Towards Understanding Perceptual Differences between Genuine and Face-Swapped Videos

Leslie Wöhler
woehler@cg.cs.tu-bs.de
Institut für Computergraphik, TU Braunschweig
Braunschweig, Germany

Susana Castillo
castillo@cg.cs.tu-bs.de
Institut für Computergraphik, TU Braunschweig
Braunschweig, Germany

Martin Zembaty
m.zembaty@cg.cs.tu-bs.de
Institut für Computergraphik, TU Braunschweig
Braunschweig, Germany

Marcus Magnor
magnor@cg.cs.tu-bs.de
Institut für Computergraphik, TU Braunschweig
Braunschweig, Germany

Figure 1: We estimate perceptual differences between genuine and manipulated videos. Towards that goal, we use face swaps as stimuli and perform three types of experiments focusing on eye tracking, different video durations, and the assessment of emotions.

ABSTRACT

In this paper, we report on perceptual experiments indicating that there are distinct and quantitatively measurable differences in the way we visually perceive genuine versus face-swapped videos.

Recent progress in deep learning has made face-swapping techniques a powerful tool for creative purposes, but also a means for unethical forgeries. Currently, it remains unclear why people are misled, and which indicators they use to recognize potential manipulations. Here, we conduct three perceptual experiments focusing on a wide range of aspects: the conspicuousness of artifacts, the viewing behavior using eye tracking, the recognition accuracy for different video lengths, and the assessment of emotions.

Our experiments show that responses differ distinctly when watching manipulated as opposed to original faces, from which we derive perceptual cues to recognize face swaps. By investigating physiologically measurable signals, our findings yield valuable insights that may also be useful for advanced algorithmic detection.

CCS CONCEPTS

- Human-centered computing → Human computer interaction (HCI); User studies; • Computing methodologies → Perception; Image manipulation.

KEYWORDS

video manipulation, human perception, eye tracking, face swapping

ACM Reference Format:

1 INTRODUCTION

Following recent technological advances, facial manipulations in videos are becoming ubiquitously integrated into our everyday live appearing in advertisements, movies, and on social media platforms. Especially face swap videos, where faces of celebrities have been exchanged, have recently received media attention [15]. While these face swaps are an extremely powerful tool in creative fields and the entertainment sector, they also pose a potential threat to society. By exchanging the original face in a video with that of a different target person, offensive actions or illicit behaviour can be attributed to arbitrary people. One field with enormous potential to be abused is politics, where face swaps are feared to increase mistrust in politicians and parties, which may drive voters towards more radical groups [18, 63, 73]. Furthermore, due to the ease of use and public availability of face-swapping frameworks [16, 17, 23], their
application is not limited to specialized users. Therefore, realistic face swaps can be generated by any individual and applied to social contacts, e.g., in the scenario of cyber-bullying. Considering these negative social implications, fundamental understanding of human perception on face swaps as well as reliable detection methods for facial manipulations are required.

The Computer Vision community is constantly in pursuit of new methods to improve the generation and detection of manipulated video content. While most of the detection systems are specifically tailored towards image analysis and the detection of artifacts [44, 67, 70], it is still unclear why and when people are misled by or successfully able to recognize face swaps. As humans are inherently very sensitive to faces and usually excel at facial recognition tasks, a deeper understanding of this phenomenon would be an important extension to current detectors.

In this paper, we aim to gain knowledge on the perception on face swap videos and investigate which features or artifacts are important for humans to recognize the forgeries. The resulting insights can help to raise the awareness of viewers and learn about common artifacts in face swaps. Moreover, we are looking into measurable physiological responses which can be used to aid automatic detection tools. To achieve this, we conduct different experiments in order to understand the perception on face swaps. First, we investigate eye tracking, which can be used to assess different types of artifacts [22, 62] and is a measurable response on various devices [28, 40, 72]. Therefore it could be integrated into face swap detection frameworks. In our experiment, we aim to obtain insights into differences in viewing behaviour between real and face-swapped videos. We further want to assess which facial areas are most important for the visual detection of manipulations as this may give hints towards noticeable and distracting artifacts. As a second step, we look at the influence of video duration on participants’ detection accuracy. We hypothesize that, the longer a video is, the more information the viewer gets, which potentially leads to a higher chance to notice manipulations. Finally, we move to a field that is still challenging to assess for computers but easy for humans: conveyed emotions and expressions [71]. We are mainly interested in whether face swaps are able to convey the same message as the corresponding real videos used to generate the facial movements. We assess this by looking at differences in the recognition as well as intensity and sincerity ratings of emotions and expressions between face swaps and real videos. Based on this data, we obtain a better estimate of the possible impact of face swaps than by just looking at their consciousness. We further investigate if participants’ ratings indicate a general mismatch between the conveyed emotion of real and face swap videos.

In summary, we contribute answers to the following research questions:
- How is gaze behaviour impacted by face swaps? Can eye tracking be used to detect facial manipulations?
- Does the length of video clips influence participants’ assessment accuracy?
- Are conveyed emotions and expressions different between face swaps and genuine videos?

2 RELATED WORK

In this section we briefly discuss current techniques for the creation and detection of facial manipulations as well as perceptual factors in the context of facial processing.

