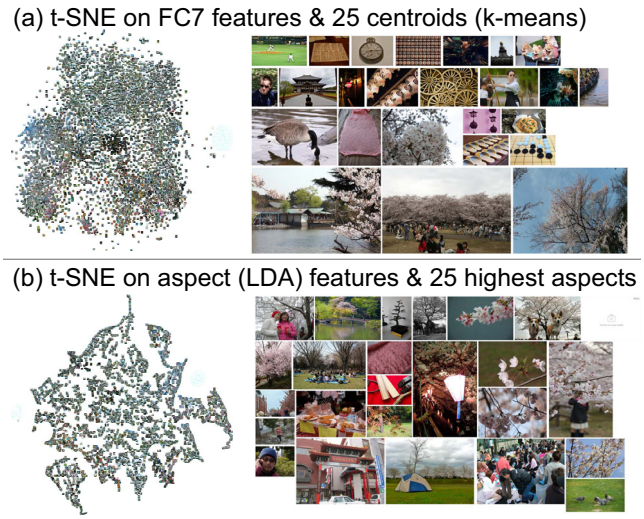


Figure 5: We apply t-SNE to visualize the effect of latent aspect discovery. (a) We observe that it might be hard to discover various aspects from the original deep CNN features. (b) However, after applying the proposed automatic aspect discovery, we can obtain more obvious aspects (topics) for the 'Hanami' event. We also provide the top-25 representative images from both features.

3.2 Temporal and Location

Consider temporal analysis and spatial analysis of an event. For temporal analysis, we found that an event may be composed of several sub-events, and the temporal distribution of the sub-events may be diverse. Some sub-events may last for only a few hours, while some sub-events may continue for several months. It is challenging to find a generic way for temporal summarization. An intuitive way is to draw timestamps of data along with a time axis. However, it might be hard to visualize the result when the distribution of the data points is too dense or sparse. As a result, we use time slots. Empirically, the size of time slots is first set to be one month. We then reduce the size to be one week for the months with peaks. For spatial analysis, we consider data title, data descriptions, and user tags of data. More specifically, for each event, we extracted the location information of a place if any of the title, description, and user tags, cover the event. Figure 3 illustrates the idea.

3.3 Scene Diversity Analysis

Scene diversity analysis can demonstrate the high diverse place information of an event and help users quickly grasp the attributes of an event. Figure 4 shows the results of some events (i.e., "Batkid", "Hanami", "Eyjafjallajokull Eruption", "Holi"). In these sets of top-10 images, we find that they can be divided into two parts, which is outdoor scene and indoor or building scene. For outdoor scene, the event "Batkid" mainly took place on *street*, "Hanami" on *sakura field*, "Eyjafjallajokull Eruption" on *iceberg/volcano*, and "Holi" on *playground*. For indoor or building scene, the event "Batkid" took place in *city hall*, "Hanami" in *washitsu*, "Eyjafjallajokull Eruption" in *train station*, and "Holi" in *religion buildings*. We also can discover other attributes such as *flavor meal* in "Hanam"' and *flag of Island* in "Eyjafjallajokull Eruption". Scene diversity analysis allows users to discriminate the semantic content in similar attributes (e.g., outdoor, indoor) and understand the event with only a few images.

3.4 Automatic Aspect Discovery

We first conduct experiments on "Hanami" event. We collect images with "Hanami" hash tag, and extract FC7 features using AlexNet described in [5]. To demonstrate the effect of the proposed latent aspect discovery, we utilize t-distributed stochastic neighbor embedding (t-SNE) [10] to reduce the dimension and visualize the relation between images. As shown in Figure 5, after LDA aspect discovery, it is more obvious to see the clustering effect. We claim LDA clustering would perform better on selecting representative images compared to k-means. Since we are discovering aspects within images related to a given event, the visual features tend to be consistent and less noise. LDA takes co-occurrence into consideration in the learning step; hence, it is more suitable to describe the common and distinctive aspects (topics).

3.5 Human Pose and Clothing Analysis

As mentioned in Section 2.3, we attempt to automatically cluster human activities for massive consumer photos. Based on the clothing and pose detection and analysis, we can roughly uncover sub-events from different dressing styles or action patterns. As shown in Figure 2, the players would have some specific actions, and wear particular clothes in the different Olympic Games. Therefore, we conduct experiments on Olympics related photos. We first search text that related to Olympics in 2012, and apply the framework described as previous. We then concatenate clothes features

