Hurricane Damage Assessment with Multi-, Crowd-Sourced Image Data: A Case Study of Hurricane Irma in the City of Miami

Haiyan Hao
University of Florida
hhao@ufl.edu

Yan Wang
University of Florida
yanw@ufl.edu

ABSTRACT

The massive crowdsourced data generated on social networking platforms (e.g. Twitter and Flickr) provide free, real-time data for damage assessment (DA) even during catastrophes. Recent studies leveraging crowdsourced data for DA mainly focused on analyzing textual formats. Crowdsourced images can provide rich and objective information about damage conditions, however, are rarely researched for DA purposes. The highly-varied content and loosely-defined damage forms make it difficult to process and analyze the crowdsourced images. To address this problem, we propose a data-driven DA method based on multi-, crowd-sourced images, which includes five machine learning classifiers organized in a hierarchical structure. The method is validated with a case study investigating the damage condition of the City of Miami caused by Hurricane Irma. The outcome is then compared with a metric derived from NFIP insurance claims data. The proposed method offers a resource for rapid DA that supplements conventional DA methods.

Keywords

Computer Vision, Damage Assessment, Disaster Management, Insurance Claims, Social Networking Platforms.

INTRODUCTION

In 2017, three catastrophic hurricanes made landfall in the United States, which caused tremendous economic losses and human suffering. In fact, statistics seemingly suggest a tendency of increased frequency and more severe outcomes of natural disasters worldwide (Coronese et al., 2019; Smith, 2019). On the other hand, huge efforts have been made by human society to minimize the negative impacts of natural disasters. Disaster management is tasked with reducing disaster risk and relieving human suffering. DA plays a critical role in the response and recovery phases of disaster management. DA collects disaster magnitude and severity information such as the number of assets damaged and people injured. The collected information increases public situational awareness, assists disaster relief operations, and aids resources allocation of official agencies.

Conventionally, DA is conducted with the field survey by official agencies and/or relevant stakeholders (FEMA, 2016; IFRC, 2008). Field survey-based methods are labor-intensive and time-consuming. Some researchers have proposed DA methods based on satellite or Unmanned Aerial Vehicle (UAV) imagery. However, both satellite imagery and the deployment of UAVs are limited by harsh environmental conditions while many natural disasters happen accompanied by adverse weather. Consequently, the satellite and UAV imagery is not always available for DA after the disaster.

The recent explosive growth of crowdsourced data from social networking platforms (e.g. Twitter, Flickr, Facebook) provides new data sources for disaster management and DA (Granell & Ostermann, 2016). Compared to conventional data sources, crowdsourced data do not require the extra deployment of labor for data collection and are less susceptible to harsh environmental conditions. The real-time nature enables the data to be collected throughout the disaster while the analysis can also be conducted in a rapid and even real-time manner. Some machine learning-based methods have been developed to estimate the damage location and extent, such as topic modeling and sentiment analysis. However, those methods are mostly based on textual contents. Social media

text messages have been criticized for their low information quality (Agarwal & Yiliyasi, 2010), which limits their accuracy and application for DA. Crowdsourced images received much less attention from the research community due to the difficulty of processing and analyzing the unstructured crowdsourced data. However, the crowdsourced images may show the real-world scenarios comparable to field assessors' perception and provide rich and objective information about damage types and severities. A few recent studies started to analyze crowdsourced images with computer vision (CV) techniques (Li et al., 2018; Nguyen et al., 2017). Those studies developed and tested models with large annotated datasets but the raw-crawled images from webs are extremely noisy and unbalanced in terms of semantic contents. Thus their models still require validations over the raw-crawled crowdsourced images.

In this study, we develop a data-driven approach to assist DA using crowdsourced images that can be consistently collected during and after disasters. The method consists of five machine learning classifiers that are organized in a hierarchical structure. Each classifier is responsible for a single task such as filtering or classification, while they jointly serve as a robust filter and extract damage type and severity information from raw-crawled crowdsourced images. We tested our method to analyze the damage caused by Hurricane Irma in the city of Miami with images collected from three crowdsourced platforms, namely, Twitter, Flickr, and National Alliance for Public Safety GIS (NAPSG) Crowdsourcing Photo App (NAPSG Foundation, 2017). We compared the method outcomes with a damage severity metric derived from the National Flooding Insurance Program (NFIP) insurance claims data as well as evaluating the method predictions with annotated raw-crawled Twitter images. The results demonstrated the effectiveness of our proposed method in identifying locations and extent of disaster damage.

LITERATURE REVIEW

In practice, DA is conducted by official agencies and stakeholders such as humanitarian organizations and insurance companies. Conventionally, human assessors are sent to disaster sites and collect detailed damage information such as the location, number, type, and severity of affected buildings (FEMA, 2016). Other DA methods such as self-reporting, fly-over, geospatial analysis, and predictive modeling are also recommended based on requirements on different time and situations (FEMA, 2016). The field survey-based methods are labor-intensive and can take up to weeks and months depending on the disaster scale. Many researchers therefore proposed rapid DA methods based on high-resolution satellite or UAV imagery (Jordan, 2015; Novikov et al., 2018). The broad-view imagery provides an overview for disaster-affected areas while the high-resolution data enable the DA conducted to individual structures. However, high-resolution satellite imagery of affected areas is not always available in the short aftermath of a major disaster (Robinson et al., 2019). The deployment of UAVs should consider technical issues such as system reliability, power supply, and physical load. Both satellite imagery and UAV data can be affected by adverse atmospheric and weather conditions such as dense clouds, heavy rain, and strong wind, while many disasters were accompanied by those harsh environmental conditions. These limitations cause barriers to the accessibility and speed of using the imagery data.

