Seeing Is Believing: How People Fail to Identify Fake Images on the Web

Mona Kasra
University of Virginia
Charlottesville, VA 22904, USA
mona.kasra@virginia.edu

Cuihua Shen
UC Davis
Davis, CA 95616, USA
cuishen@ucdavis.edu

James F. O'Brien
UC Berkeley
Berkeley, CA 94720, USA
job@berkeley.edu

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188604

Abstract
The growing ease with which digital images can be convincingly manipulated and widely distributed on the Internet makes viewers increasingly susceptible to visual misinformation and deception. In situations where ill-intentioned individuals seek to deliberately mislead and influence viewers through fake online images, the harmful consequences could be substantial. We describe an exploratory study of how individuals react, respond to, and evaluate the authenticity of images that accompany online stories in Internet-enabled communications channels. Our preliminary findings support the assertion that people perform poorly at detecting skillful image manipulation, and that they often fail to question the authenticity of images even when primed regarding image forgery through discussion. We found that viewers make credibility evaluation based mainly on non-image cues rather than the content depicted. Moreover, our study revealed that in cases where context leads to suspicion, viewers apply post-hoc analysis to support their suspicions regarding the authenticity of the image.

ACM Classification Keywords
H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

Author Keywords
Image Manipulation; Fake Images; Online Images; Credibility Evaluation; Focus Study; Image Credibility

Introduction

As the abundance of hardware and software tools continues to dramatically decrease the cost and effort required to convincingly manipulate digital images, the risks and dangers associated with ill-intentioned individuals or groups easily routing doctored images through computer and social networks to cause emotional distress or to purposefully influence opinions, attitudes, and actions have never been more severe. Unfortunately, even when doctored images are eventually exposed as forgeries, their lingering impact on viewers' emotions, viewpoints and attitudes may lead to dangerous personal and/or sociopolitical outcomes. This abstract describes results from an exploratory study focusing on how individuals react and respond to images that accompany online stories in Internet-enabled communication channels (social networking site, blogs, email), as well as their ability to identify authentic or false visual information on the Web. Our exploratory study is an initial step addressing the need to better understand how and why people trust online images, and how people are influenced by manipulated images both when they are aware or unaware of the manipulation. Our eventual goal is to lay the grounds for new technologies that aid Internet users in developing a healthy skepticism toward the mediated visual hoaxes, scams, and misinformation that they receive online.

Related Work

Numerous examples have been exposed in which manipulated images lead to significant personal or societal impact. For example, in June 2010, the Economist was criticized when it published a cover photo showing a solitary President Obama on the Louisiana beach inspecting the BP oil spill. The photo was accompanied with the headline "The damage beyond the spill," alluding to potential political problems facing President Obama as a result of the oil spill. This photograph, however, had been altered to remove two other people standing alongside the President, thus conveying the misleading impression that the President was alone and despondent [8]. More recently, political division has arisen regarding the large number of refugees of the ongoing Syrian Civil War seeking asylum in Europe. Online images with negative depictions of the refugees, including association with the extremist group ISIS, have been circulated on social media websites such as Facebook and Twitter. However many of these visuals have been substantially modified and/or presented out of context. [2].

The above examples are just a tiny fraction of the instances where deliberate photo manipulation has been exposed. Photographs can strategically be used to influence public opinion, provoke strong emotional reactions in viewers, reproduce or reinforce ideology, and shape individual and collective memory [3, 10]. Empirical evidence suggests that, by invoking a false sense of familiarity, doctored images can distort people's memory — therefore enhancing the credibility of these images — and influence their decision-making [6]. For example, Photoshopped images of political candidates can have a significant impact on potential voters' decisions [1]. To exacerbate the problem, fake images often come from reputable sources (such as mass media outlets), and are further propagated through Internet and social media websites. Even when forgeries are eventually exposed, they may leave a persistent impact on individuals' memory and attitude [9].

However, we know distressingly little about how online viewers assess digital images and make judgments and decisions about their authenticity. Most research on the credibility of online information relies on predominantly textual cues, such as websites and blogs (Twitter, Facebook, etc.), but very few studies have focused specifically on image credibility [5, 4]. The social and cognitive heuristics of information credibility and evaluation have to be tested in the context of image authenticity. Furthermore, most stud-

ies assume that individuals make credibility evaluations on their own [7] , but in fact people's decisions are heavily influenced by their social networks. Such social effects are also not well understood.

In this extended abstract, we report the findings of our exploratory study that aims to understand how people react to, respond, and evaluate the credibility of images that accompany online stories.

Study Methodology

We designed and administered an exploratory focus group study where participants were asked to evaluate and reason about a set of images paired with news and stories similar to those they might encounter on on the Web. We created a set of fake images spanning a range of topics, placed them into a variety of online stories and contexts, and presented them to our participants.

Image creation

We first collected more than 40 doctored imagery that received media attention in recent years and created a chart to sort the manipulation techniques and methods applied to them. Studying these images allowed us to identify a few common manipulation objectives ranging from fabricating scenes of natural disasters to political propaganda involving negative or positive portrayals of characters. The process also enabled us to identify four common manipulation techniques used to forge online images: *composition* whereby various elements from separate image sources are combined into one new image; *elimination* by which key elements are removed from the image, cropped out or changed; *retouching* where a part or parts of the image is repaired or improved; and *misattribution* where an image is presented in an unrelated context or under false pretense.

