EventBuilder: Real-time Multimedia Event Summarization by Visualizing Social Media

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

Anwar Dilawar Shaikh
Department of Computer Engineering, Delhi Technological University, India
anwardshaikh@gmail.com

Yi Yu
Digital Content and Media Sciences Research, National Institute of Informatics, Japan
yiyu@nii.ac.jp

Wenjing Geng
Department of Computer Science & Technology, Nanjing University, China
wjgeng@smail.nju.edu.cn

Roger Zimmermann
School of Computing, National University of Singapore, Singapore
rogerz@comp.nus.edu.sg

Gangshan Wu
Department of Computer Science & Technology, Nanjing University, China
gswu@nju.edu.cn

ABSTRACT

Due to the ubiquitous availability of smartphones and digital cameras, the number of photos/videos online has increased rapidly. Therefore, it is challenging to efficiently browse multimedia content and obtain a summary of an event from a large collection of photos/videos aggregated in social media sharing platforms such as Flickr and Instagram. To this end, this paper presents the EventBuilder system that enables people to automatically generate a summary for a given event in real-time by visualizing different social media such as Wikipedia and Flickr. EventBuilder has two novel characteristics: (i) leveraging Wikipedia as event background knowledge to obtain more contextual information about an input event, and (ii) visualizing an interesting event in real-time with a diverse set of social media activities. According to our initial experiments on the YFCC100M dataset from Flickr, the proposed algorithm efficiently summarizes knowledge structures based on the metadata of photos/videos and Wikipedia articles.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

Keywords

Event detection; event summarization; event visualization

1. MOTIVATION AND BACKGROUND

The EventBuilder system presented in this study is our solution to the ACM Multimedia 2015 Grand Challenge on automatic event summarization from a large collection of photos/videos. EventBuilder[1] is a real-time multimedia event summarization system which produces an event summary by visualizing social media such as Flickr. The number of photos/videos on social media sites has increased rapidly due to the advancement in smartphone and digital camera technologies and affordable network connectivity. However, obtaining an overview of an event from a large collection of photos/videos, such as the YFCC100M dataset (Đ), is still a very challenging task due to the following reasons: (i) the existence of much noise in Flickr metadata, (ii) the big size of such datasets, and (iii) the difficulty in capturing the semantics of photos/videos. EventBuilder leverages metadata such as user tags, descriptions, spatial information, *etc.*, of all photos/videos in Đ, and content of Wikipedia articles, using a feature-pivot approach to build the indices of event-datasets in Đ. Moreover, EventBuilder introduces a summarization system which considers aspects such as quality, diversity, coverage, and redundancy, during the building process. The summarization system consists of three steps: (i) the identification of important *concepts* which should be described in the event summary, (ii) the composition of a text digest which covers the maximal number of important *concepts* by selecting the minimal number of sentences from available texts, within the desired summary length, and (iii) the geo-graphical visualization of the event on Google Maps.

For efficient and fast processing, we compute scores of all multimedia contents in Đ for the given events and build indices for event datasets ($\mathcal{F}$) during pre-processing. We include only those photos/videos in $\mathcal{F}$ whose scores are above a threshold δ. Next, we formulate the problem and introduce a solution to produce text summaries of events by employing a sentence selection algorithm inspired by the greedy algorithm [3]. In the summarization process we consider a representative set of $\mathcal{M}$ photos/videos ($\mathcal{R}$) from $\mathcal{F}$ with the highest scores. In this way, the proposed EventBuilder system provides an overview of events through text summaries and improves the visualization of events by displaying only multimedia contents and their descriptions from $\mathcal{R}$.

[1]URL: eventbuilder.geovid.org

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

http://dx.doi.org/10.1145/2733373.2809932.

Since EventBuilder detects events from photos/videos offline rather than at search time, our system is time-efficient and scales well to large repositories. Since we use Wikipedia as background knowledge for more contextual information about the event to be summarized, our system can work well for new events by constructing feature vectors for those events by leveraging information from Wikipedia. An event summarization system can schedule event detection algorithms for newly uploaded photos/videos at regular time intervals to update event datasets.

