

Conducting disaster damage assessments with Spatial Video, experts, and citizens



Evan Lue^{a,*}, John P. Wilson^a, Andrew Curtis^b

^a Spatial Sciences Institute, University of Southern California, United States

^b Department of Geography, Kent State University, OH, United States

A B S T R A C T

Keywords:

Disaster damage assessment
Spatial Video
Expert and inexperienced assessors

Damage assessment (DA) is an important part of the disaster recovery process, as it helps organizations like the American Red Cross provide assistance and relief to those most affected. Through it, homes are characterized by the damage they have received using five classes ranging from “Unaffected” to “Destroyed”. Rapid acquisition of damage assessment data may be achieved through crowdsourcing platforms where volunteers review images of affected structures. Further, these images can be geographically enriched through Spatial Video technology, whereby a car with a GPS-enabled camera captures a video recording of homes as workers on the ground drive through neighborhoods affected by the disaster. This technology has several key benefits over field surveys, including reduced cost, the creation of a digital record, and the ability to process the images quickly through crowdsourcing. However, the quality of such user-generated content must be examined to determine its usefulness. An online survey was conducted using imagery from a Spatial Video data collection to determine the potential of using these technologies as a crowdsourcing platform. The survey was distributed to DA experts at the American Red Cross spread across the U.S. as well as individuals connected with CrisisMappers, a mailing list for people interested in the intersection between crises and geospatial technology. Groups of inexperienced and experienced users produced statistically similar results, demonstrating that such content can be trusted and useful for damage assessment. Recommendations are made for the implementation of future systems, the adoption of related methodologies, and considerations for using the resulting data.

© 2014 Elsevier Ltd. All rights reserved.

Introduction

Damage assessment (DA) is an important part of the disaster recovery process. It helps agencies and organizations provide assistance and relief to those most affected. These assessments are common following large disasters which have caused property damage to homes, such as earthquakes, tornadoes, and hurricanes. Most DA efforts will score damage to individual homes with an ordinal series of classes such as “Unaffected”, “Minor”, and “Destroyed”. Using these classes, organizations like the American Red Cross and agencies like FEMA can create an informed action plan to provide relief to individual clients and to a community as a whole.

During a Red Cross DA response, Red Cross personnel are supposed to survey homes only from their vehicles on the road. They are trained to not get out of their cars to perform inspections. From the road, they determine details of the residence such as the street address, dwelling type (e.g. single-family, mobile, or apartment) and number of floors. They also estimate damage on the following scale: “Affected”, “Minor”, “Major”, and “Destroyed” (or inaccessible if road conditions do not permit an assessment). All this information is recorded on a paper data entry form known as a street sheet (American Red Cross, 2003).

The integration of digital video technology and geographic context, also known as Spatial Video, provides value for building inventories and data collection for disaster management (Montoya, 2003). Spatial Video has further been proposed as a method for DA with enriched data capture (Curtis & Fagan, 2013). This method utilizes one or more GPS-enabled cameras mounted to a car that drives through the neighborhoods affected by the disaster. The camera captures video that can then be taken back to a computer workstation and used to do a virtual DA, reviewing the spatially

* Corresponding author. Present address: Spatial Sciences Institute, University of Southern California, 3616 Trousdale Parkway AHF B55, Los Angeles, CA 90089-0374, United States. Tel.: +1 323 393 3583.

E-mail address: elue@dornsife.usc.edu (E. Lue).

enabled imagery to assess damage levels. Aircrafts collecting oblique imagery can provide similar content and such data has also been previously proposed for disaster assessment (Kerle, Stekelenburg, van den Heuvel, & Gorte, 2005). Different methodologies for collecting geographic video may be appropriate depending on the nature of the disaster and the geographic extent of its impact. Aircraft are suited for large and inaccessible areas, but ground-based collection may be more cost-effective and provide higher resolution results.

Data collected via Spatial Video allows people to monitor recovery and better understand how a neighborhood recovers (Curtis, Mills, Kennedy, Fotheringham, & McCarthy, 2007). It provides a lasting digital record of the damage that was captured at a scale that currently cannot be matched with available aerial imagery. The effort needed to collect the data is also similar to the effort put in for current surveys being done by the Red Cross. Volunteers already drive around and perform assessments by looking through their car windows. Adding a video camera to this methodology creates value and produces digital content.

This study explores the potential to use this technology to distribute disaster assessment work and crowdsource the assessments. The inherent benefit of crowdsourcing is that it can take a large task that may take a long time and break it into tasks that are manageable by a large number of individuals working autonomously. An online survey was conducted using imagery from a Spatial Video data collection to determine the potential for the technology to support crowdsourced damage assessment. The survey was distributed to American Red Cross DA volunteers and individuals associated with CrisisMappers (<http://www.crisismappers.net>), a mailing list for people interested in the intersection between crises and geospatial technology, whether they had experience in DA or not.

Damage assessment, crowdsourcing, and the GeoWeb

Damage Assessment can be performed in a number of ways. Ground-based surveys are commonly performed by field teams which assess homes on foot or in vehicles. A more common method in recent years utilizes remotely sensed imagery to identify tornado paths (Jedlovec, Nair, & Haines, 2006; Joyce, Belliss, Samsonov, McNeill, & Glassey, 2009; Yuan, Dickens-Micozzi, & Magsig, 2002) and perform the DA (Barrington et al., 2011; McCarthy, Farrell, Curtis, & Fotheringham, 2008).