To determine 1) how the viewers make judgments about image authenticity, and 2) what social and cognitive heuristics they use in making these judgements, we made 11 compositions using found images online. The majority of our source images were selected from the image pool available under Creative Commons license on social media site Flickr. In creating content for each composition, we employed elimination, retouching, and composition techniques. Together, the final 11 images varied from fake images illustrating national disasters, propaganda, and/or negatively or positively portrayal of subjects.

The final compositions were presented to the focus study participants as mockups, showing the medium purportedly used to disseminate the images (Twitter, Instagram, Facebook, or email), and the source varying from reputable outlets such as BBC, FOX News, and CNN, to general social media users with few or many followers. Images accompanied different commentaries or stories and if applicable revealed the number of viewers, likes, shares, or retweets. The supplemental materials for this paper include the images used for the focus study.

Focus Groups

We chose the focus group format, which has been used in similar studies (e.g., Morris et al 2012), because: 1) it allows participants to explore, clarify and mutually influence their point of view in a natural collective process, interacting with others just like they would in online environments; and 2) it is more cost-effective than individual interviews yet still gives us access to multiple perspectives.

We conducted four focus group sessions to examine in detail what social and cognitive heuristics people use in making image credibility judgments. To ensure that the results of the exploratory study were minimally biased by the political climate surrounding the participants, we conducted our focus groups sessions at two separate sites, the University of California at Davis and the University of Texas at Dallas. A total of 19 participants were recruited, four of which

Figure 1: Mockup of a fake Facebook post by BBC World News showing a destroyed building. The post attributed the debris to the aftermath of a bombing strike in Syria that led to casualties. The original image was posted to Flickr depicting the earthquake that occurred in Canterbury, New Zealand in 2010.

were men. The largest session had six participants, and the smallest had four. Participants were either given a $15 gift card (Texas) or extra credit (California) for their participation in the study. The youngest participant was 19 years old and the oldest was 30.

During each focus group session, participants were first presented with the 11 image mockups and asked to individually fill a short questionnaire about their demographic background and the extent to which they think these images were credible. Afterwards, the moderator asked participants to "think aloud" about each image and to discuss how they evaluated image authenticity.

We asked participants to complete the initial credibility rating before starting the group discussions to minimize group polarization effects and prevent the group discussion from being excessively influenced by the most outspoken participants. Recording credibility evaluations prior to group discussions also allowed participants to form their individual opinions, and ensured that opinions from individuals that were counter to the group did not get obscured. Juxtaposing individual responses and the group discussion afterward provided interesting insights on how image credibility is subject to social influence, which occurs frequently in various online environments. Further, images were discussed in a random order in each focus group to avoid any ordering effect.

Each focus group session lasted approximately one hour. Audio recordings of sessions were later transcribed and analyzed for recurrent themes.

Findings

Overall, we found that participants performed poorly at identifying the fake online images presented to them. The questionnaire asked to what extent participants were confident about the authenticity of each image, with 1 = not at all confident, 5= extremely confident. The average rating for the eleven images was 2.7, even though each image had been substantially manipulated.

After completing the questionnaire, there was a discussion within each focus group. These discussion revealed some common patterns. First, we found that the participants made judgments based mostly on non-image cues. These cues included the disseminating source of the image story, the media platform used, and/or captions and commentaries accompanying the images. Image-specific cues, such as inconsistencies in lighting and shadows, were rarely mentioned.

For our participants, the single most important factor in determining image credibility appeared to be the purported source of the image. Participants made comments such as “Looks authentic. It’s BBC,” and “Seeing CNN...it’s legit” (Figure 1). Indeed, the image credibility ratings corroborated this finding, showing that the highest-rated images were those supposedly posted by CNN and BBC, suggesting that established news networks were considered authentic. The exception was Fox News, which most participants considered to be an unreliable source, with comments such as “I don’t really trust FOX News.” In contrast, participants conveyed distrust in the authenticity of images purportedly posted on Social Media sites. Mockups attributed to Twitter and Facebook repeatedly were referred by participants as lacking credibility: “You can really post anything [on] Facebook.”

Aside from the source, the textual description or commentaries accompanying the images also appeared to play a key role in credibility evaluation or detecting a forgery. In analyzing the images, participants tended to make an assessment based on predispositions and the textual information. After the initial gravitation toward the words and applying preconceptions in examining image authenticity,

Figure 2: Mockup of a Twitter post from a fake Twitter user, Svie G, depicting an image of a battered and bruised face of man. The image was composed by layering three separate images. The swollen eye and the bloody lips were separate layers added to the man's face. His image itself was a cropped portion of a separate Flickr image from a political march in DC.

they then applied post hoc analysis to look for evidence and cues that supported their assessment. In other words, if a participant wanted to believe an image was authentic, they would find ways to justify their view. For instance, one of the mockup compositions was a doctored image of a battered and bruised face of man (Figure 2). The image was tweeted by a fake Twitter user, Svie G, who wrote "he was tased, and brutally beaten by an officer from St. Louis City police http://www.kmov.com/story/2871162." The mockup also showed that the original post has been re-tweeted once. Instead of analyzing the qualities of the image, the participants mostly shifted their attention to the information provided in the post. One of the participants mentioned, "I have read stories like this, so to me it looks quite authentic," pointing to the fatal shooting of an unarmed teenager, Michael Brown, by a police officer in Ferguson, Missouri in 2014.

