Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content

Ali Khodabakhsh, Raghavendra Ramachandra, Christoph Busch
Department of Information Security and Communication Technology
Norwegian University of Science and Technology, Gjøvik, Norway
{ali.khodabakhsh, raghavendra.ramachandra, christoph.busch}@ntnu.no

***Abstract*—Facilitation of fake face generation in recent years, thanks to advancements in computer graphics and artificial intelligence, raises concerns about malicious use of these techniques for personal or political gains. Media consumers are exposed to hours of audiovisual content daily, while their vulnerability to fake audiovisual content is not yet fully studied and understood. In contrast, many recent automated fake content generation techniques are readily accessible to the public. A first step to address this vulnerability is to study the effectiveness of existing methods in passing human judgment. To this end, we examined the performance of 30 participants in the detection of 48 real and fake videos. The fake videos were sourced from six different methods of generation and were collected from a public video sharing website[1], ranging from prosthetic makeup to Deepfakes. Our results show that the participants failed to detect two different types of fake videos. However, participants' detection performance improves when they know of the displayed individual or when a biometric reference video (introducing the individual and its behavior) is available to them during the test.**

***Index Terms*—Fake Face, Subjective Evaluation, Morph-cut, Deepfake,**

I. Introduction

The consumption of audiovisual content is on the rise due to the increase in network speed and the richness and appeal of such content compared to traditional forms of media. Further, the consumption of media from free and unreliable sources such as social media channels has increased dramatically in recent years. These two factors combined can cause a massive proliferation of fake news in audiovisual format. This is alarming, because in contrast to text-based fake content detection, audiovisual fake content detection is in its infancy, and only few automatic detection methods are in place with limited applicability [1]. An important case of audiovisual content is the case of talking faces, being a usual part of online videos due to it being the most natural way of communication between humans. The generation of videos of fake faces has become possible thanks to advancements in computer graphics and more recently in artificial intelligence. Many methods claim to have video-realism, some of which are available for public use. One recent example of using fake faces is the Xinhua agency's AI presenter[2].

[1]https://www.youtube.com/
[2]https://www.bbc.com/news/technology-46136504

Humans are shown to be vulnerable to digitally manipulated images [2]. In 2012, Farid et al. [3] measured the performance of humans in detecting fake face images generated using computer graphics. Their results show above chance detection accuracy in different resolutions and compression settings. In similar studies [4], [5], authors try to pinpoint contributing factors in detection such as positioning of illumination sources and shadowing, color, and partial occlusion of the face. However, in a more recent study in 2018, Rossler et al. [6] studied the detection performance of humans on fake face images extracted from a specific fake video generation algorithm. Their results show that human detection accuracy can be as low as random guessing after video refinement and compression. This study tries to provide insights into the open question, can people distinguish real videos from fake ones? The results from this study's simulated real-life scenario will shed new light on media consumer vulnerabilities; it will also provide a review of the effectiveness of new and traditional audiovisual fake face generation methods in the current point in time.

The rest of this paper is organized as follows: Section II describes the experimental methodology and includes details on the dataset, the test protocol, and the test setup. Section III discusses the results of this study and then Section IV presents our conclusions and proposals for future work.

II. Data and Methodology

In this study, a real (a.k.a bona fide) video is defined as a continuous recording of the target individual without any modification that changes the representation or appearance of that target individual and the content of the utterance. Alternatively, a fake video is anything to the contrary and can be described as either impersonated, manipulated, or synthetic media related to the target individual. The target individual is the natural person whose appearance is used for generation of the fake video.

To reach the objective of this study, a set of videos were required that represents the status of today's technology in fake video generation, and a test setup that simulates real-life video encounters.

Fig. 1: The faces in the six categories of fake faces. Going left to right, the columns correspond to the following categories in order: Look-alike, Prosthetic Makeup, CGI, Morph-cut, Face CGI, and Face GAN.

A. Dataset

The scenario in this study is limited to continuous scenes of talking heads. As to study the effect of visual and auditory features rather than the textual content of the videos, only short utterances were considered for this study. A dataset consisting of 48 videos, each five seconds in duration, were manually collected from YouTube. The videos were selected such that they have a size of at least 640×480 pixels, and were manually screened for sufficient lighting and frontal face visibility conditions. The videos are selected such that they do not contain any meaningful uttered sentence, avoiding leakage of information about the real- or fake-ness of the video.

