A Crowdsourced Data Set of Edited Images Online

Valentina Conotter[1], Duc-Tien Dang-Nguyen[2], Michael Riegler[3], Guilia Boato[1], Martha Larson[4]
[1]DISI - University of Trento, Italy
[2]University of Cagliari, Italy
[3]Simula Research Laboratory, Norway
[4]Delft University of Technology, Netherlands
{conotter, boato}@disi.unitn.it, ductien.dangnguyen@diee.unica.it, michael@simula.no, m.a.larson@tudelft.nl

ABSTRACT

We present a crowdsourcing approach to tackle the challenge of collecting hard-to-find data. Our immediate need for the data arises because we are studying edited images in context online, and the way that this use impacts users' perceptions. Study of this topic cannot advance without a large, diverse data set of image/context pairs. The image in the pair should be suspected of having been edited, and the context is the place (e.g., website or social media post) in which it has been used online. Such pairs are hard to find, and could not be collected, due to techno-practical constraints, without the support of crowdsourcing. This paper describes a three-step approach to data set creation involving mining social data, applying image analysis techniques, and, finally, making use of the crowd to complete the necessary information. We close with a discussion of the potential and limitations of the data set collected.

Categories and Subject Descriptors

H.3 [**Information storage and retrieval**]: Content analysis and indexing, Information search and retrieval

Keywords

Human Perception; Edited Images; Test design; Data Set

1. INTRODUCTION

Large, diverse data sets are necessary in order to carry out meaningful investigations of multimedia phenomenon. However, researchers face a chicken-and-egg problem: when the phenomena under question is not yet well studied, it can be very difficult to collect a large number of examples that represent it. Without data, however, it is impossible to study the problem. This effect is dangerous because it reduces the motivation of multimedia researchers to address new topics that are interesting and important, but for which the data set is difficult to collect.

CrowdMM'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660114.2660120.

(a) (b)

Figure 1: (a) The originals of two promotional images for a real estate listing. (b) The edited versions published online. Discussion of this editing practice triggered a large public outcry, and ultimately an apology from the real-estate agent (taken from [2]).

In this work, we present a crowdsourcing approach to tackle the challenge of collecting hard-to-find data. The approach is described in practice, as it was used to create the *Edited Images Online* (EIO) data set. This data set was created within the larger context of the study of a new multimedia topic: user perceptions of deception in the use of online images. The importance of this research topic is witnessed by the circumstances surrounding the use of edited images online. Deceptive use of edited images may occur relatively infrequently, but when it occurs, the consequences are immense. Fig. 1 depicts examples extracted from a May 2013 episode of *Million Dollar Listing New York*, which showed the real estate star Luis Ortiz digitally altering some of the promotional pictures for a property. Noting that 'real photos' do not tempt people to view properties, he stated, "A little white lie isn't going to kill anybody!" However, this broadcast led to an outpouring of complaint and also an investigation, which eventually led to a public apology on the show's blog [2]. This example illustrates the importance of people's trust in the information conveyed by multimedia. The social and economic consequences of the deceptive use of edited images are huge. However, the topic is complex and challenging to study. After an initial paper [1], our work on the topic ground to a halt due to the lack of a large data set that contained not just edited images, but actually image/context pairs, which were necessary to study edited images in their context of use. The difficulty of finding an adequate number of examples due to techno-pratical contraints and the limited access to huge amount of images online lead us to solve such problem by developing the crowdsourcing approach presented here. The approach involves

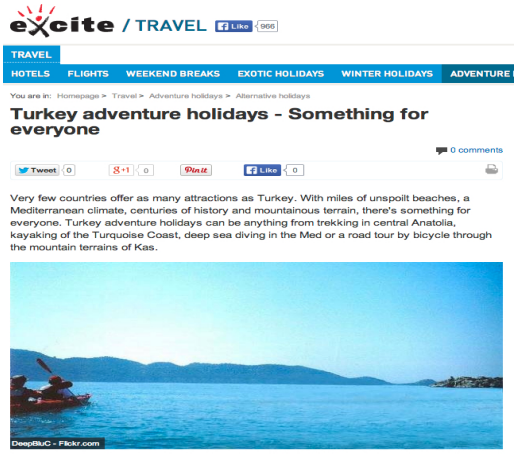

Figure 2: Example of an edited image within its context.

