Event Photo Mining from Twitter Using Keyword Bursts and Image Clustering

Takamu Kaneko and Keiji Yanai[a]

[a] *Department of Informatics, The University of Electro-Communications, Tokyo*
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585, Japan

Abstract

Twitter is a unique microblogging service which enables people to post and read not only short messages but also photos from anywhere. Since microblogs are different from traditional blogs in terms of timeliness and on-the-spot-ness, they include much information on various events over the world. Especially, photos posted to microblogs are useful to understand what happens in the world visually and intuitively.

In this paper, we propose a system to discover events and related photos from the Twitter stream. We make use of "geo-photo tweets" which are tweets including both geotags and photos in order to mine various events visually and geographically. Some works on event mining which utilize geotagged tweets have been proposed so far. However, they used no images but only textual analysis of tweet message texts. In this work, we detect events using visual information as well as textual information.

In the experiments, we analyzed 17 million geo-photo tweets posted in the United States and 3 million geo-photo tweets posted in Japan with the proposed method, and evaluated the results. We show some examples of detected events and their photos such as "rainbow", "fireworks" "Tokyo firefly festival" and "Halloween".

Keywords: Twitter, Microblog, Geotagged Image, Event Mining, Event Photo Mining, Geo-Photo Tweet

1. Introduction

Twitter is a unique microblog, which is different from conventional social media in terms of its timeliness and on-the-spot-ness. Many Twitter's users send messages, which are commonly called "tweets", to Twitter on the

spot with mobile phones or smart phones. Therefore, Twitter users can be regarded as distributed "social sensors" which report what currently happens over the world [**?** **?**]. In addition, many of tweets contain not only text messages but also photos. Then, Twitter users can be regarded as distributed cameras as well. In general, photos can explain what currently happens much more intuitively than texts. By using such distributed image sensors effectively, we can understand what kind of events happen over the world at this moment visually and intuitively. Although Twitter has been extensively studied as a distributed sensor of real-world trends and events, most of them are based on text analysis, and their outputs are usually event keywords with their locations and times, which do not explain the detail of the detected events. As distributed camera sensors, Twitter has not been explored extensively yet. This is mainly because the amount of Tweet photo data is too huge to collect and process in general. If the number of photos are very large, their visual analysis including features extraction and clustering naturally becomes computationally expensive.

In this paper, we propose a system to discover events visually from the Twitter stream. To tackle a large quantity of Tweet Photos, we adopt a two-step method consisting of event keyword burst detection based on textual analysis as the first step and clustering-based photo selection based on visual analysis as the second step. First we detect "events" with only textual analysis in the similar way as the existing Twitter event detection methods. Then we extract visual features from only images related to the detected events and carry out visual clustering to select photos associated with the detected events. Since we restricted tweet photos for visual analysis to the photos related to the detected event, the required computation is not so heavy. Thus the proposed method can be applied in a real-time event photo detection system from the Twitter stream.

To do that, we pay attention to the tweets having both geotags and photos. We call such tweets as "geo-photo tweets". So far some works on event mining which utilize geotagged tweets have been proposed. However, they used no images but textual analysis and geotag analysis. On the other hand, in this work, we detect events using visual information as well as textual information and geolocation information. To the best of our knowledge, this is the first work on Twitter event mining employing both text analysis and image analysis.

In the experiments, we analyzed 17 million geo-photo tweets posted from the United States in 2012 and 3 million geo-photo tweets posted from Japan

from February, 2011 to September, 2012 with the proposed method, and then we successfully detected various kinds of event photos such as festivals, sport games, large-scale natural phenomena and some seasonal events. In addition, we implemented a real-time event photo detection system as well, and detected event photos in the real-time way.

To summarize our contributions in this paper is as follows: (1) We propose novel event photo mining from the Twitter stream, the results of which are useful to understand what happens in the world visually and intuitively. (2) A two-step method consisting of keyword burst detection and image clustering is proposed. (3) We made two large-scale experiments on the Japan dataset with 3 million geo-photo tweets and the US dataset with 17 million geo-photo tweets to show the effectiveness of the proposed method for event photo mining from Twitter.

The rest of this paper is organized as follows: Section 2 introduces related work, and Section 3 describes the overview of the proposed method on Twitter event photo mining. Section 4 describes the detail of the proposed method. In Section 5, experimental results are presented, and finally in Section 6 we conclude this paper.

2. Related Work

In the multimedia community, an "event" is used in various contexts. Some work defined it as an activity in which people participate and take pictures such as hiking, playing sports at park and wedding party [**?**], while in the TRECVID Multimedia Event Detection task it was defined as an abstract concept of "action" or complex actions, and includes more personal activities such as making a sandwich, repairing an appliance and marriage proposal [**?**]. As another work on activity events, abnormality detection from video/image streams has been studied before [**?**]. The objective is to detect abnormal events such as invasions and accidents from fixed camera video streams. Recently, as its variant, detecting interesting events has been proposed [**?**]. They proposed a computational model which integrates multiple cues to evaluate visual interestingness of image sequences. These works focused on "event classification/recognition/detection" which was a kind of image/video recognition.

On the other hand, in case of "event detection", an "event" tends to become more public and to gather many people, since a certain number of

photos or tweets related to a certain event are needed to detect the corresponding event. The MediaEval Social Event Detection (SED) Task defined "events" strictly as public events the schedules of which were announced on the Web event database, *last.fm*, such as music events and sport events [**?**], while in some event detection works the definition of "events" was broader and they allowed more personal events such as wedding to be regarded as "events" [**?** **?**]. In our work, in addition to scheduled social events such as sport games and festivals, we regards natural phenomena as "events" such as typhoon, heavy rain/snow, and beautiful sunset which exhibit uncommon scenes and draw attention of many people.

Many works on event detection have been proposed in the multimedia community so far. Most of the works used Flickr photos and tags as a target data from which events were detected including the MediaEval SED task [**?** **?** **?**], while the number of the works on Twitter photo data is limited. Therefore, we describe some works on Flickr event detection first, and then we explain works on Twitter event detection.

Rattenbury et al. is one of the pioneer works on event detection from Flickr tags [**?**]. They proposed Scale-structure Identification which is a burst detection method with multiple time scales. They used only event frequency along temporal direction to detect event tags, and used no geotags and no visual features. Chen and Roy [**?**] extended it by taking into account spatial direction in addition to temporal direction.

On the other hand, Quack et al. [**?**] proposed an object and event photo mining method which relies on visual features, and spatial information. Since their main objective was landmark detection, temporal information was not used. The detected event photos are very similar to each other in the same event cluster, because they used the number of matched SURF keypoints [**?**] as visual similarity. Therefore, the detected events are more personal than scheduled social events.

Papadopoulos et al. [**?**] also proposed a method on landmark and event photo mining which employs graph-based clustering with hybrid similarity of both visual similarity and tag similarity. Since they employed hybrid similarity and bag-of-features [**?**], this method outperformed the method proposed by Quack et al. [**?**] as event photo detection.

The approach by Liu et al. [**?**] was different from the other works. They selected the venues where the scheduled events were regularly held in advance, and monitored the statistics of the number of photos shared to detect events. Although the method was simple, the result was promising. This indicates

that event photo detection do not always need sophisticated methods, and a simple method is enough when the sufficient number of related photos are available.

The MediaEval Social Event Detection Task [**?**] is a representative benchmark task on social event detection. Because the training data is available in the task, supervised methods are common among the participants. Reuter and Cimiano [**?**] employed SVM with temporal, geographical and textual features, while Petkos et al. [**?**] proposed multi-modal clustering with supervisory signal employing visual features as well.

There exist many works related to Twitter mining using only text analysis such as the work by Weng and Lee [**?**], although only a few works exist on Twitter mining using image analysis. Some of them tried mining events from Twitter messages.

Sakaki et al. [**?**] regarded Twitter users as social sensors which monitored and reported the current status of the places where the users were. They proposed a system which can estimate the current location of typhoons or earthquakes by detecting the geotags attached to the related tweets from the Twitter stream and analyzing them.