Half of the videos fitted the criteria of "fake", and categorized to six categories based on the technique used to generate them, meanwhile the other 24 represent the "real" video control set. Due to the very limited number of actual fake videos matching the selection criteria, the selected fake material represents an extent of videos that can be used as a fake video. Following the taxonomy introduced in [7], the fake categories are as follows:

1) Physical
   a) Look-alike: The individual in the video is a look-alike of the target individual. The voice may not match the target individual.
   b) Prosthetic Makeup: The individual in the video wears prosthetic makeup and impersonates the target individual.
2) Digital
   a) Computer Graphics Imagery (CGI): The scene has been generated using CGI. The voice may come from an impersonator or the target individual.
   b) Interframe forgery (Morph-cut): To alter the spoken audio content, the video has been cut and rejoined in a seemless manner, by using the Adobe Premiere Pro Morph-cut[3] video transition.
3) Hybrid
   a) Face CGI: This technique is similar to the CGI technique, with the difference in that only the face or a part of the face was synthesized and then overlayed on the recorded footage.
   b) Face GAN: Similar to Face CGI, only the face is replaced. Yet the synthetic face is generated by Generative Adversarial Networks (GAN) using Faceswap[4] or an alternative open-source application based on the same concept.

The selection process chose the most video-realistic examples encountered from each fake category, fitting the overall criteria of duration and quality. The chosen videos in each category were further filtered for video-realism by three colleagues in our research lab. The selected videos partially overlap with the FFW dataset [8]. The sources of the videos guaranteed their status as fake. Facial regions of all the fake videos are depicted in Figure 1. For the control set, 24 videos were randomly selected from the VoxCeleb [9] dataset after filtering those with regards to the same duration and quality criteria.

To address the effect of having a biometric reference included in the test, each video in the real and fake categories was paired with a supporting biometric reference from the target individual. The selection criteria for biometric reference videos were the same as for the "real" category and partially selected from VoxCeleb dataset. The participants' detection performance was first stabilized by using a short mock test that was based on five pairs of video and biometric references that are separate from the experimental/control datasets. The target individuals in the videos were adults who were either celebrities or political personalities of varying in age and gender.

To eliminate low-level clues that participants might use to identify the fake videos, the following metrics were measured to assure an overlapping distribution between both sets: head size, head pose, image and facial quality. In both real and fake sets the average head size was ≈ 128 pixels, average BRISQUE [10] was $\approx 36\%$, and average face quality [11] was $\approx 61\%$. The distribution of facial pose in both sets also has a high overlap. The list of videos in the dataset are made available online[5].

B. Protocol

The aim of this test is to measure the following:

Participants (i.e. media consumers) performance in the detection of the most video-realistic fake samples in each category.
Effect of presence of a biometric reference upon the detection performance.
Effect of familiarizing the participants with different categories of fake content with a guide on the shortcomings

[3]https://helpx.adobe.com/premiere-pro/using/morph-cut.html

[4]https://github.com/deepfakes/faceswap

[5]http://ali.khodabakhsh.org/fake-faces-for-subjective-testing/

of each fake face generation method on their detection performance.
Effect of prior knowledge of the target individual on the detection performance.
Possible correlations between demographic information and subjective detection performance.
Common clues used by participants.

The test aims to have a measurement corresponding to the real-life scenario and utilizes a web-based interface that participants access through their personal multimedia device (limited to devices with a large display, e.g. laptop or tablet). To make sure the participants can use both modalities, they were given guidelines for screen and audio adjustment.

The experiment sessions were split into five parts. The first part was used to briefly explain the test and also collect participants' demographic information (age, gender, education, and occupation). In addition to this, the existence of any visual deficiency is probed, along with a question regarding the expected expertise of the participant in the task.

The second part consists of a familiarization step, where fake videos are described and a set of videos depicting examples of each category is shown to the participants. To measure the effectiveness of familiarization, the familiarization page is shown before the test in half of the population, and after the test in the other half.

The third step corresponds to a mock test with a fixed order. This step tries to stabilize the performance of the participants and to reduce any inconsistency in their performance caused by the learning process. This step follows the same set of questions as the rest of the test. The answers for these videos were to be discarded in the analysis.