2. RELATED WORK

There exists significant prior work in the area of event modeling, detection, and understanding from multimedia [5, 7, 10, 11]. Earlier methods [4, 9, 15] leveraged metadata such as user tags, spatial and temporal information to detect events automatically from a large collection of photos/videos such as Flickr. Rattenbury *et al.* [9] extracted place and event semantics for tags using Flickr metadata. Raad *et al.* [8] presented a clustering algorithm for automatically detecting personal events from photos shared online on the social network of a specific user by defining an event model that captures event triggers and relationships that can exist between events. They also detected multi-site, multi-day, and multi-person events using the appropriate time-space granularities. Fabro *et al.* [2] presented an algorithm for the summarization of real-life events based on community-contributed multimedia content using photos from Flickr and videos from YouTube. They evaluated the coverage of the produced summaries by comparing them with Wikipedia articles that report on the corresponding events. Atrey *et al.* [1] presented the detection of surveillance events such as human movements and objects being abandoned, by exploiting visual and aural information. Shah *et al.* [12] presented the ATLAS system which detects events such as slide transitions in a lecture video by exploiting its visual content and audio transcription. Low level visual features are often used for event detections or the selection of representative images from a collection of images/videos [7]. Recently, Papagiannopoulou and Mezaris [6] presented a clustering approach to produce an event-related image collection summarization using trained visual concept detectors based on image features such as SIFT, RGB-SIFT and OpponentSIFT. Wang *et al.* [14] summarized events based on the minimum description length principle, which is achieved through learning an HMM from the event data. Filatova and Hatzivassiloglou [3] proposed a set of event-based features which not only offers an improvement in summary quality over words as features (based on *tf-idf* scores), but also avoids redundancies in event summaries.

3. SYSTEM OVERVIEW

Figure 1 shows the system framework of EventBuilder. First, it performs event-wise classification and indexing of all photos/videos in social media datasets such as YFCC100M (Ð). Next, it solves the problem formulation of event summarization based on a greedy algorithm using event-based features which represent *concepts* (*i.e.*, important event-related information), as described by Filatova and Hatzivassiloglou [3]. *Concepts* associate the actions described in texts extracted from user descriptions and Wikipedia articles through verbs or action nouns labeling the event itself. Hence, an event summarization can be formulated as a maximum coverage problem. However, this problem is NP-hard, as it can be reduced to the well-known set cover problem. Thus, event summarization can be solved only in polynomial time by approximation algorithms.

Notations. Let N_e, T_e, S_e, and K_e be the feature vectors for event name, temporal information, spatial information, and keywords of an event e, respectively. The keywords of e are selected from the noun phrases used in the Wikipedia article of e. Let D be the list of 1080 camera models from Flickr, which are ranked based on their sensor sizes. Let N_i, T_i, S_i, K_i, and D_i be the feature vectors of photo/video i in Ð for event name, temporal information, spatial information, keywords, and camera model, respectively, and let $\xi(N_i, N_e)$, $\lambda(T_i, T_e)$, $\gamma(S_i, S_e)$, $\mu(K_i, K_e)$, and $\rho(D_i, D)$ be their corresponding similarity scores with e. Let $\mathcal{R}$ be the representative set for e, consisting of the top $\mathcal{M}$ photos/videos with the highest scores in an event dataset $\mathcal{F}$. Let $\mathcal{T}$ be the set of all sentences which are extracted from the description of photos/videos in $\mathcal{R}$ and contents of the Wikipedia article of e, and are used to produce a text summary $\mathcal{S}$. Let $|\mathcal{S}|$ and $\bar{\mathcal{L}}$ be the current word count and the word limit for $\mathcal{S}$, respectively. Let $\mathcal{K}$ and $\mathcal{Y}$ be the set of all *concepts* (c_k) of e and the set of corresponding weights (y_k), respectively. Let $v(s)$ be the score for a sentence s, which is the sum of the weights of all concepts it covers. Let $t(i)$ be the upload time of i. Let $\omega_{(s)}$ be a binary indicator variable which indicates if s is selected in the summary or not. Let $d(i)$ be a binary indicator variable which specifies if i has a description or not. Let $p(c, s)$ and $\beta(s, i)$ be 1 if $c \in s$ and $s \in i$, respectively, or otherwise 0.