Lee et al. [**?**] proposed an event detection system for geotagged tweets. They divided target areas into small sub-regions, and monitored the number of tweets posted from each sub-region. They regarded the areas where the number of tweets rose suddenly as the event areas where some events happened. In our work, we also examine the daily changes on the number of tweets of each area to detect events.

Hong et al. [**?**] detected events taking into account the difference of event keywords depending on areas and users. They classified tweets related to events taking into account regional tendency of keywords in tweets based on user preferences on events and profile statistics of users of various areas. In the experiments, they successfully estimated the locations where tweets with no geotags were posted, and detected event keywords which are berated in the specific areas.

Li et al. [**?**] segmented tweet timelines into tweet segments regarding a specific area (In their experiment, it was Singapore.) to detect "event segments". The segments where the number of tweets were bursted were regarded as event segments, and then they calculate "newsworthiness" of the event segments to exclude the event segments the newsworthiness of which were less than a pre-defined threshold.

In these works, they used textual information extracted from tweets and

geo-location information embedded in geotags, and did not use visual information which can be extracted from tweet photos.

As works on geo-photo tweets, we have proposed "World Seer" [**?**] which can visualize geotagged photo tweets on the online map in real time by monitoring the Twitter stream. This system can store geo-photo tweets to a database as well. We have been gathering geo-photo tweets from the Twitter stream since January 2011 with this system. On the average, we gather one hundred thousand geo-photo tweets a day.

To search this database, we have already proposed a system to mine representative photos related to the given keyword or term from a large number of geo-tweet photos [**?**]. In this work, we extracted representative photos related to events such as "typhoon" and "New Year's Day", and successfully compared them in terms of the difference on places and time. However, this system needs to be given event keywords or event terms manually. Then, in this paper, we integrate a method to select representative event photos with automatic detection of event keywords.

In the above-mentioned work [**?**], we used image clustering for visualization of event photos. As a clustering method, k-means was used with the fixed number of clusters. In this paper, we perform event photo clustering for image selection as well as visualization. Instead of k-means which requires the given number of clusters in advance, we use hierarchical clustering to select relevant clusters, which is the same as our past work on Web image selection [**?**]. Because hierarchical clustering needs no fixed number of clusters, it was commonly used in work on Web image clustering [**?**].

3. Overview

To detect events visually from Twitter stream, we monitor the Twitter stream to pick up tweets having both geotags and photos, and store them into a geo-photo tweet database using the data collection part of "World Seer" [**?**]. We apply to this database the proposed visual event mining system which consists of event keyword detection, event photo clustering and representative photo selection. The processing steps of the proposed system are as follows:

(1) Detect event keyword candidates which frequently appear in the tweets posted from specific areas in specific days.

(2) Unify and concatenate the detected event keywords.

(3) Select geo-tweet photos corresponding to the event keywords by image clustering.

(4) Select a representative photo to each event.

(5) Show the detected events with their representative photos on the map

Note that the current system assumes the tweet messages are written by either English or Japanese language, since keyword extraction needs to be taken into account the characteristics of target language. In general, it should be possible to extend the proposed system to other languages, provided a morphological analyzer is available which works for Twitter messages written in the target language.

4. Proposed method

In this section, we explain the detail of each step of the proposed system.

4.1. Event keyword detection

Tweet messages are written in sentences or sets of words in general. To detect events easier, at first we extract noun words from each tweet message.

To do this, for tweets written in English, we apply the English morphological analyzer which is specialized for tweet messages, TweetNLP [**?**], while for tweets written in Japanese language, we apply the Japanese morphological analyzer, MeCab [**?**]. According to the output of the morphological analyzer, we extract only noun words as keywords from each tweet after stop-word removal.

To detect events, we search for bursting keywords by examining change of the daily frequency of each keywords within each unit area. The area which is a location unit to detect events are defined by grid cells of one degree latitude and one degree longitude as shown in Figure 1 for the United States and Figure 2 for Japan. In case that the daily frequency of the specific keyword within one grid area increases greatly, we consider that an event related to the specific keyword happened within the area in that day.

In general, the extent of activity within one grid area depends on the location of the area greatly. The activity of the Twitter users in big cities such as New York and Tokyo is very high, while the activity in countryside such as Idaho and Fukushima is relatively low. Therefore, to boost the areas with low activity and handle all the areas equally in the burst keyword detection, we need to adjust the differences of the usual activities. Then, we set up the following equation to decide if an event related to the given keyword happens in the given area. We consider that an event happens if

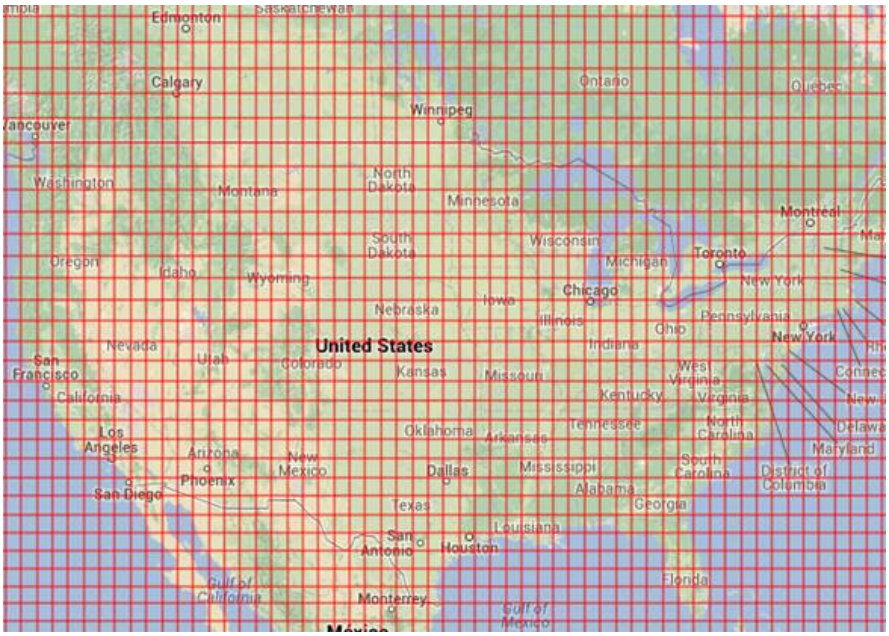

Figure 1: The grids dividing the United States. Each of them is a unit area for event detection.

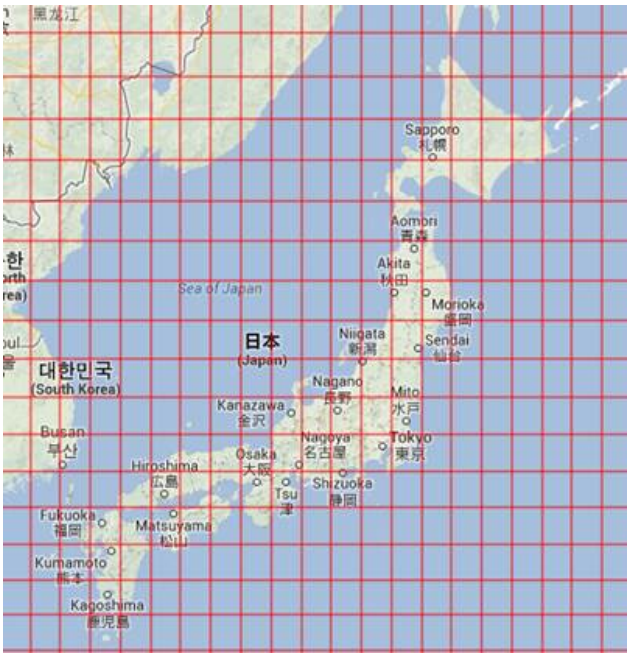

Figure 2: The grids dividing the Japanese Islands.

