Visual Analytics for Investigative Analysis of Hoax Distress Calls using Social Media

Junghoon Chae, Purdue University
Jiawei Zhang, Purdue University
Sungahn Ko, UNIST, South Korea
Abish Malik, Purdue University
Heather Connell, Purdue University
David S. Ebert, Purdue University

***Abstract*—A hoax distress call is a serious concern for the U.S. Coast Guard. Hoax calls not only put the Coast Guard rescue personnel in potentially dangerous situations, but also waste valuable assets that should be used for real emergency situations. However, conventional approaches do not provide enough information for investigating hoax calls and callers. As social media has played a pervasive role in the way people communicate, such data opens new opportunities and solutions to a wide range of challenges. In this paper, we present social media visual analytics solutions for supporting the investigation for hoax distress calls. We not only provide a set of comprehensive keyword collections, but also resolve the lack of social media data for the investigation. Our framework allows investigators to identify suspicious Twitter users and provide a visual analytics environment designed to examine geo-tagged tweets and Instagram messages in the context of hoax distress calls.**

I. Introduction

Response to false calls or false alarms wastes considerable resources [1]. The U.S. Coast Guard (USCG) receives 18 intentional false distress calls and another 121 suspected hoax calls nationally per year [2]. Hoax calls not only put Coast Guard rescue personnel in dangerous situations, but also waste valuable search and rescue assets that should be used for real emergency situations. However, the investigation of the hoax distress calls is a challenge for the USCG, because it is still very hard to identify hoax callers. The USCG has a radio communication system designed for search and rescue operations called Rescue 21 [3]. The direction-finding capability of the system has been utilized for identification of suspected hoax calls. The system, however, still does not provide enough information for finding hoax callers.

Fortunately, social medial may provide a new tool to help with this problem. Social media now plays a pervasive role in the way people communicate. In particular, some people want to emphasize and flaunt their strong identification and even criminal exploits in online social networking [4]. Therefore, we hypothesize that certain hoax distress callers may show off their activities and talk to somebody through social networking sites. However, analyzing social media for such investigative tasks presents two challenges. First, it is hard to effectively collect relevant data as social media data is usually unstructured and the user-generated content has low precision. Second, available data for the investigative process is very limited—basically we can collect only a sample of the whole Twitter stream (usually 1%) using the free Twitter streaming API.

To address these challenges, we propose a social media visual analytic solution driven by the Rescue 21 system for supporting the investigation of hoax distress calls. In this work, we use social media data from two sources: Twitter and Instagram. For effective retrieval of relevant data, we build a set of comprehensive keyword collections. In order to enrich the elementary data, we extract new relevant Twitter users from the communication networks of a basic set of Twitter users to collect additional tweets. Based on the keyword collections and the enriched data we identify suspicious Twitter users and messages. We develop a new visual analytics approach designed to identify suspicious geo-tagged tweets and Instagram posts based on the spatial and temporal information of the hoax distress calls. We integrate the approach into our existing social media analytics framework [5], [6] to provide investigators with a more comprehensive analytics environment.

II. Related Work

There are many types of threats in modern societies from natural disasters (e.g., storms) to man-made threats (e.g., crimes). In order to prepare for such threats, it is important to maintain a high level of safety and security. However, preparing, preventing, and recovering from the threats have been a non-trivial challenge due to the ever-increasing size and complexity of data generated from multiple sectors. In this section, we describe previous social media analysis and visual analytics work for mainly public and maritime safety.

Social media has gained much attention from researchers as a new data source for improving threat management (e.g., [7]). Chae et al. [5], [8] propose visual analytics approaches of social media data for disaster management and evacuation planning. They focus on detecting and examining abnormal events and finding crowded places using location-based social networks. ScatterBlogs2 [6] trains classifiers by learning historical microblog data for visual real-time monitoring and analysis of microblog feeds. Our work adds contribution to the maritime operation domain by providing a visual tool that is equipped with social media data analysis methods designed to analyze false distress calls.

