Computers & Graphics 68 (2017) 142–151

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on SIBGRAPI 2017

Humans are easily fooled by digital images

Victor Schetinger[a,*], Manuel M. Oliveira[a], Roberto da Silva[b], Tiago J. Carvalho[c]

[a] *Instituto de Informática, UFRGS, Porto Alegre, Brazil*
[b] *Instituto de Física, UFRGS, Porto Alegre, Brazil*
[c] *Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, IFSP, Campinas, Brazil*

ARTICLE INFO

Article history:
Received 26 March 2017
Revised 14 August 2017
Accepted 15 August 2017
Available online 25 August 2017

Keywords:
Digital image forensics
Behavior forensics

ABSTRACT

Digital images are everywhere, from social media to news and scientific papers. This paper describes an extensive user study to evaluate the ability of an average individual to spot edited images. By design, our study avoids lucky guesses. After observing an image, subjects were asked if it is authentic or not. Whenever a subject indicated that an image has been altered, (s)he had to provide evidence to support the answer by pointing at the suspected region in the image. We collected 17,208 individual answers from 393 volunteers, using 177 images selected from public forensic databases. Our results indicate that the average individual is not good at distinguishing original from edited images, answering correctly on 58% of all images, and only identifying the modified ones 46.5% of the time. This performance is superior to random guessing, but poor compared to results achieved by computational techniques.

1. Introduction

The idea that humans are not suited to assess the authenticity of images without the aid of tools has been widely accepted in the forensics community [1–3]. Nevertheless, there is insufficient experimental research supporting this claim. Studies on human perception of digital images have focused on very specific aspects of vision such as color [4,5], lighting [6,7], geometry [8,9], and face recognition [10–12]. However, *no extensive study has been performed to evaluate one's ability to detect editing in digital images*.

In this work, we provide evidence that supports the hypothesis that humans are not good at identifying image forgeries. For this, we performed an experiment with approximately 400 subjects. The experiment was specifically designed to avoid guessing, requiring evidence to support the subjects' answers. The results show that only 58% of the images were correctly classified as either pristine or edited, and only 46% of the edited images were identified as such, *i.e.*, *more than half of all edited images were unnoticed*. This performance is superior to random guessing, as we show in our validation, but lower than most computational forensics techniques. To make the experiment as relevant as possible for the forensics community, we used images from known public forensics datasets.

Our study differs from previous ones because it requires evidence whenever the subject believes the image has been altered (*i.e.*, (s)he should point in the suspected image region). This allows us to discard lucky guesses, and also provides insights on what subjects perceive as being suspicious in an image. To be able to gather a large amount of data, we performed an on-line experiment. Due to the uncontrolled nature of on-line tests, we apply a series of validation checks to the collected data, and discard answers containing inconsistencies. We show that our results are statistically significant.

The *contributions* of our work include:

- Experimental evidence that humans have difficulty to detect forgery in digital images, even in a context where they have been explicitly told to look for it (Section 3.1);
- Evidence that age, experience with digital images, and answering behavior of a subject, such as timing and confidence, affect one's performance when looking for forgeries in images (Section 3.2);
- A dataset of subjects' answers for real and forged images[1], with 17,208 answers over 177 images, and 8,160 image markings indicating what subjects considered to be forgeries.

In addition to these contributions, we discuss how different image features may correlate with certain types of answers (Section 3.3), and with subjects' perception of the test difficulty (Section 3.4).

* Corresponding author.
E-mail address: vschetinger@gmail.com (V. Schetinger).

[1] The dataset will be made available upon paper acceptance.

http://dx.doi.org/10.1016/j.cag.2017.08.010
0097-8493/© 2017 Elsevier Ltd. All rights reserved.

Table 1
Different answer classes for the subject study, in the notation “Image Type:Answer Type”.

Class	Meaning	Answer	Type
T:𝕋	The image is T and the subject provided a 𝕋 answer.	Correct	True negative
F:𝔽v	The image is F, the subject provided a 𝔽 answer and valid evidence.	Correct	True positive
F:𝕋	The image is F and the subject provided a 𝕋 answer	Incorrect	False negative
F:𝔽i	The image is F, the subject provided a 𝔽 answer and invalid evidence.	Incorrect	False negative
T:𝔽	The image is T and the subject an 𝔽 answer.	Incorrect	False positive

2. The user study

The goal of our study was to assess how hard it is for an average individual to determine if an image has been modified. For this, we gathered input from a large group of subjects over a large image database. Subjects are shown one image at a time and asked to provide a binary *yes/no* answer to the following question: “Is there any kind of forgery in this image?”. For simplicity, we call an *authentic* image (also referred to as *pristine* or *original*, in the forensics literature) as a T (*true*) image. Likewise, we will call a *modified* image (also denoted *forged*, *tampered*, *fake* or *edited*) as an F (*false*) image. If a subject answers *yes*, (s)he means that the image is *false*, and we call this an 𝔽 answer, as opposed to a 𝕋 (*true*) answer. In this case, the subject is asked to *provide evidence that the image has been altered*. Such evidence is given by pointing to an image region that indicates it has been altered. Different forms of evidence are considered valid, such as the altered region itself, its close surroundings, or even irregular shadows left by the forgery. *For* F *images, an answer is considered correct only if valid evidence has been provided.*