three-steps: mining social data, applying forensic analysis techniques for images, and, finally, crowdsourcing. The contribution of the paper is to offer the multimedia community an approach for using crowdsourcing to collect hard-to-find data, illustrated with a specific example. In order to achieve the goal of collecting hard-to-find data, our approach goes above and beyond commonly-used existing techniques in two important respects. First, crowdsourcing platforms make human intelligence available in abundant supply. However, it is far from an unlimited resource. For this reason, in order to address the needle-in-a-haystack problem posed by hard-to-find data, the crowd must be supported by a carefully orchestrated *combination of automatic techniques* that minimize the human effort necessary. Second, the problem we tackle here is one that requires *crowdworkers to actively search* of the 'missing pieces' of the data set, instead of easy-to-provide information, such as the validation of image labels. Active search necessitates a crowdtask that encourages crowdworkers to submit serious work, but also is sensitive to the fact that for some workers, search engine results will be restricted, and they will not be able to complete the task. In the next section, we describe how we selected an initial data set of suspicious images, using a two-step approach. Next, we describe the third step in detail: a crowdsourcing task run on Amazon Mechanical Turk. Finally, we discuss the validity of data set, exploring its usefulness to collect judgements of people's perceptions of deception.

2. COLLECTING HARD-TO-FIND DATA

In this section, we describe our approach to collecting hard-to-find data. The approach consists of three steps, first tag-based image selection, then image filtering with forensic methods, and finally a crowdtask that we publish on the Amazon Mechanical Turk platform (AMT)[1] to collect image contexts. Our aim is to create a data set consisting of image/context pairs (an edited image and the context in which it is used online) that will support our research on perceptions of deception in the use of edited images online. An example of such a pair is shown in Fig. 2.

2.1 Selecting the Initial Image Set

Our first step is to create an initial set of images that are 'editing suspect', i.e., they have a relatively high probability of having been edited after they were shot. We create this initial set by searching for social images to which users (i.e., the uploaders of the images) have assigned one or more tags suggesting that the images have been edited. For this purpose, we build a set of keywords that reflect image editing. The keyword set was built by using exploratory searches on Flickr, and also making use of a thesaurus and popular literature on photo editing. The final list of query keywords is: `manipulate | manipulation | manipulated | doctored | faked | photoshop | edited | modified | modification | doctored | retouched | enhanced`. As the basis of our image set, we use the set of Creative Commons Flickr images released by the MediaEval 2010 benchmark [4]. We are interested in using exclusively Creative Commons data so that we can release our data set publicly. This keyword-based selection process discarded images that were not associated with any of the tags in the list. This process filtered out 99.7% of the images, thus reducing the set to 8980. To further narrow the data set, we then excluded images that were associated with any one of these keywords: `original | unmanipulated | unedited | nophotoshop`. The result of this step was a set of 6881 images.

2.2 Filtering the Initial Image Set

In the next step, we further filter the images by investigating the image metadata with an analysis inspired by forensics methods. This step verifies that images we collected have indeed been edited. When dealing with multimedia trustworthiness and authenticity, forensics methods are of particular relevance since they offers solutions to directly verify whether a digital multimedia content is genuine and authentic, without any a-priori knowledge [6]. Building on [3], we extract and analyze the Exif header of each of all the 6881 images in our database. The method in [3] exploits the signature left by the camera in the JPEG header of an image, i.e., the Exchangeable image file format (Exif) metadata information, to determine if an image has been modified from its original version. The signature is shown to be highly distinct and unique to single camera models and software manufacturers and any manipulation would alter such specific signature. In particular, we are interested in the Exif entries related to the time of acquisition and manipulation of the original size of the captured image and to the software used to modify the image. Specifically, first, we check the 'original' and 'modification' time stored in the Exif metadata of each image. Any inconsistency between these two fields can be taken as a clue for manipulation, since it means that the image has been re-saved at a different time than original capture. Second, we investigated the Exif field related to the size of the original captured image. Any given camera model supports a specific standard image size; if a resizing operation has been applied to an image, its actual size will not match with the Exif information, thus providing a reliable proof for editing. Based on time and size inconsistencies we further filter our database of edited images by skimming off images that do not show evidence of being edited.

Lastly, of particular interest for our work is the 'software' field in the Exif header, where the software that has been used to modify the image is usually stored (e.g., Photoshop). Following [3], we retain all those images whose 'software' field in the Exif header contain at least one of the keywords in the following list: `photoshop | aperture | borderfix`

[1] www.mturk.com

Figure 3: Examples of the edited images collected.

| ashampoo | photo commander | bibble | capture | capture nx | capture one | coachware | copiks | photomapper | digikam | digital photo pro | gimp | idimager.com | imagenomic | noiseware | imageready | kipi | microsoft | paint.net | paint shop | photoscape | photowatermark | picasa | picnik | quicktime | watermark | hdrtist. In the end, we are left with an *Edited Images* dataset of 6069 images for which we have verified our suspicion that to be edited. Fig. 3 presents some examples.