$S_{k,d,i,j}$ is more than the pre-defined threshold, which was set as 200 and 50 in the experiments for the US dataset and the Japan dataset, respectively.

$$S_{k,d,i,j} = (N_{k,d,i,j} - N_{k,d-1,i,j})W_{i,j} \ , \tag{1}$$

where k, d, i, j and $N_{k,d,i,j}$ represent an index of a keyword, an index of date, an index of area grids, and the number users who posted tweets in the indicated day and area, respectively. $W_{i,j}$ represents a weight to adjust the scale of the number of daily tweet users, which is defined in the following equation:

$$W_{i,j} = \frac{M + s}{N_{i,j} + s} \ , \tag{2}$$

where i, j, N, M and s represents the index of grids, the number of unique users in the given grid, the maximum number of unique users among all the grids (which is equivalent to the number of New York or Tokyo area users), and the standard deviation of user number over all the grids. The adjustment weight W plays a role to boost the areas with low activity, which is always more than 1.0 and becomes bigger for the areas with low activity. With this adjusting weight, we can detect events using a fixed threshold value from the areas where tweet users are not so many as well as the areas where so many tweets are always being posted such as the New York area and Tokyo areas.

Note that because we set the threshold as 50 and 200 for Japan and US, respectively, in the experiments on the Japan data, 50 tweets are needed at least in Tokyo area where the number of the unique users is the largest, while about 10 tweets are enough at rural area where the number of the unique users are very limited. In the same way, for the US data, 200 tweets is needed at least in New York area, while about 20 tweets are enough to be detected as an event at wilderness area where almost no users are observed in the normal condition.

4.2. Keyword unification and concatenation

In the previous step, we limited an event keyword to a single noun keyword. However, since some events are represented by compound keywords, the same event are sometimes detected by several keywords independently. In such case, we unify them into a compound keyword related to the same event according to the following heuristics:

(1) In case that more than half of the tweets related to a specific event keyword overlaps with the tweets related to another event keyword, the former keywords are integrated and replaced with the latter keywords.

- E.g. “rain” and “typhoon” ⇒ “typhoon”

(2) In case that words just after or before the detected event keyword are the same in more than 80% tweets including the keyword, such words are regarded as being part of a compound event keyword.

- E.g. “Tokyo”, “sky” and “tree” ⇒ “Tokyo Sky-tree”

4.3. Event photo clustering and representative photo selection

Until the previous step, event keywords and their corresponding tweets have been selected. In this step, we carry out clustering of the photos embedded in the selected event tweets and selecting representative ones from them.

As image features, we use bag-of-features (BoF) with densely-sampled SURF [**?**] local features and 64-dim RGB color histograms. SURF keypoints are sampled every 10 pixels in the scale 5, 10 and 15. The size of the codebook for BoF was set as 1000. Both feature vectors are L1-normalized.

For clustering photos, we use the Ward method which is one of agglomerative hierarchical clustering methods. It creates clusters so as to minimize the total distance between the center of each cluster and the cluster members. It merges the cluster pairs which bring the minimum total error calculated in the following equation one by one.

$$d(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2) \tag{3}$$

In general, $E(C)$ is defined as the total square distance between the center and the members of the cluster. Since we use two kinds of visual features, we defined $E(C)$ to combine them in the following equation.

$$\begin{aligned} E(C) = \sum_{x \in C} ((x_{BoF} - \overline{x}_{BoF})^2 w_{BoF} \\ +(x_{Color} - \overline{x}_{Color})^2 w_{Color}) \ , \end{aligned} \tag{4}$$

where x_{BoF}, x_{Color}, $\overline{x}$ and w represent a BoF feature vector, a RGB color histogram vector, a vector of the center of the cluster, and the weight which is defined as a reciprocal number of the dimension of each feature vector for equalizing the effect by each visual feature. In the experiments, we used a 1000-d BoF vector and a 64-d RGB color histogram vector as visual features. Therefore, the corresponding weights, w_{BOF} and w_{Color}, are 1/1000 and 1/64, respectively.

We evaluate each of the obtained clusters in terms of visual coherence with the following equation. We designed this equation so that the score of the cluster the member photos of which are similar to each other becomes larger.

$$V_C = \frac{n_C^2}{E(C)} W_{i,j} \ , \tag{5}$$

where n_C represents the number of photos in cluster C, and $W_{i,j}$ is the adjustment weight defined in Eq.(2). When V_C is high, the corresponding cluster is expected to be strongly related to the event. On the other hand, in case that V_C is lower, the cluster is expected to be less related to the event. In the experiments, we set the threshold of V_C as 20 and 5 for the US dataset and the Japan dataset, respectively, both of which were decided based on the results in the preliminary experiments.

V_C represents the degree of visual coherence of the given cluster. $E(C)/n_c$ corresponds to the average of the square distance between the cluster center and each cluster member, which can be regarded as "variance of visual features in the cluster". Then, we can regard V_C as (area weight) * (number of cluster members) / (variance of visual features), which becomes larger with smaller variance and larger number of cluster members. Note that $W_{i,j}$ (area normalizing weight) is fixed within the same event. We regard the cluster with many visually-coherent photos as important clusters, and we regard the cluster with the largest V_c within the same event as a representative cluster. That is why we defined V_c as Eq.(5).

In addition, the cluster having the maximum value of V_C is regarded as a representative cluster, and the photo the visual feature vector of which is the closest to the cluster center is selected as a representative photo for the corresponding event. This selection is called as "near-center selection" in the experiments. As an alternative option to select representative photo, we used VisualRank [**?**] which was the photo ranking method based on random walk. The calculation is the same as PageRank [**?**]. We apply VisualRank to all the photos in the representative cluster, and select a representative photo for the detected event.

In addition, the average geotag location over all the geo-photo tweets in the representative cluster is regarded as being the center of the geographical locations regarding the detected event.

5. Experiments

5.1. Dataset and evaluation

In the experiment, we prepared two large-scale geo-photo tweet databases: The first one is a Japan geo-photo tweet database which consists of about 3 million geo-photo tweets posted from Japan from February 10th, 2011 to September 30th, 2012. The second one is a United States geo-photo tweet database, which consists of about 17 million geo-photo tweets posted from United States from January 1st, 2012 to December 31st, 2012. Note that the tweets in the prepared databases are just part of all the geo-photo tweets actually posted in the given time duration and location, since only tweets sampled from the actual Twitter stream can be collected via the Twitter Streaming API with a free of charge account.

Note that tweet photos used in the experiments include the photos posted to other image hosting services than the Twitter official photo hosting service such as Instagram, ImageShack and Twitpic as well. Although Twitter started the official image hosting service in August 2011, even after that, many users are still using third-party image hosting sites. Therefore, in the experiments, we downloaded many images as "Twitter images" from such image hosting services including Instagram in addition to the Twitter official image hosting services. In fact, a big part of the Twitter images we used in the experiments are downloaded from Instagram. To build the Japan dataset consisting of about 3 million geotagged tweet photos, we downloaded 35.7% of them from Instagram, 33.7% from the Twitter official image hosting service, 14.3% from Twitpic, 9.3% from ImageShack and 7.0% from other image hosting services. To build the US dataset consisting of about 17 million geotagged tweet photos, we downloaded 51.8% of them from Instagram, 41.7% from the Twitter official image hosting service, 2.8% from Twitpic, 1.5% from ImageShack and 2.2% from other image hosting services. Regarding the US dataset, more than half images were hosted at Instagram.

For evaluation of the experimental results, we asked three students to evaluate the results regarding evaluation on event keyword detection, visual clustering and representative photo selection. Since we had two kinds of datasets: the Japan dataset and the US dataset, totally we had six kinds of evaluations. We assigned each whole evaluation to two of the four students to keep the evaluation standard fixed. Each item in each of the six kinds of evaluations was evaluated by two independent students.

To evaluate detected event keywords, we defined "events" as regional situations of something special lasting within a limited time period. For example, sports games, musical concerts, local festivals, special seasonal events such as illumination events, and special natural phenomena such as rainbow, heavy snow, typhoon and earthquake. For evaluation, we regarded keywords related to names of events and locations of events as relevant event keywords, and we regarded the photos expressing the scenes related to the detected event keyword as relevant event photos.

5.2. Experimental results on keyword selection

As results of event keyword extraction for the given dataset, from the Japan dataset, we obtained totally 306 keywords related to natural phenomena such as "rainbow" and "typhoon" and local events related to "fireworks" and "festivals".