Event Dataset and Experimental Settings. The score $u(i, e)$ of i for e is computed by a linear combination of similarity scores as follows: $u(i, e) = (w_1\ \xi(N_i, N_e) + w_2\ \lambda(T_i, T_e) + w_3\ \gamma(S_i, S_e) + w_4\ \mu(K_i, K_e) + w_5\ \rho(D_i, D))$, where $w_{i'}{}_{i'=1}^{5}$ are weights for different similarity scores such that $\sum_{i'=1}^{5} w_{i'} = 1$. We construct the event dataset $\mathcal{F}$ by indexing only those photos/videos of Ð whose scores $u(i, e)$ are above threshold δ. We set the weights as follows: $w_1 = 0.40$, $w_2 = 0.20$, $w_3 = 0.15$, $w_4 = 0.20$, and $w_5 = 0.05$, after initial experiments on the development set for event detection. We allocate only 5% of the total score for the camera model based on the heuristic that a good camera leads to a better quality photo/video which results in a better visualization of the event. All similarity scores, thresholds, and other scores are normalized to values [0, 1]. Let $\bar{\mathcal{L}}$ and $\mathcal{M}$ be system parameters to limit the summary length (number of words) and the number of photos/videos to be considered in the event summarization, respectively.

Problem Formulation for Text Summary. An event summary $\mathcal{S}$ is produced by extracting some sentences from $\mathcal{T}$, which cover important *concepts*. With the above notations and functions, we write the problem formulation for the event summarization as follows:

$$\min \sum_{(s \in \mathcal{T}) \wedge (i \in \mathcal{R})} \omega(s)\beta(s, i) \tag{1a}$$

$$s.t. \sum_{s \in \mathcal{T}} \omega(s)\ p(c, s) \geq 1,\ \forall\ c \in \mathcal{K} \tag{1b}$$

$$v(s) \geq \eta,\ \ \forall\ s \in \mathcal{T} \tag{1c}$$

$$|\mathcal{S}| \leq \bar{\mathcal{L}}, \tag{1d}$$

$$d(i) = 1,\ \forall\ i \in \mathcal{R} \tag{1e}$$

$$t(i) \leq \tau,\ \forall\ i \in \mathcal{R} \tag{1f}$$

$$\omega(s) \in \{0, 1\},\ \forall\ s \in \mathcal{T} \tag{1g}$$

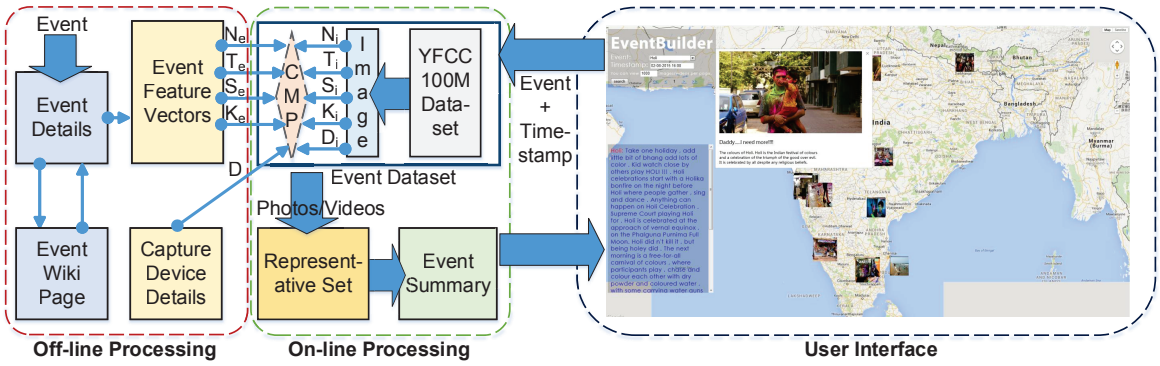

Figure 1: System framework of EventBuilder.