Recently, researchers have explored the usage of crowdsourced data as an alternative data source for DA. Increasingly emerging data from social networking platforms (e.g. Twitter and Flickr) provide free, open and real-time data sources for DA. These data also reflect the perceptions of individuals towards the damaged built environment and can be acquired easily with API calling during disasters. Some methods have been developed to extract disaster-relevant information from crowdsourced data. However, most of them are based on text mining and textual data (e.g. the textual content of social media posts) only (Alam et al., 2018a). Existing text-based DA methods mainly relied on keyword-search (Deng, et al., 2016; Eilander et al., 2016), topic modeling (Resch et al., 2018; Wang & Taylor, 2018; Yao & Wang, 2019), and sentiment analysis (Kryvasheyeu et al., 2016; Shan, Zhao, Wei, & Liu, 2019). However, unstructured short text messages from social media such as tweets are often criticized for their limited lengths, colloquial expressions, low information quality, and severe subjectivity (Agarwal & Yiliyasi, 2010). Thus, it is not surprising that the above mentioned text-based DA studies all aggregate and report their results in coarser spatial level (e.g. state-, province-, county-, and ZCTA-level).

The literature review also finds a few studies leveraging crowdsourced images for DA. Some researchers used transfer learning and fined-tuned convolutional neural networks (CNNs) to predict damage severity level with crowdsourced images (Alam et al., 2018b; Nguyen et al., 2017). Transfer learning is an approach leveraging pre-trained models to solve new problems by retraining the model with datasets annotated for new classification tasks. Alam et al. (2018b) and Nguyen et al, (2017) employed crowdsourcing platforms and annotated more than one hundred thousand images for model development. Similarly, Li et al. (2018) fine-tuned a pre-trained CNN model to locate and quantify the damage in crowdsourced images. The authors re-trained the model with tens of thousands of annotated crowdsourced images. Mouzannar et al. (2018) proposed a multi-modal model for disaster damage type classification with crowdsourced texts and images. Instead of retraining the model, the author considered two pre-trained CNNs as feature generators and fused generated text and image features for classification. Experiments showed better performance of this multi-modal method compared to its single-modal

counterparts. In another study, Ahmad et al., (2019) also took pre-trained CNNs as feature generators for detecting passable roadways during flooding from crowdsourced images. The author considered both "object-level" and "scene-level" features. The object-level features are obtained with pre-trained models developed for object (e.g. pizza, bird, and soccer) classification tasks, whereas scene-level features are obtained with models pre-trained for scene (e.g. street, playground, and bedroom) related tasks. It was shown that models with two types of features fused achieved the best performance, while scene-level features performed better than object-level features when singly used.

Previous works leveraging crowdsourced images for DA mainly rely on transfer learning or similar approaches. Those approaches use single CV models that require large annotated datasets for model development and lack the validation over noisy and unbalanced raw-crawled crowdsourced data. To address this research gap, we propose a hierarchical machine-learning method for DA. Our proposed DA method divides the raw task, DA with crowdsourced images, into sub-tasks that follow a logical sequence and use five classifiers to implement those sub-tasks. The division allows each classifier to focus on a relatively simple task and to perform on images passing preceding classifications, which are less noisy regarding the image content. Thus the classifiers can be developed with fewer annotated images. In the other sense, the hierarchical structure works like a robust filter and filtered out irrelevant crowdsourced images effectively.

METHOD DESCRIPTION

The proposed multi-sourced DA method uses a four-step approach to identify informative images and extract damage information. The four steps are, i) identifying images showing perceived outdoor environments; ii) identifying images showing damage content, i.e. flood water or uprooted tree; iii) classifying the damage type (i.e. wind and flood); and iv) classifying the damage severity level (i.e. little to minor, mild, and severe). Five distinct machine learning classifiers were developed to implement the four steps and they were organized in a hierarchical structure (Figure 1). Figure 1 demonstrates the pipeline of the proposed DA method. The method starts with aggregating raw-crawled geotagged images from three crowdsourced platforms into census tracts (i.e. a geographic region with a population between 1,200 and 8,000 people, roughly representing neighborhoods) (U.S. Census Bureau, n.d.). The aggregated images then pass through five classifiers for filtering and classification. The outcomes include the damage type and maximum damage severity level exhibited in images for each census tract. Note that in the training process, we include the false positive images predicted by the preceding classifier in the training set of its succeeding classifier. In this way, the succeeding classifier can also remove some false positive images and avoid the accumulation of misclassification to some extent.

