Heliyon 6 (2020) e05808

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

News authentication and tampered images: evaluating the photo-truth impact through image verification algorithms

Anastasia N. Katsaounidou [1], Antonios Gardikiotis [1], Nikolaos Tsipas [1], Charalampos A. Dimoulas [*,1]

Multidisciplinary Media & Mediated Communication Research Group, School of Journalism and Mass Communications, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Keywords:
Content authentication
Digital forensics
Image tampering
Misinformation
Verification assistance algorithms

ABSTRACT

Photos have been used as evident material in news reporting almost since the beginning of Journalism. In this context, manipulated or tampered pictures are very common as part of informing articles, in today's misinformation crisis. The current paper investigates the ability of people to distinguish real from fake images. The presented data derive from two studies. Firstly, an online cross-sectional survey (N = 120) was conducted to analyze ordinary human skills in recognizing forgery attacks. The target was to evaluate individuals' perception in identifying manipulated visual content, therefore, to investigate the feasibility of "crowdsourced validation". This last term refers to the process of gathering fact-checking feedback from multiple users, thus collaborating towards assembling pieces of evidence on an event. Secondly, given that contemporary veracity solutions are coupled with both journalistic principles and technology developments, an experiment in two phases was employed: a) A repeated measures experiment was conducted to quantify the associated abilities of Media and Image Experts (N = 5 + 5) in detecting tampering artifacts. In this latter case, image verification algorithms were put into the core of the analysis procedure to examine their impact on the authenticity assessment task. b) Apart from conducting interview sessions with the selected experts and their proper guidance in using the tools, a second experiment was also deployed on a larger scale through an online survey (N = 301), aiming at validating some of the initial findings. The primary intent of the deployed analysis and their combined interpretation was to evaluate image forensic services, offered as real-world tools, regarding their comprehension and utilization by ordinary people, involved in the everyday battle against misinformation. The outcomes confirmed the suspicion that only a few subjects had prior knowledge of the implicated algorithmic solutions. Although these assistive tools often lead to controversial or even contradictory conclusions, their experimental treatment with the systematic training in their proper use boosted the participants' performance. Overall, the research findings indicate that the scores of successful detections, relying exclusively on human observations, cannot be disregarded. Hence, the ultimate challenge for the "verification industry" should be to balance between forensic automations and the human experience, aiming at defending the audience from inaccurate information propagation.

1. Introduction

Audiovisual records have been always considered as a more trusted representation of reality than written stories, resulting in the documenting role of images, utilized as evidence to convince readers about news authenticity. However, the rapid evolution of Information and Communication Technologies (ICTs), along with the massive content production and distribution facilities, offered new multimedia editing and processing capabilities, which could be easily applied by ordinary individuals. In light of the above events, controversial and unverified User Generated Content (UGC) has overwhelmed Internet and Social Networking Sites (SNSs) (Jahnke and Kroll, 2018). Moreover, while the audience has always been seeking direct access to visual scenes of the events (Huxford 2001; Mitchell, 1994; Pantti and Sirén, 2015), their ability to distinguish between real and fake/doctored pictures has not been developed, thus resulting in frequent misbelief errors and, overall, disinformation vulnerabilities (Griffin et al., 2018).

* Corresponding author.
E-mail address: babis@eng.auth.gr (C.A. Dimoulas).
[1] http://m3c.web.auth.gr/.

https://doi.org/10.1016/j.heliyon.2020.e05808
Received 24 January 2020; Received in revised form 19 July 2020; Accepted 18 December 2020

To deal with the unwanted misinformation phenomenon, the domain of Digital Forensics (DF) has emerged, forming the so-called "Verification Industry". The latter is associated with a multilevel and highly multidisciplinary effort, involving debunking sites, experts among the implicated fields, specialized tools, academic individuals, institutes and consortiums that attempt to propose feasible solutions, through their collaboration in every-day practice and/or within featured research projects (Katsaounidou et al., 2018; Katsaounidou & Dimoulas, 2018a, 2018b). Clearly, the detection of tampering trails, indicating possible content manipulation, is crucial in today's cross-validation practices, which are essential to news corporations. In this direction, various approaches can be deployed, taking advantage of both human expertise and the offered algorithmic solutions, thus forming dedicated techniques for evaluating the different types of news-reporting media assets (i.e. text, images, audio, and video). Given that journalism is heading towards the era of heightened automation, machine-driven verification assistance could accelerate the involved authentication tasks, reducing the spread of manipulated articles (Katsaounidou and Dimoulas, 2018b).

As already implied, the current work focuses on the case of image documents, analyzing the ways relevant alterations can be detected. While subjective observations and related experience are very helpful in this process, the implementation of fully- and semi-automatic inspection tools is considered equally or even more important (Krawetz and Hacker Factor Solutions, 2007; Silverman, 2013; Katsaounidou and Dimoulas, 2018b). The emphasis on the visual part is justified on the extensive use of photographic news-evidence material throughout the years of journalistic cross-validation, as well as on the progress of the Digital Image Forensics (DIF) sector, which poses increased maturity over the other sub-domains, like the ones dealing with audio and video modalities (Katsaounidou et al., 2018; Vryzas et al., 2018, 2019). More specifically, automated solutions try to algorithmically indicate "abnormal information" that human observation would/should have detected, e.g. lack or presence of (un)fitting shadows and other image reflections (Farid, 2009; Johnson and Farid, 2007; Krawetz and Hacker Factor Solutions, 2007). In semantic level, the retrieval of near-duplicate photo material, used in a different time, location and/or thematic context, could reveal possible misinformation attempts. Pixel-wise, the so-called "content-based" methods examine the structural attributes of the picture data (and metadata), aiming at detecting encoding inconsistencies that could be associated with forgery attacks (Farid, 2009; Ho and Li, 2015; Thakur, 2014; Wang, 2009; Katsaounidou et al., 2018; Katsaounidou and Dimoulas, 2018b; 2018a).

Nowadays, with the advent and the vast progression of deep learning algorithms, a new category of computer-generated fake photos has emerged, taking advantage of recent Generative Adversarial Networks (GANs) and Convolutional Neural Network (CNN) architectures (Hsu et al., 2020; Hsu et al., 2018; Hulzebosch et al., 2020; Marra et al., 2018; Yu et al., 2019). The so-called deep fake images and videos have started worrying about the scientific community, in a similar way that traditional DIF methods were the subject of laborious research efforts within the last two decades (Hsu et al., 2020; Katsaounidou et al., 2018; Katsaounidou, 2020). In this context, a new era of DIF has just begun, which, according to many researchers, is significantly differentiated from traditional approaches. Hence, photos are digitally synthesized as fakes in the first place within this modern paradigm of visual forgery, which, theoretically speaking, does not actually fall into the typical image tampering chain. For similar reasons, most classical DIF tools cannot sufficiently contribute to the detection of such manipulation trails, because they exploit different inconsistency inspection mechanisms. Fortunately, the deployment of these practices requires technological knowhow that ordinary users do not pose, thus confining the production and dissemination of such misinformation streams to the experts (Gokhale et al., 2020; Hsu et al., 2018, 2020; Katsaounidou, 2020; Thakur and Rohilla, 2020). Based on the above, it is not coincidental that recent review papers, studying real-world photo tampering scenarios (and their detection), usually omit deep fake cases (Zheng et al., 2019), which are treated as a whole new approach in our view as well. The same applies for the latest GAN image generation techniques, that are not treated as standard DIF cases in related publications (Gokhale et al., 2020; Katsaounidou et al., 2018; Katsaounidou, Vrysis, Kotsakis, Dimoulas & Veglis, 2019a; Katsaounidou, Vryzas, Kotsakis & Dimoulas, 2019b; Zheng et al., 2019). Furthermore, fake and deep fake facial images are dealt as critical in various domains and multidisciplinary research activities (Hsu et al., 2018; Hulzebosch et al., 2020; Tariq et al., 2018). However, relevant works do not underrate traditional image manipulation techniques (and their tampering detection counterparts) that can be applied by broader users' categories, i.e., non-experts in machine/deep learning (Hulzebosch et al., 2020; Katsaounidou, 2020; Tariq et al., 2018; Thakur and Rohilla, 2020). Overall, without holding the maturity and stability of traditional validation practices, these latest trends cannot be treated jointly, equally, or under the same assumptions with the preceding ones, so that they are not considered in the primary focus of the current research.

No doubt, significant research progress has been observed in the DF and DIF fields during the last decades. On the contrary, a limited number of relevant tools has been developed to be used in every -day practice of media professionals, or even to assist end-users in their own informing veracity needs. For instance, the Crowd Crafting[2] and Verily[3] websites are aligned with the human factor approach to infer conclusions (Silverman, 2013). Other tools have been designed to reveal users' identity by detecting the authenticity of a profile on social networks (e.g., the WebMii[4] and Pipl[5] platforms). Findexif[6] and Jeffrey's Exif Viewer[7] are online environments, facilitating the inspection of inconsistencies in the Exif picture metadata (from the acronym **Ex**changeable **i**mage file **f**ormat, i.e. information related to features or settings of camera sensors, location-/time-tags, etc.). Content-based image analysis algorithms have also been implemented and are offered as online services to test various falsification scenarios (Katsaounidou et al., 2018, 2019a, 2019b; Katsaounidou & Dimoulas, 2018a, 2018b; Katsaounidou, 2020; Middleton et al., 2018; Zampoglou et al., 2016). Among others, the current research focuses on the evaluation of the offered capabilities and the anticipated helpfulness of such verification assistants, therefore further technical and functional insights are given next. As already implied, the progress in the availability of such tools is limited, compared to the associated research findings, which is further deteriorated when it comes to contemporary machine and deep learning automations. The latter approaches are practically available only to technologists and researchers involved in these fields. Based on conducted literature review, related papers are constantly produced and published, studying the different image tampering and manipulation detection techniques, both traditional and contemporary (Gokhale et al., 2020; Qureshi and El-Alfy, 2019; Tariq et al., 2018; Thakur and Rohilla, 2020; Zheng et al., 2019). However, little effort is given to the adoption and evaluation of the associated tools and services from users' perspectives. In many views, this hysteresis of the practice over research deteriorates the true potentials of the algorithmic approaches (Gloe et al., 2007; Katsaounidou & Dimoulas, 2018a, 2018b Katsaounidou, 2020; Williams et al., 2018). Hence, further insights concerning the perception and utilization of real-world DIF verification assistants are pursued within this work.

1.1. Digital Image Forensics software

In general, DIF approaches fall into two major categories (Gloe et al., 2007). In the first one, the purpose is to read or extract the semantics tags

[2] https://crowdcrafting.org/.
[3] https://veri.ly/.
[4] http://webmii.com/.
[5] https://pipl.com/.
[6] http://www.findexif.com/.
[7] http://exif.regex.info/exif.cgi.

of a digital photo, i.e. the capturing time, date and location. For instance, Exif viewers are used for retrieving the relevant picture metadata, if they are available in the implicated encoding format. In the second category, the goal is to automatically detect altering trails, which could be linked to potential tampering attacks. Various strategies are deployed, usually relying on image processing and multi-modal decision making (i.e., through multiple comparison rules, applied to various visual features), as well as with the help of more sophisticated machine learning algorithms. In the latter example, the aim is to classify the different patterns of data coherence or inconsistency, implying authenticity or manipulation, through the process of "learning by example" (Korus, 2017). Overall, the mortality degree of automated image verification tools is quite high; hence, applications that were available for use in the previous year are no longer in operation. Table 1 lists the consistently available services of the last four (4) years, which were considered candidates of our investigation. This selection was made, bearing in mind that the ultimate challenge for the "verification industry" should be to balance between forensic automations and the human experience, parallelly cultivating media literacy in fact-checking practices.

On the contrary, fully automated deep-learning approaches (GAN, CNN, etc.) are generally "blind" to the users, usually treated as black boxes, without truly extending visual inspection skills. Though research efforts to physically interpret the contribution of these layered architectures are currently very active, clearly such extensions are not suited in the working scenario. Accordingly, it was decided to include only first-generation DIF tools with an investigative character, concerning the interpretation of the exposed inconsistencies. The right order of assessment (i.e., you cannot proceed to evaluate the latest/unstable solutions without studying their preceding/mature technologies first) and the availability of these specific ready-to-use services, directed this real-world scenario explicitly. Therefore, except for the available web applications, software modules and code on online repositories (e.g., GitHub) were excluded because of the absence of related interfaces, which occupy a central place in the present study. Finally, the idea of jointly evaluating multiple verification assistance means (both classical and contemporary) was discarded from the early beginning of the projects. Such a configuration would have created difficulties in the participants' guidance and training capacities, weakening the significance of the observations (which was somewhat experimentally supported afterward, based on the received comments).

From the listed utilities of Table 1, the Image Verification Assistant (IVA) toolset (Figure 1) and the "clone detection tool" of the Forensically platform were chosen for serving the needs of the current experimental procedure, as discussed above. The selection of these specific environments was decided due to their online nature (they are publicly accessible to everyone through the Web) and the user-friendly character of the associated Graphical User Interfaces (GUIs). Both platforms feature multitudes of image integrity detection techniques, including the most popular evaluation principles, i.e. noise pattern analysis, inspection of compression/quantization block inconsistencies, reverse image search (to detect near-duplicate pictures), metadata retrieval (GPS/Geolocation, Exif, etc.) and others (Katsaounidou, 2016; Katsaounidou and Dimoulas, 2018b). Furthermore, it is justified the interactive way of extracting tampering heatmaps, indicating the regions with potential forgery trails, while also incorporating various reporting options to summarize the monitoring results. An overview of the available IVA services and their involvement in the different doctoring scenarios is provided in Figure 1. Based on this diagram, some methods are best suited for detecting "object-wise" operations and others for "image-wise" manipulations (Katsaounidou et al., 2018).

Figure 2 presents an example of the "Double JPEG Quantization (DQ)" algorithm in a typical copy-move scenario (object cloning) (Katsaounidou et al., 2018, p. 121). The DQ technique evaluates the statistical inconsistencies of the underlying encoding to decide whether there are parts of the image with different compression levels, revealing possible regions of pixels from a different photo. Thus, if the original document is compressed at a level A, the duplication and insertion of an object might result in different visual dynamics (i.e., the blocks of the processed area might have lower or higher quality B, depending on the applied compression thresholds). Forensics algorithms like the DQ approach can identify the processing trails, illustrating these spatial variations with the help of comparison colormaps, like the one presented in Figure 2 (Katsaounidou et al., 2018; Katsaounidou and Dimoulas, 2018b).

In the current example of Figure 2, the forensic analysis outcome offers a somewhat clear indication of the region that processing has occurred. However, in most of the cases, the interpretation of the extracted heatmaps is not such easy or self-evident, even for domain specialists that are aware of the underlying algorithmic principles (Katsaounidou, 2016; Katsaounidou and Dimoulas, 2018b). The ability of ordinary users to comprehend photo forensics and especially to read behind the DIF representations is nowadays considered very important for the successful identification of doctored images, therefore for the detection of potential forgery attacks (Katsaounidou and Dimoulas, 2018b; Nightingale et al. 2017; Schetinger, Oliveira, da Silva and Carvalho, 2017; Gloe et al., 2007; Williams et al., 2018). The systematic review of the human role in the new assisted-verification landscape will indicate the dedicated strategies for cultivating digital media literacy in the battle against misinformation. The need to equip journalists and broadly the audience with such knowledge and skillsets is close to the core of the research questions the current paper tries to answer. To the best of our knowledge, related multidisciplinary efforts are limited in the field. The lack of such featured studies might be rationalized in the marginal role of technologically-enhanced collaborative authentication practices in the existing body of media and communication literature. However, it can be foreseen that the number of such works will be increased within the next few years to fill the interdisciplinary gaps in this demanding and highly versatile domain.