Part of the keyword extraction results are shown in Table 1. In the table, "Area", "Weight" and "Score" represent the bounds of the grid in terms of latitude and longitude, the value of the area adjustment weight (Eq.(2)), and the value of the event score (Eq.(1)), respectively. Since the area where there are the largest number of unique users who posted geo-photo tweets was Tokyo, the weight value of the Tokyo area becomes 1.0. Because the other areas have less users than Tokyo, the adjusting weight value become more than 1.0.

From the US dataset, we obtained 2760 event keywords initially. Part of the results of keyword extraction on the US dataset is shown Table 2 as well.

As results of keyword unification and concatenation, the words which originally come from the same compound word such as "fireworks festival" are unified and converted into a compound keyword. Part of the results of keyword unification and concatenation are shown in Table 3 and in Table 4 for the Japan dataset and the US dataset, respectively.

After the unification and concatenation process, 306 and 2760 event words detected from the Japan and US dataset are reduced to 258 and 1676. The accuracy of the event keyword detected finally were 86.4% and 88.9%, respectively.

Note that in this work, we adopt day granularity to detect event words. We confirmed that one-day or less (several-hour) events could be detected with day granularity by the experimental results. However, there is possibility that multiple-day events such as "SXSW" (South By South West Music Festival) and "MLB World Series" are not detected except the first day. In

fact, regarding "SXSW" which is a 10-days event, the first day, the second day and the last day were detected as events in our system. Because the second day of SXSW was the first weekend day among the ten days, the number of geo-photo tweets associated with "SXSW" increased burstly compared to the first day.

Basically, from the Japan dataset, many events related to regional festival such as Gion Festival (in Kyoto), rainbow and typhoon are detected, while a lot of events related to sports games such as baseball, football and basketball and music festivals such as SXSW (in Texas) are detected from the US dataset.

Next, we examined the precision of finally detected event keywords when the threshold value to decide if the given word is related to some events was changed. We changed the threshold as shown in Table 5. As results, we obtained the graph shown in Figure 3, which showed that smaller threshold made the number of detected event smaller and the precision larger. To take into account the balance between the precision and the number of detected events, we selected 50 and 200 as the default values we used in the experiments for the Japan and US data, respectively. Note that the precision line on Japan data is lower than the line on US data in the graph shown in Figure 3. This is because sentences written in Japanese language are not separated with spaces between words unlike English, and Japanese morphological analyzer sometimes fails to separate keywords.

5.3. Comparison with EDCoW

Regarding event keyword detection, we made additional comparative experiments to EDCoW (Event Detection with Clustering of Wavelet-based Signals) proposed by J. Weng et al. [**?**] which is one of the state-of-the-art event detection methods. However, the computational cost of EDCoW is much larger than the proposed method, since it needs to calculate similarity matrix among all the pairs of the words. Therefore, we limited the temporal term and the spatial area to one month (August 2012) and one grid area (about 100km×100km) in the Bay Area in California, US where the most geo-photo tweets were collected in US in our experiment. We collected 534,141 geo-photo tweets in the given area and term.

We tested EDCoW with two kinds of the parameter settings on the window size. In EDCoW1 we used the larger window size, while in EDCoW2 we used the smaller window size. As results, in that area 27 events were

Table 1: Part of the list of the extracted keywords from the Japan dataset.

keyword	date	area	weight	score
snow	2011/2/11	34,35,135,136	1.96	135.5
earthquake	2011/3/11	35,36,139,140	1	55
fireworks	2011/8/6	34,35,135,136	1.96	149.2
festival	2011/8/6	34,35,135,136	1.96	68.7
Yodo-river	2011/8/6	34,35,135,136	1.96	72.6
dome	2011/8/10	43,44,141,142	3.96	51.5
rain	2011/8/19	35,36,139,140	1	60
typhoon	2011/9/21	35,36,139,140	1	62
Mt.Fuji	2011/9/24	35,36,138,139	3.35	67
Apple	2011/10/6	35,36,139,140	1	70
Ginza	2011/10/6	35,36,139,140	1	51
Suzuka	2011/10/9	34,35,136,137	3.94	78.8
eclipse	2011/12/10	34,35,135,136	1.96	84.4
Christmas	2011/12/24	35,36,136,137	2.9	55.2
New-Year's-Eve	2011/12/31	35,36,139,140	1	68
sunrise	2012/1/1	35,36,139,140	1	84
Meiji	2012/1/1	35,36,139,140	1	50
ski	2012/2/11	36,37,138,139	3.69	77.5
Marathon	2012/2/26	35,36,139,140	1	77
cherry-blossoms	2012/4/28	37,38,140,141	4.18	121.4
super moon	2012/5/5	35,36,139,140	1	96
firefly	2012/5/6	35,36,139,140	1	59
mother	2012/5/13	35,36,139,140	1	63
Tanabata	2012/7/7	34,35,135,136	1.96	56.9
Gion-Festival	2012/7/14	35,36,135,136	3.46	104
Tohoku-Denryoku	2012/7/14	37,38,139,140	4.4	79.2
peace	2012/8/6	34,35,132,133	4.08	77.5
Makuhari Messe	2012/8/11	35,36,140,141	3.18	168.9
Awa	2012/8/12	34,35,134,135	3.91	54.8
Daimonji	2012/8/16	35,36,135,136	3.46	83.2

Table 2: Part of the list of the extracted keywords from the US dataset.

keyword	date	area	weight	score
snow	2012/1/9	38,39,-78,-77	8.02	248.7
sunset	2012/1/13	47,48,-123,-122	10.7	290.4
Super	2012/2/5	37,38,-123,-122	10.9	251.8
Bowl	2012/2/5	37,38,-123,-122	10.9	251.8
Grammy	2012/12/12	34,35,-119,-118	6.52	208.7
Valentines	2012/2/14	37,38,-123,-122	10.9	438.1
SXSW	2012/3/8	30,31,-98,-97	9.67	464.4
Festival	2012/3/24	25,26,-81,-80	10.4	282.9
Music	2012/3/24	25,26,-81,-80	10.4	272.4
Easter	2012/4/8	33,34,-85,-84	9.13	703.1
shuttle	2012/4/17	38,39,-78,-77	8.02	577.7
Jazz	2012/4/27	29,30,-91,-90	10.3	228.2
eclipse	2012/5/20	37,38,-123,-122	10.9	1544.2
WWDC	2012/6/10	37,38,-123,-122	10.9	514.7
America	2012/7/4	33,34,-119,-118	10.9	373.5
hurricane	2012/8/26	25,26,-81,-80	10.4	241.0
rainbow	2012/9/5	37,38,-123,122	10.9	1423.7
49ers	2012/10/18	37,38,-123,122	10.9	262.8
Halloween	2012/10/31	40,41,-74,-73	1.45	375.2
vote	2012/11/6	34,35,-119,-128	6.57	345.6
Thanksgiving	2012/11/22	39,40,-85,-84	8.69	573.6
Cristmas	2012/12/24	40,41,-75,-74	3.94	1009.9
blizzard	2012/12/26	39,40,-87,-86	9.93	208.5
NYE	2012/12/31	34,35,-119,-128	6.52	436.9

Table 3: Results of keyword unification and completion for the Japan dataset.

keywords	after unification	after completion
fireworks, festival	fireworks	fireworks
forest	forest	Kodama forest
typhoon	typhoon	typhoon
Meiji	Meiji	Meiji shrine
dome	dome	Sapporo Dome
Apple, Ginza	Apple	Apple
eclipse, total	eclipse	total eclipse
Roppongi, hills	Roppongi	Roppongi Hills
firefly	firefly	Tokyo Firefly
Marathon	Marathon	Kyoto Marathon
super, moon	super	super moon
blue, moon	blue	blue moon
Makuhari, Messe	Makuhari Messe	Makuhari Messe
sea, beach, Nakamichi	sea	Nakamichi sea park
annular, eclipse	eclipse	annular eclipse
mother	mother	mother's day
sky, tree	sky	sky tree
Suzuka, circuit	circuit	Suzuka circuit

Table 4: Results of keyword unification and completion for the US dataset.