The American Red Cross performs assessments on the ground by sending volunteers to drive through the affected neighborhoods. This effort can employ many people and coordinating it can be both expensive and time-consuming. In addition, only those few volunteers will be able to offer their perspectives on the assessments. The resulting data may ultimately provide limited utility to other partner organizations since there are no universally accepted standards for damage classification, and different organizations may assess damage using their own guidelines. Capturing digital imagery or video allows just a few people to provide a versatile dataset to many, giving others the option to make their own assessments. The inclusion of GPS data for tracking collection routes also aids in the logistics of organizing large assessments. The GPS tracks provide information on what roads were covered and what roads are left to be explored, reducing the possibility of different teams duplicating each other's efforts (Curtis, Mills, Blackburn, Pine, & Kennedy, 2006).

The utility of damage assessment data is wide-ranging across multiple disaster types and is relevant wherever post-disaster recovery occurs. New technology, tools, and models can help bring this service to many places with consistent standards and quality. For example, crowdsourcing as a method for data collection after a

disaster has gained traction in recent years (Gao, Barbier, & Goolsby, 2011; Goodchild & Glennon, 2010; Roche, Propeck-Zimmermann, & Mericskay, 2013). Through crowdsourcing, large numbers of people can contribute to a project through a common workflow that centralizes and processes their user generated content.

The four general benefits of crowdsourcing are speed, cost, quality, and diversity (Alonso, 2012). The first of the two benefits relate positively to the need for rapid information collection following a disaster. A short turnaround time for data is preferable after disasters, as delays in data translate to delays in response during time-sensitive operations. Low costs help expedite implementation related to expenditure approvals, and also help in redirecting funds to direct relief.

The latter two benefits, quality and diversity, relate positively to the reliability of the information collected. Crowdsourcing as a strategy achieves quality through quantity, where repeated observations will converge toward an expected outcome. A diversity of responses will also help smooth out the effects of outliers and biased respondents.

The value of crowdsourced data can be enhanced through geography. The presence of spatial data, software, and tools has also grown sharply in recent years as the GeoWeb has developed (Crampton, 2009; Goodchild, 2009; Haklay, Singleton, & Parker, 2008; Hall, Chipeniuk, Feick, Leahy, & Deparday, 2010; Roche et al., 2013). While the technical growth in this realm is clear, a parallel cultural growth has also occurred in the democratization of map-making. These activities are no longer exclusively the purview of GIS and cartographic professionals, and this shift has been made possible by Web 2.0 technologies that offer availability, interactivity, and customizability of spatial content. The ability to customize the content to fit various needs is instrumental to the democratization of spatial information production, resulting in map mashups, which blend and combine data from various sources and formats, to serve any number of needs (Batty, Hudson-Smith, Milton, & Crooks, 2010; Liu & Palen, 2010). As offspring of Web 2.0, the GeoWeb and crowdsourcing promote and support boundless opportunities that go hand-in-hand.

Some crowdsourced damage assessments have utilized remotely sensed imagery with this new paradigm for displaying maps over the internet, such as the Virtual Disaster Viewer (VDV) created by ImageCat following the 2008 earthquake in Wenchuan, China, the GEO-CAN collaboration for the 2010 earthquake in Haiti, and the collaboration between GEO-CAN and Tomnod, Inc. to build a web-based viewer for the 2011 earthquake in Christchurch, New Zealand (Barrington et al., 2011). These DA solutions support the participation of multiple users in the assessment process using high resolution imagery taken with a bird's eye view of the affected areas. These images can be rapidly collected and processed, but do not provide perspective at ground level. These concepts of crowdsourcing and the GeoWeb can be implemented with Spatial Video, which allows for fine scale ground level imagery focusing on the streets that have been damaged (Curtis & Mills, 2012; Curtis, Mills, McCarthy, Fotheringham, & Fagan, 2009; Mills, Curtis, Kennedy, Kennedy, & Edwards, 2010).

While such imagery does not provide three-dimensional detail and there is no true substitute for observing something in person, pictures can provide a level of detail that can be adequate in conveying certain aspects of reality. The subject of the photograph under question is important, however, assessments of aspects such as scenic beauty have differed between photograph and field-based judgments (Hull & Stewart, 1992). This does not mean that photographs cannot be successful surrogates for in-person observation, as certain judgments such as presence and absence can be reliably made. Even air quality, which is determined by haze and color

gradients, has been shown to be identifiable through photographs (Stewart, Middleton, Downton, & Ely, 1984).

Methods

This study utilized footage obtained from a Spatial Video data collection effort. The footage was exported to still photographs and an online survey was built for assessing a selection of these pictures. The results of the survey were then analyzed to address the usefulness of crowdsourcing for damage assessment.

Photo collection

The tornado event that led to the imagery collection for this project occurred on Tuesday, April 3, 2012 in the Dallas Fort-Worth (DFW) area and all tornadoes were rated EF 2 on the Enhanced Fujita scale. The data collection occurred four days after the event, utilizing a vehicle mounted with three Contour GPS cameras, one on each of the rear side windows and one on the front windshield. Each camera recorded video along with a GPS track synchronized with the video playback. The rear windows were tinted, but they were lowered for the collection and did not come between the side cameras and the homes. Neighborhoods with the highest levels of reported damage were included in the survey. The collection finished at 1630 local time, making for 8 h of field collection.