	c_1	c_2	c_3	c_4	...	$c_{\|\mathcal{K}\|}$
s_1	1	1	0	1	...	1
s_2	0	1	1	0	...	0
...	...	...	...	...	...	...
$s_{\|\mathcal{T}\|}$	1	0	0	1	...	1

Table 1: Matrix model for event summarization.

$$p(c, s) \in \{0, 1\}, \forall\, s \in \mathcal{T}, \forall\, c \in \mathcal{K} \quad (1h)$$
$$\beta(s, i) \in \{0, 1\}, \forall\, s \in \mathcal{T}, \forall\, i \in \mathcal{R} \quad (1i)$$

The objective function in Eq. (1a) solves the problem of event summarization and selects the minimal number of sentences which cover the maximal number of important concepts within the desired length of a summary. Eqs. (1b) and (1c) ensure that each concept is covered by at least one sentence with a score above threshold η. Eq. (1d) assures that the length constraint of the summarization is met. Eqs. (1e) and (1f) ensure that i has a description and is uploaded before the given timestamp τ. Eqs. (1g), (1h), and (1i) ensure that a concept, sentence, and photo is either present or absent in the sentence selected for the summarization.

Algorithm 1 Event summarization algorithm.

procedure EventSummarization
  INPUT: An event e and a time stamp τ
  OUTPUT: A text summary $\mathcal{SV}$
  $\bar{\mathcal{K}} = []$, $|\mathcal{S}| = 0$, $\bar{\mathcal{S}} = 200$ ▷ initialization.
  $(\mathcal{K}, \mathcal{Y}) =$ getEventConceptsAndWeights(e) ▷ see [3].
  $\mathcal{F} =$ getEventDataset(e) ▷ pre-processed event dataset.
  $\mathcal{R} =$ getRepresentativeSet(e, $\mathcal{F}$) ▷ representative photos.
  $\mathcal{T} =$ getSentences(e, $\mathcal{R}$) ▷ user description, Wiki texts.
  while $((|\mathcal{S}| \leq \bar{\mathcal{L}}) \wedge (\mathcal{K} \neq \bar{\mathcal{K}}))$ **do** ▷ $\bar{\mathcal{K}}$ is covered concepts.
    $c =$ getUncoveredConcept($\mathcal{K}$, $\bar{\mathcal{K}}$) ▷ important c first.
    $s =$ getSentence(c, $\mathcal{Y}$, $\bar{\mathcal{K}}$, $\mathcal{T}$) ▷ $c \in s \wedge v(s)$ is max.
    updateCoveredConceptList(s, $\bar{\mathcal{K}}$) ▷ add all $c \in s$ to $\bar{\mathcal{K}}$.
    addToEventTextSummary(s, $\mathcal{S}$) ▷ add s to summary.
    for each sentence $s \in \mathcal{T}$ **do** ▷ say, $s \in$ photo/video i.
      updScr(s, $\mathcal{Y}$, $\bar{\mathcal{K}}$) ▷ v(s)= u(i, e) $\times \sum_{c \in s, c \notin \bar{\mathcal{K}}} y$.