keywords	after unification	after completion
Super,bowl	Super	SuperBowl
Golden,Gate	Golden	Golden Gate
SXSW,Convention,sxsw	SXSW	SXSW
Auditorium,Shores	Auditorium	Auditorium Shores
Rangers,Ballpark	Rangers	Rangers Ballpark
Eclipse,eclipse	Eclipse	Eclipse
Summer,Fest,Press	Summer	Free Press Summer Fest
E3,expo	E3	E3
West,WWDC,Apple	WWDC	WWDC
Festival,North	North	North Beach Festival
Dodger,stadium	Dodger	Dodger Stadium
Bowl	Bowl	The Hollywood Bowl
rainbow,Double,Rainbow	rainbow	rainbow
Theater,Greek	Theater	Greek Theater
Theater,Fox	Theater	Fox Theater
Apple,Store,iPhone	Apple	Apple Store
shuttle,Space,Endeavor	shuttle	shuttle
Hotel,Casino	Hotel	Hotel&Casino
Halloween,costume	Halloween	Halloween
Bowl,Rose,UCLA	Rose	The Rose Bowl
Square,Union	Union	Union Square
Color,Candlestick	Color	The Color Run
Oracle,Arena	Arena	Oracle Arena
Field,LP	Field	LP Field
Square,Times	Times	Times Square

Table 5: The number of finally detected events vs. precision (%) when the threshold value on the event score of each word is changed.

threshold value for the Japan data	10	30	50	70	90
number of events	8544	957	258	109	56
precision (%)	67.7	79.5	86.4	88.1	91.1

threshold value for the US data	100	150	200	250	300
number of events	8317	3312	1676	1061	695
precision (%)	79.9	84.7	88.9	90.7	96.8

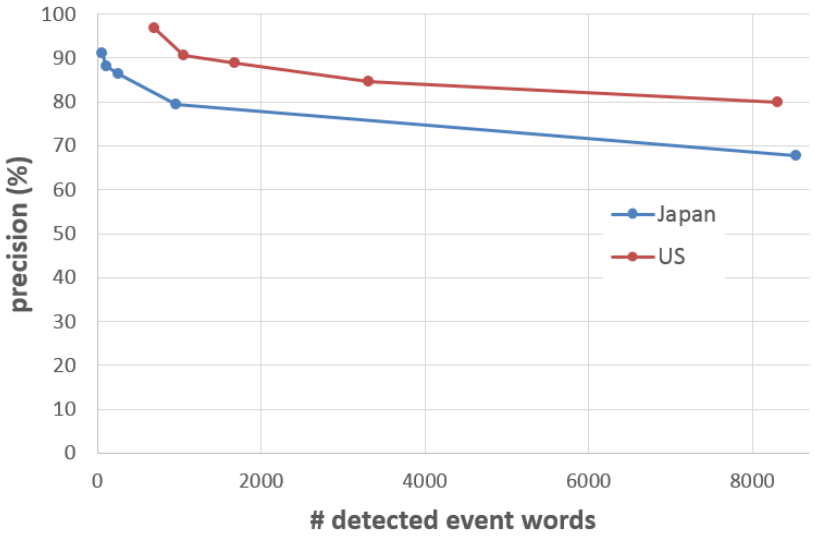

Figure 3: The number of finally detected events vs. precision (%).

detected by our method, while 25 events and 22 events were detected by EDCoW1 and EDCoW2, respectively. Figure 4 shows the relation between the number of detected events and the precision of detected events. We can see that our method is comparable to the state-of-the-arts EDCoW. Note that since EDCoW has no keyword unification mechanism, keywords are detected independently. For example, for the event that new exhibition was started as the California Academy of Science, "academy, California, sciences, and theater" are detected by EDCoW. On the other hand, by our method, "California Academy of Science" was detected as an event compound keyword. This is our advantage over EDCoW.

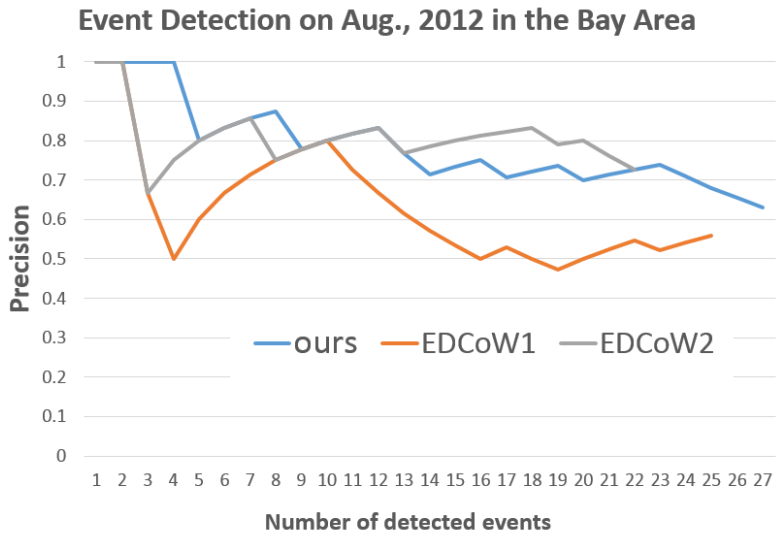

Figure 4: The number of finally detected events vs. precision (%) of the propose method and EDCoW with two kinds of window sizes.

5.4. Experimental results on photo clustering

Next, we show some example results of event photo clustering on the Japan dataset corresponding to three keywords, "fireworks", "cherry blossoms" and "firefly" (See Table 6 for the detail) in Figure 6, 18, and 19. The numbers shown on the right of each photo cluster represent cluster scores. The clusters (with red boxes) having the score which is more than 5.0 are regarded as event photo clusters, while the rest clusters (with blue boxes)

are regarded as non-event clusters unrelated to the corresponding event keyword. Within each cluster, photos are sorted in the ascending order of the distance to the cluster center. From the results, scoring of clusters worked successfully to place more visual clusters in the higher rank.

In Figure 18 ("cherry blossoms"), the first cluster consists of the photos of cherry blossoms taken in the daytime, while most of the photos in the second cluster shows cherry blossoms at night. In Figure 19 ("firefly"), the first cluster represents an illumination event of Tokyo Skytree which was called "Tokyo firefly"

In addition, as examples extracted from the US dataset, we show the clustering results on "Giants" (San Francisco Giants World Champion Parade 2012) and "Halloween" in Figure 20 and Figure 21.

We evaluated visual clustering results corresponding to events detected in August 2012 regarding both the Japan dataset and the US dataset. We regarded the photos expressing the scenes related to the detected event keyword as relevant images. The precision of detected event photos for the Japan dataset and the US dataset were 66.8% and 63.1%, respectively. The total number of detected photos and visual clusters were 2149 and 37 for the Japan dataset, and 7038 and 471 for the US dataset. Figure 5 shows the precision and the number of detected event photos for the detected event of the Japan dataset in August 2012 varying the threshold value for the visual coherence score of clusters V_c from 1 to 9. Although the default threshold value is 5 for the Japan dataset, in case that the threshold was 7 the best precision, 70.0%, would be obtained for the August 2012 data. From this graph, the sensitivity of visual clustering is not so large, although the precision depends on the setting of the threshold value.

Some detected events are shown on the map with their representative photos selected by the near-center selection in Figure 7 and Figure 8. These maps are interactive maps based on Google Maps API, and a user can see any event photos by clicking markers on the maps. Figure 9 shows an example after clicking the representative photo shown in the pop-up maker. This map-based interactive event viewing system is available via Web at http://mm.cs.uec.ac.jp/event/ for the US dataset and at http://mm.cs.uec.ac.jp/event_jp/ for the Japan dataset.