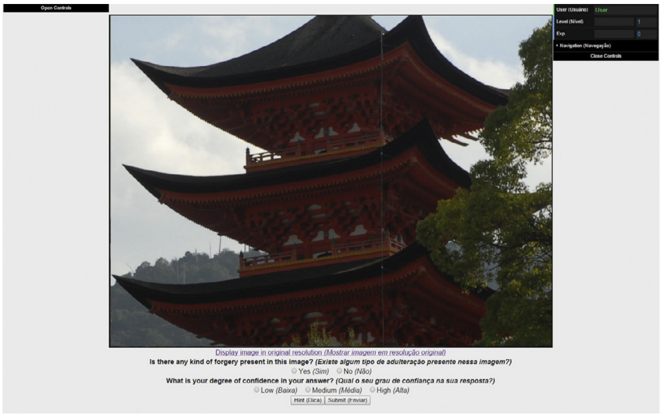
Considering all the different answer combinations, there are five possible outcomes: the image can be either T or F, the subject answer can be either 𝕋 or 𝔽, and if the subject answers 𝔽, (s)he can provide either valid or invalid evidence (Table 1).

For consistency with the forensics literature, *we treat the subjects’ answers as a binary classification problem of identifying* F *images*. Thus, a *true positive* consists of answering 𝔽 and providing valid evidence to an F image (F:𝔽v). A true negative, then, consists of answering 𝕋 to a T image (T:𝕋). A *false positive* consists of answering 𝔽 to a T image (T:𝔽). Finally, a *false negative* consists of either answering 𝕋 to an F image (F:𝕋), or answering 𝔽 to an F image, but failing to provide valid evidence (F:𝔽i) (see Table 1).

2.1. The user study

For our on-line experiment, subjects were asked to register, providing background information such as age, education, and experience level with digital images. Once registered, subjects could log in at any time to analyze and classify images, being able to interrupt and resume the classification at their convenience. The answering form consisted of a simple web page, as depicted in Fig. 1. After observing an image for at least 20 s, the subject could ask for a hint, which consists of removing a rectangular region corresponding to half of the image area not containing any editing. In the case of a T image, a randomly positioned rectangular area is used on one of the sides or the center of the image. The total area removed is always half of the image, and the image always remains contiguous. For more details about the interface of the user study, please see the supplemental material.

Each F image from the test database has an associated binary mask covering a region of the image considered as the location of valid evidence for the forgery. Such mask is called the *evidence evaluation mask* and is used for two purposes: to evaluate if the evidence provided by the subject is valid; and to determine what parts of the image can be discarded to provide a hint to the subject. The evidence evaluation masks have been created using, as source, the ground truth binary masks of pixels changed in the doctoring process of each image. To cover different kinds of evidence, the masks were edited by hand increasing the valid area. Thus, for instance, the ground truth mask on Fig. 2b only depicts the trophy added to the image. The evidence evaluation mask (Fig. 2c), on the other hand, contains a larger area. In this case, both the lack of shadows on Fig. 2b and the region around the trophy edges are also considered valid evidence. Each F image was carefully evaluated to construct its evidence evaluation mask, as this is subjective and context dependent.



Display image in original resolution *(Mostrar imagem em resolução original)*

Is there any kind of forgery present in this image? ***(Existe algum tipo de adulteração presente nessa imagem?)***

○ Yes *(Sim)* ○ No *(Não)*

What is your degree of confidence in your answer? ***(Qual o seu grau de confiança na sua resposta?)***

○ Low *(Baixa)* ○ Medium *(Média)* ○ High *(Alta)*

Fig. 1. Interface for the on-line user study. The currently evaluated image is shown at the center. Radio buttons register the subject’s answer (Yes/No) and confidence level (Low/Medium/High) for the answer. A menu (top right) displays the subject’s progress, and provides access to other options.

The data collected in the user study consists of: (1) subject answer (Yes / No); (2) evidence in the form of a click (when answered “Yes”); (3) confidence level in the answer (Low / Medium / High); (4) did the subject request a hint? (Yes/No); (5) subject observation time before asking for a hint; (6) subject observation time after asking for a hint; and (6) did the subject observed the image in its original resolution? (Yes / No).

2.2. Image database

Our image database consists of 177 images, divided into 80 true images (45%) and 97 false images (55%). The false images consist of 20 erasing images, 35 copy-and-paste images, and 42 splicing images. An *erasing forgery* consists of using brushes, blurring, or even copying some small patches to hide some portion of the original image. A *copy-paste* forgery consists of copying from, and pasting on, the same image, some region or object with or without transformations such as scaling and rotation. Finally, a *splicing forgery* consists of copying a region from an image and pasting over another image, also with the possibility of transformations (see Fig. 3).

The images used in our user study have been handpicked from three public forensics image databases: the *forensics challenge database* [13], the *splicing database* provided by Carvalho et al. [14], and Cozzolino et al. [15] *copy-and-paste* database. The total image count adding all databases is around 6000 images, with a great