Solution. A greedy algorithm is reportedly the best possible polynomial approximation algorithm to solve NP-hard problems such as the set cover problem (*i.e.*, an event summarization problem in our case). We introduce a greedy algorithm which iteratively adds sentences to the event summary $\mathcal{S}$, until it either achieves the desired length $\bar{\mathcal{L}}$ or covers all *concepts*. Hence, the maximal number of important *concepts* are covered in the summary. Every time, when a new sentence is added to the summary, we check whether it contains enough new important *concepts* to avoid redundancy. We have formulated the problem of event summarization in terms of a matrix model, as shown in Table 1. Sentences and important concepts are mapped onto a $|\mathcal{T}| \times |\mathcal{K}|$ matrix. An entry of this matrix is 1 if the concept (column) is present in the sentence (row), otherwise it is 0. We take advantage of this model matrix to avoid redundancy by globally selecting the sentences that cover the most important concepts (*i.e.*, information) present in user descriptions and Wikipedia articles. Using the above matrix, it is possible to formulate the event summarization problem as equivalent to extracting the minimal number of sentences which cover all the important concepts. In our approximation algorithm, we constrain the total length of the summary with respect to the total weight of covered concepts, in order to handle the cost of long summaries. Our adaptive greedy algorithm for the event summarization is motivated by the summarization algorithm presented by Filatova and Hatzivassiloglou [3].

Algorithm 1 presents our summarization algorithm. First, we determine all event-related important concepts and their weights, as described by Filatova and Hatzivassiloglou [3]. Next, we extract all sentences from the user description of photos/videos in the representative set and texts in the Wikipedia article of an event e. We compute the score of a sentence by multiplying the sum of weights of all concepts it covers and the score of the photo/video which this sentence belongs to. Since each concept has different importance, we cover important concepts first. We consider only those sentences that contain the concept with the highest weight that has not yet been covered. Among these sentences, we choose the sentence with the highest total score and add it to the final event summary. Then we add the concepts which are covered by this sentence to the list of covered concepts $\bar{\mathcal{K}}$ in the final summary. Before adding further sentences to the event summary, we recalculate the scores of all sentences by not considering the weight of all the concepts that are already covered in the event summary. We continue adding sentences to $\mathcal{S}$ until we obtain a summary of the desired length $\bar{\mathcal{L}}$ or a summary covering all concepts.

4. EVALUATION

Dataset. The YFCC100M [13] (Yahoo! Flickr Creative Commons 100M) dataset, consisting of a total of 100 million photos/videos (approximately 99 million photos and 1 million videos) from Flickr, were provided with several metadata annotations such as user tags, spatial and temporal information, *etc.*, for the Yahoo-Flickr Event Summarization Challenge. To evaluate the proposed automatic summarization systems, the organizers have provided several events and timestamps, as listed in Table 2.

Event Name	Timestamp @ 12:00am UTC	Informative	Experience	Acceptance	Event Name	Timestamp @ 12:00am UTC	Informative	Experience	Acceptance
Holi	8 March 2009	3.93	4.13	4.00	Olympic Games	25 June 2012	3.00	2.60	2.80
	1 April 2009	3.80	3.80	3.66		1 Oct 2012	3.33	2.93	3.06
	1 April 2011	3.93	4.00	3.93		1 Jan 2015	3.27	3.13	3.13
	Avg. Rating	3.87	3.98	3.86		**Avg. Rating**	3.20	2.87	3.00
Eyjafjallajökull Eruption	1 April 2010	3.93	3.73	3.80	Batkid	15 Nov 2013	3.93	4.00	3.80
	18 April 2010	3.73	3.73	3.73		16 Nov 2013	3.80	4.00	3.73
	1 Jan 2014	4.06	4.13	4.0		1 Jan 2014	3.73	3.80	3.67
	Avg. Rating	3.91	3.86	3.86		**Avg. Rating**	3.61	3.93	3.76
Occupy Movement	19 Sep 2011	2.46	2.53	2.20	Byron Bay Bluesfest	1 June 2011	3.53	3.53	3.40
	1 Nov 2011	3.60	3.53	2.93		1 Jan 2015	3.53	3.60	3.60
	19 Sep 2012	3.67	3.67	3.67		-	-	-	-
	Avg. Rating	2.40	3.24	2.93		**Avg. Rating**	3.53	3.57	3.50
Hanami	1 Feb 2013	3.60	3.60	3.80	**All Events**	**Avg. Rating**	**3.46**	**3.61**	**3.54**
	1 June 2013	3.80	4.00	3.93					
	Avg. Rating	3.70	3.80	3.87					

Table 2: Evaluation results for the Yahoo-Flickr Event Summarization Challenge (YFCC100M dataset).