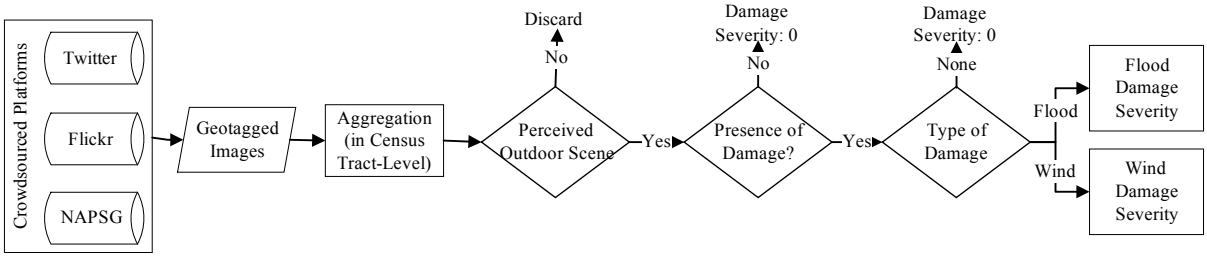


Figure 1. The Pipeline of Damage Assessment based Multi-, Crowd-Sourced Image Data

Two researchers cooperated for the image annotation. Communication is allowed to resolve any labelling ambiguities. The images are collected from Twitter during hurricane events i.e. Hurricane Florence, Hurricane Mathew and the YFCC100M dataset (Thomee et al., 2016). Table 1 summarizes the amount of annotated images and labels for distinct tasks.

Table 1. Images Used in Image Classifier Development

Classification Task	Labels	# of Images
Perceived outdoor environment classifier	Total	1,230
	Positive (images showing perceived outdoor environment)	585
	Negative (other images including maps, selfies, indoor scenes, screenshots of texts and etc.)	645
Damage presence classifier	Total	1,089
	Positive (images showing damage content):	577
	Negative (images showing normal environment views)	512
Damage type classifier	Total	655
	Wind damage	273
	Flood damage	269
	Wind and flood damage	35
	None	78
Damage severity level classifier	Wind damage	468
	Little to none	160
	Minor	121
	Severe	187
	Flood damage	483
	Little to none:	179
	Minor:	139
	Severe:	165

Feature Extraction with Deep Learning Models

Feature extraction represents images with feature vectors, while the following machine learning models are developed with extracted feature vectors. In this study, we considered both object-level and scene-level features. We used three deep learning models for feature extraction. Specifically, (1) an Inception-v3 CNN trained on the ILSVRC dataset (Russakovsky et al., 2015; Szegedy et al., 2016); (2) a ResNet18 CNN trained on the Places365 dataset (Zhou et al. 2016); and (3) an Autoencoder (AE) trained on the ADE20K dataset (Zhou, Zhao, et al., 2016). The ILSVRC dataset is for object recognition tasks whereas the Places365 and ADE20K are for scene classification and segmentation tasks. Thus, the Inception-v3 CNN extracts the object-level features while the ResNet18 CNN and AE are responsible for the scene-level feature extraction in this study. The features are extracted as the output of the penultimate layer of two CNNs and from the bottleneck layer of the AE. The Inception v3 CNN and AE return features of 2,048 dimensions and the ReNet18 CNN yields features of 512 dimensions. We used these three feature generators to extract feature vectors for all annotated images. These features were only extracted once and repeated for use when developing different image models in the pipeline (Figure 1).

Classification of Images Showing Perceived Outdoor Environment

The raw-crawled crowdsourced images can vary enormously regarding the image contents. In this study, we restricted the "informative" images as the ones that exhibited perceived outdoor environments (Figure 2b). Other images such as cartoons, maps, posters (Figure 2a) were not considered even their content may be related to the disaster. The first image classifier sorts out those "informative" images. Images showing the indoor environment may also reveal the damage condition for an individual building. They vary a lot with the outdoor environment regarding the exhibited objects and backgrounds. Also, we did not collect enough images to train a separate classifier for it. Thus, we excluded them from analyses. The annotated images are divided into 80% training set and 20% testing set, which is stratified according to classes (the same setting applied to the development of the remaining classifiers). We tested both scene-level and object-level features extracted with the method described in the feature extraction section, and experimented with logistic regression (LR) models and support vector machine (SVM) with a linear kernel. Default parameters were used for model training except the class weights

are balanced. Table 2 shows the classifier performance in the test set for different combinations. Generally, all classifiers achieved high classification accuracies of 90%. The linear SVM with the ResNet18 scene-level features performed slightly better than other combinations.

(a). Non-Informative Images (b). Informative Images

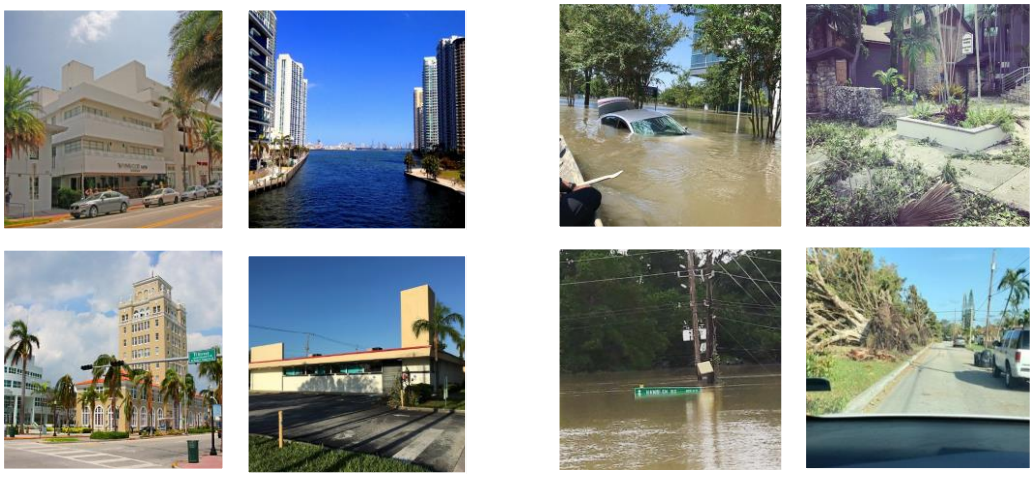
Figure 2. Examples of Non-Informative (a) and Informative Images (b)