Crime data is important due to its relevance to the public's daily activities. Much research has been performed for visual analysis of crime data for crime prevention. A visual crime analysis approach focuses on projecting the crime data onto a map to enable geo-spatial crime analysis [9], [10]. Malik et al. [11] design a law enforcement toolkit that utilizes heatmap and clustering algorithms to visualize crime hotspots and other auxiliary data in multiple coordinate views. Their

approach has been extended to a mobile crime visual analytics environment [12]. Chen et al. [13] focus on categorizing crime types and concerns and they present a general framework that is equipped with data mining techniques. Their framework also allows users to explore networks of criminals based on the assumption that criminals develop networks for carrying out illegal activities. Xu et al. [14] concentrate on enhancing the safety in small communities (e.g., university campus) by designing a visual analytics system to improve the safety level through CCTV cameras.

Since the operation coverage of USCG spans across wide areas, maintaining maritime domain awareness is imperative to effectively prevent maritime threats and a waste of resources [15]–[17]. To enhance the awareness in the maritime domain, multiple approaches have been proposed. For example, Anthony et al. [18] propose an approach for detecting radiological and nuclear threat by using an imaging technology. Lavigne [19] designed interactive visualization applications for detecting and analyzing anomalies of vessels. Malik et al. showcase a visual analytics system for efficient resource allocation and risk assessment in USCG's search and rescue operations [20].

III. Background: Rescue 21

Rescue 21 [3] is the USCG's radio communication system for search and rescue operations. The Rescue 21 records transmission time and audio and provides lines of bearing (LOB) from radio towers to the source of any VHF radio calls that it receives. The provided LOB is used for locating boaters who are not able to specify their locations. Figure 3 shows examples of LOB (dashed lines) and the radio tower location on a map. Even though the direction-finding capability of the system provides a rough idea of the location of the distress call for search and rescue operations, it often insufficient to help investigators identify hoax callers. In order to mitigate this limitation, we design a visual analytics framework that is able to help investigators better identify hoax callers by utilizing social media data.

IV. Analysis Process

Figure 1 shows the overview of our analysis process. We first build keyword collections to exclude data unrelated to USCG and hoax distress calls (Section V-A). Then, we enrich the initial set of social media corpus using communication relationships, such as retweet and reply (Sec. V-B). Next, we identify suspicious users based on the temporal aspects of their online communication activities (Sec. VI). For geo-located messages, we provide visual means to explore and examine suspicious geo-located social media data based on the spatial and temporal clues driven by the USCG search and rescue system (Sec. VII).

V. Keywords Collection and Dataset Enrichment

Retrieval of highly relevant social media data through keyword filtering can potentially reduce the analysis space and enable effective analysis tasks. We use social media

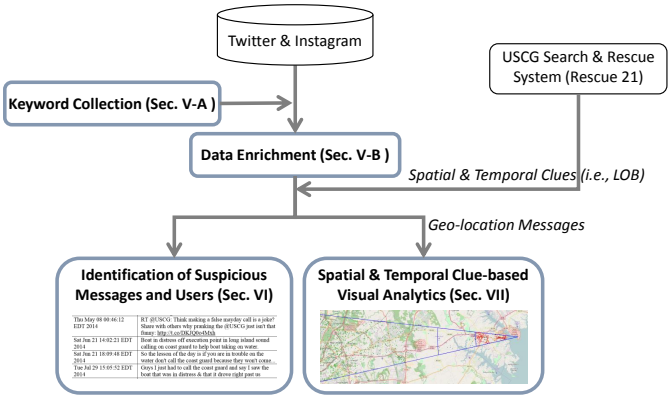

Fig. 1. Analysis Process Overview

users' communication activities to resolve the lack of data, as well as identify potentially suspicious users. The following sub-sections describe how we build comprehensive keyword collections and how we enrich the elementary data.