1.2. Background and related work: literature review and theoretical justification

1.2.1. The human factor

Extending the above remarks, it is vital for the forensics communities to comprehend the level at which ordinary users can identify digital

Table 1. Automated Digital Image Forensics online services.

Image Forensics Toolsets		Launch year
FotoForensics	http://fotoforensics.com/	2012
Amped Authenticate	https://ampedsoftware.com/authenticate	2013
JPEG Snoop	https://sourceforge.net/projects/JPEGsnoop/postdownload	2014
Ghiro	http://www.getghiro.org/	2014
Forensic Image Analyzer	http://www.forensic-pathways.com/forensic-image-analyser/	2015
Forensically	https://29a.ch/photo-forensics/	2015
PhotoDetective	http://metainventions.com/photodetective.html	2015
Image Verification Assistant	http://reveal-mklab.iti.gr/reveal/	2016

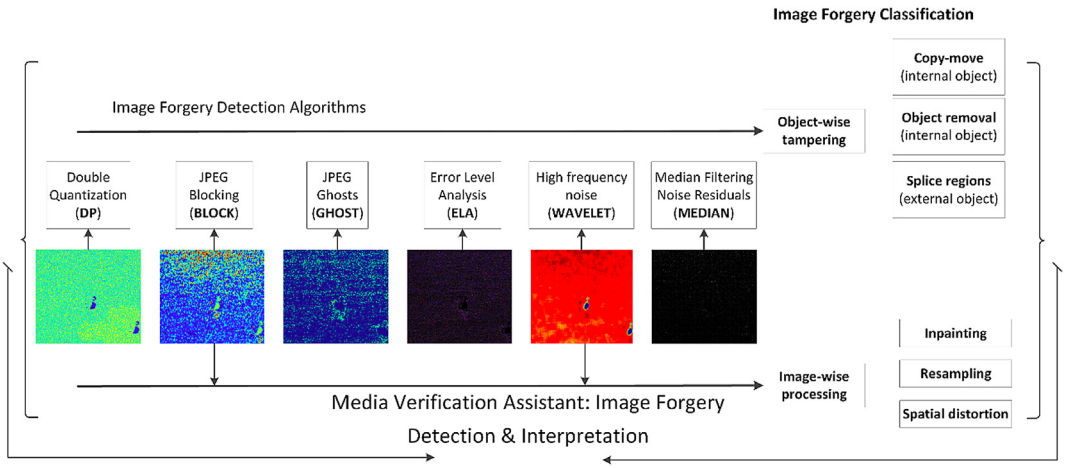

Figure 1. "Image Verification Assistant" algorithms and their use in the most common image manipulation operations (Katsaounidou et al., 2018 p. 121 p. 121).

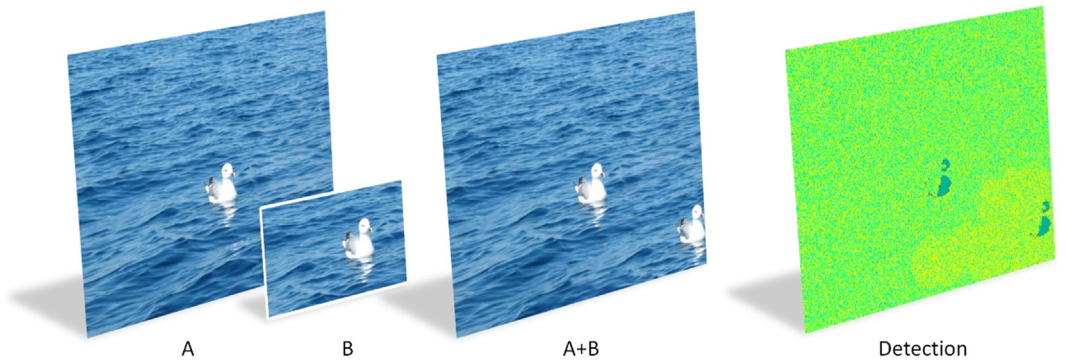

Figure 2. A step by step scenario of image forgery and the indications of the "Media Verification Assistant", while using "Double JPEG quantization (DQ)" (Katsaounidou et al., 2018, p. 121, p. 121).

image forgeries (Katsaounidou and Dimoulas, 2018b). A related study has been conducted by Schetinger et al. (2017) and found that 47% of the participants were able to spot visual modifications. Given that the specific test was entirely relying on simple computer-generated graphical interventions, someone can anticipate even poorer results for the detection of more complex geometrical inconsistencies within a natural scene. Indeed, Nightingale et al. (2017) principally confirmed the above expectations with their research on real-world photos, resulting in an approximate fake recognition score of 45%, but with noteworthy variations among the different manipulation types. Pantti and Sirén (2015) published a related survey of semi-structured face-to-face interviews with 19 Finnish journalists from leading newspaper publishers and television broadcasters, asking them to appraise the value of amateur images during breaking-news coverage. The research outcomes provided support to the central role of journalists' commitment to accuracy and truthfulness, but with many of the participants to distant themselves from having responsibility in the verification of the available UGC captures. The rest of the involved news professionals claimed that they attempt to perform or find some suitable form of authentication, indicating the need for some guidance or related practices.

Previous studies assessed the human ability to detect and locate picture manipulations, relying explicitly on subjective observation, without any algorithmic assistance. Gloe, Kirchner, Winkler & Böhme (2007) evaluated the efficiency of DIF tools, wondering whether their outcomes can be unconditionally trusted. Particularly, they attempted a critical view on the reliability of the forensic techniques for the authentication of visual documents, examining two specific cases: the resampling detection approach and the digital camera identification strategy. At the same time, the necessity of "explanation interfaces" for comprehending the DIF-outputs was highlighted, focusing on fitting User eXperience (UX) design practices that would enhance the experience (and the trust) between humans and machines. Williams, Sherman, Smarr, Posadas & Gilbert (2018), for instance, analyzed the different ways that individuals perceive information coming from other people or machines (algorithms), thus aiming at identifying implicit biasing reactions. More particularly, a dataset of authentic and tampered images was evaluated by experts (Expert Graphic Artist) and machine (Image Processing Algorithm). Participants were then asked to answer to what extent they agree with the evaluations mentioned above on a 5-point Likert scale. The researchers contacted a paired t-test to identify that no significant differences were found regarding the time needed for the subjective responses in each evaluation case. The second level of the analysis revealed that humans did not show specific trusting preferences in each of the judge types (perceptual vs. algorithmic). In specific, participants seemed to evenly agree with the analysis of the Algorithm and the Experts, when the images were untouched. In the case of altered pictures, the participants were more likely to agree with the Image Processing Algorithm than the Expert Graphic Artist. Overall, the conclusions of the reviewed research converge in the unique role of the human factor, which is considered vital in the new assisted-verification era. Appreciating technology is very important for users to accept that digital services will complete veracity tasks for them, as the present work also tries to stress, thus pointing in the direction of developing the necessary media literacy.

1.2.2. Crowdsourced image verification

Algorithms may classify a tremendous amount of content when it comes to breaking news events, but only human beings can sift through

and make sense of material efficiently (Thorne and Vlachos, 2018; Graves, 2018). Indeed, the "wise" crowd is an essential component for the verification practices, though there are continuously arising automated solutions in the related research field (Katsaounidou et al., 2018a). The idea of validating events using people's justification advantages (rational judgment, criticism, physical presence, etc.) is not new. The audience always had a strong impact on the published stories regarding their formulation and perception (Katsaounidou and Dimoulas, 2018a). It is no coincidence that the term "collective intelligence" has been used for many years (Lévy, 1997), referring to the exploitation of all the individual (cognitive) skills within a group to successfully deliver a task. Indeed, examples have shown that the best procedure for information authentication is a network of trusted sources, focusing on a specific topic area or a physical location (Silverman, 2013). Broadly speaking, "the crowd" has always been a crucial part of how news-reports are shaped and perceived. Thus, the concept of crowdsourced fact-checking on incidents and emergencies concerning the daily agenda was always there. Todays' social media technologies, like Twitter, Facebook, YouTube, and others, enable users to engage in this kind of shared decision-making on a much larger, broader, and faster scale. While many flaws can be detected in this process, still, obvious benefits are suggesting that it is better now, within this networking pipeline, rather than before, without considering the targeted audience's opinions and preferences. In this context, Krawetz and Hacker Factor Solutions (2007) argues that the most straightforward image forensic evaluations rely on subjective observations, which can identify forgeries or misclassification errors, even without the use of sophisticated analysis tools. Yet, the incorporation of easy to operate (and interpret) automated solutions could further extend the potentials of machine-assisted verification, which the current work experimentally stresses.

2. Material and methods

2.1. Hypotheses and research questions: formulation of the studies

As already stated, the current work investigates visual tampering detection through the setup and execution of two complementary studies. The focus of the first level of research is to evaluate the degree at which ordinary individuals are able to identify parts of a picture that have been doctored. An online cross-sectional questionnaire was formed to validate the following Hypotheses (H):

H_1: Ordinary users can identify some parts of images that have been manipulated, based on their subjective/visual inspection.

H_2: The analysis of the collected/crowdsourced answers would be useful to verify the originality of pictures, used as evidence documents in news-reporting.

The second study relies on the incorporation of selected image verification assistance tools, that could help in recognizing specific forgery attacks. A repeated measures experiment was chosen this time, conducted by means of personal interviews into two different small-sized groups of experts in the field (i.e., journalists and digital photography specialists). This testing procedure was carefully set-up to answer the following Research Questions (RQ):

RQ_1: Are any differences observed in the subjective evaluation of image authenticity with and without the help of DIF tools (i.e., before and after the experimental treatment)?

RQ_2: Do DIF algorithms help in the choice of the correct answer (concerning visual tampering detection)?

RQ_3: Are there correlations between decision updates/modifications and specialty?

This second study was further supplemented with an online experiment (N = 301), combining and extending the two above mentioned studies. In this case, both the ability of individuals' in identifying manipulated visual content and the usefulness of image verification algorithms were put into the core of the analysis procedure to examine their impact on the authenticity assessment task.

Overall, the two studies were designed to investigate whether professional and ordinary individuals are aware of the available image verification assisting services and their proper use. Also, the evaluation of the practicality of the tested tools (in their present form) would provide useful insights towards necessary updates, both in terms of the offered utilities and their usability, including their support and broader media literacy strategies.

2.2. Data organization: creation of a dataset with genuine and tampered images

With the advent of deep neural systems, the creation of publicly available repositories has gained much attention, including the case of doctored pictures. Although the DIF community provides a number of datasets for evaluating visual tampering detection, they differ significantly in quality and diversity of the implicated forgeries (Korus, 2017; Thakur and Rohilla, 2020; Zheng et al., 2019). Their main purpose is to train smart systems (machines) through Machine/Deep learning techniques (ML/DL), but without adapting to the needs of human visual examination. Specifically, images from datasets, recommended for the associated forensic tools, were initially regarded, such as "the--wild-web-tampered-image-dataset"[8] (Zampoglou et al., 2016), the CASIA[9] (Dong et al., 2013) and CASIA v2.0[10] datasets (Pham et al., 2019), the "Image Manipulation Dataset"[11] (Christlein et al., 2012) and the "Deutsche Welle Image Forensics Dataset"[12] (Zampoglou et al., 2016). However, in most of these repositories, numerous examples seem to be irrelevant as news-reporting documents (i.e., they cannot be used to document specific news-stories, which was the main focus of the current work). Moreover, the levels of difficulty in verifying those images (both authentic and tampered) are entirely random and not rated, which would probably cause difficulties during the interpretation of the results (also given the vast amount of the involved samples). Finally, several photos of the "Deutsche Welle Image Forensics Dataset" and "the-wild-web-tampered-image-dataset" were utilized while demonstrating the IVA platform tools. Hence, they had to be excluded to eliminate the possibility of previous knowledge for some of the candidate participants. This choice also helps to avoid similar future inconveniences, since we would like to encourage targeted audience visiting such sites and the offered help information, to train themselves in the interpretation of the DIF maps.

For this reason, it was decided the selection of representative samples containing both real and fake instances, with the latter being grouped into scalable levels of treatment, from relatively easy-to-reveal to quite demanding. The formation of this human-centric database allows us to record and weight the individual scores, occurred in each example, and also to facilitate the understanding of the underlying processing/altering operations and the associated algorithmic DIF visualizations, from the perspective of the average user. Since the task of observing images can quickly become tedious and underwhelming after a few sessions, a small-sized dataset was decided, initially consisting of 23 records in total (17 tampered and 6 authentic). The purpose of this accommodation was to help participants in fully completing the surveys, i.e. not quitting the activities/questionnaires without answering to all challenges (Katsaounidou, 2016). The formation of the specific set of samples was made after a careful and elongated validation procedure, in which, as much as three times the number of the final image samples were involved. Hence, while classical image manipulation processes were employed, the selection of the photos and the associated forgeries were meticulous, aiming at incorporating the under-examination news-documentation character.

[8] https://mklab.iti.gr/results/the-wild-web-tampered-image-dataset/.
[9] https://www.kaggle.com/sophatvathana/casia-dataset.
[10] https://github.com/namtpham/casia2groundtruth.
[11] https://www5.cs.fau.de/research/data/image-manipulation/.
[12] https://revealproject.eu/the-deutsche-welle-image-forensics-dataset/.

For instance, most of the authentic images are not very common in terms of probability of occurrence, rising difficulties in related story-verification tasks. Likewise, doctoring attacks are also linked to visual evidence that increases the viral aspect of the underlying stories, posing difficulties in the implicated veracity needs. Overall, the finally assembled testing samples were carefully picked and validated as representatives of common/real-word image forgery attacks, requiring careful inspection. In the end, the ultimately formed set is the result of a conscious and laborious process, which is considered an advantage compared to larger-size online datasets and their unrealistic instances (as long as human inspection and evaluation are concerned).

Specifically, 11 out of 23 photographs were derived from the Internet (5 authentic and 6 doctored, Figure 3), while 5 additional original pictures were downloaded and subjected to typical tampering operations, thus forced to the falsifications presented in Figure 4. The dataset was completed with the capture of 7 more photos using two digital cameras, which were again processed to generate pairs of real and manipulated instances (Figure 5). As far as manipulation kinds are concerned, splicing (i.e., synthesize a new object by multiple other photos) and cloning (i.e., object removal/insertion) processes were deployed, along with object duplication (within the same image), resampling and inpainting (Figures 3, 4, and 5). Thus, Copy-Paste operations and particularly the employment of the "cloning tool" were employed in the Adobe Photoshop CS6 environment, while resizing/spatial transforming and blurring filtering were also partially applied. The choice of the mild level of visual processing is a deliberate decision that represents an essential element of the adopted procedure. The utmost goal is to facilitate the assessment of the traceability of typical operations, as they are produced by plenary users within broadly available photo-editing software utilities. The concept behind this practice was to simulate the way common visual alterations are created, primarily as the result of splicing multiple different sources. This context also fits with the selection of the proposed DIF tools, as justified in section 1.1. Overall, the aim is to evaluate forensic tools and services that an average individual can use to fact-check potentially falsified news-photos, caused by non-experts. As already intimated, neither deep fakes nor algorithms or sophisticated code are part of this reality, since they do not belong to the majority of cases encountered in the daily agenda.