Facial Manipulation and Detection Techniques. Among the existing facial manipulation techniques, this paper focuses on face-swapping. This technique applies a face from one video to another video while keeping the original body and expressions. Many algorithms have been proposed for face-swapping using deep learning techniques. While most approaches require a training on both target and source subjects [39, 54], recent work also introduced a face agnostic method [53].

The fast progress of face-swapping techniques, has led to a high interest in facial manipulation detection methods. Many methods focus on detecting errors and artifacts produced by deep learning-based facial manipulation techniques [2, 27, 44, 68, 70]. Another line of work has analyzed the specific properties of human facial motion and physiological measurements to detect mismatches in tampered videos. This way, unnatural blinking behaviour [43] and the heart rate of actors were analyzed to detect manipulations [24]. Further, it was proposed to learn person-specific facial motion cues to reliably detect manipulations of specific subjects [3]. While all these methods are based on human behaviour and biology, none of them directly uses feedback from the observer of the manipulated videos. Thus, in this paper we aim to incorporate perceptual insights into the detection pipeline using eye tracking.

Perception of Faces. Faces gather most of our attention in social situations, as we rely heavily on them to recognize and assess information and emotions. Therefore, humans are highly specialized in processing and analyzing faces [60].

In order to gain insights into the underlying processes of facial recognition and exploration, eye tracking has been applied to facial images and portrait videos. Early research showed that the eye, nose, and mouth regions are fixated in facial images [34, 48]. Especially the eyes were found to draw the viewer’s attention [6]. The gazing behaviour in facial images is, however, influenced by various factors like the gender [46], the presence of artifacts [7] or the familiarity with the shown face [4, 64]. Similarly, the emotions displayed on a face [11, 20, 42] and the task can influence the viewing behaviour of participants [8, 9, 41, 56]. It was additionally found that viewing behaviour is aligned to motion [50] and varies based on the performed actions like talking or establishing eye contact [61]. Finally, with the availability of speech and audio, more fixations occur on the mouth region, in contrast to muted videos [65].

Eye tracking. Analyzing gaze has not only been employed in facial processing research, but also to assess artifacts in videos. Thereby previous research [10, 13, 22, 62] found that artifacts attract the gaze of observers. As face swapping introduces artifacts in different areas of the face, we hypothesize that participants will look at artifacts and their gaze differs between real and swap conditions. Recently, the idea of exploring participants’ gaze in deepfake videos has arisen indicating general interest in this topic [29]. In contrast to our work, the authors only use partial face swaps in unrestricted environments. Furthermore, their analysis focuses only on eye
tracking statistics, like the number of fixations and their duration, without consideration of the fixated facial areas.

Considering the highly specialized viewing behaviour for faces and previous insights in the detection of artifacts via eye tracking, we investigate whether original and manipulated videos evoke differences in gaze which could be used in facial manipulation detection.

**Emotions and Expressions.** Next to general recognition of faces, humans are also very good at recognizing emotions and expressions. During conversations, information is not only conveyed through speech but largely through facial expressions [47]. If the message conveyed by the words and the perceived emotion do not match, more emphasis is put on the expression of the speaker [12]. In the case of face swaps, this means that it is of high importance that the expressions and emotions in the original video are retained to evoke the desired effect in the viewer. However, emotion perception is not only based on facial movements but also heavily influenced by contextual cues like the situation, body language, or cultural factors [5]. Moreover, people base their assessment of emotions and expressions on previous knowledge of the speaker [57]. For face swaps, this could produce a strong mismatch for viewers watching manipulated videos of people they know. As a first step, we set out to investigate whether facial expressions and emotions are perceived differently between unaltered videos and their corresponding fake counterparts for unknown actors.

**Face Swap Datasets.** In order to unify research on the detection of AI-synthesized face swap videos, numerous facial manipulation datasets have been introduced. Among these, one contains short clips (< 10 seconds) in a controlled environment [39], one uses publicly available videos of celebrities [45], and others focus on news shows [26, 58, 59]. As an interesting concept for perceptual research, Jiang et al. [35] included some recordings in a controlled environment. However, this dataset is still clearly oriented towards the training and evaluation of neural networks for manipulation detection. In contrast to these works, the PEFS dataset [69] was specifically designed for perceptual research and contains partial human annotations towards the realism and artifacts of the stimuli. The videos from this dataset were recorded in a controlled environment with three camera angles, feature various induced emotions and expressions as well as long video durations and different quality levels for the obtained face actors. Therefore, we mainly use stimuli from the PEFS dataset in our experiments.

## 3 Experimental Design

We first formulate our research hypotheses to then detail the design and procedure for our three experiments.

### 3.1 Hypotheses

Based on our initial research questions, we formalize several hypotheses regarding the perception of face swaps:

- **H1:** Even without knowing about the occurrence of manipulations, participants are able to pick up artifacts in state-of-the-art face swap videos.
- **H2:** Eye tracking data differs between real and manipulated videos.
- **H3:** The length of videos has an effect on the recognition accuracy as manipulations are easier to detect in longer videos.
- **H4:** While face swaps can retain the recognizability of emotions, their intensity and sincerity can differ.

Based on these hypotheses, which we discuss in detail in the following sections.