A. Collection of Hoax-Call-Related Keywords

To retrieve relevant social media data, we collaborate with Coast Guard personnel to build a keyword corpus related to USCG and hoax distress calls. The keywords are divided into 6 categories as shown in Table I. We then use WordNet [21] that provides words with semantic relationship to the initial keywords in order to get a more comprehensive keyword corpus. Furthermore, we use USCG related online articles, RSS feeds, and historical social media data to find related words and hashtags that appear frequently with these keywords.

The initial keyword list contains around 100 words. The final keyword corpus contains 650 keywords, which are used for keyword filtering of social media messages in our analysis process. A partial result of the final keyword collection is shown in Table I.

Category	Initial Keywords	Extended Keywords
Coast Guard	USCG, vessel, ship, boat, ferry, shore, coastline...	#cg, #coastguard, maritime, navy, fleet, lifeboat, craft, waterfront, barge ...
Hoax Call	prank call, distress call, mayday, crank call, hoax call, bogus call...	siren, save, help, call for help, prankster, alert, alarm...
Search and Rescue	helicopter, seek, search, rescue, flare, flashing, spark, signal...	#sos, copter, airliner, navigate, 911, warning, investigation, drift...
Accident	explosion, accident, sink, crash, damaged, burning, emergency, drown...	disaster, hazard, collision, shot, smack, hit, break, smoke, injured, flood...
Radio Channel	broadcast, radio, transmit, communicate...	cable, on the air, airwaves, wireless, phone, report...
Illegal Immigration	immigrant, immigration, panga, smuggling, drug...	foreigner, exodus, migration, crossing, resettlement, relocation, pirate...

TABLE I
A SNAPSHOT OF THE INITIAL KEYWORDS AND EXPANDED KEYWORDS.

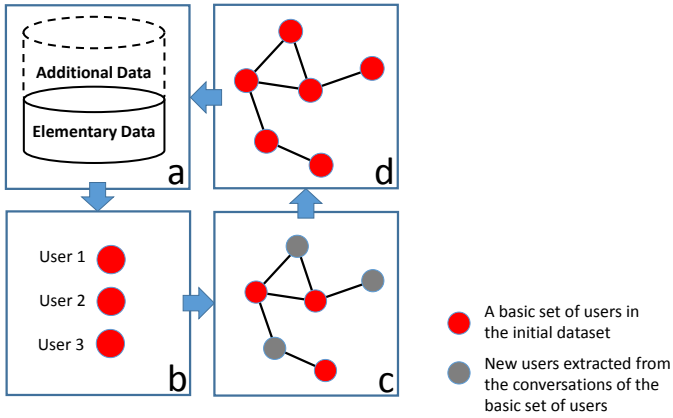

Fig. 2. Process of data enrichment through users' conversation networks

B. Enrichment of Social Media Data through User Networks

In this work, we utilize the free Twitter[1] and Instagram[2] APIs that provide a random sample of the whole dataset. To overcome the data insufficiency limitation, we propose a method using conversation networks of Twitter users. The enrichment process is illustrated in Figure 2. We prepare the elementary dataset—tweets with one of the words in the pre-defined keyword collections (Figure 2a). Next, we extract a basic set of users from the initial dataset (Figure 2b). Then, we build a larger user list with new users extracted from the basic users' communication activities (retweets and replies). Finally, we retrieve the messages posted by those newly added users using the Search API[3], and add up the new data into the existing dataset (Figure 2d).

The aforementioned approach is an iterative process. We can identify new users and add their messages into the existing data corpus repeatedly. However, as the number of iterations increase, the scale of the data grows dramatically, thereby introducing false positives. To resolve this, we extract only the new users whose conversations are relevant to our interests during the process. In other words, their messages should contain one of the words in the pre-defined keyword collections. The keyword filtering is able to effectively avoid increasing the amount of noisy data while retaining the data of interest.