As Figures 3, 4, and 5 present, there was an attempt to include both physically implausible (e.g. *cow_fake, balloons_fake)* and generally acceptable/believable manipulations *(street_fake)*. The "impossible" scenario, for example, might depict an outdoor scene, in which some items were removed (e.g., a balloon, a bird, etc.) but their reflections remained. The way and purpose of the processing are to investigate whether users will understand the visual paradoxes. On the contrary, it is quite realistic/convincing when an object (car) is retouched in a photo in its natural environment (road). It is worth noting that a pilot test was conducted to improve validity and replicability of the stimuli used. In this pilot, multiple/alternative examples were created and reviewed within internal/formative validation procedures before the final dataset was formed (Katsaounidou, 2016). Regarding the red and yellow color frames surrounding the images in Figures 3, 4, and 5, the former indicates the samples that have been used in the cross-sectional study (n = 16) and the latter in the repeated measures experimental procedure (n = 8). It should be pointed out that the purpose of Figures 3, 4, and 5 is not to provide high-resolution views for the reader to inspect, but only an indicative overview of the testing data and the kinds/levels of tampering involved in the process. Apart from the comments that are given in the next sections, concerning the utilization and the rationale behind the selection of each file, further information with original quality images are provided (Katsaounidou, 2016, pp. 63–66, 70–88), with the respective references/sources of the initially downloaded pictures.

It has to be pointed out that the main purpose of this research is not to monitor how subjects respond to the authenticity evaluation of multiple news-photos (especially for the first part of the repeated measures experiment conducted through the interviewing test of small groups of experts). Other studies deal with this matter, in various perspectives (Gloe et al., 2007; Nightingale et al., 2017; Pantti and Sirén, 2015; Schetinger et al., 2017; Williams et al., 2018). In the current setup, stimulating images are treated as control variables that exhibit specific informatory and viralness features that are strongly linked to the studied authentication processes. Hence, these samples needed to maintain some set attributes, without risking unwanted variations due to the diverse nature of the different visual stimulus. Besides, multiple images can be easily added to control various statistical reliability measures (as it happens with the second part of the repeated measures survey, described

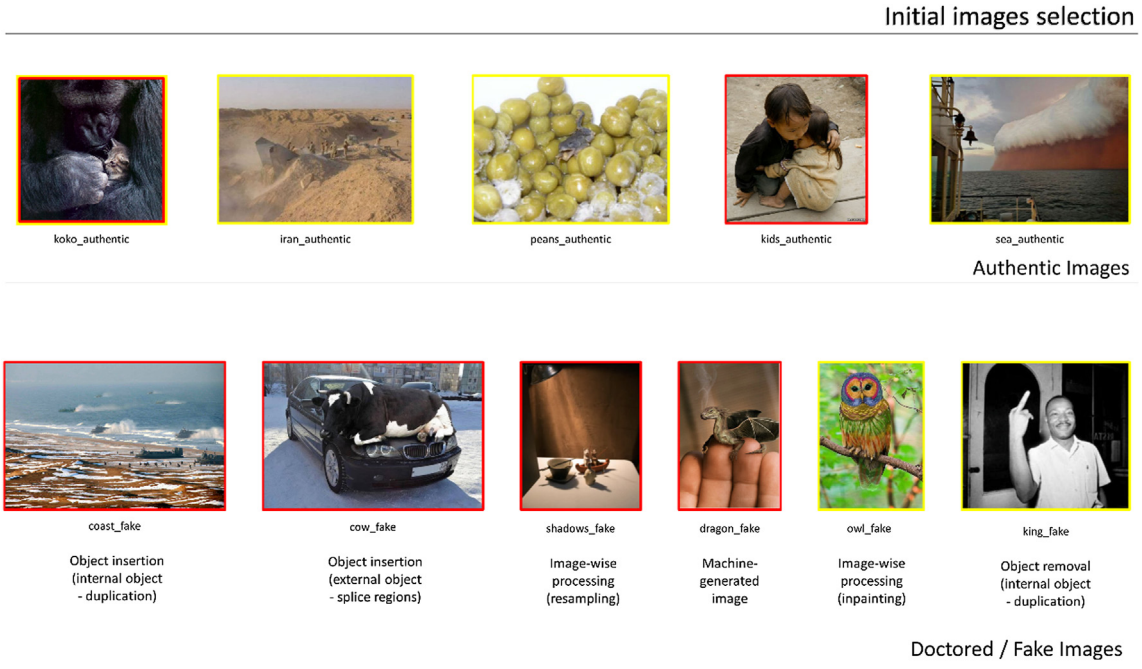

Figure 3. Initial data-set preparation: Authentic and fake photos derived from the Internet (red-framed pictures were used at the cross-sectional study, while the yellowed ones were part of the experimental treatment).

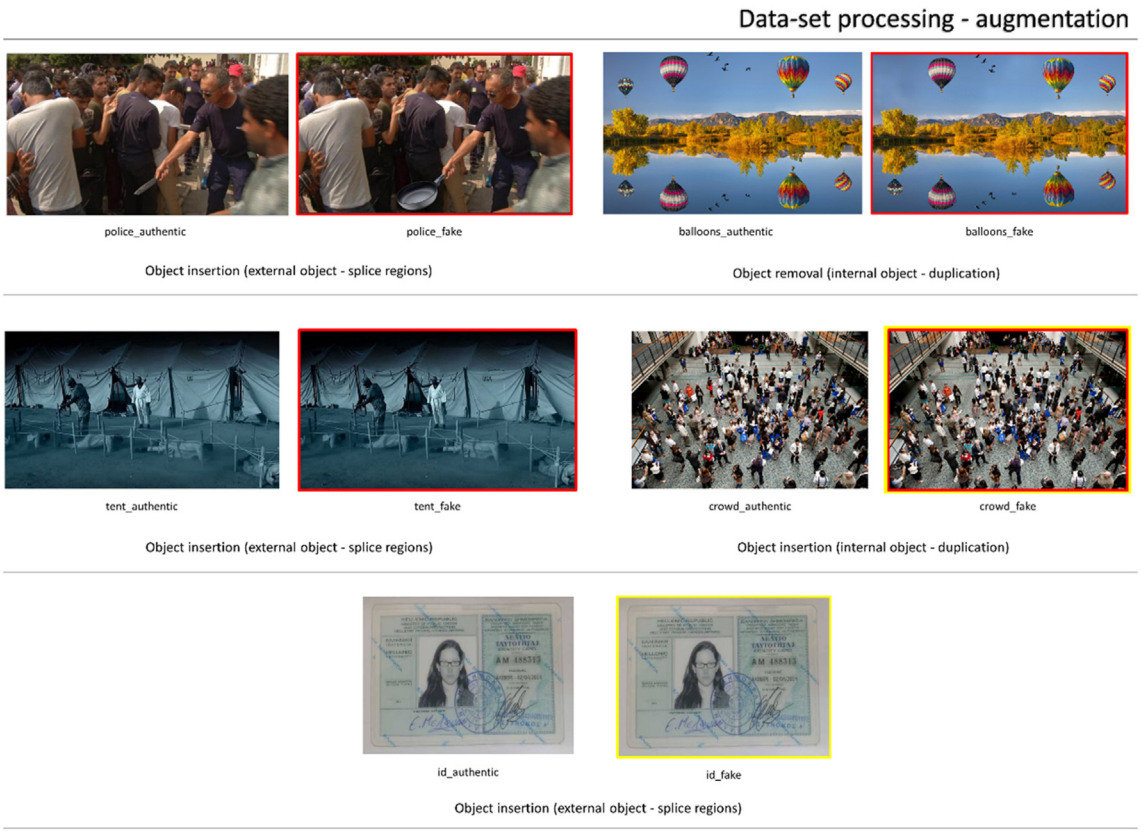

Figure 4. Dataset processing - augmentation: Photos derived from the Internet and edited for the experimental purposes (red-framed pictures were used at the cross-sectional study, while the yellowed ones were part of the experimental treatment).

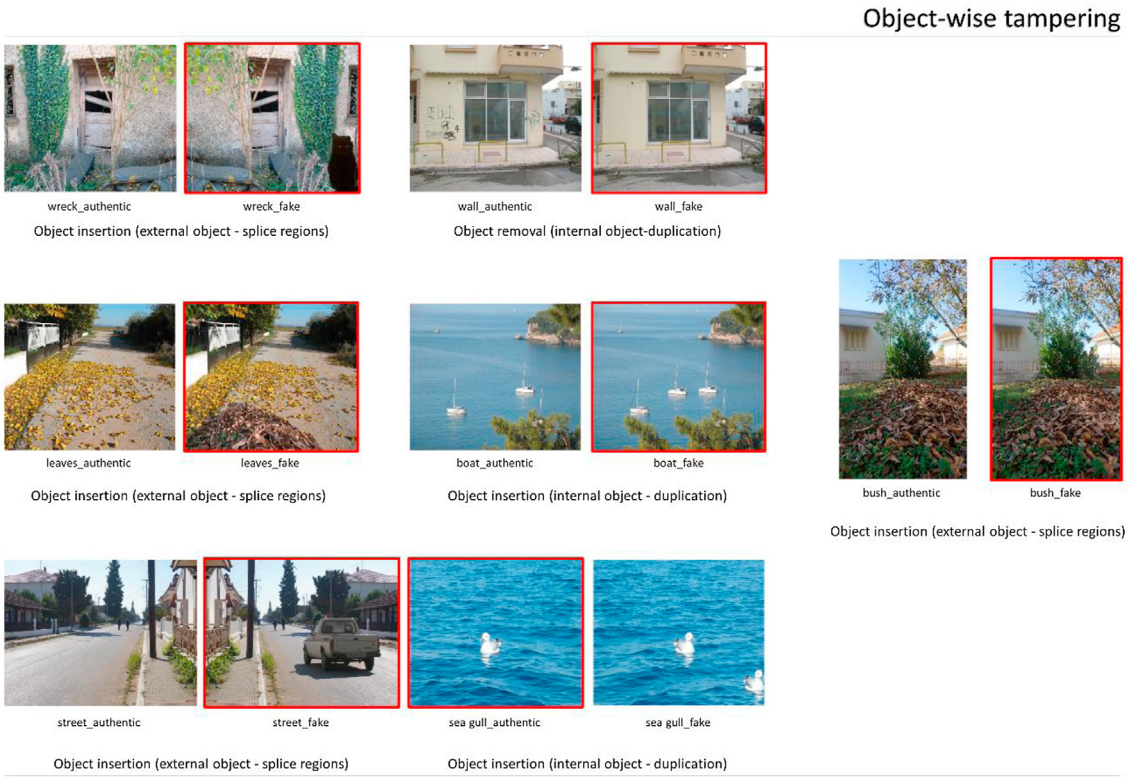

Figure 5. Dataset extension - augmentation: Additional photos, captured and processed (altered) to serve specific needs of the implemented evaluation procedure (red-framed pictures were used at the cross-sectional study).

next). Another difference is located in the employment of verification-assistance algorithms, with the prerequisite that users had to be appropriately guided first. All these procedural overheads create duration limitations, with associated constraints regarding the number of images used per session. In this context, no further alterations/variations were tested (e.g., color, resolution, etc.), as other studies do (Hulzebosch et al., 2020), investigating only specific forgery inconsistencies (and their detection mechanisms). In all cases, adequate resolutions were preferred to expedite the monitoring process, ranging from 257 × 374 pixels for small size depicting items (i.e., dragon) to larger values of 3000 × 2000 pixels (i.e., crowd)[13]. While these values were selected based on the availability of the real-world pictures (following the real-world character of the study), resizing was possible for practical reasons (i.e., in the crowd sample to accommodate image representation within the web-browser window). Apparently, users had the option to increase/decrease the viewing scale of the browsed photos, as explained in the associated experimental sections.

2.3. Cross-sectional study: image forgery detection by humans

As already implied, the aims of this study were to build a "consensus" around what kind of forgeries can be detected with the aid of human observation. Based on the content produced by the process described above, a questionnaire of 16 images (14 tampered-2 authentic) was created with the use of the widely known Google Forms platform, allowing to initiate, manage, share and complete the crowdsourcing needs of the projects.

Apart from the goal to measure people's accuracy in determining whether a photo had been manipulated or not, their ability to identify the doctored areas were also examined. In this context, human skills in recognizing common types of visual altering that are frequently applied in real-world scenarios were monitored. Participants were properly introduced to this evaluation procedure, regarding the integrity and authenticity of the testing samples, and were informed that some of them were intact while others had been edited. They were asked to respond to two different answers: *(a) Yes, the image is a product of processing, (b) No, the image is not a product of processing*. In case the (a) response was selected, a brief justification was required and the identification of the point that the tampering is located. The judgment would be relying explicitly on the inspection of the available (sole) picture, without the ability to compare between original and manipulated versions. There was no time limitation or restriction for inspecting and deciding for the status of each one of all the 16 photos.

In addition, subjects had to fill common forms of demographics and answer questions about their daily Internet browsing habits, their skills and knowledge regarding the use of personal computers and digital photography. As presented in Table 2, most participants were between the ages of 25 and 44 years old, which was somewhat expected, considering the online nature of the survey. Before the distribution of the test, a rehearsed session took place among the collaborating researchers, who completed the questionnaire to sense the level of engagement for the average user, understanding potential unwanted difficulties. While images randomization would be helpful, this option was not possible within the Google forms platform, thus it was abandoned for practical reasons. However, it was tested within the validation process, indicating that the order of the photos did not affect the received answers. After this validation procedure, the querying forms were made public, followed by focused invitations and dissemination actions, to crowdsource an adequate amount of responses for further analysis.

2.4. Repeated measures experiment: machine-assisted image forgery detection

The second experiment was designed to test the potentials of machine-assisted forgery detection and its dependence on the skills of the involved users. Two different groups of specialties were formed: a) The Image Experts, who are knowledgeable of digital photography due to education and profession, and b) the Journalists, who are/should be familiar with fact-checking/cross-validation procedures, given their experience in verifying news-events (and the associate photo documents). The aim of this arrangement is two-folded. First, to examine whether news-reporters and overall media professionals are capable of properly operating and interpreting the offered IVA tools. Second, to identify potential base-skills, possessed by image specialists, facilitating the inspection process.

Participants were requested to observe eight pictures (m = 1:8) and judge their originality (i.e., if they are considered to be authentic or the product of processing), before and after training in the use and interpretation of successive forensic evaluation algorithms (Figure 6). Following this initial step, subjects were introduced to the selected IVA services, to become familiar with the interfaces and learn how to interpret the associated forensic analysis outcomes. After this guidance, they had to re-examine the eight photo-samples with the help of seven different algorithms (ℓ = 1:7) and their extracted evaluation heat-maps. Again, randomization of images and forensic visualizations were principally considered in this experiment, but they were not applied for practical reasons. Apart from the convenience of this choice, the preparatory validation session showed that the order of the presented samples had an insignificant effect on the responses. Nevertheless, such randomization aspects were incorporated in the second/supplemental part of the repeated measures experiment, presented next. The primary goal of this testing procedure was to find out if the employed tools were helpful, reinforcing (or not) people's original impression about the veracity of the images. The block-diagram with the dataflow describing the successive inspection sessions of the repeated-measures experiment is depicted in Figure 6. Table 3 provides a high-level/basic overview of the utilized IVA techniques.