### 3.2 Experiment E1: Eye Tracking

In experiment E1, we recorded participants’ eye movements while watching real videos and face swaps in order to assess their viewing behaviour. After each trial, we additionally asked participants to report whether they noticed something regarding the video quality like artifacts. As we are interested in their unbiased impressions and viewing behaviour, we did not inform them about the face swaps. In order to investigate the effect of artifacts and their conspicuousness, we used high quality and low quality face swaps.

**Stimuli.** For this experiment, it is crucial that the stimuli have uniform background and illumination – to keep the attention of the viewer on the actor –, as well as that all used stimuli are consistent enough to compare viewing behaviour between videos. As previously mentioned, the restricted setup used in the recordings from the PEFS dataset [69] ensures the satisfaction of both conditions, in contrast to other available datasets. This dataset consists of muted
video portraits taken in a controlled environment where the actors are seated, one at a time, in front of a white wall and recorded while talking. We used the same 11 annotated videos with 60 seconds length (25 Hz) as in the experiments in the original paper consisting of 2 female–female, 2 inter–gender and 7 male–male face swaps. An overview for the selected face swaps can be seen in Fig. 2. We used the real videos as well as their corresponding high-quality and low-quality manipulations. Some differences between the quality of the stimuli are highlighted in Fig. 3.

**Apparatus.** We conducted the experiment using an EyeLink 1000 eye tracker by SR Research Ltd. with a sampling frequency of 1000 Hz which was placed at 65 cm distance to the participants and performed monocular tracking of the right eye. The videos were displayed on a 47-inch screen (100 Hz, 1920 × 1080 pixel) positioned at a distance of 90 cm from a chin rest where participants placed their head. Before each session, the eye tracker was calibrated using a 9-point calibration and adjusted via drift correction between trials. The participants sat behind black curtains which prevented direct contact between them and the experiment conductor and ensured a darkened environment avoiding external distractions.

**Participants.** We invited 40 participants to the experiment. The participants consisted of 22 females and 18 males with ages between 18 and 35, an average age of 23.15 with a standard deviation (SD) of 3.98. All of the participants were university students and came from the fields of computer science, psychology, engineering, and business studies. They could choose to either receive one course credit or 10 EUR as compensation for their participation. Participants were mostly of German nationality with one Russian, one South African and two Vietnamese nationals. Every participant reported normal or corrected to normal vision.

**Procedure.** The experiment used a counterbalanced design with full randomization, the real-manipulation pairs and the video quality as the in-between participant factor.

First, participants filled out a demographic questionnaire and an informed consent form. Afterwards, they were instructed about the general experimental setup. We briefly explained the eye tracker workings and that videos would be displayed during the experiment. Participants were informed that we were especially interested in observations regarding the video quality. They were, however, not informed of the purpose of the experiment and the usage of face-swapping.

Participants were instructed to sit on a chair with as little movement as possible to not compromise the eye tracking. Additionally, we asked them not to speak while viewing the videos. Before a video was displayed, a fixation cross appeared in the center of the screen for three seconds. After each video, the question ‘Did you notice anything in the video?’ was displayed, which the participants answered via oral free description and was transcribed by the conductor. The procedure was repeated for 11 trials randomly selecting either five real and six swapped videos or vice-versa. During randomization, we made sure that no participant would see a manipulated video and its original counterpart.

After all trials, we performed a debriefing with the participants. We first asked about their general impressions of the faces in the videos in order to gain further knowledge about suspected manipulations and conspicuous artifacts. Finally, we explained the concept of face-swapping, asked whether they knew about this concept, and
informed them about the purpose of our experiment. The average duration of the experiment was around 30 minutes.

3.3 Experiment E2: Video Length

In our next experiment, we focused on the accuracy of our participants at detecting face swap videos with different durations.

Stimuli. Given the more general and less restrictive nature of the research question we want to address in this experiment, we increased the number of stimuli and considered several sources. This allows a more general look at the quality of state-of-the-art face swaps. Next to the 11 face swaps used in E1, we selected four more swaps from PEFS [69]. For each swap, we included sequences from both frontal and right viewing angle. For this experiment, we only used high-quality manipulations and their matching real videos. We further included 16 face swaps from FaceForensics [58] which contain short clips of news shows with one news anchor. We chose these stimuli manually, aiming to include only face swaps of high quality. Example frames for the stimuli are shown in Fig. 2. Finally, we cut each video to the length of either 3, 5, 10, 30 or 60 seconds to obtain a good sample of time spans. Hereby, the same length is used for each face swap and the corresponding real video.

Participants. We performed the experiment using Amazon Mechanical Turk. Participants were compensated with 1 USD and were from the United States, India, and Brazil. Overall, 40 participants (age range: 21-53) took part in this experiment.

Procedure. Before the experiment, participants were educated about face swaps and their task of detecting the face-swapped videos. Each trial consisted of a video stimulus and one question with two possible answers. The video playback started automatically and could not be paused or repeated. Afterwards, the video was removed and instead the question “Was this video manipulated?” was displayed. It appeared together with two answer options real/manipulated as a 2 alternatives forced choice task (2AFC). We performed this experiment using a counterbalanced design with full randomization and the real-manipulation pairs as the in-between participant factor, making sure that no participant saw both a face swap and its corresponding real version. Each participant completed a total of 44 trials which were selected at random. The experiment took around 15 minutes to complete.