We show some correctly detected representative photos in Figure 10 and some incorrectly detected representative photos in Figure 11 from the Japan dataset with the near-center selection. Figure 10 shows the representative photos of historical festival, the new year's sunrise in 2012, "Tokyo Firefly"

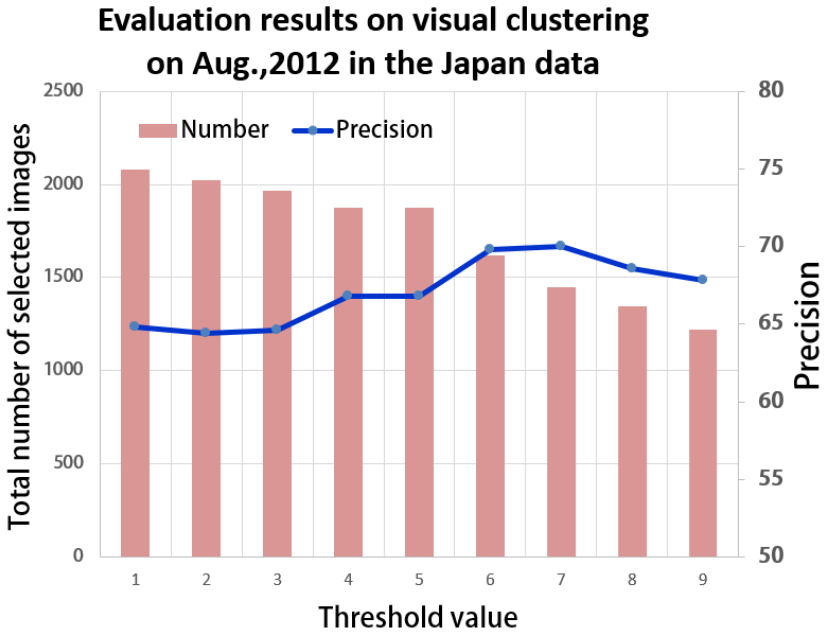

Figure 5: Evaluation results on visual clustering of event photos in August 2012 in the Japan dataset.

illumination event in the Tokyo Skytree tower, the eclipse in May 21, 2012, the clear sky after typhoon and an event held at Makuhari Messe. Figure 12 shows the representative photos selected by VisualRank corresponding to three events where the representative photo selection by the near-center failed, which shows two of three representative photos are correctly selected. Only "fireworks" representative photo was irrelevant.

We show some correctly detected representative photos in Figure 13 and some incorrectly detected representative photos in Figure 14 from the US dataset. Figure 13 show the representative photos of beautiful sunset, the Hollywood Bowl, Dodger Stadium, rainbow, Balloon Fiesta, and Christmas. Figure 15 shows the representative photos selected by VisualRank corresponding to three failed events shown in Figure 14. All the representative photos were selected correctly. This indicated that VisualRank is expected to be better than the near-center selection. To examine it, we compared their precision for all the detected events.

Finally, 258 and 1676 event keywords were detected in this experiments

Table 6: Summary for the example results.

event keyword	date	grid (lat,lng)	area	# photos
fireworks	2011/12/23	35,36,139,140	Tokyo	91
tree	2011/12/23	35,36,139,140	Tokyo	91
cherry blossoms	2012/04/21	34,35,135,136	Osaka	57
rainbow	2012/05/04	35,36,139,140	Tokyo	93
firefly	2012/05/06	35,36,139,140	Tokyo	93
Giants	2012/10/31	37,38,-123,-122	San Francisco	179
SXSW	2012/03/16	30,31,-98,-97	Austin	232
Halloween	2012/10/31	34,35,-119,-118	Los Angels	275

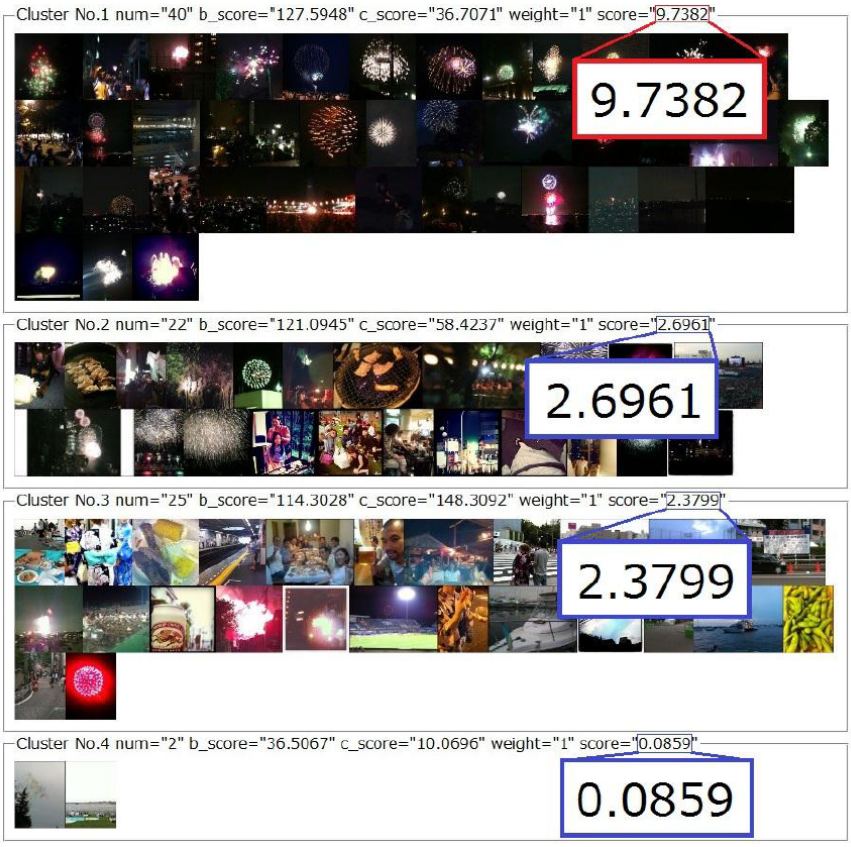

Figure 6: "Fireworks" photo clusters.

Figure 7: Some detected events in Japan are shown on the map with their representative photos.

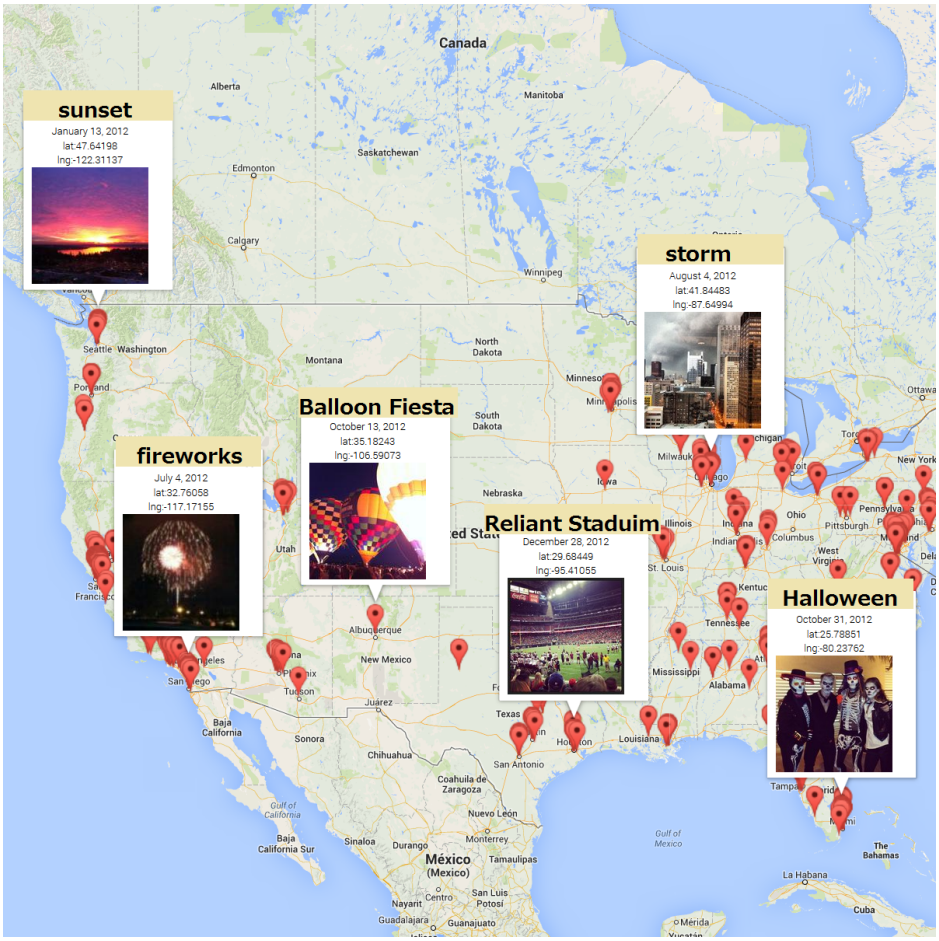

Figure 8: Some detected events in US are shown on the map with their representative photos.