Note that starting with a larger number of Flickr images would have resulted in a larger *Edited Images* set. However, our goal was to ultimately arrive at a set of ca. 1,000 image context pairs. This number of images would fulfill our future goal of studying the impact of edited image on user's perception. We also point out that we cannot be 100% sure that the images in the *Edited Images* set have indeed been edited. For example, someone could have artificially manipulated the Exif data without actually changing the image.

2.3 Calling on the Crowd to Collect Contexts

Having collected a set of *Edited Images* we carry out a final step in which we collect the contexts in which these images have been used online. This step yields the final *Edited Images Online* (EIO) data set. Here is where human intelligence is needed. Initially, we assumed this task could be approached automatically with reverse image look up. However, the costs of using the TinEye API were prohibitive (AMT was cheaper), and Google Images (which provides no look up API) was shown to give better results in exploratory tests.

For this reason, we decided to ask crowdworkers to search for images. An added advantage of this approach is that it gave us experience with the kinds of limitations that AMT workers may experience accessing search engines and websites, which will inform our future work. Following the crowdtask design principles described in [5], we designed our context collection crowdtask using an iterative process. The crowdtask was designed as an AMT Human Intelligence Task (HIT). First, the initial design was created in close discussion with the entire team, then a pilot HIT was carried out in the AMT Sandbox with colleagues in order to catch places in which the description of the task was unclear, finally the HIT was published to AMT. The purpose of the context collection HIT is to gather the contexts (URLs) in which the images have been used. In the HIT, we present the workers with seven images asking them to search the Internet for a place (i.e., a website) where the images can be found. We ask the workers to use the Google Image search engine and look for an image (and the corresponding context of use) that is exactly the same as the one provided (only differences in size are accepted). The HIT included

(a) (b)

Figure 4: Examples of collecting contexts for edited images (a) negative: no context for the image can be found; (b) positive: a context can be found.

a link to an extra instruction page with detailed directions for how to search using Google Image Search. We provide workers two boxes ('yes' and 'no') to indicate if they were able to find the image (or not), and a free text box where they must copy and paste the link to the place where the image has been used (i.e., the URL) if they have found it. Fig. 4 illustrates the process of finding images.

The main design challenge of this HIT is to implement a quality control mechanism that is fair to the crowdworkers. A highly effective quality control mechanism is to include control questions in the HIT, i.e., questions for which the answer is already known. If the control questions in a HIT assignment are not answered correctly, we assume that the worker is not carrying out serious work, and we reject the HIT assignment. The control questions presented a specific issue that made this HIT unique with respect to other typical crowdtasks used, for example, for multimedia annotation. Specifically, we could not be 100% sure about the correct answer to the control question from the crowdworker's point of view. The reason is that we are not 'omniscient' about the behavior of the Google Image search engine from the perspective of the worker. Workers may search for a specific image and find nothing, because, for example, they are located in a country that blocks certain websites.

In order to address this challenge, we use two types of control questions. First, a negative control image, an image that we know is not findable, and, second, a positive control image, an image that we know is findable. The negative control image helps to control for unserious work and the positive control image helps to identify workers with a different 'view' of the Internet via Google Images. Because of the presence of the positive control image, under normal circumstances, the worker is guaranteed to be able to find at least one image of the seven images in the HIT. In order to determine if the worker's 'view' of the Internet is creating issue, the HIT included a final question asking workers to report their location if they are unable to find any of the images. We also request workers who found no images in a HIT assignment to stop working on the HIT. This was to prevent people unable to carry out the HIT due to limited internet access from wasting their time. The result of

Iteration	# New images	# Found by both	# Found by one	# Not found
Sandbox	140	34	40	66
AMT 1	150	26	29	95
AMT 2	300	54	71	175
AMT 3	600	87	145	368
AMT 4	1200	196	307	697
AMT 5	2400	324	535	1541

Table 1: Summary of contexts gathered with each iteration to show the rate of harvesting positive control images. '# Found by both': number of images for which both the workers found the same context; '# Found by one': number of images where the workers found a context, but not the same one; '# Not found': number of images for which no context was found. The '# Found by both' images were used as positive control images in the next iteration.

the crowdtask was the final EIO data set consisting of 1801 edited images for which the crowdworkers were successful in identifying context online. In the next section, we describe the practical aspects of how we ran the context collection HIT in greater detail to allow other researchers to be able to repeat our process.