Online survey

Using the Qualtrics survey software, an online survey was produced that presented 32 images of homes in or near the tornado's path. Respondents assigned each picture a damage score. If the home is not damaged, it would be classed as "Unaffected". Otherwise, it was assigned one of four damage classes consistent with the American Red Cross terminology: "Affected", "Minor", "Major", or "Destroyed" (Table 1).

The results were stratified by the respondents' experience with damage assessment efforts. Those with previous experience were classified as experts and included both Red Cross volunteers and others from the CrisisMappers community who had worked on similar efforts with other organizations. The images were also split into two groups: ordered images and random images. Ordered images show homes in a sequence, where it is clear that the houses being assessed are located next to one another (and thus likely to introduce proximity bias in assessment). The random images showed homes that were not adjacent to each other. The 16 ordered images were presented first, followed by the 16 randomized images. Six images from the first set of images were repeated in the second set to test for discrepancies in re-scoring.

Photo selection

The videos were parsed into individual screen captures at a rate of one frame per second. The images chosen for the survey were selected to represent all possible damage levels with an emphasis on the middle classes where the distinction between two successive classes is more difficult to determine. An "Unaffected" house and a "Destroyed" house are less similar than houses with "Minor" and "Major" damage. A street segment from the data collection with 16 consecutive houses was chosen to represent the set of ordered pictures with at least one house in each damage class (according to the primary author's assessment of damages). The random set of pictures was composed of only "Affected", "Minor", and "Major" damage levels (again, according to the primary author's assessment). For each of those levels, two images from the ordered set were repeated (Table 2).

When selecting the neighborhood for the 16 ordered pictures, the first step was to find an image of a "Destroyed" home. Then adjacent homes were examined to determine the surrounding level of destruction and ensure diversity of damage classes (Fig. 1). The resulting set included four "Unaffected" homes, which were useful in setting respondent expectations; knowing that there are "Unaffected" homes among the pictures would help remove bias from an incorrect expectation that all homes would be damaged.

Aside from the repeated images in the random picture set, pictures were selected by first classifying damage for approximately 100 random photos. Once those photos had been placed into their damage classes, random photos were selected from each to fulfill the quotas for the random picture set.

Comparing respondent groups and inter-rater reliability (IRR)

After the survey period closed, respondents' answers were categorized into groups based on the experience levels of the respondents. Responses by the experienced and inexperienced groups were compared for each picture using the Mann–Whitney–Wilcoxon test (MWW), a non-parametric test where the null hypothesis is that two populations are the same. This test was chosen over other comparison tests such as a *t*-test because the data have non-normal distributions. Another reason for the use of this test is that the data are ordinal, which is appropriate for an MWW test.

Krippendorff's α is a statistical test to determine IRR within an individual group and was used on each of the survey respondent groups. It was chosen for its acceptance of multiple observers (i.e. respondents) and missing data, and was deemed more appropriate than other tests such as Cohen's kappa, which is used to gauge IRR for a pair of respondents. The value of α represents the percentage of the data that is coded better than it would be if the data were coded randomly. Suggested guidelines for interpreting α are that data should be considered reliable when $\alpha \geq 0.800$ and can be considered for drawing tentative conclusions between 0.667 and 0.800. However, interpretations of α values can vary depending on the datasets used and their implications, and there is more leeway for accepting lower α values in the social sciences than there is in the physical sciences (Krippendorff, 2013).

Results

Respondent summary

The survey was taken by a total of 108 valid respondents consisting of 23 members of the general public who were inexperienced in disaster assessment, two who did not reveal any information on prior DA experience, 13 who did have experience but were not affiliated with the Red Cross, and 70 Red Cross volunteers and employees (Table 3). Of the Red Cross respondents, eight did not indicate their position within the Disaster Assessment activity and the remainder did, with this group comprising of 37 service associates, 13 supervisors, and 12 managers.

For most of the following analyses, respondents were arranged into several non-exclusive groups, the largest being Group A (all 108 viable responses). Two more groups – Group I (the 23 inexperienced respondents) and Group E-All (the 83 experienced) – were compared with one another. Other groups include Group E-M (the 12 managers), Group E-S (the 13 supervisors), Group E-SA (the 37 service associates), and Group E-O (the 13 with experience outside the Red Cross). Group U (the two who did not indicate any experience) and Group E-U (the eight with Red Cross experience but did not provide DA experience information) were used only in analyses that ignored experience level.

Table 1

Four damage assessment classes as defined on street sheets (American Red Cross, 2003). These class definitions were reproduced in the damage survey, including a fifth class, "Unaffected".





Class	Picture example	Description
Affected		Some shingles and/or siding missing Debris against or around dwelling Structure damage considered to be nuisance Dwelling is livable without repairs
Minor		Minor structural damage Damage to small sections of roof Numerous broken windows Large portions of roofing material and/or siding missing Penetration damage where it is believed no structural damage has occurred
Major		Large portions of roof missing or debris penetration One or two walls missing
Destroyed		Total collapse Shifted on foundation Not economically feasible to repair

Table 2

Number of pictures used in the online survey belonging to different damage scores as determined by the primary author.