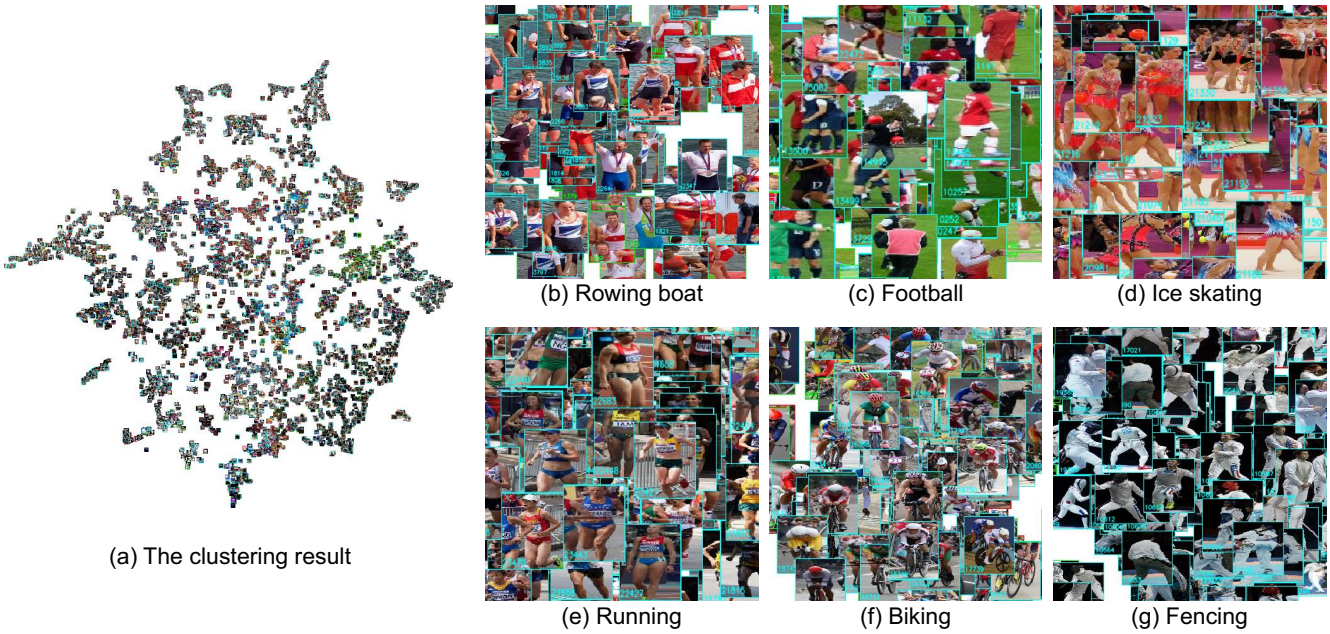


Figure 6: Based on the clothing and pose detection and semantic feature extraction, we can automatically uncover various sub-events for Olympic Games without specifying any target sub-event names. The idea is that people tend to wear similar clothing (e.g., fencing) or act similar poses (e.g., running, biking) for those sub-events. Therefore, we can provide various perspectives for a given event and generate more diverse event summarization.

and action pattern features to perform unsupervised image clustering by LDA. As shown in Figure 6, it is interesting to find that some specific clothes, including fencing suit, running sportwears, and biking are easy to recognize. Furthermore, we also notice the special backgrounds, such as football field (green grass) and fencing venues (black background), help to distinguish different sub-events using CNN features. Based on the summarized results, we analyzed what people in the images are doing. Generally, people participating in different games have different poses and different clothes. We detected human poses and extracted clothing features of visual content from user-contributed photos. Based on our experimental result, we concluded that the analysis of human poses and clothes is useful for the events that many people are participating in. As a result, we can detect some sub-events accordingly.

4. CONCLUSIONS AND FUTURE WORK

To tackle the challenges of event summarization, we investigate various perspectives for a given event, and propose unsupervised sub-event discovery to uncover latent aspects from text, visual content, and human activities. Compared to previous works on event detection, our work can not only detect events, but also sub-events, based on the given dataset. Experiment results demonstrate that the automatically discovered aspects can provide more distinctive and representative images for sub-events. Meanwhile, by analyzing human activities, we are able to realize the clothing styles and action patterns for those specific sub-events. In the future, we attempt to integrate the results from multimodality and further associate them together for achieving a more comprehensive event summarization.

5. REFERENCES

[1] R. Arora and B. Ravindran. Latent Dirichlet allocation based multi-document summarization. In *AND*, 2008.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[3] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. In *MM*, 2014.

[4] M. Koskela and J. Laaksonen. Convolutional network features for scene recognition. In *MM*, 2014.

[5] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[6] J. C. Niebles et al. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[7] B. C. Russell et al. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[8] T. Sakaki et al. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, 2010.

[9] B. Thomee et al. The new data and new challenges in multimedia research. *arXiv preprint:1503.01817*, 2015.

[10] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-SNE. *JMLR*, 9:2579–2605, 2008.

[11] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[12] B. Zhou et al. Learning deep features for scene recognition using places database. In *NIPS*, 2014.