Following the interview sessions with the selected experts and their proper guidance in using the tools, the repeated measures experiment was also deployed on a larger scale through an online survey (N = 301). A related web service was developed, taking advantage of the offered web-API of the IVA platform, utilizing algorithms 1–6 of Table 3. While additional DIF tools are currently offered within this environment, it was a conscious choice to focus on these basic algorithms for two main reasons. First, to avoid confusing participants with excessive DIF data and related help information (the feedback received during the previous actions validated that the number of tools and images should remain at low rates). Second, to have the possibility of matching the observed measures against the one received in the first part of the experiment. The above choices (and especially the second one) are directly coupled with the scope of this complementary investigation. Specifically, the new survey allows testing the formed set of images and the implicated algorithms within a more statistically reliable set, thus strengthening the validity of the observations. It also offers the opportunity to include new scales for measuring participants' confidence and perceived decision difficulty per image, with or without algorithmic assistance. In this context, it was decided to incorporate additional testing photos that could control the whole process, serving the requested comparisons. The case of facial pictures was thought as a fitting choice, given their popularity in relevant authentication tasks and contemporary competitions, their usefulness in a variety of news-documentation needs, falling into the terms of the

[13] Resolution of images used at **A) cross-sectional study**: 1. wall: 1000 × 750; 2. boat: 1000 × 750; 3. gull: 1000 × 750; 4. crowd: 3000 × 2000; 5. wreck: 1000 × 750, 6. street: 1000 × 667; 7. leaves: 1000 × 750; 8. bush: 1000 × 1500; 9. cow: 600 × 399; 10. coast: 2000 × 1328; 11. tents: 620 × 356; 12. police: 599 × 348; 13. balloons: 1920 × 1200; 14. shadows: 433 × 533; 15. dragon: 257 × 374; 16. kids: 403 × 403, and **B) repeated-measures experiment**: 1. owl: 346 × 484; 2. peans: 578 × 433; 3. koko: 489 × 480; 4. sea: 591 × 435; 5. king: 570 × 500; 6. iran: 592 × 447; 7. id: 810 × 608; 8.crowd: 1280 × 853.

Table 2. Participants' demographics for gender, Age, Education, Daily Internet Usage, Devices (cross-sectional study).

Factors	Answers	Frequency	Percentage (%)
Gender	Female	64	53
	Male	56	47
	Total	120	100
Age	16–24	31	26
	25–44	75	63
	45–64	14	11
	Total	120	100
Education	Elementary	2	2
	High School	20	17
	Graduate	59	49
	Master's Degree Student	39	32
	Total	120	100
Daily Internet Usage:	0–1	11	9
	1–3	36	30
	3–5	42	35
	5–10	3	26
	Total	120	100
Device	Mobile Phone	48	40
	Personal Computer	67	56
	Tablet	5	4
	Total	120	100

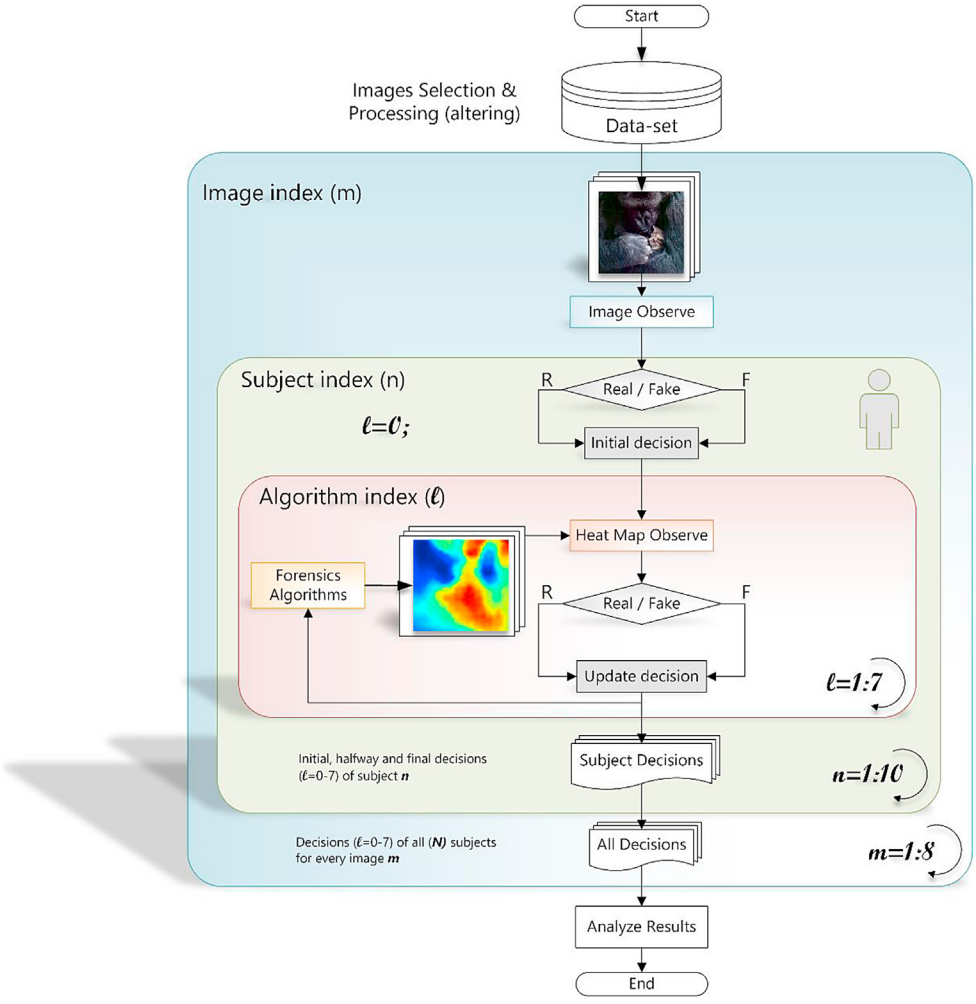

Figure 6. Block diagram with the dataflow of the repeated-measures experimental evaluation procedure.

Table 3. Overview of the seven algorithms utilized at the current experiment (Katsaounidou et al., 2018) (Retrieved by the Image Verification Assistant and the Forensically websites: http://reveal-mklab.iti.gr/reveal/, https://29a.ch/photo-forensics/).

No	Name	Description
1	Double JPEG Quantization Inconsistencies	The DQ algorithm inspects the traces left by consecutive JPEG compressions on an image. When a spliced region from one photo is inserted into another with different compression history/levels, the discrepancy may be detected by this algorithm.
2	High-Frequency Noise	This algorithm decomposes the image into multiple scales/levels of approximations and details using the 2D Wavelet Transform. The presence of significant/localized visual variances indicates differentiated photo histories, i.e. dissimilar patterns of invisible high-frequency noise (due to different capturing devices, add-on image filters, other processing, compression, etc.).
3	JPEG Ghosts	This algorithm recompresses the input JPEG image at all possible qualities (65–100%) and subtracts the produced results from the original data. If an area has been inserted from a different photo (with different compression), a quality gap (i.e., Ghost) should appear at the level that the splice block was originally compressed.
4	JPEG Blocking Artifact Inconsistencies	This algorithm searches for artifacts related to JPEG block inconsistencies, indicating traces that have been left during the procedures of splicing, copy-moving or inpainting (again, due to the different compression quality/level).
5	Error Level Analysis	This algorithm removes a recompressed version of the JPEG image from the image itself. The process is similar to "JPEG Ghosts" but only a single version/quality level of the image is subtracted (75%) to display areas with different noising patterns that might pinpoint the potential falsification.
6	Median Filtering Noise Residue	This algorithm also relies on the detection of altered high-frequency noise patterns due to the enforced tampering. Median filtering is applied to estimate the de-noised image, which is subtracted from the original one. If areas of similar content exhibit different intensity residues (i.e., noise), it is likely that the specific region has been originated from a different image source. Due to the stochastic nature of noise, this method is considered as an unreliable tampering estimator. Hence, it should be rather used in combination with other algorithms and not as an independent detector.
7	Clone Detector (Forensically)	This algorithm highlights similar regions within an image, which can be a good indicator that a picture has been manipulated with cloning operations.

studied scenario, and the availability of related datasets (Hsu et al., 2018; Hulzebosch et al., 2020; Marra et al., 2018; Tariq et al., 2018).

Based on the above, the "Real and Fake Face Detection dataset"[14] was finally chosen, provided by the Computational Intelligence and Photography Lab in the Kaggle online community of data scientists and machine learning practitioners, which hosts/promotes related competitions. Among the conveniences (and advantages) of the specific dataset, it was justified its relatively small size (i.e., 2,041 photos in total, 1,081 real and 960 fake), facilitating the quick inspection and validation of all samples. Moreover, fake photos are rated in three difficulty levels (easy, mid, hard), which could be a useful indication for our formed images to compete against. Another important aspect is the relatively high image resolution (600 × 600), which is comparable to the ones of the initial "8 photos". Finally, "*the dataset contains expert-generated high-quality photoshopped face images,*" which aligns with the current approach, making the selected IVA algorithms applicable. Moreover, the dataset creators support that manual face tampering was preferred over automatic computer-generated fake images (i.e., through a GAN model). They claim that this expert-level fake face formation was a conscious choice, because the automated GAN "*patterns can be futile in front of human experts, since exquisite counterfeits by experts are created in completely different process*". Apart from this argument, to which we totally agree, the incorporation of fully automated visual falsification and its algorithmic authentication seems to be more vulnerable to counter-forensic or anti-forensic techniques, that aim at concealing the forgery tracks, thus making harder the proper forensic investigation (Katsaounidou et al., 2018; Qureshi and El-Alfy, 2019). As already inferred, such deep-fake creation and verification approaches are blind, inappropriate for cultivating related media literacy to humans, therefore unfitting to the current research goals.

The deployed web-service functionalities allow the random selection of multiple images and their corresponding DIF heatmaps, which users had to evaluate. Apart from the eight (8) formed photos (4 authentic, 4 tampered), represented as yellowed-framed thumbnails in Figures 3, 4, and 5, 200 Kaggle face pictures were also randomly selected (100 real and 100 fake: 33 easy, 34 mid and 33 hard). Based on the experience of the in-person repeated measures with the 5 + 5 interviewed experts (first part of the experiment) and after a laborious validation procedure, it was decided the selection of five (5) pictures per session (i.e., each subject would have to respond on five images). More specifically, each test is assembled with three (3) samples of the initial "8-p dataset" and two (2) face photos (1 real and 1 fake), all randomly selected. The raw image size was, in most cases, around 600 × 400; however, all visual items were presented equally sized on the user interface of the survey environment to simplify the evaluation process. Participants had to vote for the authenticity of each image before and after the algorithmic assistance, using the following six-levels scale {-3, -2, -1, 1, 2, 3}. The sign of the response denotes the authenticity of the tested photo (negative/- means fake, positive/+ means authentic), while the absolute value indicates the confidence of the evaluators. In the end, an overall difficulty level for each picture was provided on a five-point Likert scale (i.e., 1–5). Apparently, it was not possible to guide the participants in the proper use of the DIF algorithms (the reason for which the first part of the experiment was initially developed). Instead, access to the IVA directions/help files was enabled (both at the beginning of the process and during the testing), imitating the real-world situation. The duration of each evaluation step was recorded as an indication of users' activity to be self-guided and comprehend the principles of the involved algorithms. The survey was disseminated to under-graduate Journalism students and the page of the Greek debunking site Ellinikahoaxes[15], thus targeting users (generally) involved in the news-verification processes. Six (6) simple questions were initially asked to test the expertise, a) in digital images (inquiring knowledge of the terms block, EXIF and experience on photo-editing software) and b) in media (checking

[14] https://www.kaggle.com/ciplab/real-and-fake-face-detection.

[15] https://www.ellinikahoaxes.gr/.

Table 4. Participants' demographics for gender, Age, Education and Profession (supplemental repeated-measures experiment).

Factors	Answers	Frequency	Percentage (%)
Gender	Female	159	53
	Male	121	40
	Decline to respond	21	7
	Total	301	100
Age	<18	2	1
	18–28	162	55
	29–39	67	21
	40–50	52	17
	>50	18	6
	Total	301	100
Education	Elementary	6	2
	High School	93	30
	Graduate	136	46
	Master's Degree Student	53	18
	Doctoral Degree	13	4
	Total	301	100
Profession	Unemployed	22	7
	Student	151	51
	Employee	89	30
	Employer	34	11
	Retired	4	1
	Total	301	100

familiarity and previous experience on Gatekeeping, frequency of fact-checking and related attitude on social media). Accordingly, subjects were classified to Image and Media Experts (to have an analogy with the first part of the experiment), as well as to Average Users (if the received answers were not supporting any of the two expertise). Participants also filled common forms of demographic information synopsized in Table 4 and representing the survey identity.

3. Results

3.1. Cross-sectional study results

The purpose of the cross-sectional study was to evaluate the human perception in detecting and spatially locating possible image forgery attacks. A number of 120 subjects in total completed the online questionnaire that was described in the previous section, evaluating the authenticity of the 16 pictures. Participants provided (as asked) some basic rationalization of their judgment, in cases that tampering was

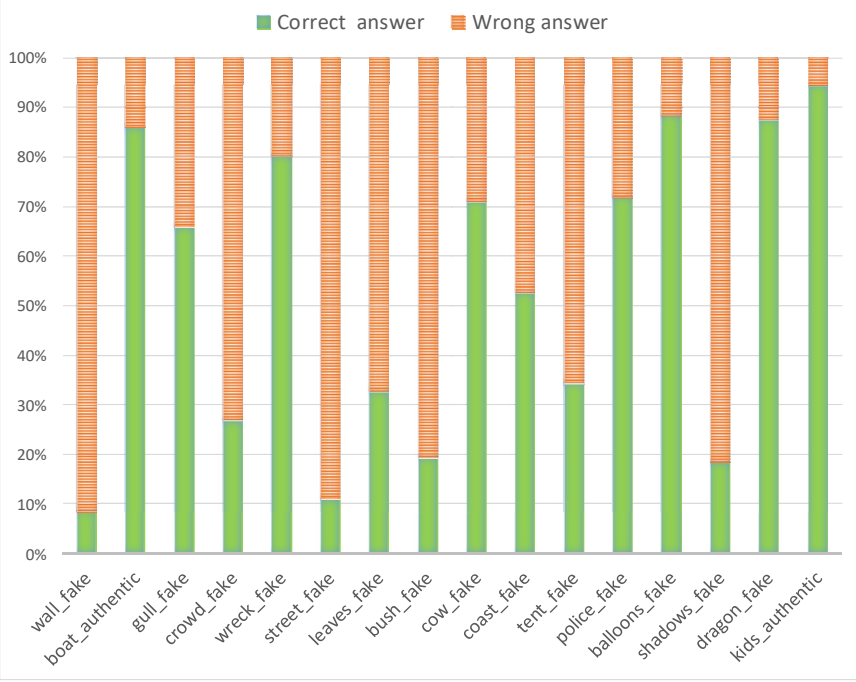

Figure 7. Authenticity recognition accuracy per image.

decided, indicating the potential areas that editing was applied. The percentage scores of correct and wrong answers are depicted in the diagram of Figure 7, using the indexes with the sample-names of Figures 3, 4, and 5.