Estimated damage score	Ordered set (16 total)	Random set (16 total)
(1) Unaffected	4	0
(2) Affected	3	5 (1 repeat, 3 new)
(3) Minor	6	6 (1 repeat, 4 new)
(4) Major	2	5 (1 repeat, 3 new)
(5) Destroyed	1	0

Damage scores

The response data for damage class were ordinal and were coded for analysis. Damage classes were assigned numeric codes from 1 (Unaffected) to 5 (Destroyed). Modes and medians were calculated as measures of central tendency. In most cases, mode and median values were the same, although the results presented here focus on modes since this method is not susceptible to skewing by a few outliers.

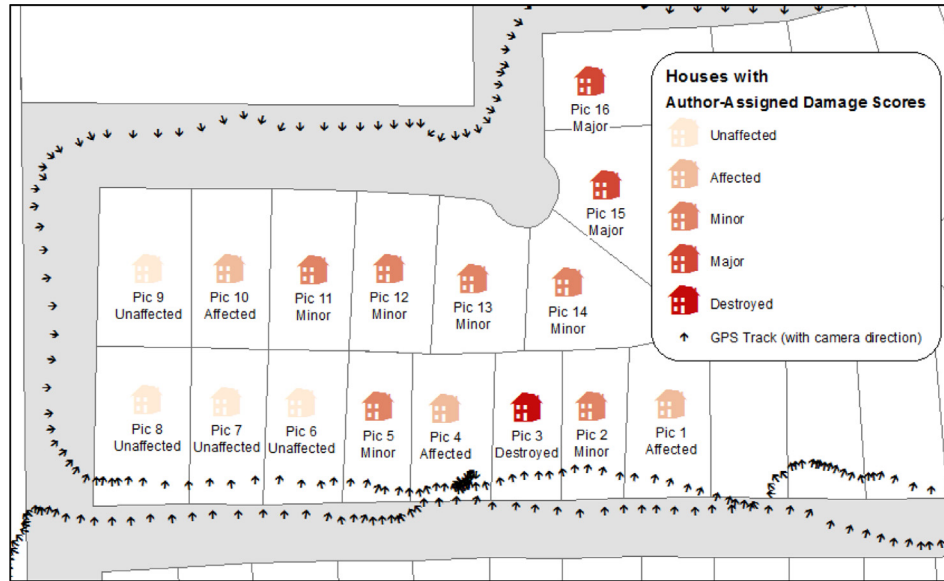


Fig. 1. Map of the neighborhood used for the ordered picture set.

For Group A, only three pictures in 32 had different values for medians and modes (Table 4). In only one picture did the range of values received span all five damage classes (Picture 29). The range spanned four classes in 16 cases, three classes in 10 cases, and two classes in five cases. Of those five cases, four occurred between the highest two damage classes and one occurred between the lowest two damage classes. This makes sense given that pictures of “Unaffected” and “Destroyed” homes are more likely to have one or two scores due to more visibly clear damage states. However, range size did not necessarily correlate with the percentage value of the mode score. Of the four pictures where the mode consisted of over 80% of responses (Pictures 6–9 and 32), three had ranges with three damage classes (Pictures 6–8).

Agreement in damage scores was similar between the inexperienced and experienced groups. The mode of damage scores for each picture is shown in Fig. 2 in three series: as scored by all respondents, by the experienced respondents only, and by the inexperienced respondents only. The mode of the experienced response disagreed with the mode of the inexperienced response in three cases.

Group E-All was expected to demonstrate more precision in its scoring as a group than Group I given its prior experience with DA, but this trend was not observed. Precision was measured by comparing the two groups in terms of how frequently the mode score was chosen for each picture. For Picture 8, for example, 96% of Group E-All chose the mode score whereas only 86% of Group I

chose it, suggesting greater precision in terms of scoring by Group E-All (Table 5). The two groups were within 5% points of each other in 11 cases. Of the remaining 21 cases, Group E-All chose its mode score more frequently than Group I 11 times while the reverse was true in 10 cases. In this sense there is no compelling evidence that Group E-All demonstrated any more certainty than Group I. The same conclusion would be reached if the range size of damage classes were the measure of precision. Between the two groups, the pictures’ range sizes were nearly identical (six pictures where Group E-All’s range was 1 smaller than Group I, nine where the reverse was true, and 17 where the range sizes were the same).

Table 4

Percentages of scores for each picture as assigned by Group A (n = 108). Cases of “no response” are ignored for these calculations. Bold and shaded values designate the mode while italicized scores designate the median. An asterisk (*) indicates a picture where the median and mode were not the same.

Score	Pic 1	Pic 2*	Pic 3	Pic 4	Pic 5	Pic 6	Pic 7	Pic 8
1	12			8	6	90	81	94
2	51	8	1	56	38	8	15	3
3	36	45	9	34	50	2	4	3
4	1	46	25	1	6			
5		2	65					
Score	Pic 9	Pic 10	Pic 11	Pic 12	Pic 13*	Pic 14	Pic 15	Pic 16
1	88	17	4	21				
2	13	57	43	24	10	10		
3		26	51	52	42	58	1	
4			2	4	47	30	72	23
5					1	2	27	77
Score	Pic 17	Pic 18	Pic 19	Pic 20	Pic 21	Pic 22	Pic 23	Pic 24
1					13	9	4	1
2		16		32	60	49	6	23
3		64	1	58	26	42	66	56
4	17	20	43	10	1		24	20
5	83		56					
Score	Pic 25*	Pic 26	Pic 27	Pic 28	Pic 29	Pic 30	Pic 31	Pic 32
1	17				3	2	3	
2	36		8		59	17	39	
3	46	1	63		26	65	50	
4	1	57	27	22	4	16	8	80
5		42	2	78	8			20

Table 3

Number of respondents organized into groups and subgroups.