The fourth step is the main part of the test, and was organized as follows: a video is shown to the participant, sometimes along with a biometric reference, and the participant is asked to answer a set of multiple choice questions about the video in question. The questions address the decision of the participant on the video being real or fake and ask if the participant knows of the target individual. Furthermore, the participant is asked about the main clue that led to their decision to be selected from a list of clues, with the option of mentioning additional clues in a comment box. This process is then repeated for the remaining 47 videos. To avoid any effect of ordering in the test, the videos were shown in a randomized order, and the biometric reference video appeared randomly in half of the videos. The participants were also allowed to have a *no answer* choice in the questions, if they were uncertain of their response.

Finally, a feedback page is presented to the participants that provides a visualization of their performance to reward them by increasing their awareness of their mistakes and vulnerabilities.

C. Test Setup

The test was implemented using the online survey tool Limesurvey [12]. The participants were invited by an email that included a token which limited each participant to a single

Fig. 2: The test interface for a sample video with a biometric reference based on the Limesurvey tool.

test trial. The participants were able to stop the test at any point and resume later. The participants were asked to take the test on a large display with adequate brightness at arms distance, and have their audio on, and to be connected to a high-speed internet connection.

The first page of the test included the previously described demographic questions. International Standard Classification of Education (ISCED) 2011 was used to measure the participants' highest completed level of education and Standard Occupational Classification (SOC) System was used to classify their occupation. The participants were asked if they have any deficiencies in their vision, defined as any deficiency that had not been corrected (e.g. by corrective lenses) at the time of the test. They were also asked about their level of expected expertise in the task and given a choice between none, very low, moderate, quite high, and very high. The parameters corresponding to the ordering of videos, biometric reference, and familiarization page was randomly initialized and saved for the analysis step.

The familiarization page is now available online[6]. It contains seven videos illustrating the different fake content generation methods used in the test, along with a description of fake video categires used in this study, and the artifacts they typically create.

A typical test survey page with a biometric reference is shown in figure 2. The playback quality of videos were set to "medium"[7], corresponding to 30fps, 360p videos in VP9

[6]http://ali.khodabakhsh.org/research/fake-faces-and-fake-face-detection/
[7]https://developers.google.com/youtube/iframe_api_reference

format for video and Opus for audio. The mock test included five videos similar to the actual test, with two real and three fake videos of which two had a biometric reference while three were presented alone.

The mock test was followed by the main test, where the 48 videos were presented in a randomized order, 24 with biometric reference and 24 without biometric reference. The time taken to answer each question set is also recorded.

D. Performance Evaluation

The performance evaluation metrics used in the experiment are from the ISO/IEC 30107-3 standard [13], they include: Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). APCER measures the proportion of fake (i.e. presentation attack) videos incorrectly classified as real (i.e. bona fide), while BPCER measures the proportion of real videos mistaken for fake.

In addition, to evaluate the confidence intervals for detection accuracies, Clopper-Pearson method was used on the binomial distribution of decisions with a 95% confidence interval. For evaluating the significance of difference between distributions, two-tailed student's t-test was used with a significance threshold of 0.05 (specified otherwise). Lastly, Pearson's correlations were reported along with their confidence interval using a Student's t distribution for a transformation of the correlation, and p-values below 0.05 were considered significant [14].

III. Results and Discussion

The results presented in this work are based on the participation of volunteers affiliated with our campus, as well as acquaints who were interested in taking the test. During four weeks 30 people participated in the test. 60% of the participants have a master degree, while 23% have a doctorate. 77% of the participants self-identified as male while the remaining self-identified as female. 67% were employed in Computer and Mathematics, while 13% were variously employed in Education, Training, and Library services. The average age was 31.2 with a standard deviation of 7.5. The participants' average time to complete the test was 39 minutes; this corresponds to an average of 37.6 seconds per video and 4.5 minutes for familiarization.

Out of 30 participants, five had vision deficiencies; but their performance was not statistically significantly different from the performance of the rest of the experimental cohort, so their data has been included (p-value of t-test is 0.58, 0.29, and 0.66 for correct, uncertain, and incorrect choices. $n = 5$ for with and $n = 25$ for without vision deficiency.). Out of 30 participants, 18 expected to have moderate expertise and six expected to have very low expertise. The remaining six participants were equally distributed between *quite high* and *none*. The participants were of different nationalities, with 93% from Eurasia. The participants, when asked, did not mention any mismatch in presentation or low-level patterns useful for distinguishing between real and fake videos.