Table 2. Classifier Performance for Classification of Perceived Outdoor Environment

Machine Learning Models + Features	Classification Accuracy
SVM (linear) + Inception v-3 object-level features	91.46%
SVM (linear) + ResNet18 scene-level features	**94.31%**
SVM (linear) + AE scene-level feature	92.28%
LR + Inception v-3 object-level features	91.46%
LR + ResNet18 scene-level features	90.00%
LR + AE scene-level feature	91.87%

Classification of Images Showing Damage Content

The next image classifier finds out images showing damage content. Figure 3 shows example images with and without damage contents. Similarly, we trained and tested models with three extracted features and experimented with both the LR model and linear SVM. According to Table 3, the combination of the LR model and ResNet18 scene-level features achieved the highest classification rate of 88.53% among all combinations.



(a). Images Shown Non-Damage Scene (b). Images Shown Damage Scene

Figure 3. Examples of Images Showing Non-Damage (a) and Damage Scene (b)

Table 3. Classifier Performance for Classification of Damage Presence

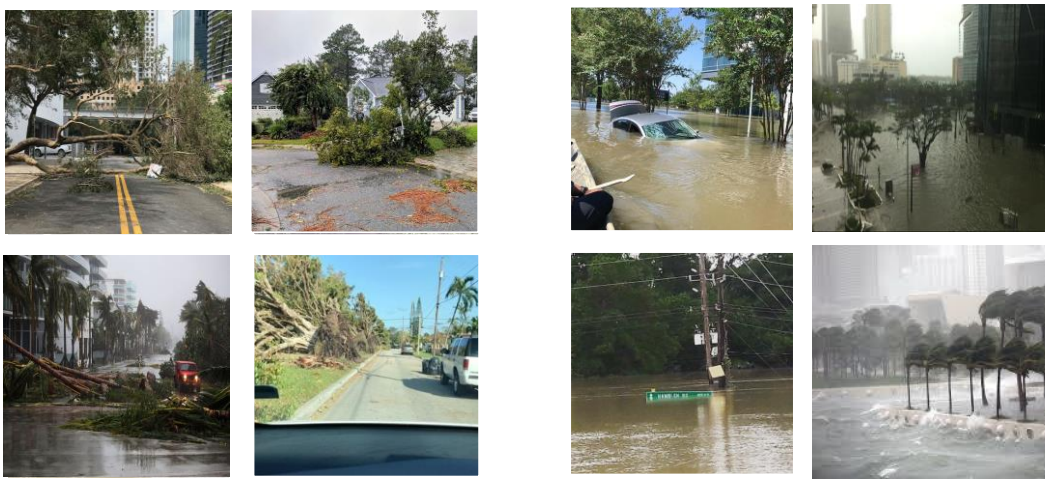
Machine Learning Models + Features	Classification Accuracy
SVM (linear) + Inception v-3 object-level features	83.41%
SVM (linear) + ResNet18 scene-level feature	87.55%
SVM (linear) + AE scene-level feature	75.12%
LR + Inception v-3 object-level feature	84.79%
LR + ResNet18 scene-level feature	**88.53%**
LR + AE scene-level feature	75.58%

Classification of Damage Types

When examining crowdsourced images showing evident disaster damage, we found highly disparate contents for different damage types. For example, an image showing flood damage usually contains the water body while the wind damage is generally represented with uprooted trees, roofs with missing tiles Figure 4. Thus we added a classifier to detail the damage type.

The classification of damage type is a multi-label question as an image can include either wind damage or flood damage or both. We used an artificial neural network (ANN) model for this classification task as ANN considers the possible correlations between labels. We annotated 655 images for the classifier development. Each image is assigned with a multi-label showing whether there is evidence of wind and/or flood damage. We tested three types of features as well as their concatenations. Table 4 presents the classifier performance.

The neural network is learned with backward propagating the loss during the training process, while the loss is determined with the model predictions and annotations. For a simple illustration, the correct prediction leads to a loss of 0 while the wrong prediction leads to a loss of 1. In this study, we considered the correct prediction of a single damage type contributing to the loss with a factor of 0.5 in the training process. We found this measure could maximize the use of collected data and increase the prediction accuracy by 5 to 10 percent. Note that the accuracies presented in Table 4 are determined as the percentage of accurate predictions for **both** damage types.