Group description	Group symbol	No. of respondents
Red Cross managers	E-M	12
Red Cross supervisors	E-S	13
Red Cross service associates	E-SA	37
Red Cross, unknown position	E-U	8
Other experience	E-O	13
Total experienced	E-All	83
Unknown	U	2
Inexperienced	I	23
All respondents	A	108

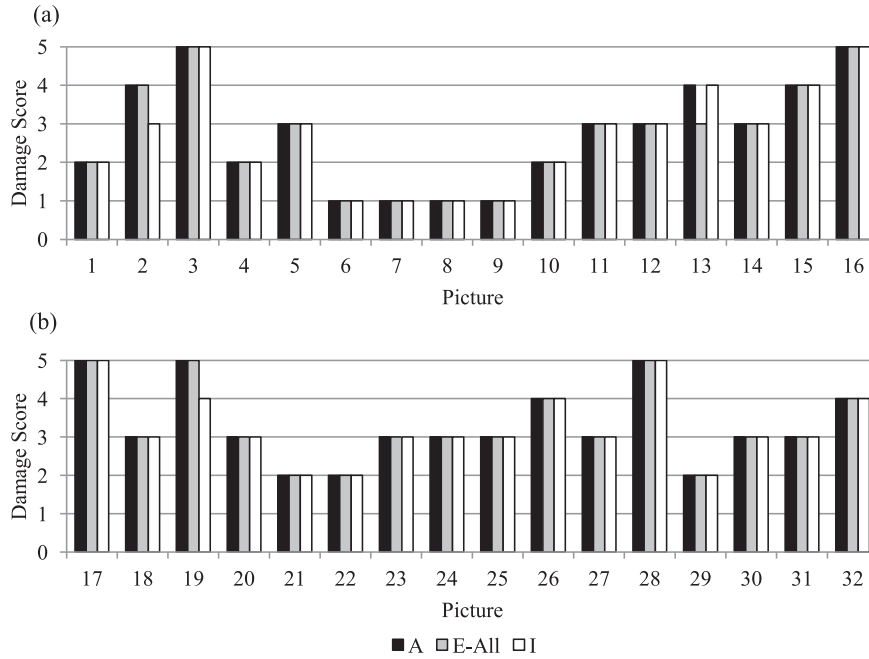


Fig. 2. Mode of damage scores for the ordered (a) and random picture sets (b) as scored by Groups A, E-All, and I.

The MWW statistical test was used to test whether or not Group E-All's scores differed from Group I's scores. This test invokes the null hypothesis that two populations are the same while the alternative hypothesis is that they are different. The test results show that the null hypothesis could only be rejected in the case of Picture 19 (Table 6). The home in this image was debated by the respondents as either having sustained "Major" damage or having been completely "Destroyed".

While median damage scores are generally consistent between these two groups just as the mode is, it should be noted that some scores by a respondent group were sometimes split between two or more classes. In one case, the counts of damage classes chosen were evenly split across three damage classes. Group E-O gave Picture 21 (shown as the "Affected" example in Table 1) an equal number of damage scores across the "Unaffected", "Affected", and "Minor" damage classes. Twelve people were evenly split between the three

Table 5

Percentages of scores for each picture as assigned by Group E-All (left; n = 83) and Group I (right; n = 23). Cases of "no response" are ignored for these calculations. Bold and shaded values designate the mode while italics designate the median. An asterisk (*) indicates a picture where the median and mode were not the same.

	Group E-All								Group I							
	P01	P02*	P03	P04	P05	P06	P07	P08	P01	P02	P03	P04	P05	P06	P07	P08
S1	10			9	6	90	82	96	19			6	5	90	78	86
S2	52	7	1	54	38	10	16	3	43	9		61	33		11	5
S3	38	43	10	37	49		2	1	33	55	5	28	57	10	11	9
S4		48	25		6				5	36	27	6	5			
S5		2	64								68					
	P09	P10	P11*	P12	P13	P14	P15	P16	P09	P10	P11	P12	P13	P14	P15	P16
S1	87	18	5	20					89	14		20				
S2	13	55	46	27	9	10			11	62	32	15	14	9		
S3		27	47	50	48	60				24	64	60	18	55	5	
S4			1	3	42	29	73	20			5	5	68	32	68	32
S5					1	1	28	80						5	27	68
	P17	P18	P19	P20	P21	P22	P23	P24	P17	P18	P19	P20	P21	P22	P23	P24
S1					13	9	3						14	10	9	5
S2		15		37	63	49	6	22		14		14	48	48	5	24
S3		66	1	53	24	42	67	61		59		76	33	43	64	43
S4	15	19	37	11			24	18	27	27	68	10	5		23	29
S5	85		62						73		32					
	P25*	P26	P27	P28	P29	P30	P31	P32	P25*	P26	P27	P28	P29	P30	P31	P32
S1	18				4	1	3		15					5	5	
S2	35		10		62	16	40		40				52	23	36	
S3	45	1	63		24	68	51		45		62		33	55	45	
S4	1	55	26	22	4	15	6	81		64	33	23	5	18	14	73
S5		43	1	78	6		19			36	5	77	10			27

Table 6

The p -values from MWW tests between Group E-All and Group I's damage scores. Only Picture 19, in bold, had a p -value within a 95% confidence limit.