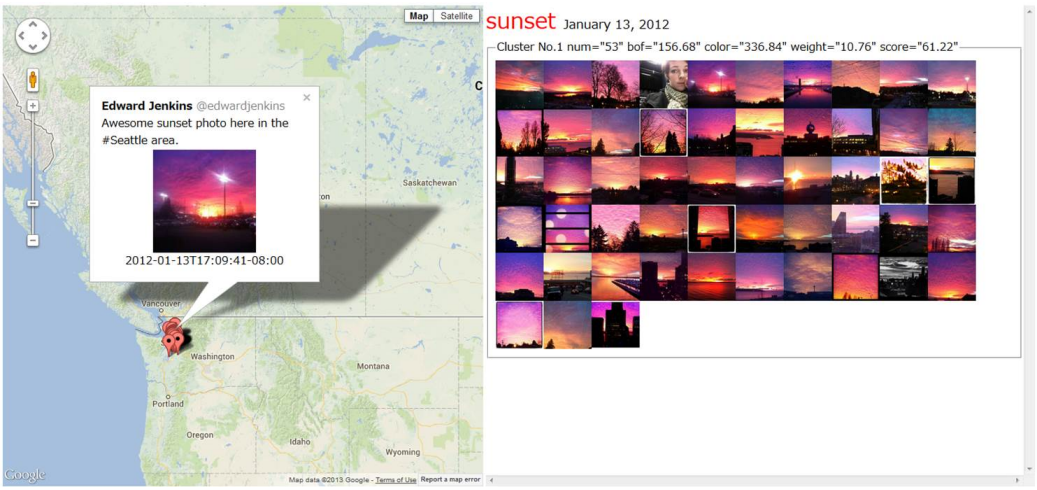

Figure 9: "Sunset" photos after clicking the representative photo shown in the pop-up maker.

from the Japan dataset and the US dataset, respectively. Among them, 224 and 1490 keywords were judged as being related to some of various kind of actual "events" including weather condition, natural events, festivals and sport games. Then, we evaluated if each of the representative photos selected by two methods, near-center selection and VisualRank [**?**], corresponding to the true event keywords are relevant to the event or not subjectively by hand. The criterion to evaluate relevancy of representative photos is if the given representative photo reminds an evaluator of the corresponding event. If so, the photo is regarded as relevant. If not, it is considered as irrelevant.

As results, the precision of the representative photos by the near-center selection were 72.9% and 52.7%, while the precision by VisualRank were 80.9% and 59.6%, for Japan and US data, respectively. The detail of the results are shown in Table 7. This result confirmed VisualRank is more suitable to select representative event photos.

5.5. Extension for real-time event photo detection

With the proposed method, we implemented a real-time system. Because our method requires relatively light computation, the proposed method can be used as a method of the real-time event detection with multi-thread processing.

Figure 10: Examples of representative photos selected correctly from the Japan dataset.

Figure 11: Examples of representative photos selected incorrectly from the Japan dataset.

Figure 12: Examples of representative photos selected by VisualRank from the Japan dataset.

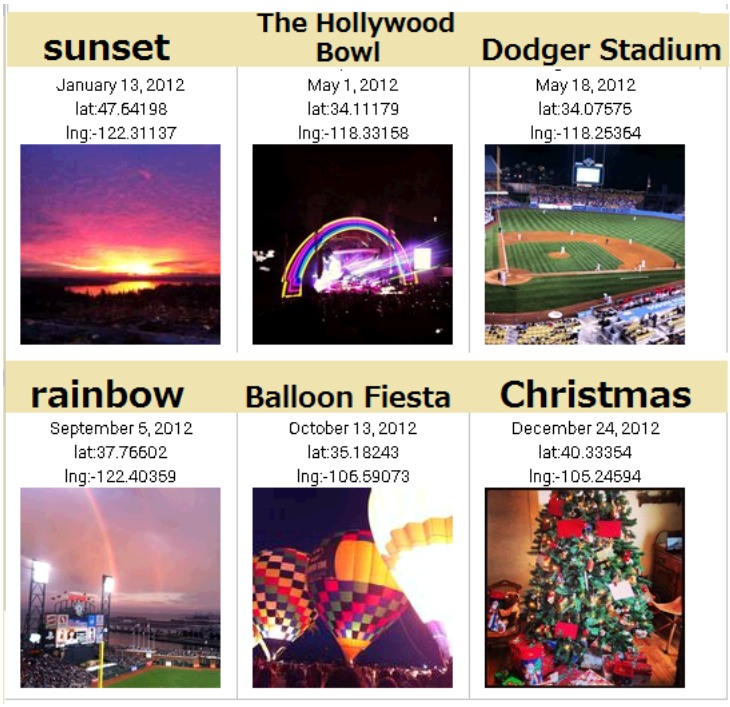

Figure 13: Examples of representative photos selected correctly from the US dataset.

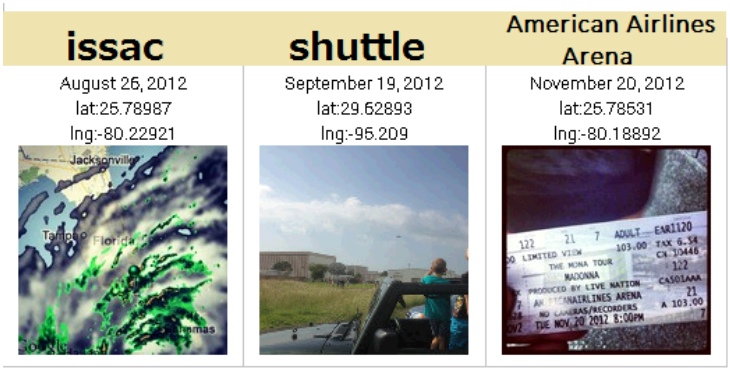

Figure 14: Examples of representative photos selected incorrectly from the US dataset.

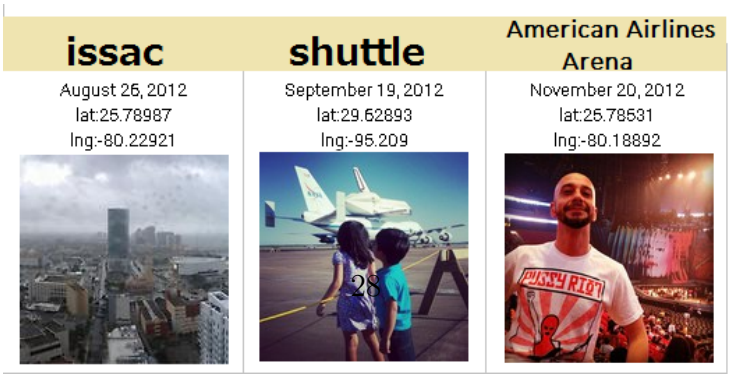

Figure 15: Examples of representative photos selected by VisualRank from the US dataset.

Table 7: The number and precision rate of the selected representative event photo for the JAPAN and US dataset by two methods: near-center selection and VisualRank. Note that the number of relevant photos are counted by two independent evaluators, and the averaged numbers are shown in this table.