(a). Images Shown Wind Damage (b). Images Shown Flooding Damage

Figure 4. Examples of Images Showing Wind Damage (a) and Flood Damage (b)

Table 4. Classifier Performance for Classification of Damage Type

Features	Classification Accuracy
Inception v-3 object-level feature	80.30%
ResNet18 scene-level feature	82.01%
AE scene-level feature	69.70%
Inception v-3 object-level features + AE scene-level feature	81.06%
Inception v-3 object-level features + ResNet18 scene-level feature	**84.84%**
ResNet18 scene-level features AE scene-level features	41.67%
Inception v-3 object-level features + ResNet18 scene-level features + AE scene-level features	79.54%

Classification of Damage Severity Levels

We used two machine learning models to determine the severity level of wind and flood damage respectively. We assigned each training/testing image with one of the three severity levels (see Figure 5 for examples):

- Little to None: images show adverse weather conditions and/or little damage that does not cause economic loss or impact human activities;
- Minor: images show minor damage that results in certain economic loss or partially affect human activities; and
- Severe: images show severe damage that is accompanied by extreme environmental conditions, significant economic loss or severely impact human activities.

In addition to the LR and SVM models, we also tested the ordinal LR model considering the ordered nature of damage severity level. The classifier performances of different combinations are shown in Table 5.

Figure 5. Examples of Wind Damage and Flood Damage of Different Severities

Table 5. Classifier Performance of Damage Severity Classification

Machine Learning Models + Features	Flood	Wind
LR (Multinomial) + AE scene-level feature	68.09%	58.12%
LR (Ordinal) + AE scene-level feature	58.51%	52.14%
SVM (RBF) + AE scene-level feature	64.89%	54.70%
SVM (Linear) + AE scene-level feature	68.09%	58.97%
LR (Multinomial) + ResNet18 scene-level feature	71.28%	58.97%
LR (Ordinal) + ResNet18 scene-level feature	62.77%	61.54%
SVM (RBF) + ResNet18 scene-level feature	**73.40%**	**71.79%**
SVM (Linear) + ResNet18 scene-level feature	70.21%	63.24%
LR, Multinomial + Inception v-3 object-level feature	68.09%	68.38%
LR, Ordinal + Inception v-3 object-level feature	69.15%	64.96%
SVM, RBF + Inception v-3 object-level feature	60.64%	59.83%
SVM, Linear + Inception v-3 object-level feature	69.15%	64.96%

CASE STUDY AND METHOD VALIDATION

Case Description

We tested the proposed method with a case study investigating the damage condition of the City of Miami during Hurricane Irma. Hurricane Irma was a category 5 hurricane that made two landfalls in Florida with a category 3 and 4 intensities on September 10 respectively. It caused tremendous impact to the whole Florida state including the city of Miami. During the affected period, Miami experienced estimated wind gusts of 63-73kt and a sustained wind of 45-55kt (Cangialosi et al., 2018). The combined effect of storm surge and tide produced an inundation level of 3 to 5 ft. along Biscayne Bay shoreline in Miami. The heavy rainfall and urban runoff caused significant flooding in downtown areas (Cangialosi et al., 2018).

We collected geotagged images from three crowdsourced platforms in the city boundary of Miami and during the affected period of Hurricane Irma. The output is compared with the Average Asset Value Loss Ratio (AAVLR), a metric derived from NFIP insurance claims data, for validation.

Data Collection

Images were collected from three crowdsourced platforms, namely, Twitter, Flickr, and NAPSG Crowdsourcing Photo App (NAPSG Foundation, 2017). The former two are well-known social media platforms that have already received intensive attention from researchers. The NAPSG Foundation is a non-profit organization dedicated to solving emergency management and public safety problems with technology and data. NAPSG volunteers collected and mapped online photos showing disaster impacts to indicated locations on the NAPSG Crowdsourcing Photo App during hurricane events. We collected the Twitter images from the links carried by 3,929 raw-crawled tweets that are posted in Miami during Hurricane Irma. One tweet may include none or more than one images. Both Twitter and Flick images are collected through API calling and NAPSG crowdsourced images are downloaded from its online web portal. We totally collected 3,593 images in the city boundary of Miami during the affected period of Hurricane Irma (Table 6).

Table 6. Crowdsourced Images Collected for the City of Miami during Hurricane Irma

Source	# of images collected	Temporal and spatial span
Twitter	2,883	09/01/2017 – 09/13/2017, Miami city
Flickr	659	09/01/2017 – 09/30/2017 locations within 15 miles from the center of Miami
NAPSG	51	09/10/2017 – 09/18/2017, Miami city

Application and Results

The geotagged images were aggregated in census tracts for analysis. The collected crowdsourced images are distributed in 112 census tracts in the region of Miami. For each census tract, the combined set of crowdsourced images are input to the pipeline shown in Figure 1. For the five classifiers in the pipeline, we generally used the ones yielding the highest classification accuracies in the testing set (marked as the bold underlined ones) in Table 2-5. Figure 6 presents the spatial distribution of wind and flood damage severities obtained with our method. The green patches in Figure 6 suggest no evident wind or flood damage found from crowdsourced images collected for that census tract.