Picture	p -value	Picture	p -value	Picture	p -value	Picture	p -value
P01	0.81	P09	0.83	P17	0.17	P25	1.00
P02	0.25	P10	1.00	P18	0.48	P26	0.61
P03	0.59	P11	0.07	P19	0.02	P27	0.15
P04	0.95	P12	0.45	P20	0.12	P28	0.92
P05	0.66	P13	0.13	P21	0.32	P29	0.20
P06	0.97	P14	0.57	P22	0.99	P30	0.61
P07	0.57	P15	0.75	P23	0.64	P31	0.70
P08	0.11	P16	0.26	P24	0.90	P32	0.39

scores while one person did not score it. Interestingly, this picture is a repeat of Picture 10. When it was scored the first time, five people in the group called it "Unaffected", three people called it "Affected", and four called it "Minor" damage. As there appears to be no apparent damage to the house other than debris in the front yard and a tarp on the roof, this disagreement is likely due to assessors' interpretations of what a tarp meant for damage scoring.

In most cases, the two most common scores for a picture were sequential (i.e. the most common scores for a single picture were 1 and 2 as opposed to 1 and 3). This indicates that split decisions tended to be between one damage class and the next highest or lowest class. The only cases where the two most common damage scores were not adjacent were Group I's assessment of Picture 12 (though the 2nd and 3rd largest counts for damage scores were only one respondent apart) and Group E-O's assessments of Pictures 5, 10, 12, 22, and 25. It should be noted that Group E-O is the group with the least amount of group cohesion, as it is composed of a variety of experienced damage assessors, but with no specified common assessment framework such as the one used by the American Red Cross.

The Krippendorff's α tests were initially run for Groups A, E-All, E-M, E-S, E-SA, E-O, and I three times each, once with all 32 pictures as subjects, once with only the ordered set of 16 pictures, and once with the random set of 16 pictures (Table 7). The results indicate that there is general inter-rater agreement in the data. Using Krippendorff's (2013) guidelines, the data for all groups and subject sets can be used at least for drawing tentative conclusions, with the exception of the inexperienced group rating the random

picture set where $\alpha = 0.639$. Each group of experienced assessors demonstrated more agreement over all three sets than the inexperienced group, with the American Red Cross DA supervisors showing the most agreement. Each group also demonstrated more agreement in the set of ordered pictures than in the set of random pictures.

Each group observed the highest α values for the 16 ordered pictures rather than the 16 random pictures. At first, the likely explanation for this difference in IRR is that there actually is some bias when the pictures are of houses near each other, leading to more convergence toward agreement. However, this result is contrary to the hypothesis that responses begin to converge after respondents get practice with assessing pictures; in this case, IRR should be higher in the later picture set. The more likely explanation for the higher α values is that the ordered picture set contained several homes that were completely unaffected and were easy to evaluate. The random picture set only had images of damaged homes. A fourth run of Krippendorff's α was conducted on the ordered picture set data with the four unaffected homes removed, with the results indicating smaller α values (Table 7).

Repeated images

The repeated images were inserted into the survey to explore whether or not respondents could reliably repeat their ratings. Exploring consistency in rating is independent of a group's ability to agree on its ratings. In this sense, if a rater looked at a picture of a 4 and assigned it a 2 the first time and then a 4 the second time, that is a worse result for consistency than assigning it incorrectly as a 2 both times.

The mode damage scores were the same within each pair of repeated images. While these results at a group level may suggest that the group consistently rated the pairs of pictures, an examination of how individuals did at repeatability was showed less encouraging results. More often than not, a picture that was repeated within the survey was given the same score, but there were still many who changed their minds the second time around. Of the people who changed their minds, there was no particular trend in whether that change had a specific direction. In most cases, that change was only up or down by one damage class, and few cases showed a change in two damage classes or more.

Table 7

Krippendorff's α values for IRR and the damage scores assigned by different respondent groups to multiple subject sets. Green values are reliable, black values are acceptable for tentative conclusions, and red values should not be accepted as indicators of agreement.

Group	Total Raters	All 32 Pictures	16 Ordered Pictures	16 Random Pictures	12 Ordered & Damaged Pictures
A	108	0.754	0.787	0.691	0.669
E-All	83	0.764	0.795	0.706	0.676
E-M	12	0.739	0.754	0.701	0.619
E-S	13	0.833	0.864	0.785	0.783
E-SA	37	0.770	0.812	0.703	0.672
E-O	13	0.757	0.774	0.701	0.695
I	23	0.712	0.751	0.639	0.632

Contingency tables were calculated for each repeated picture, with the score assigned the first time recorded as rows and the second score recorded as columns (Fig. 3). In each contingency table, the diagonal series of boxes from (1, 1) to (5, 5) represent repeat damage scores. Boxes above the diagonal represent pictures that were assessed the second time as being more damaged. Boxes below the diagonal represent pictures that were assessed the second time as being less damaged. The repeat ratings ranged from 60% (P05 and P31) to 84% (P16 and P28) and the unequal sums between tables are due to missing data.