Results. The same algorithm was used for all summaries. No event-specific tuning was performed. All event summaries in our evaluation cover the maximal number of important concepts and their summary length is within the desired word limit. Since the summary evaluation is subjective in nature, we evaluated the performance through a user study. We asked fifteen evaluators to assess EventBuilder and provide scores from 1 to 5, with a higher score indicating better satisfaction. We defined three perspectives that evaluators should consider: (i) informativeness, which indicates to what degree a user feels that the summary captures the essence of the event, (ii) experience, which indicates if the user thinks the summary is helpful for understanding the event, and (iii) acceptance, which indicates if a user would be willing to use this event summarization function if Flickr were to incorporate it into their system. The default event summary length $\bar{\mathcal{L}}$ was set to 200 words during evaluation. Experimental results (see Table 2) indicate that users generally think that the summaries are informative and can help them to obtain a quick overview of an event.

5. CONCLUSIONS

The proposed EventBuilder system is a real-time automatic event detection and summarization system. First, it computes event-wise scores for each photo/video in social media datasets such as YFCC100M for the given events and builds event-wise indices. Next, it produces event summaries based on the given events and timestamps in real-time and facilitates efficient access to a large collection of photos/videos. In our future work we plan to extend the approach by exploiting visual contents, in addition to the available texts, using machine learning methods.

ACKNOWLEDGMENTS

This research has been supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office through the Centre of Social Media Innovations for Communities (COSMIC).

6. REFERENCES

[1] P. K. Atrey, A. El Saddik, and M. S. Kankanhalli. Effective Multimedia Surveillance using a Human-centric Approach. *Springer Multimedia Tools and Applications*, 51(2):697–721, 2011.

[2] M. Del Fabro, A. Sobe, and L. Böszörmenyi. Summarization of real-life events based on community-contributed content. In *MMEDIA*, 2012.

[3] E. Filatova and V. Hatzivassiloglou. Event-based Extractive Summarization. In *ACL Workshop on Summarization*, pages 104–111, 2004.

[4] C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Bringing Order to Your Photos: Event-driven Classification of Flickr Images based on Social Knowledge. In *ACM CIKM*, pages 189–198, 2010.

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

[6] C. Papagiannopoulou and V. Mezaris. Concept-based Image Clustering and Summarization of Event-related Image Collections. In *Workshop on HuEvent at ACM Multimedia*, pages 23–28, 2014.

[7] G. Petkos, S. Papadopoulos, V. Mezaris, R. Troncy, P. Cimiano, T. Reuter, and Y. Kompatsiaris. Social Event Detection at MediaEval: a Three-Year Retrospect of Tasks and Results. In *Workshop on SEWM at ACM ICMR*, 2014.

[8] E. J. Raad and R. Chbeir. Foto2Events: From Photos to Event Discovery and Linking in Online Social Networks. In *IEEE BdCloud*, pages 508–515, 2014.

[9] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *ACM SIGIR*, pages 103–110, 2007.

[10] A. Scherp and V. Mezaris. Survey on Modeling and Indexing Events in Multimedia. *Springer Multimedia Tools and Applications*, 70(1):7–23, 2014.

[11] A. Scherp, V. Mezaris, B. Ionescu, and F. De Natale. HuEvent '14: Workshop on Human-Centered Event Understanding from Multimedia. In *ACM Multimedia*, pages 1253–1254, 2014.

[12] 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 Multimedia*, pages 209–212, 2014.

[13] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The New Data and New Challenges in Multimedia Research. *arXiv preprint arXiv:1503.01817*, 2015.

[14] P. Wang, H. Wang, M. Liu, and W. Wang. An Algorithmic Approach to Event Summarization. In *ACM SIGMOD*, pages 183–194, 2010.

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