Statistically, 9 of the 16 photos were evaluated correctly and the remaining seven (7) samples were mostly misclassified. The images with the highest recognition rates (80% and above) are the "*boat_authentic*" (103/120 correct answers, 85.8%), the "*wreck_fake*" (*objects insertion*, 96/120 correct answers, 80%), the "*balloons_fake*" (*object removal*, 106/120 correct answers, 88.3%), the "*dragon_fake*" (*machine-generated,* 105/120 correct answers, 87.5%), and the "*kids_authentic*", which displayed the highest rank (113/120 correct answers, 94.1%). In contrast, the pictures with the lowest rates are the "*wall_fake*" (*object removal,* 10/120 correct answers, 8.3%), the "*bush_fake*" (*object insertion*, 23/120 correct answers, 19.2%), and the "*shadows_fake*" (*object transform/rescale,* 22/120 correct answers, 18.3%). It is worth noting that the associated false rates in these occasions range from 80-81%. Finally, the "*street_fake*" image (*object insertion*, 13/120 correct answers, 10.8%) proved to be the most challenging case, for which the inability to identify the tampering was 89% (107/120) incorrect answers.

Based on the analysis conducted at the end of section 2.2, regarding the necessity of a small set of pictures with very particular and differentiated, among them, forgery or authenticity attributes, the interpretation of the results on a per-sample basis was considered central and rather indispensable. In this context, it is also further justified the applicability and complementarity of the assembled photo collection. Overall, it proved that people's ability to identify tampered photos depends on the types of encountered manipulation, combined with the informatory and viralness dynamics of the showing news documents. Particularly, in the "*boat_authentic*" case, only a small percentage of the participants (14.2%) misjudged the picture as tampered, justifying their decision to the uniformity of the sea, the light and sharpness of the image, and its vivid colors. The fact that seventeen users (17 of 120) classified this sample as edited verified previous research findings, according to which, when an illustration gathers the aforementioned impressive features, it is very likely to be considered as processed (Schetinger et al., 2017). In most cases, this unwanted behavior (or even habit) emanates from bias-generating factors associated with the lack of confidence in inspecting visual content.

Likewise, in the special case of the "*wreck_fake*" picture, an item (cat) has been inserted in a clumsy way. The above editing has led 80% of users to its successful detection. However, none of the participants realized that the photo had a more complex tampering history. Specifically, apart from the main alteration (the added cat), two more treatments were also applied in a more sophisticated way (i.e., the two cushions were overlaid by a different frame capture, with a different camera). The above behavior proves that subjects, after founding that the image was corrupted, they were not interested in observing it further.

It seems that the philosophy of processing the "*balloons_fake*" image led the participants to respond correctly on an 88.3% rate, with only the remaining 11.7% marking this photo as authentic. More specifically, the vast majority of subjects were very successful in detecting the illogical violations in the reflections of the duplicated area. However, the overall vivid/flawless nature of the image, with the very intense colors and the perfect harmony, also played an instrumental role in the "fake" judgment (which is not necessarily aligned with the right decision, i.e., in case that object removal was not performed).

The "*dragon_fake*" is a machine-generated image with weird (science fiction) content, which led 87.5% of the sample to mark it as fake. However, it worth noting two very interesting comments, made by the participants. The first one supports the authenticity of the picture in case the depicted dragon would be the capture of a statue/art installation. The second points out the possibility of the photo-shooting of an actual small-sized living lizard, in which the tobacco smoke was naturally inserted, i.e. with the air of a cigarette blow-out.

The "*kids_authentic*" sample refers to a well-known photo of the Vietnamese photographer Na Son Nguyen, which has been shared by hundreds of thousands of Facebook users with inaccurate descriptions[16]. While it was unexpected that the majority of subjects would recognize it as fake (113/120 correct answers), this also proves that participants were aware of this famous hoax.

Regarding the cases that gathered the highest erroneous responses, the "*wall_fake*" image holds the lowest score. Subjects failed to detect the image distortion, probably due to the absence of visual indication for removing the painted parts of the wall (presented in Figure 5). A small percentage of 8.3% came to the right conclusion but for the wrong reasons. Based on user justifications, the two front yellow bars (and their small differences) led some people to respond that this is a treatment product. The above proves that over perceptive inspection can lead to unsuited or even compulsive observations that will eventually result in erroneous decisions.

The "*bush_fake*" sample also garnered a high percentage of wrong responses (80%), despite the existence of its strong visual indications (i.e., a new bush-object has been inserted next to the existed/authentic one, in a rather obvious manner). Moreover, only one (1) of the twenty-three (23) subjects who detected the treatment, also perceived the careful splicing of a second section (leaves) that came from a different photo.

The "*shadows_fake*" example depicts a scene with many small objects having processed shadows, which only 18.3% of the participants perceived. The remaining 81.7% believed that the image was not treated, probably due to the fine alterations located in the visual details.

In the "*street_fake*" case, only thirteen users (13 out of 120, 10.83%) managed to detect the rather "strange"/unfitting nature of the inserted car, which was derived from a different photo. As already argued, when an object (car) is retouched in a new image, which is close to its natural environment (road), it is quite plausible and often difficult to raise suspicions.

Except from the manipulation types, that proved to play an instrumental role in shaping common decisions, associations were also observed among specific sessions, in which users' characteristics revealed a strong influence. In the case of gender, positive correlations were observed with the responses in *police_fake"* ($r = 0.25$) and "*kids_authentic"* ($r = 0.23$). A related interpretation is that women presented a higher percentage of correct judges on these two occasions. Furthermore, education level did not appear to statistically affect the answers, except for the "*balloons_fake*" case ($r = -0.20$). Hence, the higher the subjects' degree the greater the chance of detecting tampering in this example, which seems quite reasonable since the latter is justified on the geometry of the involved objects-shadows.

Another variable that worths further investigation is the device that subjects used during the testing procedure. As shown in Figure 8, in which all the discussed correlations are graphically plotted, related positive correlations were found in the cases of "*police_fake*" ($r = 0.33$), "*balloons_fake*" (0.23), "*tent_fake*" ($r = 0.20$), and "*wreck_fake*" ($r = 0.19$). These findings imply that users who filled out the survey utilizing mobile phones (therefore small screen) had less correct responses.

Finally, Age and Daily Internet Usage do not affect individuals' ability to inspect and assess image authenticity, except for the case of "*wreck_fake*" ($r = -0.19$). The analysis showed that the more time someone spends online, the more likely he/she is to answer this question correctly. As already mentioned, all these correlation findings are depicted in Figure 8.

[16] **Nepal Earthquake Photo** Fauxtography: Photograph purportedly shows two children holding each other for comfort in the aftermath of the 2015 Nepal earthquake. Source: https://www.snopes.com/fact-check/nepal-earthquake-photo/

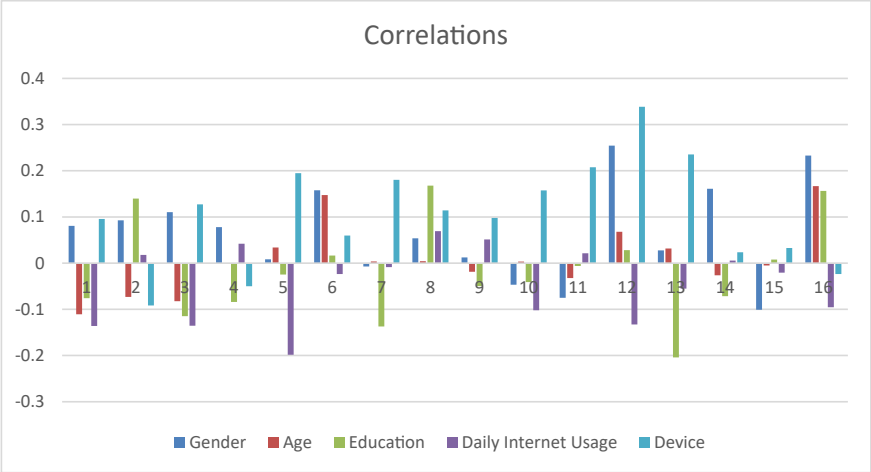

Figure 8. Correlations between the Authenticity Evaluation Results and the variables Gender, Age, Education, Daily Internet Usage, and Device in the different testing sessions (1–16).

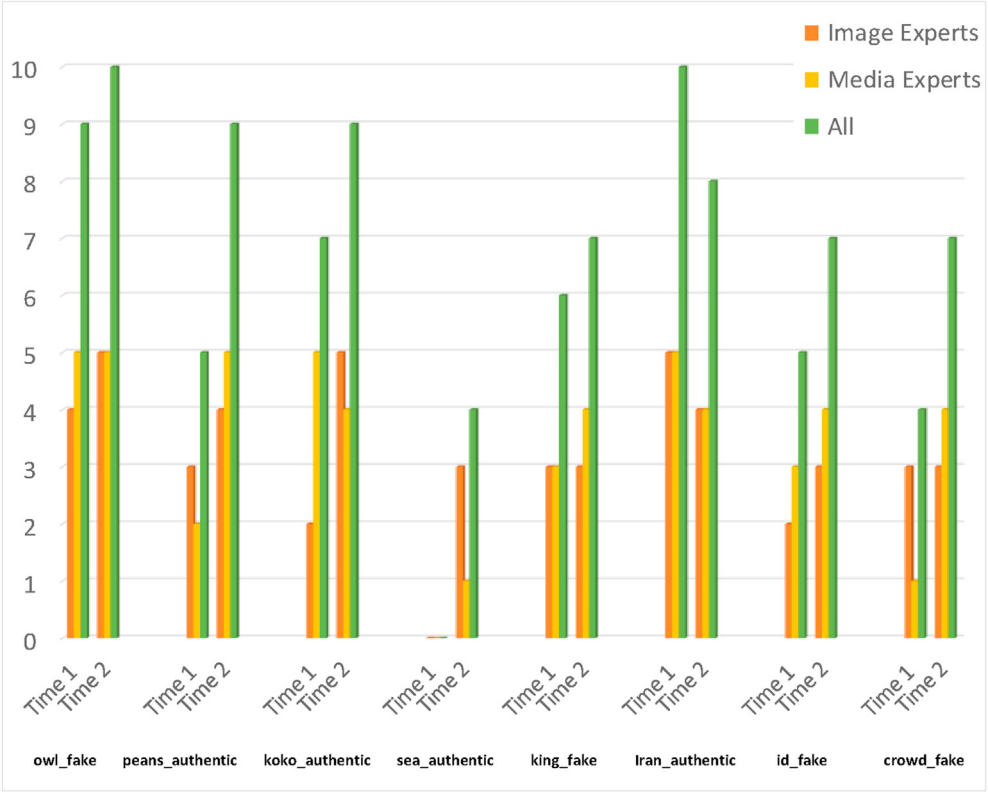

Figure 9. Authenticity evaluation correct answers before (Time 1) and after (Time 2) the experimental treatment (i.e., machine-assisted evaluation with the help of IVA tools).

3.2. Repeated measures experimental results

As already explained, the repeated-measures testing was conducted to weight the contribution of selected IVA tools. For this purpose, a small number of experts (n = 10) were organized in the two different specialty groups: Image (N = 5) and Media Experts (N = 5). Initially, participants were asked about their prior knowledge of DIF algorithms and specifically in the IVA toolset (3/10 were aware). Then, they were involved in the successive inspection sessions of identifying possible manipulated images. The whole process was carried out in the form of short interviews with a supervising researcher to provide necessary guidance on the proper use of the algorithms. Figure 9 depicts the initial (Time 1/T1) and final (Time 2/T2) authenticity estimations before and after the experimental treatment (i.e., with the help of the algorithmic solutions) for the three formed categories: Image Experts (orange bars), Media Experts (yellow bars), All (green bars). It can be noticed that machine-assistance improved the evaluation results since an overall boost in the T2-scores is observed. An exception to this rule is seen in the "*iran_authentic*" case, which in general displayed the worst evaluation results, as it is further analyzed below. Another minor remark is that the correct T2-answers are fewer (by one) than the corresponding T1 ones for the "*koko_authentic*" picture, only in the media group (it seems that one of the journalists mistakenly altered the early vote, falsely driven by the extracted forensic colormaps).

Further analysis in the entire group of participants (green bars) can also provide some interesting insights, considering both the type of the image (and its evaluation difficulty) and the rationalization of the answers, provided by the interviewees. Following the argumentation conducted for the cross-sectional study, regarding the careful qualitative interpretation of the observed outcomes, a per-sample analysis basis was also adopted here. Specifically, in the first "*owl_fake*" sample, 9/10 users initially responded (T1) that this picture is edited, with a high level of confidence. They justified their judgment to the vivid colors and the recognition of two quite distinctive layers (foreground and background), which was commented by all interviewees. While the single wrong vote also located this finding, it erroneously associated it with lens focusing issues. The extracted forensic analysis heatmaps further highlighted the boundaries of the two superimposed regions, so that all T2-votes converged that the owl object was inserted from a different photo.

In the second case (*peans_authentic*), the T1-responses were equally distributed between positive and negative answers (50-50%). The utilization of the algorithmic guidance was catalytic, since 4 of the participants alter their initial decision, boosting the T2-score to 90% successful assessment. Despite this evaluation shift, there was not a single remark supporting that the observed colormaps indicated evident traces of tampering.

The third case (*koko_authentic*) started with a high T1-score of correct answers (70%), with seven users (7/10) identifying that this picture was manipulated. The recognition accuracy was further boosted in the T2-session to the almost perfect level of 90%. The single mistake was made by a media expert who erroneously altered the initial response due to the absence of strong inconsistency boundaries in the extracted error/noise analysis heatmaps (ELA, Median Filter). This unexpected turn can be explained by the lack of knowhow concerning the underlying algorithmic principles and the interpretation of the extracted outcomes.

The next image (*sea_authentic*) failed to collect a single correct answer before the experimental treatment (0% score at Time 1). All subjects were affected by the visual content, i.e., the impressive and unusual appearance of such intense natural events (i.e., the huge sea-wave), justifying their decision primarily on the appeared reddish cloud. While reviewing the forensic analysis maps, most participants were rather disoriented, focusing their attention on a different point (the bell), which seemed to be cloned from a different file (based on the interpretation of the IVA indications). However, four (4) users altered and actually corrected their responses, even without particular reasoning (just based on cumulative intuition). Again, the results showed that machine-assistance is not such self-evident or entirely credible, and, in all cases, further effort is needed to train people on the proper interpretation of the algorithmic outcomes.

The "*king_fake*" photo depicts Martin Luther King in an unappropriated gesture, initially dividing the audience into their responses. Thus, in the T1 session, four participants (4/10) judged the image as authentic (erroneously), and the remaining 6 considered it as a product of photo-editing. It seems that different political beliefs and prejudice directed the individual answers, as the tailed discussion also indicated. A substantial portion of 40% neglected to count the weird nature of the image, so that the authenticity evaluation was probably driven by instinct. Even in the view of the quite clear forensic maps, revealing rather strong boundaries/edges of potential inconsistencies around the contentious hand/finger area, only one (1) of the four (4) subjects corrected the T2-vote to "non-authentic" (Katsaounidou, 2016). A simple interpretation of this specific example points out the difficulty of changing solid/well-shaped opinions, even with the use of sophisticated algorithmic solutions.

A worth mentioning remark was found in the "*iran_authentic*" case. Initially (T1), all interviewees voted correctly that the picture was authentic. However, in the view of the IVA heatmaps, two (2) of the subjects erroneously updated (T2) their decision to non-authentic. Interestingly, this vote shifting was present in both participating groups (Image/Media experts), without being able yet to pinpoint this incident to random or specific causes.

The "*id_fake*" example has some analysis similarities with the "*king_fake*" case. The answers were divided into the T1 session (50-50%), whereas the comprehension of the algorithms led to a slightly boosted score during T2 (70%).