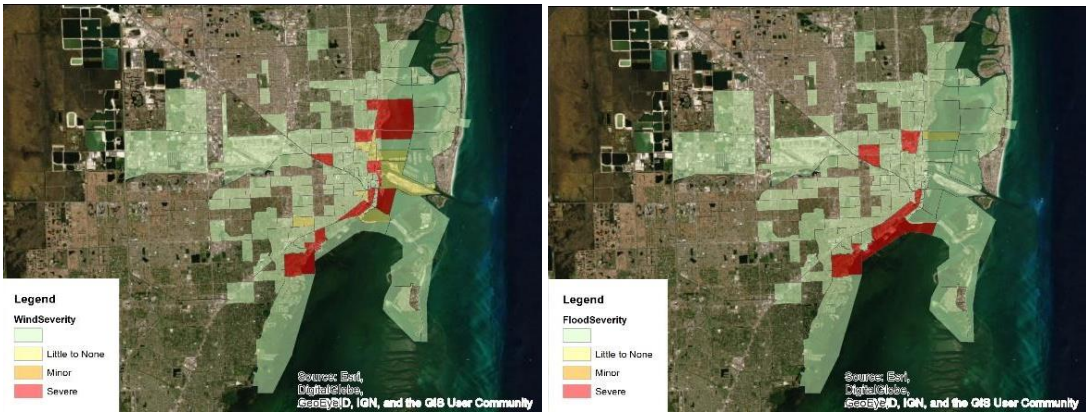


Figure 6. Wind and Flood Damage Severity in Census Tracts

The wind and flood damage shown in Figure 6 is generally distributed along the Biscayne Bay shoreline, which conforms to the disaster report (Cangialosi et al., 2018). With respect to the disaster severity level, it appears that most areas with damage identified show severe damage. This may be because we use the maximum severity level to represent the damage severity experienced by that census tract. Another reasonable explanation is that social media users are more likely to post images shown severe situations.

Evaluation of Classifier Performance

To evaluate the performance of the proposed pipeline on raw-crawled crowdsourced images, we annotate the 2,883 images collected from Twitter and inspect how they are classified by the five image classifiers (Figure 1). Figure 7 shows the process. When applied to the noisy and unbalanced raw-crawled Twitter data, the proposed method successfully identified 26 and 28 images showing flood and wind damage respectively, but also missed 2 images showing flood damage and 13 images showing wind damage. In addition, it also wrongly predicted 10 and 11 images that show no damage as exhibiting damage contents. In Table 7, we present comprehensive performance metrics of the proposed pipeline on identifying wind and flood damage images. Based on Table 7, the ordered structure is especially effective in filtering out irrelevant images or reduce the Type I error. Flood damage is easier to be identified compared to wind damage with associated recalls of 92.9% and 68.3%. The method shows close precision for the two damage types. Examples of the false positive images for flood and wind damage include scenic sea surfaces and forests, which are common scenes for Miami.

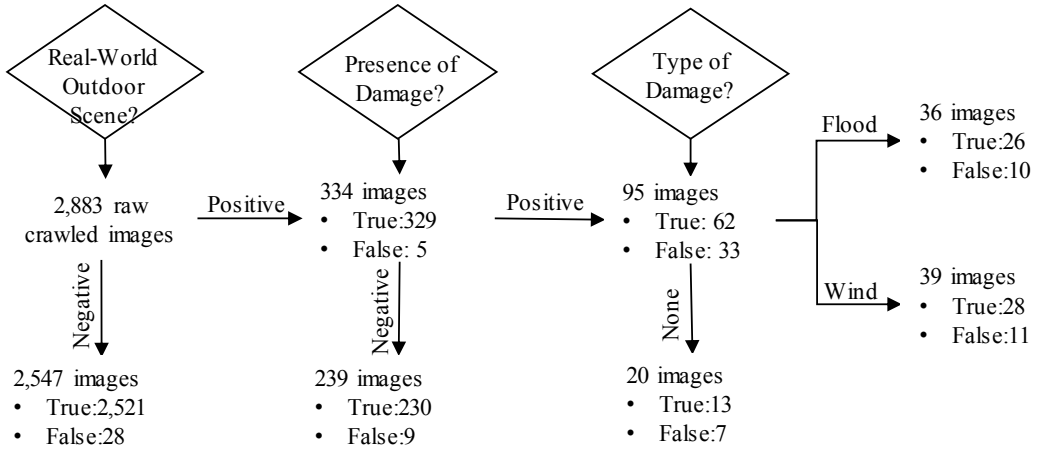


Figure 7. The Filtering and Classification of Raw-Crawled Twitter Images

Table 7. Performance Statistics of the Method on Identifying Wind and Flood Damage

Damage Type	Precision	Recall	Specificity	Accuracy
Wind damage	71.8%	68.3%	99.6%	99.2%
Flood damage	72.2%	92.9%	99.6%	99.6%

Validation with NFIP insurance claims data

We further assessed the validity of the proposed method by comparing its outcome with a disaster severity metric derived from NFIP insurance claims data. NFIP is a program aimed to provide affordable federal flooding insurance to the public. FEMA released historical NFIP claims data in July 2019, which includes more than two million claims from 1978 to the present. The raw data are redacted in the census tract level to protect the privacy of policyholders (FEMA, 2019). We used the NFIP claims from September 1, 2017 to December 31, 2017 in the 112 census tracts with crowdsourced data collected, which identified 710 insurance claims.