	Japan	United States
(A) # detected event keywords	258	1676
(B) # relevant event keywords	223	1490
precision (B/A) (%)	*86.4*	*88.9*
(C) # relevant representative photos by near-center selection	162.5	785.0
precision (C/A) (%)	*63.0*	*46.8*
precision (C/B) (%)	*72.9*	*52.7*
(D) # relevant representative photos by VisualRank	180.5	888.5
precision (D/A) (%)	*70.0*	*53.0*
precision (D/B) (%)	*80.9*	*59.6*

By using the Twitter Streaming API, we monitor geo-photo tweets in the same way as [**?**]. We count the frequency of each of the extracted words every 100 seconds, and calculate an event score of each extracted word. If the event score exceeds the pre-defined threshold, the corresponding keyword is regarded as an event keyword. Every time a new event keyword is detected, a new thread is created to process unification and concatenation of event keywords, visual clustering and representative photo selection in the background. By using multi-threading, the system always monitor the Twitter stream by the main monitoring thread. Because the frequency of a keyword continuously increases even if the keyword was extracted as an "event" keyword once. Then, every time the frequency increases by the pre-defined count value, visual clustering and representative photo selection are re-performed.

In the experiment, we set the threshold as 30 and set the incremental count value as 10. Figure 16 and Figure 17 show examples of event photos detected by the real-time event photo detection system. The first example shows "snow" in Nagoya area at March 10th, 2014, and the second example shows "fireworks" in Tokyo area at July 26th, 2014.

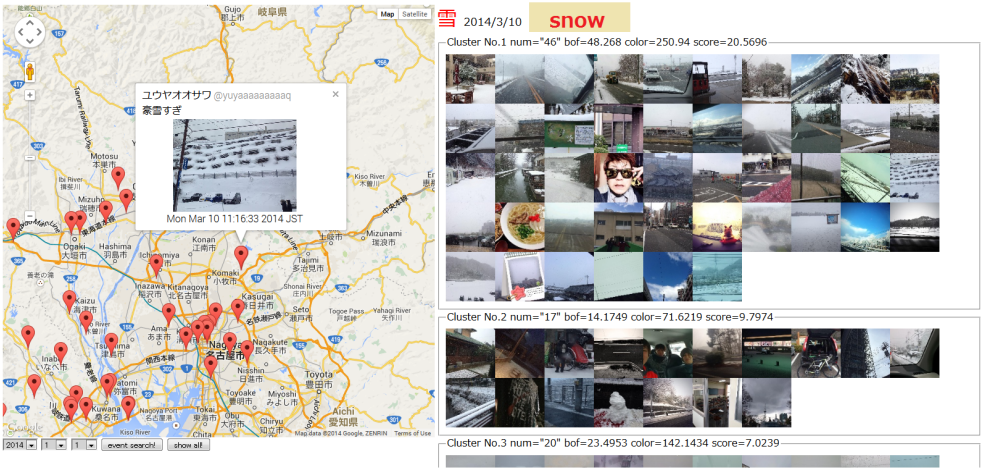

Figure 16: The event keyword, “snow”, detected by real-time event photo detection.

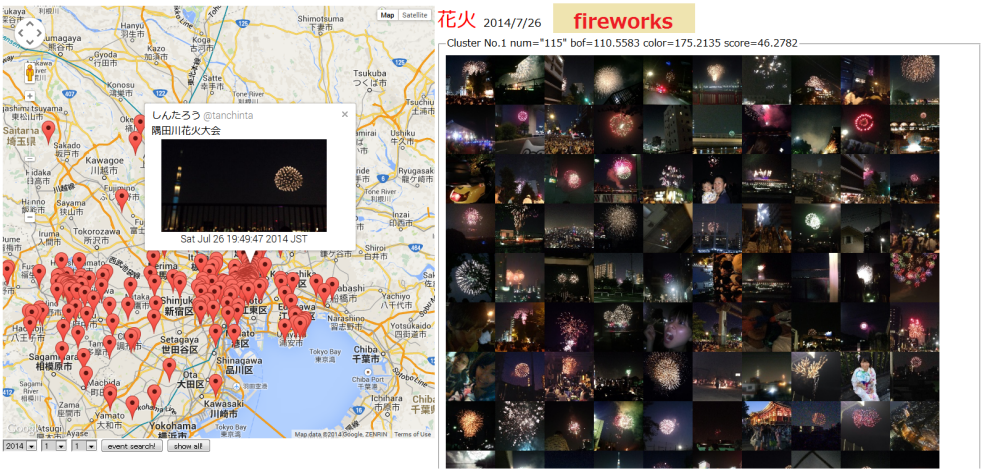

Figure 17: The event keyword, “fireworks”, detected by real-time event photo detection.

6. Conclusions

In this paper, we proposed a visual event mining system from the Twitter stream using visual information as well as textual and location information. The system enables us to discover and understand events visually, which is the novel contribution of this work. By integrating the proposed system with the Twitter Streaming API, it can be expanded into a real-time event photo detection system. To the best of our knowledge, this is the first system on visual event detection from the Twitter stream data.

In the experiments, we made visual event detection experiments on two large-scale datasets: the Japan geo-photo tweet dataset containing about three million tweets and the US geo-photo tweet dataset containing seventeen million tweets.

For future work, we plan to propose more sophisticated visual event mining methods which integrate visual, textual and location information more closely and more comprehensively. In the current system, the grid size and the term to extract events are fixed to one degree and one day, respectively. We will extend the system so that the grid size and the time unit for detecting events are adjusted automatically depending on events.

Currently we use only geo-photo tweets for visual event mining. To increase the number of detected events and event photos, we plan to use both geo-tweets without photos and photo tweets without geo-information as well as geo-photo tweets.

In addition, we plan to analyze the difference between Tweet photos and Flickr photos in terms of their characteristic.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24300036.

Figure 18: "Cherry blossoms" photo clusters.

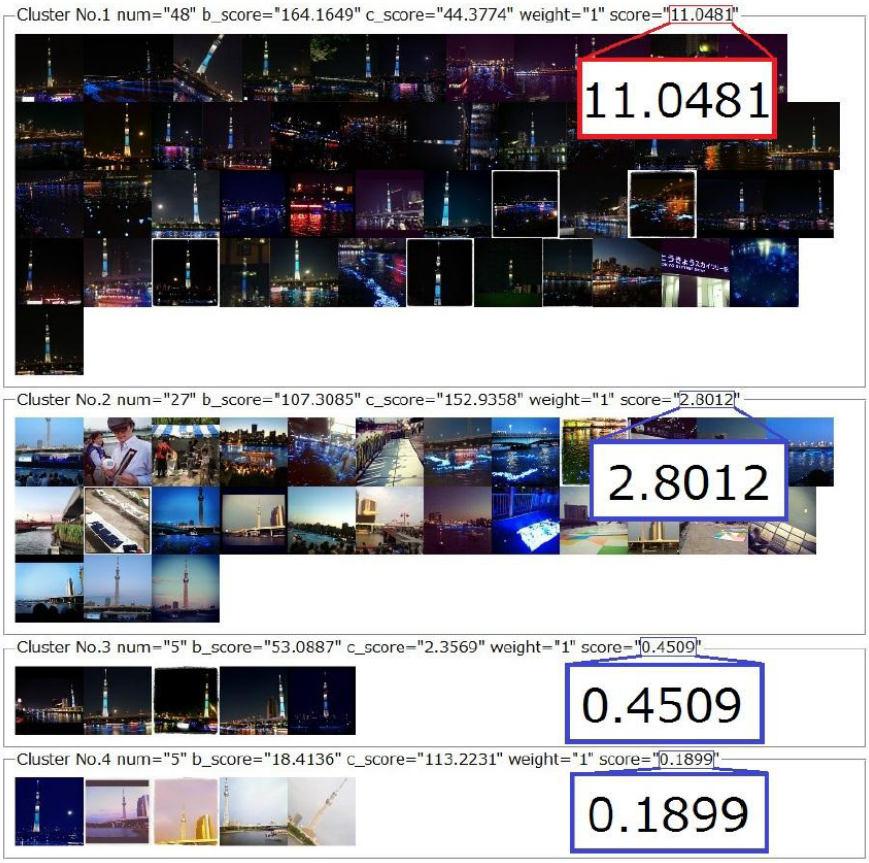

Figure 19: "Firefly" photo clusters.

Figure 20: "Giants" (San Francisco Giants World Champion Parade 2012) photo cluster.

Figure 21: "Halloween" photo clusters.