VI. Identification of Suspicious Users

In some cases, the Rescue 21 system receives multiple hoax distress calls by one caller during different time frames. To this end, we propose a method that identifies suspicious users based on the temporal aspects of their online communication activities. Algorithm 1 illustrates the procedure of identifying suspicious users based on the timestamps of each hoax call. The intuition behind this approach is that users who post tweets within a time window when a hoax call was received are potentially more suspicious than others. So, we assign higher scores to such users, even though the volume of their messages is not significantly large. In contrast, we assign relatively low scores to the users who post tweets out of the time frames, even though they post a large number of tweets. Hence, this method is able to identify specific users of our interest based on their temporal activity patterns.

[1]https://dev.twitter.com/streaming/overview
[2]https://www.instagram.com/developer/
[3]https://dev.twitter.com/rest/public/search

Algorithm 1: Identification of Suspicious Users based on Multiple Timestamps

Input : A user list of size n: u_1, u_2, ..., u_n.
A timestamp list of size m: t_1, t_2, ..., t_m
Output: The array $score$ that contains the suspicious scores for each user.

```
// Initialize the scores for each user.
for each user u_i do
    score[u_i] ← 0
end
tolerance ← t // Initialize temporal tolerance.
// Loop the user list and calculate the suspicious scores.
for each user u_i do
    for each timestamp t_j do
        if u_i posted at least one message in the time range [t_j ± tolerance] then
            score[u_i] ← score[u_i] + 1
        end
    end
end
```

VII. Combination of Visual Analytics and a Search and Rescue System of USCG

As mentioned in Section III, the Rescue 21 system provides LOB information for each hoax distress call. The LOB information can be utilized in searching for relevant geo-tagged social media data. We have taken a visual analytics approach that provides investigators with scalable and interactive social media data analysis and visualization for the examination of geo-tagged suspicious Twitter and Instagram data using this LOB information. Given the LOB information including the radio tower location, the direction of the signal, and the time for each hoax call, our system provides an initial visual context of suspicious tweets and Instagram data. Based on the given spatial and temporal clues, our system displays a regionalized sector (blue) on a map as shown in Figure 3. For this case, multiple lines of bearings can be given for a call and the sector contains all the lines of bearings to avoid missing data. We split the sector into multiple regions based on its distance to the radio tower that received the call; 0 - 5 miles, 5 - 10 miles, and more than 10 miles. As shown in Figure 3, we show the corresponding location of the geo-tagged Twitter and Instagram messages generated within the time window of a given hoax call. The color of each dot represents the corresponding region. Red is within 5 miles to the tower, blue