The individual insurance claim data does not represent the damage severity level directly because households and asset values vary among census tracts, i.e. more claims and higher insurance payouts may be attributed to less damaged areas with high-value properties. In order to diminish the impact of the asset number and values for each census tract, we collected the asset value and counts data from the Florida Department of Revenue (FDOR). FDOR maintains the asset value data as assessed by property appraisers and updates them annually. These data are published with GIS layers that are mapped in the parcel level. We did not find the historical data in 2017 and used the 2019's data instead in this study. The asset value and counts were also aggregated in the census tract-level. The metric, AAVLR, is determined as the ratio of the average insurance payout and the average asset value for that census tract.

$$AAVLR = \frac{\text{Total Insurance Payouts} \times \text{Number of Assets}}{\text{Number of Insurance Claims} \times \text{Total Asset Value}}$$

In fact, we tried different metrics derived from insurance claims and asset value data. The AAVLR proves its plausibility as it conforms to the geographic characteristics of Miami, that higher severity level is found along the Biscayne Bay shoreline and Miami River (as Figure 7 shown). We do not show other metrics here due to the word limit.

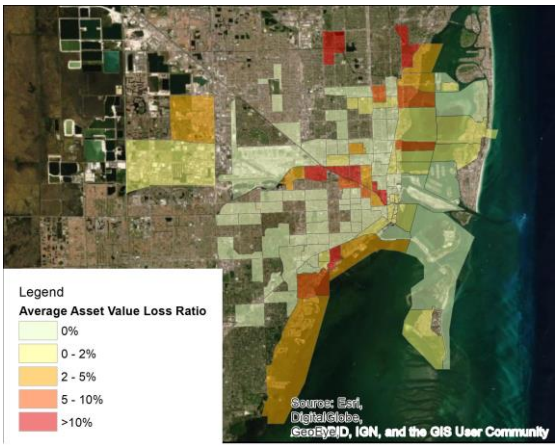


Figure 8. Spatial Distribution of AAVLR

We binned the AAVLR into five ordinal groups: 0%, 0-2%, 2-5%, 5-10%, and >10% (Figure 7), and then compared the estimated wind and flood damage severity (shown in Figure 6) with AAVLR using three correlation tests: the Pearson, Spearman, and Kendall's tau correlation tests. Table 7 shows the correlation results. The Pearson correlation test is suitable for numeric and dichotomous variables such as the presence of wind and flood damage, while the Spearman and Kendall's tau correlation analysis works better for ordinal variables such as wind and flood damage severity. According to Table 7, our estimated disaster damage, including the wind and flood damage severities and the presence of flood damage, correlates with the disaster severity metric (AAVLR) significantly (p-value < 0.05), though the obtained correlations are around 0.25. The statistical analysis results suggest that these correlations between image-derived damage level and the ratio of average issuance are significant, but more impact factors should be considered in the future to decide the damage level, such as alternative data for census tracts without damage images (see Figure 6 and Figure 7).

Table 8. Correlation Analysis of Damage Severity and Mined Damage Information

Damage Information	Pearson Correlation	Spearman Correlation	Kendal's tau Correlation
Presence of Wind Damage	0.1539 (0.1052)	0.1477 (0.1201)	0.1389 (0.1196)
Wind Damage Severity	**0.2438 (0.0096)**	0.1655 (0.0812)	0.1532 (0.0805)
Presence of Flood Damage	**0.2598 (0.0057)**	**0.2596 (0.0057)**	**0.2441 (0.0062)**
Flood Damage Severity	**0.2286 (0.0153)**	**0.2575 (0.0061)**	**0.2404 (0.0069)**

* Numbers in brackets are the p-value that suggests the statistical significance.

DISCUSSION AND CONCLUSION

In this study, we proposed a method for DA (damage assessment) with multi-, crowd-sourced images. The method uses five machine learning classifiers organized in a hierarchical structure, which takes the raw-crawled crowdsourced images and outputs as the damage types and severity levels. We applied this method on a case study investigating damage conditions of Miami during Hurricane Irma. The performance evaluation shows that our method is effective in identifying images showing wind and flood damages. When compared the method outcome with flood insurance data, the significant correlation results demonstrate the usefulness of the method in identifying damage at census tract-level.

Previous studies found difficulties in leveraging crowdsourced images for DA due to loosely-defined forms of damage, unstructured raw-crawled image data, lacking large annotated datasets for model development and so forth. Our method addressed these challenges with the proposed hierarchical-structure and multiple-classifier design. The hierarchical structure filters irrelevant and less informative images at the beginning. As succeeding classifiers work on the less noisy images that pass preceding classification, they can better focus on learning the difference between positive and negative samples with defined semantic contents. Thus, classifiers may be constructed with less annotated training and testing samples while still achieving satisfactory accuracies. In this study, we used around 1,800 annotated images. The hierarchical structure also allows this method to be applied to the unbalanced and noisy crowdsourced images directly without human sorting. Note that only a very small portion (71 out of 2,883, or 2.5%) of Twitter images are related to disaster damage.

We will improve this preliminary work in the following aspects in the future study. First, the method performance can be improved with more annotated images when data become available, though retrieving relevant data is very challenging. Second, the case study uses about 3,500 images from three platforms, but only a small portion of images are related to wind or flood damage. The method may yield better performance with more crowdsourced images collected for the investigated case. Our future work will focus on larger-scale geographic regions with more data sources to test the method. Third, we will implement the data-driven method within a real-time setting for processing streaming multi-modal data.