Finally, in the "*crowd_fake*" case, four (4) users judged the picture as fake (correctly) and the rest six (6) as original (mistakenly). The challenging part of this example lies in the plurality of items, directing the participants in the detection of object-based manipulation. However, the utilization of the IVA tools doubled the score of the correct answers (from 40 to 80%).

One could argue that after getting acquainted with the proper usage of the machine-driven techniques and the interpretation of the associated forensics colormaps, users would improve their responses. Update knowledge (Ku) is a related parameter that expresses the improved accuracy in the after-treatment session, which can be compared to the corresponding base knowledge (Kb), i.e., before the experiment (Kalliris et al., 2011; Kalliris et al., 2014a, 2014b). According to Figure 9, Ku values are consistently higher than Kb, with their difference being increased as (testing) time passes by, meaning that the familiarization with the IVA processes advances. It is rather not accidental that the second-best positive evaluation shift (30%) is observed in the last example (*crowd_fake*). On the contrary, cumulative tiresome of the whole testing procedure and the quite complicated content of this image (containing many visual details) may be the causes for falling below the best transition (40%), which was discerned in the quite easier (concerning picture elements) "*peans_authentic*" sample.

As far as the intrinsic specialty-rates are concerned, i.e., the scores within each group, it seems that media experts overall gathered higher scores (according to Tables 5, 6, 7, and 8). While their first correct answers were fewer than those of the image specialists, the updates after the experimental treatment led to better and more stable reactions. Based on these indices, someone can argue that forensic assistance was beneficial to the journalists, even to a small degree. Though this finding might look strange, it can be justified on the basis of excessive trust that technologists tend to display when dealing with algorithmic solutions. Overall, according to the data of Table 5, an equal number of 11 wrong to correct vote-changes were observed in each group, driven by the machine guidance. However, a rather paradox discovery is that image experts conserved their erroneous T1-responses to a higher degree than the media specialties (7 versus 5). Someone would expect the exact opposite effect for people with a better understanding of the DIF techniques and their results.

Similar indications and with stronger confidence intervals have been produced from the statistical Multivariate Analysis of Variance (MANOVA) that was conducted for the collected answers in each different image sample. Various criteria were tested to estimate the weight and correlation of the different variables (Pillai's Trace, Wilks' Lambda, Hotelling's Trace, Roy's Largest Root). Statistical indexes were extracted for the estimation of the effect size, i.e., to measure the validity of the experimental procedure (p-value<0.0001), but also to assess the statistical significance of the adopted grouping (unfortunately, with not quite optimistic results). The next step was to examine the impact of the independent variable (Specialty) to each of the dependent parameters (i.e., initial authenticity evaluation/experimental treatment). The latter task was accomplished by conducting multivariable tests for each image (within-subjects/between-subjects' effects). The analysis led to useful findings, implying a significant relationship between the examined factors, the details of which are given in Table 6. Specifically, it was discovered that the IVA-training experimental treatment affected four (4) of the total seven (7) picture samples, while the Specialty parameter proved to be significant in two (2) of the seven (7) cases (highlighted in bold, in Table 6).

The "Algorithm" column of Table 6 presents the impact (i.e., p-value) of the IVA-training experimental treatment and the "Specialty" column

Table 5. Crosstab specialty experimental treatment for all the algorithms and all image samples.

Specialty			Time 2		Total
			Correct	Wrong	
Image Experts	Time 1	Correct	19	3	22
		Wrong	11	**7***	16
		Total	30	10	40
Media Experts	Time 1	Correct	20	4	16
		Wrong	11	**5***	24
		Total	31	9	40
Total	Time 1	Correct	39	7	16
		Wrong	22	12	24
		Total	61	22	40
				Total	80

* denotes the highest values that are further commented in the text.

Table 6. Multivariate analysis within subjects.

Case	Algorithm	Specialty
owl_fake	.347	.347
peans_ authentic	**.035***	.242
koko_authentic	.242	**.035***
sea_authentic	**.035**	.242
King_fake	1	1
iran_authentic	.197	.879
id_fake	**.020***	.580
crowd_fake	**.040***	**.040***

* denotes the highest values that are further commented in the text.

Table 7. Decision changes - specialty crosstabulation (image: "crowd_fake").

Time 1 * Time 2 * Specialty Crosstabulation
Image: *"crowd_fake"*

Specialty			Time 2		Total
			Correct	Wrong	
Image Experts	Time 1	Correct	3	0	3
		Wrong	0	2	2
	Total		3	2	5
Media Experts	Time 1	Correct	1	0	1
		Wrong	3	1	4
	Total		4	1	5
Total	Time 1	Correct	4	0	4
		Wrong	3	3	6
	Total		7	3	10

the corresponding Specialty influence. If p is less than 0.05 ($p < .05$), the one-way repeated measures MANOVA is statistically significant. Alternatively, if $p > .05$, the one-way repeated measures MANOVA is not statistically significant. With respect to the criteria laid above, in the "*peans_authentic*" case (p-value = 0.035), the forensic-assistance treatment led four of the ten participants (4/10) to alter their first subjective impression, ascribing the authenticity of the image correctly. The same applies to the "*sea_authentic*" picture (p-value = 0.035), with the experimental impact to be even more prominent in the "*id_fake*" example (p-value = 0.020). In this last case, four (4) of the subjects alter their early evaluation decision correctly, while two (2) others were driven to change their estimation from doctored photo to original, falsely. Finally, three (3) initially wrong answers were corrected during the "*crowd_fake*" session (p-value = 0.040), as Table 7 also presents.

On two occasions, the experimental treatment exhibits a strong relationship with the specialty. The first refers to the "*koko_authentic*" case (p-value = 0.035, Table 8), in which three (3) subjects altered their votes in the right direction, all coming from the image-experts group. At the same time, a media specialist followed the exact opposite course, shifting the correct answer to wrong. As already commented, journalists probably had prior knowledge of this example, which has bothered the daily news agenda quite a few times in the past. Hence, the rest of the results seem reasonable. In the second instance ("*crowd_fake*", p-value = 0.040), all the participants who accurately changed their first estimation belonged to the media category (Table 7). Someone could rationalize this positive result by the machine-assisted verification, combined with the superior subjective inspection of experienced users and their increased suspiciousness in algorithmic solutions. Another general conclusion from

Table 8. Decision changes - specialty crosstabulation (image: "koko_authentic").

Time 1 * Time 2 * Specialty Crosstabulation
Image: *"koko_authentic"*

Specialty			Time 2		Total
			Correct	Wrong	
Image Experts	Time 1	Correct	2	0	2
		Wrong	3	0	3
	Total		5	0	5
Media Experts	Time 1	Correct	4	1	5
		Wrong	0	0	0
	Total		4	1	5
Total	Time 1	Correct	6	0	3
		Wrong	3	1	7
	Total		9	1	10

this session is that, the more the number of items and the visual details of a picture, the less likely to gain high confidence levels by news-reporters.

Another interesting finding refers to the instances that algorithmic support had a negative impact, resulting in the falsification of initially correct answers. Two such occurrences were spotted in the "*koko_authentic*" and "*iran_authentic*" cases, with a single error observed on both occasions within the media group. In the latter case (*iran_authentic*), a related incident was also noticed within the image experts' group. Overall, the whole experimental procedure is susceptible to the nature and the content of the original pictures, as well as to the type and the level of manipulation, making it even harder to provide reliable generalizations (as it is further analyzed in the subsequent section).

The second part of this experiment was then decided to emphasize on the quantitative aspects of the attempted examination. Although the discussion part of the interviews seems to be irreplaceable when it comes to qualitative remarks, the deployed online survey offers alternative statistical scales. In this perspective, the value of a per-sample analysis is rather downgraded here (in contrast to the previous evaluations) given the lack of control in quite a few aspects. For instance, while subjects' expertise was accurately adapted to the first testing needs, the corresponding mechanism is not so straightforward or precise in this supplemental approach. As already mentioned, simple familiarity questions were created, from which previous related experience had to be estimated. Specifically, the average values of the three-question "image and media skills measures" were computed and compared to the mid-range threshold (2.5 in the five levels Likert scale), to decide "specialty class" of the participants (110 image experts, 220 media experts, and 53 average users). As expected, most individuals were closer to journalism and media disciplines than imaging and their algorithmic perspectives (mainly due to the dissemination of the questionnaire to students of Journalism and members of the Ellinikahoaxes Facebook group). Moreover, the above categorization enables the registration of subjects mutually in the two specialty groups, i.e., the ones that outreached the set threshold to both acquaintance queries. A third category of "average users" was assembled for those that did not meet the criteria, neither for media nor for image authorities, therefore they do not possess prior knowledge associated with the task of image authentication. Although multiple different thresholds could have been tested for the formation of different experts' groups/distributions, this was considered out of our primary research aims, so it was not further investigated in this paper. Besides, the involved questions are quite simple for providing just an indication of the potential skillsets, adapting to the limitations of an online survey.

Another limitation of this online examination is the lack of a rightly guided training process in the use of the IVA algorithms, which played a central role in the design of the repeated measures experiment. For

Figure 10. Authenticity evaluation correct answers before (Time 1) and after (Time 2) the experimental treatment (in the second/supplemental part of the repeated-measures experiment), for the initially 8-photos set (% scores are given for the machine-assisted evaluation using the IVA tools). While images were randomly presented to the subjects, the order of the first part of the experiment was adopted here (as in Figure 9), to facilitate related comparisons.

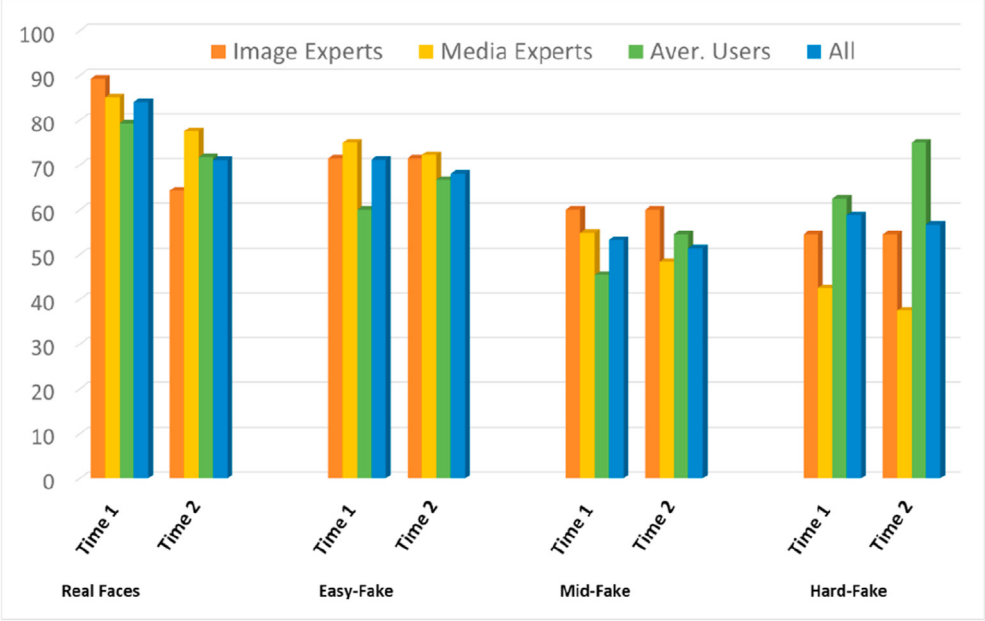

Figure 11. Authenticity evaluation correct answers before (Time 1) and after (Time 2) the experimental treatment (in the second/supplemental part of the repeated-measures experiment), for the Kaggle real and fake faces dataset (% scores are given for the machine-assisted evaluation using the IVA tools). Images are grouped according to their authenticity status and the subjective difficulty of the tampered samples, as provided by the dataset creators.

instance, there is no assurance that all subjects have adequately operated the given directions/examples to understand the underlying algorithmic principles, therefore, to interpret the associated forensic visualizations accurately. In all perspectives, the value of a help file cannot be measured against the effectiveness of a face-to-face/in-situ training session. The above is further deteriorated by the admitted complexity in the comprehension of DIF techniques, as interviewees had notably expressed in the first part of the experiment. Hence, it is not accidental that many relevant comments were received in the online survey, too, validating the initial decision to restrict the analysis to a small number of algorithms. These difficulties were also projected in the received authenticity evaluation results, before and after the algorithmic assistance, as further commented in the following paragraphs. Overall, without underrating the individual vulnerabilities of the two experimental parts, their supplemental nature aims at enlightening all the combined qualitative and quantitative views of the attempted repeated measures investigation.

Following the organization and intention of Figure 9, Figure 10 depicts the initial (T1) and final (T2) authenticity estimations before and after the online experimental treatment (i.e., with the help of the IVA toolset) for the four, this time, formed categories: Image Experts (orange bars), Media Experts (yellow bars), Average Users (blue bars) and All users (green bars). It can be noticed that machine-assistance does not improve the results in most cases, a finding that can be justified in the absence of a carefully guided training session, as already explained. Indicative online sessions with parallel teleconferencing support were conducted to imitate the conditions of the first experimental part, which

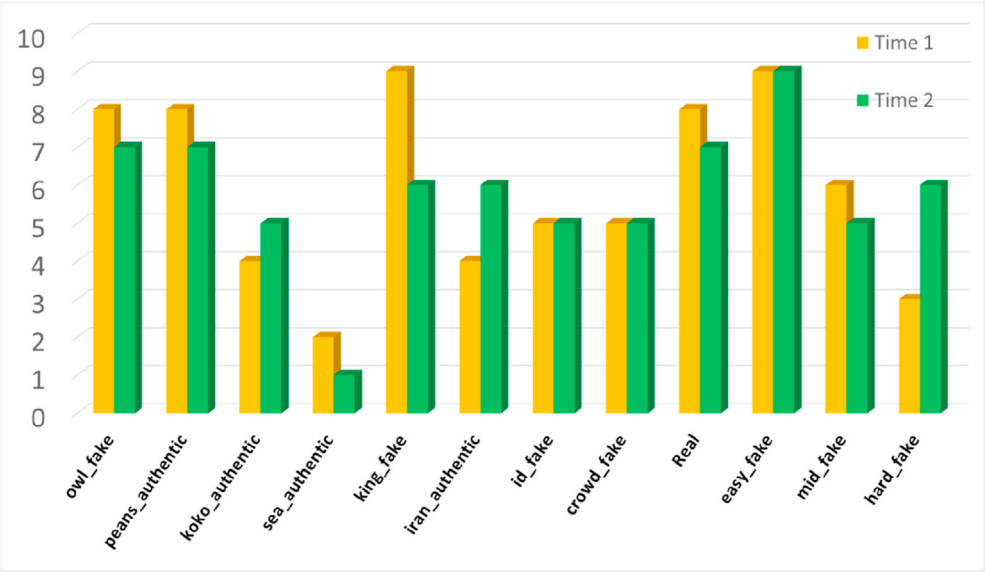

Figure 12. Authenticity evaluation correct answers before (Time 1) and after (Time 2) the experimental treatment (in the second/supplemental part of the repeated-measures experiment). The 10 answers that took the longest time in the corresponding sessions are shown, for both the initial 8-photos set and the Kaggle real and fake faces dataset (absolute-value scores are given for the machine-assisted evaluation using the IVA tools).

seemed to validate this remark (i.e., the results were closer to the view of Figure 9). However, looking into the details of Figure 10, scores are somewhat improved for the average users, with a clear positive impact of the DIF algorithms in two of the samples (*peans_authentic*, *sea_authentic*). This sign of progress is more clearly depicted in Figure 11, representing the same results for the used real and fake facial photos. In specific, a significant contribution to the decisions of average users is observed in all fake categories and especially in the hardest group. A possible explanation is that people with zero or little background can get maximum machine assistance, which is also enhanced by their smaller biases and expectations (related comments pointed to that direction). On the contrary, the lack of familiarity with the DIF tools and the subsequent insecurity may cause decreased algorithmic gain in the authentic (face) photos. Hence, colormaps will show differences that look suspicious, even in authentic instances, requiring more apt and knowledgeable interpretation. While the reliability of these remarks relying on the received feedback can be questioned, they still represent significant findings that deserve further investigation.