Another finding was that when participants purposefully looked for clues to dismiss the authenticity of an image due to existing predispositions, they tended to fail to identify the elements of the image compositions that had actually been modified. In many cases, the participants were unable to identify any manipulation, and their solution was to assume misattribution rather than forgery.

Many of the participants found the idea of image doctoring feasible when done by experts. As the topic of image forgery was brought up by others during the discussion, many became more skeptical about the authenticity of the images presented and wanted to adjust down their initial ratings. This suggests that initial credibility evaluation is often a very hasty process and requires minimal cognitive processing. Talking to others and exposure to the possibility of doctoring images may lead them to question their initial decision and invoke a more effortful judgment. Still, at the end of the study, the participants expressed surprise when we informed them that the images were all fabricated mockups, They also conveyed that the group discussion raised their awareness of the issues surrounding image credibility.

It should also be noted that, even though we conducted focus group studies in both California and Texas, the majority of the participants described their political views as liberal, with an average of 4.79 on a 7-point scale (1 = extremely conservative and 7 = extremely liberal). We found an interesting correlation between political views and image credibility ratings: the more liberal, the less credulous (r=-.34, n.s.). However this correlation was not significant, possibly due to small sample size.

Discussion and Conclusion

Preliminary analysis of the focus group results supports the assertion that people generally perform poorly at making credibility assessments of online images. Further, non-image factors, such as the source of the image and its accompanying story, appear to play a much more significant role in participants' credibility judgment than image-specific factors such as inconsistencies in lighting and shadows.

These preliminary findings have important implications. For image creators and publishers, the most productive ways to increase credibility ratings lie elsewhere, namely, in non-image features such as the source, story content, and online media interface. For example, images posted on Twitter can be perceived credible if they have a large number of retweets, favorites and followers, regardless of the actual image content. For image consumers at large, it is advisable not to assume every image on the web is authentic. Paying close attention to image source, online interface and the congruity between the image and the accompanying story contributes to a more accurate judgment.

We chose to use only manipulated images in this small-scale exploratory study because we specifically wanted

to observe the process used by participants to determine which parts of an image had been altered. However, participants mostly failed to realize that the images had been manipulated. Further, even when told that an image had been manipulated, they failed to identify the modified regions. By using doctored images, we also wanted to avoid the possibility of a participant having previously encountered one of the images.

This focus study was limited by its small sample size and lack of demographic diversity. Moreover, the fake images analyzed in the study provided limited combinations of image content, source, and other contextual factors, making it difficult to isolate each factor's specific effect on image credibility perceptions. These limitations will be addressed in the next stage of our work, a larger-scale online experiment on Amazon Mechanical Turk. This approach will allow us to recruit a reliable, inexpensive, and demographically diverse sample. We will design the study as a series of between-subjects factorial experiments that randomly assign participants into a condition which presents them with real and forged images. Participants will have the opportunity to indicate to what degree they think the image is authentic based on various image-related (such as lighting) and non-image related (such as the source of the image) features.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS- 1444840.

REFERENCES

1. Jeremy N. Bailenson, Shanto Iyengar, Nick Yee, and Nathan A. Collins. 2008. Facial Similarity between Voters and Candidates Causes Influence. *Public Opinion Quarterly* 72, 5 (2008), 935–961.
2. Lizzie Dearden. 2015. The fake refugee images that are being used to distort public opinion on asylum seekers. *The Independent* (September 2015).
3. Robert Hariman and John Louis Lucaites. 2007. *No caption needed: Iconic photographs, public culture, and liberal democracy*. University of Chicago Press.
4. Linda A. Henkel and Mark E. Mattson. 2011. Reading is believing: The truth effect and source credibility. *Consciousness and Cognition* 20, 4 (2011), 1705–1721.
5. Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. (2012).
6. Robert A. Nash, Kimberley A. Wade, and Rebecca J. Brewer. 2009. Why do doctored images distort memory? *Consciousness and Cognition* 18, 3 (2009), 773–780.
7. Sophie J. Nightingale, Kimberley A. Wade, and Derrick G. Watson. 2017. Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications* 2, 1 (2017), 30.
8. J. W. Peters. 2010. On The Economist's cover, only a part of the picture. (July 5 2010).
9. Dario L. M. Sacchi, Franca Agnoli, and Elizabeth F. Loftus. 2007. Changing history: doctored photographs affect memory for past public events. *Applied Cognitive Psychology* 21, 8 (2007), 1005–1022.
10. Barbie Zelizer. 1998. *Remembering to forget: Holocaust memory through the camera's eye*. University of Chicago Press.