3.4 Experiment E3: Emotion Assessment

In experiment E3 we aim to assess whether the conveyed emotions and expressions of face swaps differ from their real counterparts by looking at their recognition accuracy and ratings for intensity and sincerity.

Stimuli. In contrast to other datasets, one part of the recordings in PEFS [69] focuses on evoking emotions in the actors using a method acting protocol [37]. We used a set of these recordings in which the emotions and expressions have been isolated to short clips. Each of the clips is labeled with the emotion which was evoked in the actor. The set consists of 9 face swaps from the PEFS dataset (2 female-female, 2 inter-gender, 5 male-male), see Fig. 2. Our selection of stimuli was based on primary emotions and conversational expressions using 5 emotions proposed by Ekman [21] (Happiness [Hap], Sadness [Sad], Anger [Ang], Disgust [Disg], Surprise [Sur]) and 4 conversational expressions [14] (Agree [Agr], Disagree [Disa], Thinking [Thi] and Clueless [Clu]). We showed these together with Neutral [Neu] for reference, totaling in 10 expressions. This way, our main consideration was on expressions and emotions likely to occur in everyday conversations [14].

In this experiment, we only used high quality face swaps as these are highly relevant due to their presence in modern entertainment media or usage potential for defamatory content. Additionally, the
high quality stimuli contain less distracting artifacts allowing participants to focus on the expressions and movements in the videos. Overall, our stimuli set consisted of a variety of 10 emotions/expressions per video for 9 face swaps and 9 corresponding real videos, leading to a total of 180 stimuli with an average duration of 3 seconds per video. Example frames of the stimuli can be found in Fig. 4.

Participants. We performed the experiment online, while gathering participants via university mailing lists. We recruited 21 students who received a compensation of 10 EUR for the completion of the experiment. Participants were between 19 and 33 years old with an average age of 25 years (SD = 3.81). One participant was of Vietnamese nationality, the rest were German. The balance of genders was nearly equal with 10 male and 11 female participants.

Procedure. The experiment began with an explanation of the task of assessing emotions in video portraits. We further explained that facial manipulations were applied in some of the videos, however, this should not be taken into account when rating sincerity. After participants read the instructions, we asked for their nationality, gender, and age. During the main experiment, participants started the playback of a video by the press of a button. The video was only played once and no interaction (pausing/rewinding) was possible. In order to avoid the analysis of a still frame, we removed the video after playback. The participants had to choose one option from a list containing all emotions and expressions included in the experiment (10AFC task). Afterwards the participant had to rate the intensity and sincerity of the shown emotion/expression on a 7-point Likert-scale (1 indicating extremely low and 7 extremely high). All trials were chosen in a fully randomized manner. In contrast to the previous experiments, here we used a full within design regarding stimuli. Overall, the experiment took around 45 minutes.

4 ANALYSIS

In this section we evaluate our hypotheses based on the data obtained in the experiments. In the following we refer to the different conditions as follows: HQ for high-quality stimuli, with RealHQ and SwapHQ for the real videos and manipulated videos; and LQ for the low-quality stimuli, with RealLQ and SwapLQ. Note that, the real videos are identical, however, participants’ responses may be biased by the difference in manipulated stimuli and, therefore, we analyze them independently.

4.1 H1: Detection of Artifacts in Face-Swapped Videos

In our first hypothesis, H1, we pose that participants are able to spot artifacts in state-of-the-art face swap videos, even if they are not informed about the applied manipulations beforehand. We used a free description task in E1 asking participants to report anything they notice about the video quality. Therefore, we avoided biasing the participants and can explore whether uninformed participants suspect facial manipulations and which kind of artifacts are noticeable without pointing the participants towards them. Please note that the nature of the task determines that there are no fixed answers, and no fixed number of possible occurrences for an answer — aside from the maximum: numberTrials \times numberParticipantsPerCondition = 110.

For our analysis, we exclusively consider artifacts reported in the facial region as only these could be introduced due to the utilized face-swapping approach. Participants sometimes stated which areas of the face are affected by artifacts as shown in Fig. 5 (right). The most commonly reported artifacts were blur, facial manipulations (including face swaps, partial face alterations, and beauty filters), unnatural expressions or eye movements, and contour artifacts, see Fig. 5 (middle). Two participants did not report any artifacts on the faces.

Using this data, we compute the rate of correctly assessed videos for each condition as shown in Fig. 5 (left). Real videos were seldom reported to contain artifacts and therefore have a high correct assessment rate. Meanwhile, it is noticeable that the correct assessment rate of SwapHQ trials is under 50% of the videos (Mean = 44.44%, standard error of the mean (SEM) = 4.78). As participants therefore only reported artifacts in about half of the videos, face swaps created by publicly available frameworks seem
to mislead many viewers. In contrast, 83.49% (SEM = 3.56) of the SwapLQ trials were reported to contain artifacts demonstrating their noticeable lower quality. This is further confirmed by an analysis on the correct reports yielding significant differences between conditions (ANOVA F(3, 2636) = 189.934, p < 0.000001).