Excluding, for a moment, the algorithmic assistance contribution, an attempt to compare the results between the first interviewing part (Figure 9) and the online experiment (Figure 10) shows that the degree of the perceived difficulty is somewhat maintained between samples. "*Sea_authentic*" still remains the most demanding case followed by "*crowd_fake*". The largest alteration is observed in the "*king_fake*" example, which presented better results in the online survey. A reasonable explanation may be found on the photo individualities (already discussed) and the journalistic background of most subjects. Furthermore, an alternative effort was conducted to measure the influence of the experimental treatment. Specifically, the time spent in each of the testing sessions was considered as an estimate of the self-training dynamics. The hypothesis behind this settlement is that the more time spent, the strongest the experimentation and familiarization with the DIF help files and the associated DIF techniques. Figure 12 depicts the ten (10) slowest responses for all the different samples/categories, facilitating a direct comparison with the results of Figure 9. Apart from the "*king_fake*" example that exhibits the vastest divergence (as already commented), neutral machine assistance is observed, which turns to have a clear positive impact in the cases of "*iran_authentic*" and "*hard_fake*" faces, that proved to be the most demanding ones. These findings somewhat agree to the average users' discussed behavior, i.e., the algorithms are helpful when they are rightly used to offer missing visual inspection insights. The last remark also approves the selection of the used real and face photos as a control dataset. Hence, while facial pictures are very common in news authentication tasks, the specific samples are completely random content-wise, without holding any of the discussed info-documenting or viralness aspects. Consequently, the requested authentication tasks entirely rely on the visual inspection of pixels and heatmaps. Thus, doubling the score after the experimental intervention is considered very important for the presumably hardest facial case (*hard_fake*). In this direction, the subjective treatment and ratings of the associated dataset are also positively validated, aligning with the aims and requirements of the research. Once again, it is important to underline the speculative nature of these remarks, due to the already commented limitations. However, the combination of all conducted studies provides quite exciting and significant pilot research findings.

There are also some clearer/undisputed indications, derived by comparing the two parts of the repeated measures experiment. It is evident that the verification task within the initial collection of 8 photos is statistically more demanding, covering a broader region of correct answers. Specifically, scores range among 14% and 74% in Figure 10, with the respective values for the face photos to vary between 38%-89% in Figure 11. The results proved that the most challenging cases of our assembled pictures (8p) are quite more difficult to evaluate than the hardest category of fake faces. This is also visible in Figure 13, which illustrates the Confidence and Difficulty levels received for the associated samples. It can be seen that both variables have smaller dispersion in the 8-photos set, with the confidence levels to be statistically lower than those of the control images. Likewise, the perceived difficulty covers a broader area, with some samples to be appraised as harder than the "*hard_fake*" face photos and others to be considered as easier. The above observations seem to be amplified when only the correct answers are accounted, which turns out to be the most reliable (i.e., the given difficulty represents a more objective and reliable measure if a subject has evaluated the image authenticity right to the case of an incorrect answer).

Another view of results is given in Table 9, using the typical evaluation metrics Precision, Recall, F1-score, and Accuracy, which are typically employed in relevant authentication tasks (Hsu et al., 2020; Thakur and Rohilla, 2020; Zheng et al., 2019). Accuracy calculates the ratio of the correctly recognized instances to the total number of samples, expressing the overall ability of a classification model. However, this metric does not provide the full picture alone, especially if the classes are unevenly distributed within the set. Precision estimates the proportion of correct identifications, i.e., in our case, the number of authentic images correctly classified as true-positives to the total positive results (true and false). Likewise, Recall extracts the ratio of actual positives that were identified correctly, i.e., the number of true-positive authentic images to the total sum of all actual positives, either classified correctly (true--positives) or not (false-negatives). These metrics express the robustness of a system concerning its true-positive and false-positive rates. Furthermore, F1-score combines Precision and Recall measures through their harmonic mean to provide a composite estimate. Apparently, the higher the values appear in these metrics, the more accurate the results are.

Returning to the values of Table 9, it is clear that all statistical ratings of the initial 8-photos collection are constantly (and in some cases considerably) lower than those of the control facial dataset. These findings verify the demanding character of the formed set, a prerequisite from the early beginning of this research. It is essential to mention that the conducted comparison is quite tough, considering that the selected "Real and Fake Faces" dataset is intended for contemporary completion tasks on image authentication. Moreover, the extracted results seem to be significantly lower than those presented in similar research works (Hulzebosch et al., 2020), even when using lower resolution pictures. As in the cases of Figures 10, 11, 12, and 13, data in Table 9 cannot provide a clear indication concerning the impact of the experimental treatment. The lack of control in the training procedure and the associated (online) limitations seem to be the causes of the problem, as already explained. In this perspective, the initially deployed repeated-measures experiment, conducted with a small number of selected experts, seems to be preferable, despite its reduced sample size and the statistical reliability consequences. Hence, a recommended direction would be to deploy multiple such short sessions, either via face-to-face interviews or through online surveys. In all cases, proper experimental supervision is critical to ensure control of both the participants (and their background) and mostly the training procedure in the use of the DIF tools.

4. Discussion

This work examined the ability of both average users and targeted experts in detecting photo manipulation traces, with and without the assistance of suitable image forensic assistance tools. The study deployed two different but complementary monitoring approaches to address the stated questions: a cross-sectional survey and a repeated-measures experiment. The relatively small sample size (especially for the first part of the experimental treatment case) makes generalizations difficult (at least, without pointing out specific limitations and precautions). However, the supplementary online survey dealt with some of the stated limitations, mostly the ones linked with the quantity and statistical reliability of the collected data. On the contrary, control and qualitative insights were sacrificed when the experiment was massively deployed online. Based on the justification of the adopted research configuration,

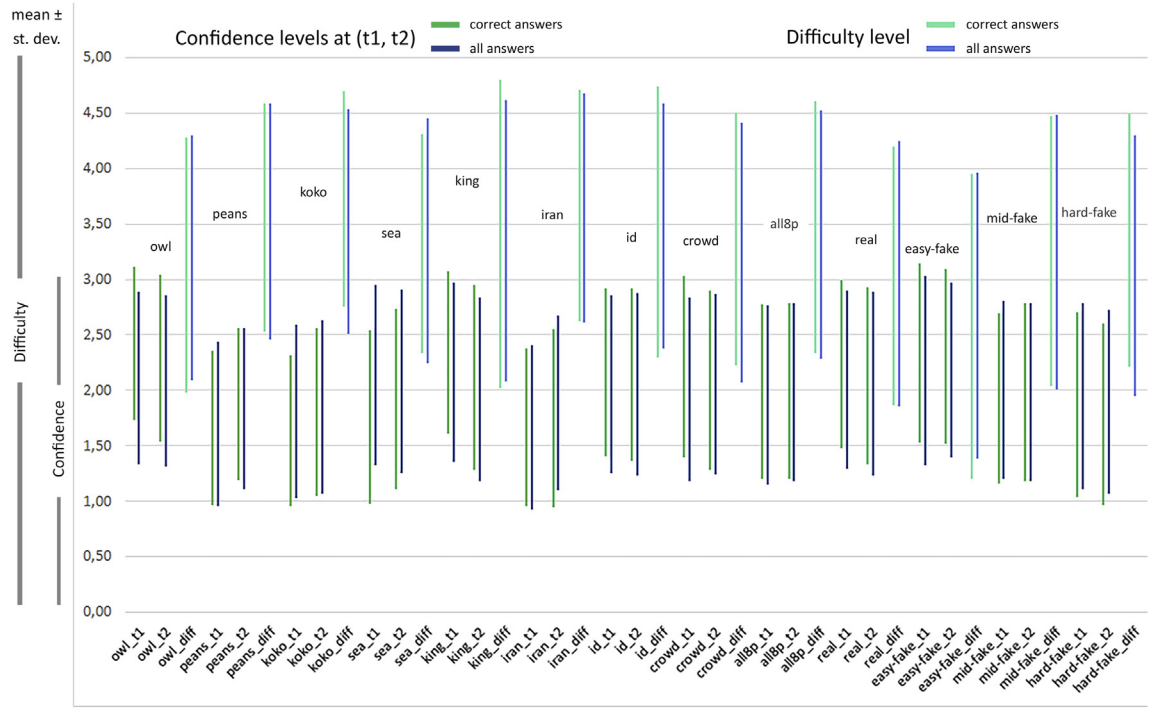

Figure 13. Representation of the voted Confidence (1–3) and Difficulty (1–5) levels before (Time 1) and after (Time 2) the experimental treatment (in the second/supplemental part of the repeated-measures experiment), for both the initial 8-photos set and the Kaggle real and fake faces dataset, divided in the groups of correct and all answers (mean ± standard deviation values are given).

Table 9. Estimated evaluation scores (Precision, Recall, F1_score, Accuracy) in the second/supplemental part of the repeated-measures experiment.

	Precision		Recall		F1-score		Accuracy	
	t1	t2	t1	t2	t1	t2	t1	t2
Initial 8 photos	0.444	0.421	0.433	0.445	0.439	0.432	0.462	0.434
Kaggle faces dataset	0.650	0.614	0.840	0.710	0.733	0.659	0.694	0.632
All	0.542	0.504	0.599	0.553	0.569	0.527	0.555	0.513

the selected material and methods were carefully designed in their present form, aiming at enlightening different views and perspectives of the problem under study. Overall, the study provided some very interesting preliminary findings, also indicating a plan for future research and more targeted insights, as explained below.

Specifically, all the stated hypotheses were confirmed convincingly. The results of the first test indicated that ordinary users could identify image manipulation attacks based on their subjective/visual inspection (H_1). The data presented in Figure 7 confirm that most of the inquired authenticity samples were accurately marked, with a total of 53% correct answers and 47% wrong. These ratings were also validated in the second part of the repeated measures experiment, conducted online, resulting in 55,5 % recognition accuracy to all tested photos. While someone can support that the achieved performance is marginally above average, this can be justified to the vast heterogeneity of both the participants and the used dataset, which was formed that way to cover a broad range of the studied phenomenon. Hence, several subjects were utterly ignorant about the "fake news" issue and the extent of the contemporary photo-editing and doctoring capabilities. Likewise, the organized test-set included some quite challenging pictures with increased authentication difficulties, to stress the limits of the human inspection. Reasonably, both the factors above had a substantial impact on diminishing the overall level of the measured veracity rates. The successful recognition becomes even higher when specialists are accounted for, as the outcomes of the second experiment revealed. Thus, a 56.3 % of right replies was monitored (with 43.7% incorrect votes), despite the incorporation of some yet more peculiar cases (i.e., the "sea-authentic" example, the verification of which was a complete failure in the Time-1 session, according to Figure 9). The stated findings additionally confirmed the second hypothesis (H_2), i.e., that the analysis of the collected/crowdsourced answers would be useful to verify the originality of pictures, used as evidence documents in news-reporting. Indeed, an authenticity decision based on the previously presented statistics would be mostly positive (i.e., the overall majority of subjective responses lies in the correct site), even within the discussed constrains.

Providing additional insights, the repeated-measures experiment indicated substantial differences in the subjective evaluation of image authenticity with and without the help of the DIF tools, i.e., before and after the experimental treatment (RQ_1). As already analyzed, the use of forensic utilities overall improved the authentication results, pointing out the positive impact of the algorithmic assistance (RQ_2). Furthermore, the interpretation of Figure 9 and Tables 4, 5, 6, and 7, conducted in section 2.4, brought forward the way that the different groups altered their votes in T2 sessions, detecting specific correlations between decision updates/modifications and specialty (RQ_3). Notably, despite the small difference in the overall accuracy shift, it was confirmed that image experts (and generally technologists) tend to display excessive trust when dealing with algorithmic IVA solutions. This fact makes them more willing to employ the offered assistance in updating their initial responses, even in the wrong direction. On the contrary, journalists and media professionals display higher confidence and trust in their subjective judge, experience, and instinct, so they are becoming more cautious when it comes to the

adoption of automated machine suggestions. Both attitudes need to be stressed in future planning and releases/updates of the associated methods, aiming at getting the maximum verification potentials by all disciplines.

Apart from the findings that provided support to the stated hypotheses and questions, additional interesting findings were identified, especially during the interviews within the repeated-measures experiment. For instance, it was confirmed that only a small percentage of the participants had prior awareness or experience of IVA solutions. On multiple occasions, the analysis detected weaknesses concerning the proper interpretation of the machine-extracted forensic visualizations. Thus, proper instructions/guidance and familiarization in the use of the offered automations proved to be beneficial for the veracity assessment task. Indeed, the difficulty for journalists and broad audiences to comprehend the underlying algorithmic principles requires substantial interfacing improvements to enhance the understandability of the operations, clarifying the reading of the extracted heatmaps. According to the interviewees, the whole process could be profoundly facilitated through more featured examples, rule-of-thumb tips, and suggestions, associating the utilization of specific tools to the different types of image tampering.

The second online survey, which supplemented the repeated-measures experiment, validated the desired attributes for the formed set of pictures, covering an adequately broad range of tampering and verification means. Furthermore, it was verified the demanding nature of this photo-collection, which proved to feature increased difficulty dynamics, compared to related contemporary datasets. As already explained, this latest test and its online character diminished the control over the IVA training sessions and the participants' expertise, as initially foreseen. Hence, the unsupervised treatment configuration does not favor the extraction of stable and reliable conclusions concerning the assistance of the DIF tools. Nevertheless, additional scales, measures and insights were enlisted, such as the time needed to complete the different experimental sessions and the estimated specialty classification. In these directions, valuable qualitative and quantitative examination remarks were discovered. Hence, interesting (in many aspects) research findings were detected for interpreting and reasoning the underlying facts behind the numerical results. While these justifications seem entirely adequate within the aims of the current research, they also provide a fertile ground for an extended investigation of the available data, which could be further reinforced in multiple targeted directions.

It is worth noting that related research efforts are continually elaborating on the technological domain, with the design and the advancement of more sophisticated algorithmic and interfacing solutions. For instance, further IVA tools have been released since the beginning of this study. Unfortunately, this kind of progress is missing in the associated training and guiding actions, which are considered even more critical, in alignment with corresponding theories and research findings of the current work. It is vital to launch and integrate such supporting actions, pinpointed with the aimed feedback of featured surveys, like the ones examined in this paper. Apparently, the number and the diversities of both the participants and the used photo-samples should be further enlarged to increase the statistical reliability of the testing procedures. However, based on the already stated preliminary results, a small set of inspecting images per session is necessary to avoid tiresome and bias of the subjects. Similarly, the interviewing feedback that proved very informative and useful within the repeated measures experiment cannot be replaced with larger-scale questionnaires of any form (i.e., vast numbers of queries and involved participants). A careful interpretation of the conducted analysis would point in the direction of elaborating and reorganizing the presented tests to many focusing groups. The latter would help in enlightening more perspectives of imaging falsification operations and their perception by various additional specialties. Apart from the targeted integration and the assessment itself, the configuration of such systematic evaluation reviews would offer valuable multidisciplinary knowledge, including the formation of dedicated tampered-images repositories with multiple uses.