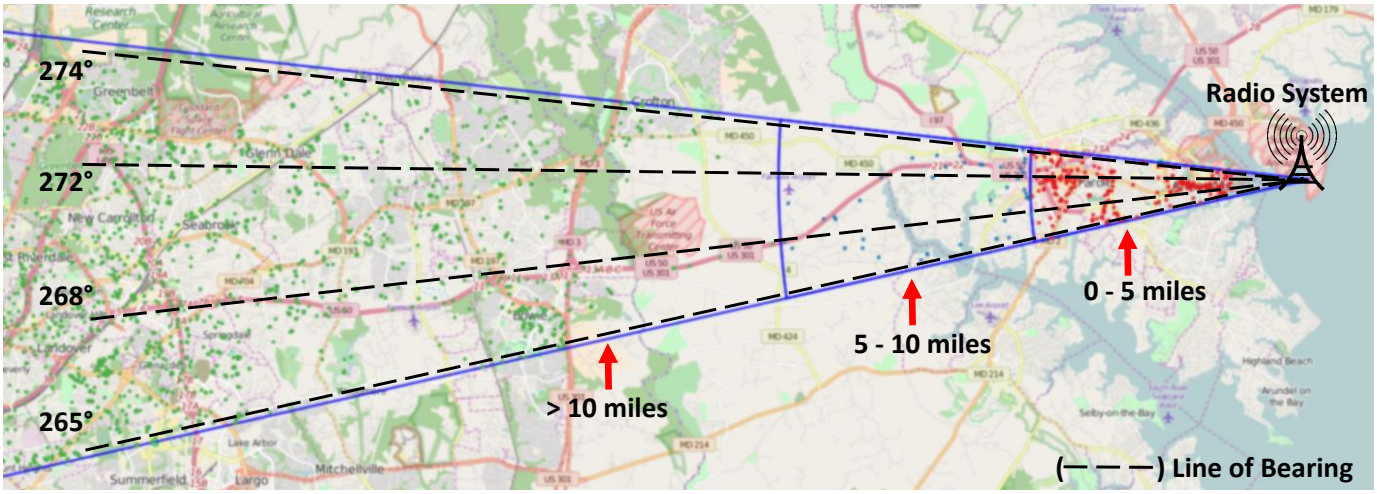

Fig. 3. Analysis of suspicious geo-tagged social media messages based on spatial and temporal clues driven by a Rescue 21 system—a Coast Guard radio communication system. The system filters messages based on the LOB and the time slot for the hoax call and provides the spatial and importance context of the suspicious messages to support investigation of the hoax call.

is between 5 and 10 miles, and green is more than 10 miles away from the tower. Here, we assume that data points that are spatially closer to the radio tower are more relevant to the hoax call. Analysts can select and view specific messages of interest by using the pre-defined keyword collections described in Section V-A and a spatial filter, thus enabling a more detailed analysis.

VIII. Case Study

In this section, we provide a case study that demonstrates how social media data can be used to facilitate the investigation of hoax distress call cases. Table II provides details regarding several distress calls received by the USCG that were determined to be a hoax. An analysis performed by the USCG analysts suggested that many of the calls were related and could be grouped into three main cases. For each of these three cases, we were provided with different LOBs information. For each case, we first extracted a set of relevant messages using the provided time windows and the geographical LOBs of interest and utilized the prebuilt keyword collections shown in Table I. We then enriched the initial datasets using communication relationships described in Section V-B. For example, our system identified around 100 users for Case 3 before the data enrichment process, and the number increased to around 500 after the data enrichment process. Next, we identified suspicious users using our frequency scoring method described in Section VI. Table III presents the distribution of users with respect to the frequency scores of the Twitter and Instagram datasets for Case 2 that had with 10 hoax calls. We found several users with a high score, and one of these Twitter users who had a frequency score of 6. In other words, the message posting times were close to the times of 6 out of the 10 hoax distress calls.

We then utilized our visual analytics approach described in Section VII to interactively examine the messages and the users. For example, for Case 2, multiple lines of bearings

Case	Time	Number of Calls
1	January, 2014	1
2	July and August, 2014	10
3	May, 2014	3

TABLE II
Information of the hoax distress call cases

	Frequency Scores						
Data Type	6	5	4	3	2	1	**total**
Twitter	1	4	4	15	60	269	353
Instagram			2	3	20	312	337

TABLE III
Distribution of the users with respect to the frequency scores of each dataset for case 2.

(black dash line) are given, and our system displays a regionalized sector (blue) on a map as shown in Figure 3. The sector contains all the lines of bearings. We split the sector based on its distance to the radio tower that received the calls, and each sector presents the corresponding location of Twitter and Instagram messages by different colors: red (0 - 5 miles), blue (5 - 10 miles), and green (more than 10 miles). Figure 4 shows our entire social media analytics system. The detailed analysis procedure of the components are described in our previous work [5], [6]. The system provides the USCG analysts with a list of relevant messages (See Figure 4 (B)) obtained using the pre-defined keyword collections (See Figure 4 (A)) and keywords extracted from the messages (See Figure 4 (C)) in order to investigate more detailed information about the messages and the users. Therefore, we conclude that the system provides analysts with the ability to utilize social media data to further corroborate the suspicious hoax calls.