Respondents' comments

The survey provided an option to include comments on the survey itself or on the images that were used. The majority of feedback given in this section regarded the concept of a 3-point view method of damage assessment. A total of 34 people (41% of all experienced respondents) suggested that multiple points of view were needed for a single house in order to determine damage: one point of view for the front of the house and two for each of the adjacent sides. Nearly half of the experienced assessors mentioned it, and there were surely some others who thought it but just did not leave a comment. The inclusion of side pictures was considered during the survey design, but these pictures were intentionally left out to simplify the questions and manage the length of time needed by respondents to provide answers. In hindsight, the inclusion of these additional views may have reduced the number of respondents willing to sit through the survey, but it would have likely added confidence to peoples' assessments.

In a similar vein, 13 respondents commented that they felt they needed to see more to make better assessments. Suggestions included the option to zoom into a picture (one respondent), see images from a higher vantage point (e.g. a camera tower mounted to a car; one respondent), or employ image processing to remove shadows (four respondents). Five respondents recognized the importance of seeing the roof, and would have liked an aerial image

to complement the street-level image. Two respondents commented that scores were difficult to determine where blue tarps concealed roof damage, but this limitation would also be observed by someone at the damage site since most levels of damage assessment would disallow the removal of protective tarps. Only one respondent observed that these pictures did not give a person the same amount of detail that he or she would get in person, but no explicit explanation for that observation was provided.

Discussion and conclusions

One of the most promising features of crowdsourcing is the ability to collect accurate information by aggregating responses from a large number of people. This study showed that similar results were achieved using inexperienced and experienced damage assessors. This result does not suggest that experience adds little value to the assessments, but rather that inexperienced assessors can be tapped for this work if needed and the content they generate can be trusted as tentative or preliminary results.

Even the most experienced experts can disagree on how a house should be scored, especially if a more refined damage scheme with more classes is utilized. For this reason, no analysis was performed to learn how "right" or "wrong" any of the groups were in their assessments; hence, there was no answer key or baseline to compare these responses with. However, an organization that participates in damage classification that wishes to employ a similar crowdsourcing platform may benefit from creating a tutorial showing visual examples of the damage classes used. Such a tutorial would surely improve IRR.

Similarly, such a tutorial could be used for disaster assessment training and testing. The results of this study do not imply that field crews are not necessary. Rather, the results emphasize that such rapid information collection can be deemed tentatively reliable. A thorough damage assessment might employ a number of techniques, such as field crews who assess houses on the ground while capturing imagery and video which can be referenced at a later

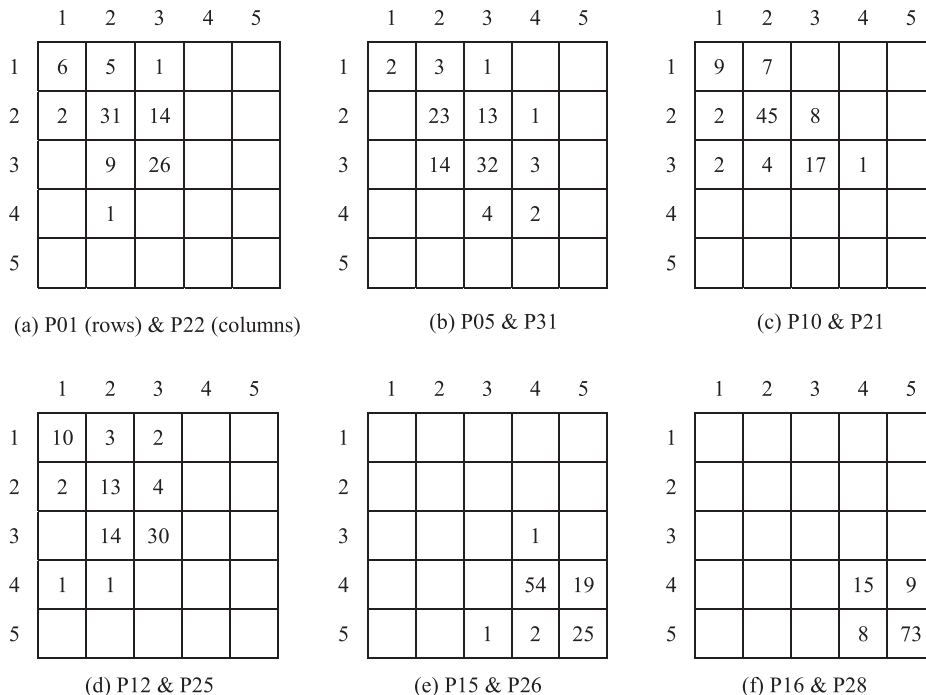


Fig. 3. Contingency tables for each pair of repeated pictures as rated by Group A.

time. As mentioned, aerial imagery can be used for damage classification and integrated into a multi-method approach to data collection, providing different views of affected structures.

Performing similar surveys on images collected from other tornado events would further refine the definitions of each class. Images of houses affected by more extreme tornadoes may show a greater variety of damage. Such data has been collected for the EF4 2011 Tuscaloosa–Birmingham tornado, EF5 2011 Joplin tornado, and the EF5 2013 Moore tornado. Expanding this survey to include a larger set of images would improve confidence in the determinations of damage scores. Inclusion of images from a variety of events would help in exploring whether or not a specific event, disaster type, or geography that may influence scoring.