5. Conclusion

The current paper attempted a diverse investigation of the photo-truth impact in news verification, analyzing both subjective and machine-assisted evaluations through real-world forensic tools and scenarios of pictures, used as proof documents. The presented literature review and the subsequent argumentation reveal a rather significant hysteresis between the massively deployed research on the implementation of DIF techniques and the associated cognitive studies, focusing on the perception and real use of such practically available services. This remark is rather evident in most related (and recent) review papers and surveys (Gokhale et al., 2020; Katsaounidou, 2020; Qureshi and El-Alfy, 2019; Tariq et al., 2018; Thakur and Rohilla, 2020; Zheng et al., 2019), in which, such featured works are seldom or entirely missing. Nevertheless, limited publications nearer to the current approach do exist, such as the work of Gloe et al. (2007) that examined DIF tools to detect tampering traces associated with resampling and/or identification of sources patterns in the origin of the images, using a pull of 300 initial photos. Likewise, Williams et al. (2018) investigated the human trust factor concerning the assistance of a graphics expert or an algorithm, using a set of 80 pictures. Schetinger et al. (2017) used 393 volunteers and 177 images to collect 8,160 markings indicating what subjects considered as forgeries. While the above studies are methodologically interesting (in some respects, pioneering) and lead to useful remarks and conclusions, none of them deals with the proofing character of the visual documents, i.e., the real utilization of photos in news authentication.

Nightingale et al. (2017) moved a step forward, questioning the identification of real-scene original and manipulated photos, though there were not all firmly pinpointed to news stories. A considerable number of users were involved in two online surveys (707 and 659 participants), inspecting 40 images in total that were produced from an initial pull of 10 images, using 5 different manipulation ways (with an internal limit of 100 responses per photo). They concluded that people's ability to detect manipulated photos of real-world scenes is extremely limited. Pantti and Sirén (2015) focused exclusively on the news-proofing character of the visual documents, interviewing 19 journalists on the topic of newsroom value of non-professional UGC images. All participants admitted the central role of verification in the journalistic profession, with many of them distancing themselves from having responsibility or skills in performing the implicated forensic investigation. While most of the above remarks were also addressed in the associated review sections, the attempted summarization is vital to display the current situation in the field. Thus, none of the above attempted to test the usefulness of multiple real-world DIF services (at least, to this extent), neither emphasized in such degree on the news authentication aspects (and the associated informatory and viralness dynamics of the used datasets) nor combined equivalently qualitative and quantitative insights trough the deployment of both face-to-face interviews and online questionnaires. Overall, the current study investigated more than 3,000 pictures to end up on the use of 224 images (from a pull of approximately 2, 113 samples[17]) and431[18] volunteers to collect a total of 13,095 markings[19], which seems to outreach previous records.

Elaborating on the above, this work managed to combine most of the perspectives analyzed in related studies. The formed dataset proved to be quite demanding and representative of the various manipulation attacks, incorporating additional news-verification features that are considered very important in the current research. A high number of evaluations

[17] ~**3 x (16 + 8)** [initial samples to form the 16-photos and 8-photos sets] + **2,041** [Kaggle real and fake photos dataset].

[18] **120** (cross-sectional study) + **10** (repeated-measures experiment, part A) + **301** (repeated-measures experiment, part B).

[19] **16** images **x 120** users + **8** images **x 10** users **x (7 + 1)** views [7 IVA] + **5** images/session x 301 users **x (6 + 1)** views [6 IVA].

were received for this set of photos in the repeated measures experiment, i.e., 10 in the first part and other 903 in the second part (115 owl; 106 peans; 111 koko; 112 sea; 110 king; 109 iran; 115 id; 125 crowd). The adequacy of these numbers, valued against the ones of the associated research, provides a reliable and solid validation of the appropriateness of the data used here. This remark is further strengthened by the results of the indirect comparisons with the control face images. As pointed out in the discussion section, it is essential to continue extending the research within smaller groups and in a more supervised manner (either distantly, via online tools and remote audiovisual conferencing guidance, or through face-to-face interviews). In this direction, such firm and featured datasets seem to be fitting in the context, so they need to be updated with additional samples and their multiscale ratings. Another positive sight is that relevant works adopted similar categorization of subjects, i.e., experts in graphics and artificial intelligence (Hulzebosch et al., 2020; Williams et al., 2018). Hence, the adopted categorization of participants appears to be on the right track. To the best of our knowledge, it is the first time that such a systematic review has attempted to impact photos in news authentication, combined with an extended set of DIF tools. Overall, the complementarity of the deployed methods allowed to enlighten most of the considered evaluation perspectives and to draw the necessary future elaboration plans.

As already implied, the conducted evaluation leaves a slightly bitter aftertaste because automations tend to have a non-linear effect due to their complex nature. In all cases, the progress in real-world utilization does not align with the advancement of associated research, notably in the sub-domains of technology and machine-driven authentication. Nevertheless, the extracted results showed that DIF tools can offer valuable help under specific conditions concerning users, types of manipulation, and specific content attributes. Hence, journalists need to get familiarized with how they could use DIF services. The same applies to average people involved in the daily informing processes as UGC-contributing or news-consuming individuals. Specifically, the audience should comprehend the basic underlying principles of forensic operations, realizing that not all tools are suitable for all cases. Once again, the vital goal for the "verification industry" should be to make forensic automations and their interpretations more transparent and more accessible for everyone, especially the newsroom professionals. Entering the era of deep fakes, it is vital to have some fundamental skills and know-how on machine-assisted forensic investigation through traditional DIF techniques and practices, before the upcoming deep-learning authentication automations create additional inconveniences and insecurities. In this context, future planning of focused testing on narrower DIF families is considered essential, before one gets involved in the latest algorithmic solutions, which are principally blind to the users. Overall, trust is crucial in this field and can only be built upon proper training and feedback, with the use of applicable explanation interfaces that follow the human experience.

Declarations

Author contribution statement

A. N. Katsaounidou, N. Tsipas: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

A. Gardikiotis: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

C. Dimoulas: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

C. Dimoulas is a member of the Advisory Board of the Heliyon Computer Science.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors would like to acknowledge the valuable help of Dr. S. Papadopoulos and the Media Verification team (mever.iti.gr) for their support in the use of the Image Verification Assistant Web API. They also acknowledge the valuable contribution of the Greek debunking site Ellinikahoaxes (www.ellinikahoaxes.gr) in the dissemination of the questionnaire used in the second part of the repeated measures experiment.

References

Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E., 2012. An evaluation of popular copy-move forgery detection approaches. IEEE Trans. Inf. Forensics Secur. 7 (6), 1841–1854.

Dong, J., Wang, W., Tan, T., 2013. July). Casia image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing. IEEE, pp. 422–426.

Farid, H., 2009. Photo fakery and forensics. Adv. Comput. 77, 1–55.

Gloe, T., Kirchner, M., Winkler, A., Böhme, R., 2007, September. Can we trust digital image forensics?. In: Proceedings of the 15th ACM International Conference on Multimedia. ACM, pp. 78–86.

Gokhale, A., Mulay, P., Pramod, D., Kulkarni, R., 2020. A bibliometric analysis of digital image forensics. Sci. Technol. Libr. 39 (1), 96–113.

Graves, D., 2018. Understanding the Promise and Limits of Automated Fact-Checking (Reuters Institute for the Study of Journalism Factsheets). Reuters Institute for the Study of Journalism.

Griffin, D.S., Muhlbauer, G., Griffin, D.O., 2018. Adolescents trust physicians for vaccine information more than their parents or religious leaders. Heliyon 4 (12), e01006.

Ho, A.T., Li, S. (Eds.), 2015. Handbook of Digital Forensics of Multimedia Data and Devices. John Wiley & Sons.

Hsu, C.C., Zhuang, Y.X., Lee, C.Y., 2020. Deep fake image detection based on pairwise learning. Appl. Sci. 10 (1), 370.

Hsu, C.C., Lee, C.Y., Zhuang, Y.X., 2018, December. Learning to detect fake face images in the wild. In: 2018 International Symposium on Computer, Consumer and Control (IS3C). IEEE, pp. 388–391.

Huxford, J., 2001. Beyond the referential: uses of visual symbolism in the press. Journalism 2 (1), 45–71.

Hulzebosch, N., Ibrahimi, S., Worring, M., 2020. Detecting CNN-generated facial images in real-world scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 642–643.

Jahnke, I., Kroll, M.M., 2018. Exploring students' use of online sources in small groups with an augmented reality-based activity–group dynamics negatively affect identification of authentic online information. Heliyon 4 (6), e00653.

Johnson, M.K., Farid, H., 2007. January). Exposing digital forgeries through specular highlights on the eye. In: Information Hiding. Springer Berlin Heidelberg, pp. 311–325.

Kalliris, G., Matsiola, M., Dimoulas, C., Veglis, A., 2014a. Emotional aspects and quality of experience for multifactor evaluation of audiovisual content. Int. J. Monit. Surveill. Technol. Res. 2 (4), 40–61.

Kalliris, G., Matsiola, M., Dimoulas, C., Veglis, A., 2014b. Emotional aspects in quality of experience and learning (QoE & QoL) of audiovisual content in mediated learning. In: Proceedings of the IEEE 5th International Conference on Information, Intelligence, Systems and Applications (IISA 2014), pp. 198–203, 7-9 July 2014.

Kalliris, G., Dimoulas, C., Veglis, A., Matsiola, M., 2011. Investigating quality of experience and learning (QoE & QoL) of audiovisual content broadcasting to learners over IP networks. In: Proceedings of the Sixteenth IEEE Symposium on Computers and Communications (ISCC11), pp. 836–841 art. no. 5983946.

Katsaounidou, A., 2016. Content Authenticity Issues: Detection (And Validation) Techniques of Untruthful News Stories from Humans and Machines. Unpublished Master Thesis (In Greek.). Post-Graduate Program of the School of Journalism and Mass Communications. Aristotle University of Thessaloniki online: http://ikee.lib.auth.gr/record/282833.

Katsaounidou, A.N., Dimoulas, C.A., 2018a. Integrating content authentication support in media services. In: Encyclopedia of Information Science and Technology, fourth ed. IGI Global, pp. 2908–2919.

Katsaounidou, A.N., Dimoulas, C.A., 2018b. The Role of media educator on the age of misinformation Crisis. In: Proceedings of the EJTA Teachers' Conference on Crisis Reporting, Thessaloniki, Greece, 18–19 October 2018.

Katsaounidou, A., Dimoulas, C., Veglis, A., 2018. Cross-media authentication and verification: emerging research and opportunities: emerging research and opportunities. IGI Global.

Katsaounidou, A., Vrysis, L., Kotsakis, R., Dimoulas, C., Veglis, A., 2019a. MAthE the game: a serious game for education and training in news verification. Educ. Sci. 9 (2), 155.

Katsaounidou, A., Vryzas, N., Kotsakis, R., Dimoulas, C., 2019b. Multimodal news authentication as a service: the "true news" extension. J. Educ. Innov. Commun. (JEICOM) 1, 11–26.

Katsaounidou, A., 2020. Interactive and Collaborative Environments to Support Digital Content Authentication. Unpublished Ph.D. Dissertation (In Greek.). School of Journalism and Mass Communications. Aristotle University of Thessaloniki available online (last access on July 11, 2020). http://ikee.lib.auth.gr/record/318930/.

Korus, P., 2017. Digital image integrity–a survey of protection and verification techniques. Digit. Signal Process. 71, 1–26.

Krawetz, N., Hacker Factor Solutions, H.F., 2007. A Picture's Worth. Hacker Factor Solutions, p. 6.

Lévy, P., 1997. L'intelligence collective: pour une anthropologie du cyberspace. La découverte, Paris.

Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L., 2018, April. Detection of gan-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, pp. 384–389.

Middleton, S.E., Papadopoulos, S., Kompatsiaris, Y., 2018. Social computing for verifying social media content in breaking news. IEEE Int. Comp. 22 (2), 83–89.

Mitchell, W.J., 1994. The Reconfigured Eye: Visual Truth in the post-photographic Era. Mit Press.

Nightingale, S.J., Wade, K.A., Watson, D.G., 2017. Can people identify original and manipulated photos of real-world scenes? Cogn. Res.: Prin. Impl. 2 (1), 30.

Pantti, M., Sirén, S., 2015. The fragility of photo-truth: verification of amateur images in Finnish newsrooms. Digit. J. 3 (4), 495–512.

Pham, N.T., Lee, J.W., Kwon, G.R., Park, C.S., 2019. Hybrid image-retrieval method for image-splicing validation. Symmetry 11 (1), 83.

Qureshi, M.A., El-Alfy, E.S.M., 2019. Bibliography of digital image anti-forensics and anti-anti-forensics techniques. IET Image Process. 13 (11), 1811–1823.

Schetinger, V., Oliveira, M.M., da Silva, R., Carvalho, T.J., 2017. Humans are easily fooled by digital images. Comput. Graph. 68, 142–151.

Silverman, C. (Ed.), 2013. Verification Handbook. European Journalism Centre.

Tariq, S., Lee, S., Kim, H., Shin, Y., Woo, S.S., 2018, January. Detecting both machine and human created fake face images in the wild. In: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, pp. 81–87.

Thakur, M.K., 2014. Tampered Videos: Detection and Quality Assessment. Unpublished Doctoral Dissertation. Jaypee Institute of Information Technology, Department of Computer Science Engineering & Information Technology. downloaded on March, 1, 2016, from. http://shodhganga.inflibnet.ac.in/handle/10603/44603.

Thakur, R., Rohilla, R., 2020. Recent Advances in Digital Image Manipulation Detection Techniques: A Brief Review. Forensic Science International, p. 110311.

Thorne, J., Vlachos, A., 2018. Automated fact checking: task formulations, methods and future directions. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING '18 (2018), pp. 3346–3359.

Vryzas, N., Katsaounidou, A., Kotsakis, R., Dimoulas, C.A., Kalliris, G., 2018, May. Investigation of audio tampering in broadcast content. In: Proceedings of the 144th Audio Engineering Society Convention. Milan, 23-26 May 2018.

Vryzas, N., Katsaounidou, A., Kotsakis, R., Dimoulas, C.A., Kalliris, G., 2019, March. Audio-driven multimedia content authentication as a service. In: Proceedings of the 146th Audio Engineering Society Convention. Dublin, 20-23 March 2019.

Wang, W., 2009. Digital Video Forensics. Ph.D. dissertation. Department of Computer Science, Dartmouth College, Hanover, New Hampshire.

Williams, A., Sherman, I., Smarr, S., Posadas, B., Gilbert, J.E., 2018. July). Human trust factors in image analysis. In: International Conference on Applied Human Factors and Ergonomics. Springer, Cham, pp. 3–12.

Yu, N., Davis, L.S., Fritz, M., 2019. Attributing fake images to gans: learning and analyzing gan fingerprints. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7556–7566.

Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., Spangenberg, J., 2016. April). Web and social media image forensics for news professionals. In: Tenth International AAAI Conference on Web and Social Media.

Zheng, L., Zhang, Y., Thing, V.L., 2019. A survey on image tampering and its detection in real-world photos. J. Vis. Commun. Image Represent. 58, 380–399.