3. THE CROWDTASK IN PRACTICE

The practical considerations in running the context collection HIT are related to the challenges mentioned in the previous section. First, we needed to create a HIT that is fair to crowdworkers who might, possibly unknowingly, see a restricted version of Google Images. Second, we needed a large number of control images (i.e., images for which we already know the correct response), so that the crowdworkers could not conceivably memorize which images were the control images, and answer them selectively. Addressing the first challenge required careful inspection of the crowdworkers' responses. After an initial period of leniency, which we used to confirm that our rejection criteria were reasonable and well understood, we rejected those assignments in which one of the control images received an incorrect answer. However, we did not reject a HIT assignment in which no images were found, if the worker had put a comment in the box asking for information on location. While the HIT was running, we observed cases of the workers making use of the box, indicating that the mechanism had been successful. If for a given image, one worker found a context and another did not, we investigated the case by hand in order to determine which was correct. The process had to be carried out with care, since it was possible that an image would disappear from online between iterations. In order to tackle the second challenge, we ran the HIT iteratively. The contexts that were collected in a given iteration fed the next iteration with additional control questions. The evolution of the number of control images collected can be seen from Table 1.

Each line of the table reports an iteration of the HIT. We discuss the first two iterations in detail, as an aid to understanding the table. The first iteration was carried out in the AMT Sandbox by volunteers from among the authors and their colleagues in order to understand the workers experience doing the HIT. This iteration consisted of 20 HIT assignments, which were carried out by 13 volunteers. Table 1 shows that this iteration let to the contexts of 34 images (i.e., URLS) being found by both workers, and 66 found by neither. These were taken forward as positive and negative control images into the next iteration. All subsequent iterations ran on AMT. For 'AMT' iterations, each HIT was carried out by two workers. The 'AMT 1' iteration collected contexts for 150 new images. Since five of the seven images in each HIT are new images, this iteration consisted of 30 HITs. For this purpose we needed a total of 60 control images (30 positive and 30 negative). These control images were available from the previous HIT. As the iterations got larger, we occasionally used the same control images more than once in an iteration. However, we did this only when the total number of HITs in the iteration was large enough so that a connection could not be easily established between repetition and control questions. After all iterations, we obtained 530 'Found by both' and 1271 'Found by one' images, summing up to a total of 1801 images.

4. DISCUSSION AND CONCLUSION

We have described an approach for using crowdsourcing to collect hard-to-find data, illustrated with a specific example. The main design challenge faced is to carry out quality control even though we cannot be sure that crowdworkers are in a position to answer the reference questions correctly. The value of our contribution is practical: we demonstrate that crowdsourcing provides critical support in the collection of hard-to-find data and describe our solutions for tackling the challenges that arose during implementation. Note that we do not claim that our data set is generally representative of the use of edited images in context online. Rather, it is a selection of cases that are process has been able to uncover. The data set's worth lies in two aspects. First, it is larger than any existing data set of the kind, and, second, it was built without relying on any assumptions on our part about the types of contexts in which edited images are used. We hope that the approach that we describe here will support other multimedia researchers in pursuing research for which they originally assumed the data would be too hard to find.

Acknowledgments: This work has received funding from EC FP7 project CUbRIK (grant agreement No. 287704) and by the iAD center for Research-based Innovation (project number 174867) funded by the Norwegian Research Council.

5. REFERENCES

[1] V. Conotter, D.-T. Dang-Nguyen, G. Boato, M. Menéndez, and M. Larson. Assessing the impact of image manipulation on users' perceptions of deception. In *SPIE, HVEI XIX*, volume 9014, 2014.

[2] FourAndSix. Photo tampering throughout history. Online at http://www.fourandsix.com/photo-tampering-history. 2014.

[3] E. Kee, M. K. Johnson, and H. Farid. Digital image authentication from JPEG headers. *IEEE TIFS*, 6(3):1066–1075, 2011.

[4] M. Larson et al. Automatic tagging and geotagging in video collections and communities. In *ACM ICMR*, pages 51:1–51:8, 2011.

[5] M. Larson, M. Melenhorst, M. Menéndez, and P. Xu. Using crowdsourcing to capture complexity in human interpretations of multimedia content. *Fusion in Computer Vision*, page 229, 2014.

[6] A. Piva. An overview on image forensics. *ISRN Signal Processing*, 2013:1–22, 2013.