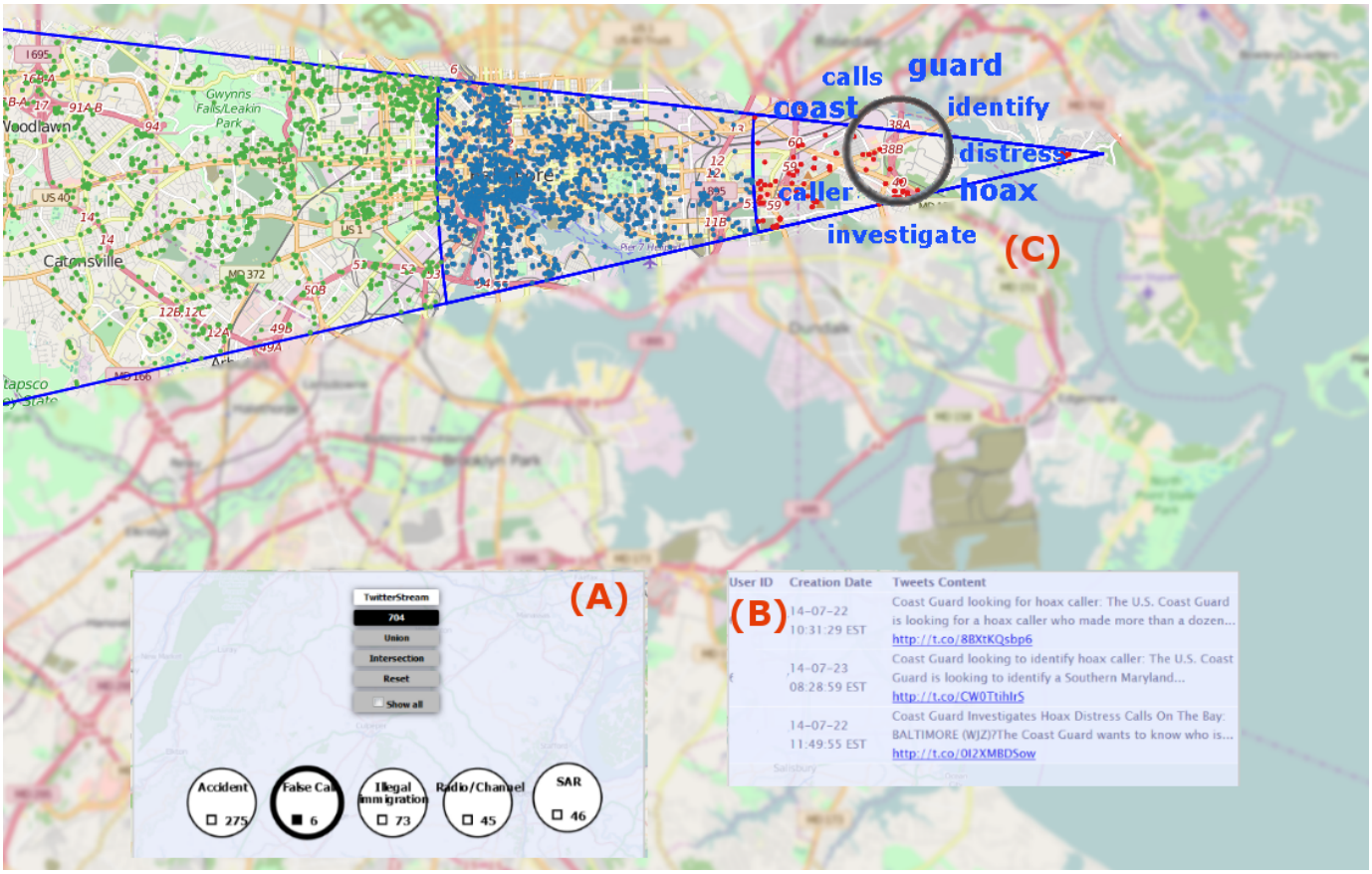

Fig. 4. The entire social media analytics system [5], [6]: Classifier view (A), Message table (B), and Content lens (C).

IX. Conclusion

We have proposed a visual social media analytics solution driven by the USCG search and rescue system (Rescue 21) for supporting the investigation for hoax distress calls. We described how our system helps investigators find suspicious hoax callers through the interactive visual analysis of social media data. In particular, we presented a visual analytics system to identify suspicious geo-tagged Twitter and Instagram messages based on the spatial and temporal clues from the USCG system. Utilizing this interactive visual environment for investigation offers great promise to combine the domain expert's knowledge with traditional data sources and emerging social media data sources.