These results are especially interesting in comparison with the annotations of the PEFS stimuli [69]. In their experiment, the authors showed participants the same 11 face swaps and real videos used in our experiment, but informed them beforehand of the existence of face swaps and had them decide whether a video was real or manipulated (2AFC task). While the assessments of their participants are overall similar to ours, they differ in the SwapLQ condition. In their experiment 47% of participants reported these videos to be face swaps, but in our experiment 83.49% (SEM = 3.56) noticed facial artifacts. This may indicate that even though participants consciously perceive artifacts in low quality face swaps, they still believe them to stem from the video quality and do not attribute them to face-swapping. In our experiment, full or partial exchange or manipulation of facial features were suspected only seven times for SwapHQ and three times for RealHQ videos. In contrast, manipulated faces were reported 31 times for SwapLQ and six times for RealLQ (out of 110 possible occurrences).

These results indicate that participants are able to detect facial artifacts in high quality face-swapped videos, however, they only seldom become suspicious of the face swaps.

4.2 H2: Differences in Viewing Behaviour between Real and Swapped Videos

In our second hypothesis, H2, we posed that the viewing behaviour of participants differs between originals and face swaps as gaze is affected by artifacts. As these are not bound to a specific location [69], H2 is exploratory.

We base our evaluation of the eye tracking data on areas of interest (A0Is). Similar to previous research [30], we generate the A0Is automatically, using facial landmark detection. We first extract 68-facial landmarks for each frame using Python’s Dlib library [38] and afterwards group these landmarks to facial areas (Eyes, Mouth, Nose, Contour), see Fig. 6. Each fixation was then assigned to the nearest landmark, to the background, or to the face outside of the A0Is. Afterwards, we compute the cumulative duration of fixations that fall inside each AOI, i.e., the time spent looking at each area. A visualization of the average time spend looking at each facial area, over all videos and participants, is shown in Fig. 7 (left).

During our statistical analysis, we use a repeated measures ANOVA (RMANOVA) to compare fixations within one condition with the A0Is as the within subject factor. We check for sphericity using Mauchly and correct the results with Greenhouse-Geisser. Afterwards, we use pair-wise Bonferroni corrected t-tests for the post hoc analysis. For comparisons between conditions, we use a Welch’s ANOVA as the assumption of homogeneity of variance was violated (Levene) and Tukey as a post hoc test to investigate all pairwise differences with Bonferroni-corrected p-values.

Gaze Behaviour based on Areas of Interest. We first analyze the distribution of fixations on eyes, mouth and nose within each condition, as these are most important for facial processing in images and videos [34, 48, 65]. For both types of real videos as well as for SwapLQ, the attention of participants seems to be equally balanced between mouth, nose, and eyes with no significant differences in their distribution (RMANOVA all F’s < 0.5, all p’s > 0.05).

In contrast to this, fixations are not equally distributed for the SwapHQ condition. In this condition, participants focused more on the mouth and nose while less on the eyes (RMANOVA F(2,175) = 3.56, p = 0.04). Testing all pairs of fixation areas for the SwapHQ condition using a pair-wise Bonferroni corrected t-test, yields no significant differences (all p’s > 0.087). Based on these results, we can conclude that the distribution of fixations in videos differs between face swaps and original videos. Especially, a lower amount of fixations on the eyes indicate manipulations in high quality videos.

As a next step, we want to assess whether participants inhibit a different viewing behaviour between manipulated and original videos. For this, we include fixations on the contour as we observed artifact reports for this area, see Fig. 5 (right). In this scenario, the video conditions as well as the A0Is are dependent variables, therefore, we first assess the data with a multivariate ANOVA which yields a significant result (F(15, 1302) = 2.48, p = 0.0013). To afterwards analyze differences between the conditions on A0Is, we perform a Welch’s ANOVA. We find significant differences for the fixations on the mouth between all conditions (F(3, 240) = 3.25, p = 0.022). This also holds true for the fixations on the contour region (F(3, 242) = 3.86, p = 0.01).

This result is also visible in Fig. 7 (left) where contours are focused more for SwapLQ and the mouth for SwapHQ trials. The bias of fixation towards the facial contours in SwapLQ could be attributed to contour artifacts (example shown in Fig. 3), which were reported more often in this condition than in the others. While contour artifacts were less often reported than eye and mouth artifacts, they seem to be more salient. This may indicate, that not all types of artifacts are equally fixated. Therefore, the saliency of artifacts occurring in face swaps may not only be based on their visibility but also on their spatial occurrence.

Viewing behaviour based on reported artifacts. We next look at the fixation times based on the self-assessments in order to analyze which regions contribute most to the report of artifacts, see Fig. 7. The right plot shows trials without artifact reports.
Following the results on the analysis of fixations, we compare the change in fixations between trials with and without artifact reports with a Tukey test. For the RealLQ condition, if artifacts were reported participants focused more on the face ($p = 0.0142$; $\text{Mean}_{\text{report}} = 5297.26$, $\text{SEM} = 495.89$; $\text{Mean}_{\text{noreport}} = 860.0$, $\text{SEM} = 256.9$). This may indicate that the face was analyzed more as participants were looking for artifacts previously encountered in the low quality manipulations. For the SwapLQ condition, trials with artifact reports show more fixations on the contours ($p = 0.04$; $\text{Mean}_{\text{report}} = 15531.3$, $\text{SEM} = 1121.56$; $\text{Mean}_{\text{noreport}} = 9740.0$, $\text{SEM} = 2631.97$) and mouth ($p = 0.045$; $\text{Mean}_{\text{report}} = 11848.26$, $\text{SEM} = 663.48$; $\text{Mean}_{\text{noreport}} = 8368.89$, $\text{SEM} = 1894.8$). This seems reasonable, as artifacts were reported often for these areas. The eyes are not focused more in reported trials, even though they were as often reported to show artifacts as the mouth region. Furthermore, participants generally looked less onto the face for unreported trials, indicating that they missed artifacts while watching different areas of the video ($p = 0.000225$; $\text{Mean}_{\text{report}} = 3651.3$, $\text{SEM} = 352.98$; $\text{Mean}_{\text{noreport}} = 7062.22$, $\text{SEM} = 919.96$). In contrast, no significant difference between reported and unreported trials is found for the SwapHQ condition (all $p$’s $> 0.09$). This is aligned with the overall reduced number of reported artifacts.