In summary, the proposed method supplements conventional data acquisition methods for rapid DA such as satellite and UAV imagery, especially when the equipment/data of conventional methods fail to be deployed or accessed. Our method performs satisfactorily in Hurricane Irma and can also be adapted for DA tasks in other disaster scenarios with other crowdsourced visual data. The incremental growth of user-generated data and advances of data analytical methods provide unique opportunities for the crisis informatics research. In light of these opportunities, we present a data-driven approach for studies on dealing with the noisy, unstructured, multi- and crowd-sourced image data.

REFERENCES

Agarwal, N., & Yiliyasi, Y. (2010). Information quality challenges in social media. In *Proceedings of the 2010 International Conference on Information Quality,* ICIQ 2010. Little Rock, Arkansas, 2010.

Ahmad, K., Pogorelov, K., Riegler, M., Ostroukhova, O., Halvorsen, P., Conci, N., & Dahyot, R. (2019). Automatic detection of passable roads after floods in remote sensed and social media data. *Signal Processing: Image Communication*, *74*, 110–118.

Alam, F., Ofli, F., & Imran, M. (2018a). CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Twelfth International AAAI Conference on Web and Social Media,* 465-473.

Alam, F., Ofli, F., & Imran, M. (2018b). Processing Social Media Images by Combining Human and Machine Computing during Crises. *International Journal of Human-Computer Interaction*, *34*(4), 311–327.

Cangialosi, J. P., Latto, A. S., & Berg, R. (2018). *Hurriane Irma IRMA*. Retrieved from https://www.nhc.noaa.gov/data/tcr/AL112017_Irma.pdf.

Coronese, M., Lamperti, F., Keller, K., Chiaromonte, F., & Roventini, A. (2019). Evidence for sharp increase in the economic damages of extreme natural disasters. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21450–21455.

Deng, Q., Liu, Y., Zhang, H., Deng, X., & Ma, Y. (2016). A new crowdsourcing model to assess disaster using microblog data in typhoon Haiyan. *Natural Hazards*, *84*(2), 1241–1256.

Eilander, D., Trambauer, P., Wagemaker, J., & Van Loenen, A. (2016). Harvesting Social Media for Generation of Near Real-time Flood Maps. In *Procedia Engineering* (Vol. 154, pp. 176–183). Elsevier Ltd.

FEMA. (2019). FIMA NFIP Redacted Claims Data Set. Retrieved November 24, 2019, from https://www.fema.gov/.

FEMA. (2016). *Damage Assessment Operations Manual*. Retrieved from https://www.fema.gov/.

Granell, C., & Ostermann, F. O. (2016). Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Computers, Environment and Urban Systems*, *59*, 231–243.

IFRC. (2008). *Guidelines for Assessment in Emergencies*. 1. Retrieved from https://www.icrc.org/.

Jordan, B. R. (2015). A bird's-eye view of geology: The use of micro drones/UAVs in geologic fieldwork and education. *GSA Today*, 50–52. https://doi.org/10.1130/gsatg232gw.1.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3).

Li, X., Caragea, D., Zhang, H., & Imran, M. (2018). Localizing and quantifying damage in social media images. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018* (pp. 194–201). Institute of Electrical and Electronics Engineers Inc.

Mouzannar, H., Rizk, Y., & Awad, M. (2018). Damage Identification in Social Media Posts using Multimodal Deep Learning. In *15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Rochester, NY, May, 529–543. ISCRAM.

NAPSG Foundation. (2017). Hurricane Irma Photo Map. Retrieved November 26, 2019, from https://www.arcgis.com/.

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017* (pp. 569–576). Association for Computing Machinery, Inc.

Novikov, G., Trekin, A., Potapov, G., Ignatiev, V., & Burnaev, E. (2018). Satellite imagery analysis for operational damage assessment in emergency situations. In *Lecture Notes in Business Information Processing* (Vol. 320, pp. 347–358). Springer Verlag.

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362–376.

Robinson, T. R., Rosser, N., & Walters, R. J. (2019). The Spatial and Temporal Influence of Cloud Cover on Satellite-Based Emergency Mapping of Earthquake Disasters. *Scientific Reports*, *9*(1), 1-9.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Shan, S., Zhao, F., Wei, Y., & Liu, M. (2019). Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—A case study of Weibo (Chinese Twitter). *Safety Science*, *115*, 393–413.

Smith, A. (2019). 2018's Billion Dollar Disasters in Context | NOAA Climate.gov. Retrieved from https://www.climate.gov/.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2016-December, pp. 2818–2826). IEEE Computer Society.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., ... & Li, L. J. (2016). YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2), 64-73.

U.S. Census Bureau. (n.d.). Glossary-Census Tract Definition. Retrieved March 17, 2020, from https://www.census.gov/glossary/

Wang, Y., & Taylor, J. E. (2018). Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake. *Natural Hazards*, *92*(2), 907–925.

Yao, F., & Wang, Y. (2019). Tracking urban geo-topics based on dynamic topic model. *Computers, Environment and Urban Systems*, *79*, 101419.

Zhou, B., Lapedriza, A., Torralba, A., & Oliva, A. (2017). Places: An Image Database for Deep Scene Understanding. *Journal of Vision*, 17(10), 296.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, *127*(3), 302–321.