Further research in validating user generated content can employ datasets where houses have already been assessed in the field. A crowdsourcing effort can then be performed to compare computer-based assessments to the field assessments. Of the respondents in this study who had prior DA experience, 30% felt that their assessments would have differed if they had been in the field. Studies utilizing existing field assessments can further explore such claims.

Acknowledgments

We would like to thank Joseph Martin and Eric Klein of the Dallas chapter of the American Red Cross as well as Greg Tune and Jim Dooley at the National Headquarters.

References

- Alonso, O. (2012). Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, 16, 101–120.
- American Red Cross. (2003). *On-Site Detailed Damage Assessment Worksheet (Street Sheet)*. Washington DC: American Red Cross.
- Barrington, L., Ghosh, S., Greene, M., Har-Noy, S., Berger, J., Gill, S., et al. (2011). Crowdsourcing earthquake damage assessment using remote sensing imagery. *Annals of Geophysics*, 54, 680–687.
- Batty, M., Hudson-Smith, A., Milton, R., & Crooks, A. (2010). Map mashups, Web 2.0 and the GIS revolution. *Annals of GIS*, 16, 1–13.
- Crampton, J. W. (2009). Cartography: Maps 2.0. *Progress in Human Geography*, 33, 91–100.
- Curtis, A., & Fagan, W. F. (2013). Capturing damage assessment with a spatial video: an example of a building and street-scale analysis of tornado-related mortality in Joplin, Missouri, 2011. *Annals of the Association of American Geographers*, 103, 1522–1538.
- Curtis, A., & Mills, J. W. (2012). Spatial video data collection in a post-disaster landscape: the Tuscaloosa Tornado of April 27th 2011. *Applied Geography*, 32, 393–400.
- Curtis, A., Mills, J., Blackburn, J. K., Pine, J. C., & Kennedy, B. (2006). Louisiana State University geographic information system support of Hurricane Katrina recovery operations. *International Journal of Mass Emergencies and Disasters*, 24, 203–221.
- Curtis, A., Mills, J. W., Kennedy, B., Fotheringham, S., & McCarthy, T. (2007). Understanding the geography of post-traumatic stress: an academic justification for using a spatial video acquisition system in the response to Hurricane Katrina. *Journal of Contingencies and Crisis Management*, 15, 208–219.
- Curtis, A., Mills, J. W., McCarthy, T., Fotheringham, A. S., & Fagan, W. F. (2009). Space and time changes in neighborhood recovery after a disaster using a spatial video acquisition system. In P. Showalter, & Y. Lu (Eds.), *Geospatial technologies in urban hazard and disaster analysis* (pp. 373–392). Springer.
- Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26, 10–14.
- Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services*, 3, 82–96.
- Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3, 231–241.
- Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: the neogeography of the GeoWeb. *Geography Compass*, 2, 2011–2039.
- Hall, G. B., Chipeniuk, R., Feick, R. D., Leahy, M. G., & Deparday, V. (2010). Community-based production of geographic information using open source software and Web 2.0. *International Journal of Geographical Information Science*, 24, 761–781.
- Hull, I. V. R., & Stewart, W. (1992). Validity of photo-based scenic beauty judgments. *Journal of Environmental Psychology*, 101–114.
- Jedlovec, G. J., Nair, U., & Haines, S. L. (2006). Detection of storm damage tracks with EOS data. *Weather and Forecasting*, 21, 249–267.
- Joyce, K. E., Belliss, S. E., Samsonov, S. V., McNeill, S. J., & Glassey, P. J. (2009). A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 33, 183–207.
- Kerle, N., Stekelenburg, R., van den Heuvel, F., & Gorte, B. (2005). Near-real time post-disaster damage assessment with airborne oblique video data. In P. van Oosterom, S. Zlatanova, & E. M. Fendel (Eds.), *Geo-information for Disaster Management* (pp. 337–353). Berlin.
- Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: SAGE Publications.
- Liu, S. B., & Palen, L. (2010). The new cartographers: crisis map mashups and the emergence of neogeographic practice. *Cartography and Geographic Information Science*, 37, 69–90.
- McCarthy, T., Farrell, R., Curtis, A., & Fotheringham, A. S. (2008). Integrated remotely sensed datasets for disaster management. In U. Michel, D. L. Civco, M. Ehlers, & H. J. Kaufmann (Eds.), *SPIE 7110, Remote Sensing for Environmental Monitoring, GIS Applications, and Geology VIII*.
- Mills, J. W., Curtis, A., Kennedy, B., Kennedy, S. W., & Edwards, J. D. (2010). Geospatial video for field data collection. *Applied Geography*, 30, 533–547.
- Montoya, L. (2003). Geo-data acquisition through mobile GIS and digital video: an urban disaster management perspective. *Environmental Modelling & Software*, 18, 869–876.
- Roche, S., Propeck-Zimmermann, E., & Mericskay, B. (2013). GeoWeb and crisis management: issues and perspectives of volunteered geographic information. *GeoJournal*, 78, 21–40.
- Stewart, T., Middleton, P., Downton, M., & Ely, D. (1984). Judgments of photographs vs. field observations in studies of perception and judgment of the visual environment. *Journal of Environmental Psychology*, 283–302.
- Yuan, M., Dickens-Micozzi, M., & Magsig, M. A. (2002). Analysis of tornado damage tracks from the 3 May tornado outbreak using multispectral satellite imagery. *Weather and Forecasting*, 17, 382–398.