Considering the distribution of fixations on nose, mouth, and eyes for SwapHQ separately for reported and unreported artifacts, yields only a significance for artifact reports (Welch’s $F(2, 95) = 3.71$, $p = 0.028$). This means, we can only detect a shift in viewing behaviour in trials where artifacts were reported consciously.

4.3 H3: Higher detection accuracy for longer videos

We formulated the directed hypothesis H3 stating that a video’s length and correct assessment should be positively correlated as the amount of artifacts increases with the length of face swaps. This is based on previous research stating that the duration of stimuli affects both attention [19, 31] and performance on artifact detection [49].

Therefore, we visualize the assessment accuracy of E2 based on the length of stimuli in Fig. 8. The plot shows that the overall recognition rate for original videos of both datasets is rather constant, while face swaps show slight variations. (Please note that FaceForensics does consist of short clips under 60 seconds and was therefore excluded from the condition with that length.) However, we find no significant differences in correct assessments between the different video lengths (Welch’s $F(3, 15.6) = 1.1$, $p = 0.379$). This may indicate that participants decide early whether a video is real or not.

As this was an interesting observation for us, we also decided to look into the eye tracking data of E1 for different time spans. To investigate if there is a specific gaze pattern early in the videos, we look at the fixations in the first 3, 5, 10, 30 and the full 60 seconds based on AOIs. The bias we found in the SwapHQ condition (see Sec. 4.2), where participants focused more on mouth and nose and less on the eyes, is present in all of these time frames (RMANOVA 3s: $F(2, 202) = 6.78$, $p = 0.002$; 5s: $F(2, 195) = 5.2$, $p = 0.008$; 10s: $F(2, 189)= 5.64$, $p = 0.006$; 30s: $F(2, 176) = 4.02$, $p = 0.027$; 60s: $F(2, 175) = 3.56$, $p = 0.04$). This indicates that the bias may be independent of the video length.
Figure 9: E3: Correct recognition percentages (left), intensity (middle) and sincerity (right) rates, all averaged among participants and videos. Error bars represent the SEM, the chance line for recognition is drawn in black, and the color legend is common to all graphs.

4.4 H4: Recognition, Intensity and Sincerity Differences

Our fourth hypothesis (H4) is that the recognition accuracy of emotions and expressions is the same for real videos and face swaps, however, the perceived intensity and sincerity differ.

For face swap videos, the face of a target is applied to a video of another individual, while keeping the second person’s movements and facial expressions. Therefore, a general goal of face swaps is to stay as close as possible to the targeted facial expressions in order to convey the same message as the unaltered video. Another area that focuses on how the conveyed message is impacted by facial alterations is the stylization of videos. We follow a well known analysis in this field and assess emotions and expressions by examining the recognized emotion along with its perceived intensity and sincerity [66]. To check the results for significance, we use Welch’s ANOVA (as inhomogeneous variances were detected by Levene) and do the post hoc analysis with Tukey tests with Bonferroni-corrected p-values.

The first step is to look at how well our participants were able to recognize the emotions and expressions and which of them were confused easily. Looking at the overall recognition accuracy in Fig. 9 (left), we see that participants were overall able to recognize the emotions similarly between real videos and face swaps. We further visualize the recognized emotions as heatmaps in Fig. 10. As seen from these figures, participants mainly confused clueless and thinking in both conditions. For face swaps, there was also some confusion between disgust and sadness. During our analysis we find that the recognition is significantly different in real and swap videos (Welch’s F(1,3778) = 11, p = 0.000898). A post hoc test only reveals differences for recognising the emotion disgust (p = 0.000028, MeanReal = 0.64, SEM = 0.04, MeanSwap = 0.40, SEM = 0.03). This indicates that overall the recognition of emotions and expressions in face swaps and real videos is similar.

Next, we perform the same analysis for the intensity and sincerity ratings of participants as shown in the middle and right plot of Fig. 9. As a first impression from these plots, real videos are generally rated with higher intensity and sincerity, however, the assessments do not differ too strongly. Assessing the ratings between conditions we do obtain a highly significant results for intensity (Welch’s F(1,3773) = 29.4, p = 0.000000613) and sincerity (Welch’s F(1,3776) = 12.8, p = 0.000352). A post hoc test reveals high significance for intensity ratings on clueless (p = 0.000005; MeanReal = 5.18, SEM = 0.16; MeanSwap = 4.46, SEM = 0.19) and disgust (p = 0.00004; MeanReal = 5.56, SEM = 0.1; MeanSwap = 4.91, SEM = 0.1) and significance for surprise (p = 0.04; MeanReal = 5.62, SEM = 0.11; MeanSwap = 5.32, SEM = 0.15). Differences in sincerity ratings are significant for disagree (p = 0.025; MeanReal = 4.71, SEM = 0.09; MeanSwap = 4.31, SEM = 0.09) and thinking (p = 0.004; MeanReal = 4.59, SEM = 0.1; MeanSwap = 4.18, SEM = 0.1). This indicates that certain emotions and expressions are rated less intense and sincere in face swaps than in real videos.