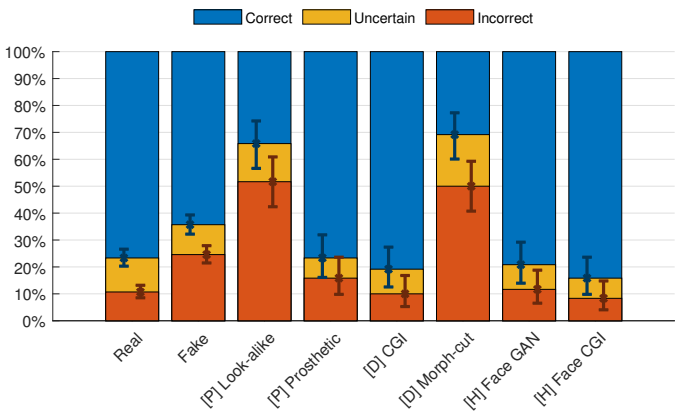

Fig. 3: Overall choice percentages with 95% confidence intervals of the participants in the real and fake categories, along with each subcategory of fake videos. The letter before the fake category names correspond to the classification of fake videos ([P] Physical, [D] Digital, and [H] Hybrid) [7].

The participants had a below 30% BPCER and APCER in detecting real and fake videos respectively, except for the look-alike and Morph-cut categories. No statistical significance (with a 95% confidence) was observed between the other categories of the fake and average time taken to answer each question per category. Figure 3 shows the percentage of correct, uncertain, and incorrect identification of real and fake videos, along with the performance in each fake category separately.

Figure 4 shows the percentages of correct, uncertain, and incorrect identification for every single video sorted by detection accuracy, along with their corresponding category. The look-alike and morph-cut videos are gathered around the left-hand side, while the other four categories are distributed in-between the real videos. A close inspection of outliers in each category shows these videos having special lighting conditions. For example, the most misclassified sample of prosthetic makeup is the face on the fourth row, second column, in Figure 1. The most misclassified example of CGI and Face GAN are the faces at row one column three and column six respectively. It is also interesting to observe that the percentage of uncertain answers per video never exceeds 25% even when the percentage of incorrect reaches above 50%. This implies that the participants were on average, confident of their decision in all the videos. The three videos that were classified correctly 100% of the time were of well-known political personalities (presidents of the united states) and are shown in row one column two, row two column five, and row four column six in Figure 1.

The most common main clues used is shown in Figure 5. The difference in usage shows participants relied mostly on Head/Face compared to other clues. It is also interesting to see that the distribution of clues is different in fake and real videos. In addition, some clues resulted in different performance across classes. For example, when participants mentioned movements as the main clue, BPCER was 95%

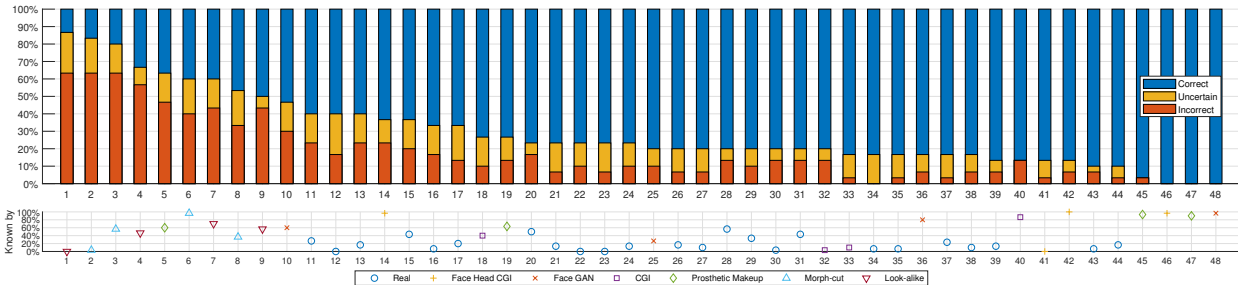

Fig. 4: The percentage of correct, uncertain, and incorrect choices per video, sorted by the percentage of correct from low to high. The category of videos is shown in the plot below with colored markers along with the percentage of population that knew the subject in the video. Look-alike and Morph-cut samples are concentrated in the left side of the graph.