Acknowledgment

This work was funded by the U.S. Department of Homeland Security VACCINE Center under Award Number 2009-ST-061-CI0003.

References

[1] E. A. Blackstone, A. J. Buck, and S. Hakim, "The economics of emergency response," *Policy Sciences*, vol. 40, no. 4, pp. 313–334, 2007.

[2] NavyTimes, "Coast guard says hoax distress calls a problem," Retrieved December 29, 2015, http://www.navytimes.com/story/military/coast-guard/2015/05/21/coast-guard-says-hoax-distress-calls-a-problem/27711439/, 2015.

[3] U. C. Guard, "Uscg: Rescue 21," Retrieved December 29, 2015, http://www.uscg.mil/acquisition/rescue21/default.asp, 2011.

[4] C. Morselli and D. Décary-Hétu, *Crime Facilitation Purposes of Social Networking Sites: A Review and Analysis of the" cyberbanging" Phenomenon*. Public Safety Canada., 2010.

[5] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *IEEE Symposium on Visual Analytics Science and Technology*, Oct., pp. 143–152.

[6] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl, "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2022–2031, 2013.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. ACM, 2010, pp. 851–860.

[8] J. Chae, D. Thom, Y. Jang, S. Kim, T. Ertl, and D. S. Ebert, "Public behavior response analysis in disaster events utilizing visual analytics of microblog data," *Computers & Graphics*, vol. 38, pp. 51–60, 2014.

[9] J. Eck, S. Chainey, J. Cameron, M. Leitner, and W. Ronald, "Mapping crime: Understanding hot spots," National Institute of Justice, Technical report, 2005.

[10] A. Luc, J. Cohen, D. Cook, W. Gorr, and G. Tita, "Spatial analyses of crime," *Criminal Justice*, vol. 4, no. 2, pp. 213–262, 2000.

[11] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert, "Visual analytics law enforcement toolkit," in *IEEE International Conference on Technologies for Homeland Security*, 2010, pp. 222–228.

[12] A. M. Razip, A. Malik, S. Afzal, M. Potrawski, R. Maciejewski, Y. Jang, N. Elmqvist, and D. S. Ebert, "A mobile visual analytics approach

for law enforcement situation awareness," in *IEEE Pacific Visualization Symposium (PacificVis)*, 2014, pp. 169–176.

[13] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: A general framework and some examples," *IEEE Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[14] H. Xu, J. Tay, A. Malik, S. Afzal, and D. Ebert, "Safety in view: A public safety visual analytics tool based on cctv camera angles of view," in *IEEE International Symposium on Technologies for Homeland Security*, 2015, pp. 1–6.

[15] V. Lavigne, D. Gouin, and M. Davenport, "Visual analytics for maritime domain awareness," in *IEEE International Conference on Technologies for Homeland Security*, 2011, pp. 49–54.

[16] M. Balci and R. Pegg, "Towards global maritime domain awareness-" recent developments and challenges"," in *International Conference on Information Fusion*, 2006, pp. 1–5.

[17] M. Glandrup, "Improving situation awareness in the maritime domain," in *Situation Awareness with Systems of Systems*. Springer, 2013, pp. 21–38.

[18] A. Hutcheson, B. Phlips, E. Wulf, L. Mitchell, W. Johnson, and B. Leas, "Maritime detection of radiological/nuclear threats with hybrid imaging system," in *IEEE International Conference on Technologies for Homeland Security*, 2013, pp. 360–363.

[19] V. Lavigne, "Interactive visualization applications for maritime anomaly detection and analysis," Defence Research Reports, Technical report, 2014.

[20] A. Malik, R. Maciejewski, B. Maule, and D. S. Ebert, "A visual analytics process for maritime resource allocation and risk assessment," in *IEEE Conference on Visual Analytics Science and Technology*, 2011, pp. 221–230.

[21] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.