Figure 10: E3: Confusion matrices for the recognition of expressions in real (left) or swap (right) videos. Rows index the shown expression while columns indicate the voted ones.

5 DISCUSSION OF FINDINGS

In this section, we first answer our initial research questions formulated in the introduction. Afterwards we discuss interesting observations, limitations and new ideas for future research.

5.1 Answers to Our Research Questions

Based on our analysis, we are now able to address our initial research questions.

How is gaze behaviour impacted by faces swaps? Can eye tracking be used to detect facial manipulations? Our results indicate that
viewing behaviour is impacted by face swaps. We find that fixations are more prominent on the nose and mouth than on the eyes for high quality face swaps in trials where participants reported artifacts. These artifacts, however, do not necessarily lead the participants to assess a video as a face swap. As eye tracking is easily measurable and integrable with common displays [28, 40, 72], it could be a more reliable and less intrusive substitute for self reports in the debunking of facial manipulations. As we find significant differences within the distribution of fixations of high quality face swaps, the corresponding real video would not be necessary for the classification of a video as a face swap.

Furthermore, we find that the location of artifacts in face swaps influences their saliency. Especially artifacts on facial contours, which were reported less often than artifacts on eyes and mouth, still influenced the fixation behaviour. Additionally, we find that artifacts are generally more fixated, when they are also reported by the participants indicating that participants may consciously explore artifacts when they notice them.

Does the length of video clips influence participants’ assessment accuracy? We find no significant difference in manipulation recognition even for clips as short as three seconds. This contrasts our initial hypothesis as we assumed that longer videos would give participants more time to explore the face and notice unnatural expressions and artifacts. However, the results indicate that the conscious debunking of the video, if to happen, occurs rather early and thus, the first impression has an impact on the decision of participants.

Are conveyed emotions and the expressions different between face swaps and genuine videos? While the recognition accuracy, as well as intensity and sincerity ratings for emotions and expressions, are similar in both conditions, real videos generally seem to obtain higher ratings. We further found significant differences for disgust, surprise, disagree, and thinking which therefore do not match the target video. Despite this, the assessment of emotions and expressions in face swaps already matches the corresponding real videos surprisingly well. This indicates that face swaps are overall able to convey the intended emotions and expressions making them even more powerful but also potentially more dangerous.

5.2 General Discussion

In the following we discuss interesting observations based on our data and analysis.

Discussion on the Fixation Distribution. We observe a nearly equal amount of fixations on the eye, nose and mouth regions for real videos and low quality face swaps. While eye tracking experiments for static images often found a higher number of fixations on the eyes than on the mouth [6], previous research on facial videos suggest that the mouth is more often fixated if the actor is talking [41, 65]. A high number of fixations on the nose can be attributed to a central bias as participants tend to fixate on a region that allows them to quickly change their attention to other areas [32]. Thus, the general distribution of fixations in our experiment aligns to previous research.

However, we do detect a difference in this distribution for high quality swaps. Here our participants looked less often on the eyes and instead focused more on the mouth and nose. As the actors in the videos speak and therefore constantly move their mouth, the increased focus on this area in high quality face swaps may be further amplified by unnatural movements which may not match those of natural speech. This mismatch may trigger the interest of observers and lead to more fixations.

Quality of State-of-the-Art Face Swaps. Overall, the high quality stimuli of the PEFS dataset [69] were often mistaken for real videos in our experiment focusing on the influence of video length (see Sec. 4.3). They were reported to be real nearly as often as the genuine videos (74% vs. 71%). This contrasts not only the number of artifact reports in our eye tracking experiment (see Sec. 4.1) but also the realness assessment reported in the PEFS paper (Real 80%, Swap 65%) [69]. Therefore, the ratings of our participants are probably influenced by the varying length and diversity of the stimuli. As this reduces the ability to successfully distinguish between real videos and face swaps, it may also mean that in real world scenarios when users are leisurely watching video clips from different sources and with varying length, their ability to detect face swaps may be even further reduced. We believe that, soon, the visual difference between original and manipulated videos will disappear, as classifiers can directly be used to improve face swaps [25]. Thus, it is crucial to also discuss ethical and legal regulations on the handling of face swaps as they may become ubiquitous.

Conspicuousness of Face Swaps. In our experiments, we find several hints towards conspicuous artifacts and features of face swaps. We highlighted some artifacts reported by participants in Fig. 3. To our surprise, the contour artifacts that appear when the shapes of two faces do not match, were only seldom reported for high quality manipulations (3 out of 20 participants). In contrast, those seem to be a good cue for low quality manipulations, where 10 of the 20 participants noticed them. Additionally, visible blur on the faces was often attributed to other types of manipulations like beauty filters. Instead of clear artifacts, participants often discussed the unnatural movements of eyes and mouth as noticeable for high quality face swaps. Some even stated that the eyes of face swaps seemed to be lifeless or suspected the actors to be blind. These findings indicate, that for high quality manipulations artifacts are less notable, however, viewers can use behavioural cues like unnatural expressions to debunk face swaps.