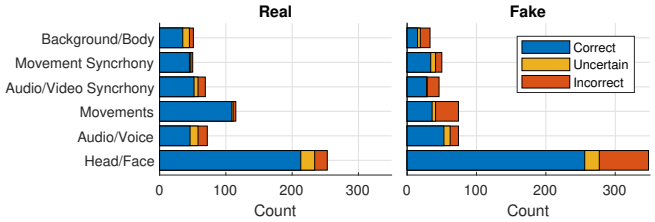

Fig. 5: The number of correct, uncertain, and incorrect choices, given the most common main clue selected by the participants. The difference in distribution and accuracy of clues in real and fake categories are visible.

while APCER was only 49%. No statistical significance (with a 95% confidence) was observed between the accuracy of detection given a specific clue due to small sample size. To measure clue diversity per participant compared to the clue diversity in the whole group, the clue entropy is calculated. The average participant entropy was measured to be 2:36 while the total entropy was 3:12, showing that participants tended to focus on a smaller set of clues in comparison to the population.

The presence of familiarization was accompanied by a shift in the distribution of incorrect percentage towards lower values (t-test $p = 0{:}07$, $n = 12$ for with and $n = 18$ for without familiarization) and reduced the inter-participant variability for incorrect and uncertain responses as shown in Figure 6a. Yet this reduction only caused an insignificant increase in uncertain and correct responses (t-test $p = 0{:}90$ and $0{:}60$ respectively at aforementioned sample sizes). This shows that the provided familiarization oriented their decisions, yet was not effective in increasing their overall accuracy. As shown in Figures 6b and 6c, Having a biometric reference shifted the distribution of incorrect percentages to lower values (t-test $p = 0{:}05$, $n = 30$ for both conditions), while knowing the target individual mostly shifted the distribution of uncertain percentages in the same direction (t-test $p < 0{:}01$, $n = 30$ for both conditions). The distribution of correct percentages was shifted towards higher values when the target individual was known (t-test $p < 0{:}01$).

The following were observed between the demographic information and the performance of individual participants: Due to the small population size no significant correlation was observed comparing the level of education and gender to performance. Level of expected expertise in the task had a positive trend in comparison to the number of correct responses, yet the 95% confidence intervals for these values were overlapping. A moderate positive correlation was observed between the age and the number of incorrect answers ($p = 0{:}07$), simultaneously a moderate negative correlation existed between the age and the number of uncertain ($p = 0{:}02$), canceling the overall effect on the number of correct, as shown in Figure 7.

IV. Conclusion and Future Work

We evaluated the performance of 30 participants in distinguishing fake videos from real ones using a web-based platform. 48 pair of videos were collected from an online video sharing website, 24 of which could fit the definition of fake and were generated using six different methods ranging from prosthetic makeup to Deepfakes.

The results suggest the vulnerability of participants to the traditional methods more than the new methods, specifically to look-alikes and interframe forgery. This aligns well with the long history of use of look-alikes as fake faces, especially as political decoys. Interframe forgery, on the other hand, has a limited footprint as it only affects a part of the video. The footprint is further covered using the morph-cut technique for smoothing the transition in jump cuts. Yet both these techniques are expensive in practice, due to the difficulty of finding look-alike impersonators, and of finding long videos depicting consistent scenes of the target person to be used in the morph-cut setting.

It can also be concluded that the selected fake videos from CGI, Face CGI, Face GAN, and Prosthetic Makeup techniques had not yet reached convincing video-realism. The results also suggest special lighting setups to be effective in resulting in more errors in the population, obfuscating the artifacts caused by the generation method.

The existence of a biometric reference reduces the number of errors, while knowing of the target individual reduces the uncertainty, contributing to a higher number of correct classification. The presented familiarization was not effective in increasing the accuracy of participants, yet it caused a lower number of incorrect choices which was in turn compensated with a higher number of uncertain ones. Furthermore, it is observed that individuals rely on a small set of clues for their decision, and the main clue supporting the participants' decision is in the head/face area.