Quality Analysis After Action Report For The Crowdsourced Aerial Imagery Assessment Following Hurricane Sandy

Robert Munro
Idibon
rob@idibon.com

Tyler Schnoebelen
Idibon
tyler@idibon.com

Schuyler Erle
Idibon
schuyler@idibon.com

INTRODUCTION

We present an after-action report for a large-scale damage assessment project that followed Hurricane Sandy's landfall on the Eastern seaboard of the USA in 2012. The Civil Air Patrol (CAP) took over 35,000 GPS-tagged images of damage-affected areas, as part of their mandate to provide aerial photographs for disaster assessment and response agencies, primarily FEMA, who used the aggregate geolocated data for situational awareness.

The scale of the destruction meant that there was a relatively large amount of photographs for a single disaster. As a result, it was the first time that CAP and FEMA used distributed third-party information processing for the damage assessment, with 6,717 public volunteers evaluating the level of damage present in the images via an online crowdsourcing system. The volunteers saw one image at a time using an online *MapMill* system run by (author) Schuyler Erle, giving a three-way judgment: little/no damage; medium damage; or heavy damage. This report is quality of the damage assessment evaluating the volunteer workers' performance in three ways:

1. Inter-annotator agreement: how often did different volunteers agree with each other?

2. Comparison with experts: 11 expert raters from the GIS Corps assessing the same images.

3. Ground-truthed ratings: comparison to ratings made by FEMA at the same grid locations.

INTER-ANNOTATOR AGREEMENT

For inter-annotator agreement, we find that the public volunteers had majority agreement for almost 95% (93.54%) of all images they assessed, and a super majority for about 80% of images, indicating a high level of agreement. However, there was unanimous agreement on less than 50% of the images, showing that complete agreement was relatively rare.

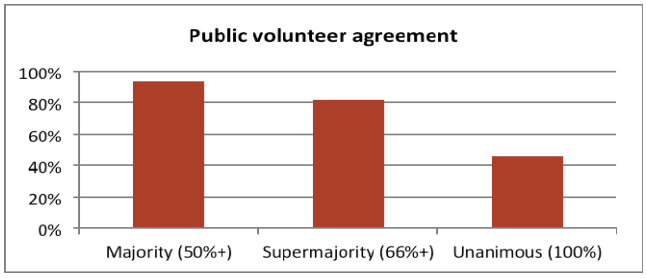

Figure 1: Three different levels of quality assessment on 17,070 images (limited to images with three or more ratings per image), showing that public volunteers generally agreed with each other on how to classify an image.

COMPARISON WITH EXPERTS

For the comparison with experts, 720 of the most problematic images were assessed by 11 GIS Corps experts, using the same platform and instruction set. 81% of the images had a supermajority agreement among the experts, compared to just 37% for public volunteers, showing that the volunteers were not as accurate. The main

area of disagreement between the groups were for images that the volunteers said showed no real damage and which the experts said showed some damage (9% of the images). Only 11% of the ratings were dramatically off (where one group said there was no damage and the other group said there was severe damage).

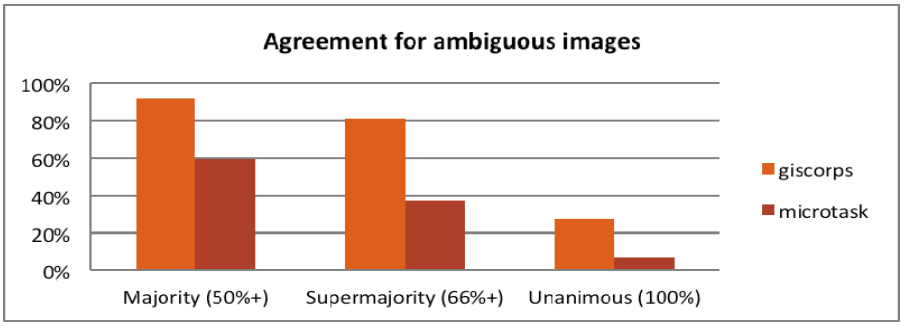

Figure 2: Tough-to-classify images (count = 720) rated by experts. They received pretty consistent ratings by the experts but note that these images were chosen *because* non-experts had substantial disagreement.

GROUND-TRUTHED RATNGS

The third evaluation produced a negative result, as we were not able to find a strong correlation between the aerial evaluations and FEMA's ground-reports. We can identify some grids where this is due to timing: the presence of flood-water was typically marked as high damage, but it had receded before the FEMA assessments. In other places, there was a mismatch between aerial photographs and grid-references and only a small area needed to be damaged according to FEMA ratings, while CAP ratings were over the whole area.

There are no previous reports comparing damage assessment from CAP imagery and FEMA ground-truth reports (that we are aware of), so this disparity may not be specific to the context of a crowdsourced workforce. We conclude that the rating systems need to be investigated in more detail and that different correlation/aggregation methods should be tested to ensure compatibility between the assessment methods.

DISCUSSION

In general, a volunteer's inter-annotator agreement goes up the more experience they have with the task. Using overall agreement per worker, there is 95% confidence on an image's rating once five workers have seen it:

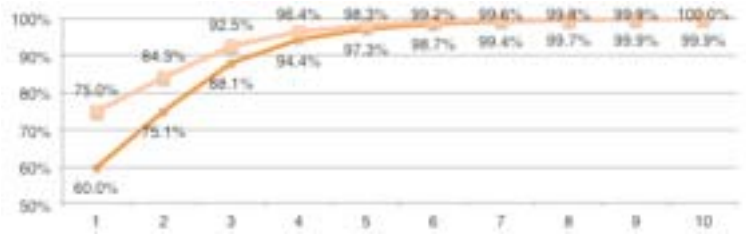

Figure 3: The more workers assess an image, the more confident we can be about the assessment. If you have experienced crowdsource workers available (the lighter, upper line), fewer raters are required.

Depending on the quality of assessment required, you would choose between four and six and judgments to ensure accuracy, and increase the number of judgments where disagreement occurs, or back off to experts.

Most disasters are not as prominent as Sandy and might struggle to find a large enough volunteer community. Crowdsourcing is typically paid, so we also surveyed 20 professional crowdsourced workers to establish a price-point for paid, crowdsourced damage assessment. The results varied from $0.001 to $0.02 per judgment depending on worker expertise. This would come to a maximum of US$3,000 for the entire operation if paid workers were used, which is less than the cost to manage volunteers and on par with a single aerial survey.

We conclude that it is possible to deploy the information processing strategies that we used for Hurricane Sandy aerial image assessment for future disasters, while also addressing some of the quality and reliability concerns that arise from using crowdsourced workforces.