In one of our experiments (E1), we did not inform the participants about the manipulated videos or face swaps beforehand. However, we explained the manipulations to them in a debriefing and asked them whether they knew about the concept of face swaps and whether they had some suspicion towards this technique which they did not mention before. During the debriefing, 11 of our 40 participants stated that they thought some form of manipulation was applied to the videos but were not able to exactly tell what kind of manipulation. Six other participants would not have suspected face swaps, but rather thought videos were post-processed by techniques like beauty filters. All other 23 participants did not suspect this level of manipulation or post processing and rather attributed artifacts to the recording, limited video quality or compression. This is interesting, as it shows how convincing face swaps are and that, even when participants report artifacts, they usually do not assume to be confronted with a strongly manipulated video.
Finally, even though the participants of the experiment were rather young and mostly university students (including nearly 50% computer science students) eight of them never heard about face swaps. This emphasizes not only the need to establish reliable detection methods but also the urgency to educate people about possible manipulations and modern face-swapping techniques.

Observations on the Assessment of Emotions. Our findings from E3 do not indicate a strong mismatch between intended and conveyed message in face swaps. The assessment of participants is, however, not matching real videos for disgust, surprise, disagree, and thinking. Previous research found that these expressions are some of the ones relying the most on a perfect balance of rigid head motion, body language, and the coordinated movement of several facial areas. If either of these factors is slightly off, the recognition rate drops instantly [55]. Current face swaps, therefore, seem to reproduce the overall expressions of body and face correctly, however, the balance between movements is not fully consistent to the original video, as some of the micro-expressions do not transfer perfectly. This can impact the recognition as well as intensity and sincerity impression for more complex expression.

Demographic Considerations. Eye tracking and emotion recognition literature considers age-dependent factors [1, 33, 51]. However, we assume young people are more likely to come into contact and concern themselves with face swaps, making them a challenging demographic. Nevertheless, a first analysis on participants’ background points towards familiarity with technology not being a deciding factor for face swap recognition. Following that line, it could also be interesting to analyze the impact of gender-based differences (e.g., [46]), to further extend our comprehension of gaze patterns on face swaps. We refer the reader to the supplemental material for our preliminary research in those directions (see Supplement Material, Sec 4). Overall, fully analysing the effect of face swaps on different demographic groups may be an interesting line for future work.

Limitations. Our main goal was to investigate differences between face swaps and original videos, therefore, we used stimuli recorded in a controlled environment which shows only one actor at a time in a seated pose and without audio. This way we could make sure that the responses of our participants mainly stem from the face swaps and that there was no distraction by other objects in the scene. However, real world videos will only seldom be this restricted. In order to transfer our findings to real world scenarios, further experiments focusing on more varied scenes could be conducted. Possibilities for this include videos with unrestricted environments, free movements, more than one person, or audio.

Furthermore, the participants throughout all our experiments are from a rather young demographic. Therefore our results may not reflect the impact of face swaps on society as a whole. As the stimuli of the used dataset only contain young actors [69], the results may also be influenced by the age of the actors.

Actionable Insights. Current state of the art on forgery detection relies on specific artifacts inherent to CNN-based generator models [67]. While currently performing well, researchers foresee this changing in the near future, requiring new perspectives [52]. Humans excel in facial processing, thus we believe that insights into the perception of face swaps can be a new perspective to detect forged videos. Following our findings, we could use the gathered eye tracking data to design a face swap classifier, similarly to frameworks for saliency prediction [10, 36, 68]. Such a perceptually driven classifier could outperform the human eye, being able to detect those artifacts on the pixel level. This is necessary as face swaps are already very close to be visually indistinguishable from real videos.

6 CONCLUSION

In this paper, we investigated the perception of face swaps using three experiments. We found that participants are able to recognize artifacts in state-of-the-art face swaps but only seldom attribute them to manipulations when they are not previously informed about them. Furthermore, we investigated eye tracking patterns and found significant differences in fixation behaviour, with participants focusing less on the eyes for high quality swaps. Interestingly, the length of video clips did not influence participants’ assessment accuracy. Finally, conveyed emotions and expressions in face swaps are not yet completely indistinguishable from reality. However, these generated results are already very close to be visually non-detectable.

Our findings yield valuable insights towards better understanding the perceptual differences between genuine and face-swapped videos. We found valuable indications of physiologically measurable signals to debunk face swap videos that could assist and guide future algorithmic detection tools.

As future work, it seems interesting to investigate whether our findings are also applicable to other types of facial manipulation techniques and to assess familiarity effects in face swap videos.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding by the German Science Foundation (DFG MA2555/15-1 “Immersive Digital Reality”), and the L3S Research Center, Hanover, Germany.

REFERENCES


The gaze 

Kjell 

DeepFaceLab. business-51204954 

Computer fun?

Deepfake Towards 

Hoboken, in 

Guo, (1994), 

and 

topics.

Emotion 

based expression. 


Towards Understanding Perceptual Differences between Genuine and Face-Swapped Videos


