Defending Low-Bandwidth Talking Head Videoconferencing Systems From Real-Time Puppeteering Attacks

Danial Samadi Vahdati , Tai Duc Nguyen and Matthew C. Stamm
Drexel University
Philadelphia, PA
{danial.samadivahdati, tdn47, mstamm}@drexel.edu

Abstract

Talking head videos have gained significant attention in recent years due to advances in AI that allow for the synthesis of realistic videos from only a single image of the speaker. Recently, researchers have proposed low bandwidth talking head video systems for use in applications such as videoconferencing and video calls. However, these systems are vulnerable to puppeteering attacks, where an attacker can control a synthetic version of a different target speaker in real-time. This can be potentially used spread misinformation or committing fraud. Because the receiver always creates a synthetic video of the speaker, deepfake detectors cannot protect against these attacks. As a result, there are currently no defenses against puppeteering in these systems. In this paper, we propose a new defense against puppeteering attacks in low-bandwidth talking head video systems by utilizing the biometric information inherent in the facial expression and pose data transmitted to the receiver. Our proposed system requires no modifications to the video transmission system and operates with low computational cost. We present experimental evidence to demonstrate the effectiveness of our proposed defense and provide a new dataset for benchmarking defenses against puppeteering attacks.

1. Introduction

Talking head videos are a type of video where the main focus is on a speaker being filmed from the shoulders up and directly addressing the camera. Advances in AI have allowed for the development of systems that can synthesize realistic talking head videos [7, 44, 48]. In recent years, researchers have made significant progress in creating "one shot" talking head synthesis networks [20, 42, 43, 46, 47]. This new technology enables the creation of realistic talking head videos of a speaker using only a single image of that speaker. As a result, synthetic talking head videos can now be easily created for positive uses ranging from virtual asisstants and movie production, to potentially malicious uses such as deepfakes

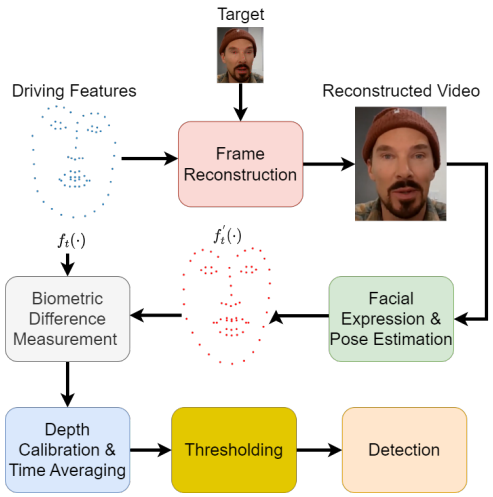

Figure 1. High-level overview of our proposed defensive system.

Recently, researchers have proposed low bandwidth talking head video systems for use in applications such as videoconferencing and video calls [1, 14]. These systems are based on one-shot talking head synthesis networks. In these systems, a single frame or latent space representation of a speaker is sent from the sender to the receiver. After this initial transmission, the sender only transmits information related to facial expression and pose features to the receiver. Using this information and the initial face representation, the receiver synthesizes each frame of the video in real-time. As a result, the speaker on the sender side can "drive" the actions of the synthetic version of themselves on the receiver side in real time. This method significantly reduces the bandwidth needed for videoconferencing, as only the vectors of facial pose and expression information are transmitted rather than the entire frame itself. This advancement has the potential to greatly improve the quality and

accessibility of videoconferencing, especially in low bandwidth or remote areas.

Talking head videoconferencing systems are unfortunately susceptible to real-time puppeteering attacks, in which the synthetic video generated at the receiver side does not match the person driving the video. To carry out such attacks, the attacker first sends an image of the target speaker to the receiver during the initialization phase of the video. The system then receives the attacker's facial expression and pose information, which is used to create a synthetic video of the target speaker. This allows the attacker to control a realistic version of the target speaker in real-time, potentially deceiving the viewer on the receiver side.

The ability to puppeteer a target speaker in real-time using talking head videoconferencing systems poses significant risks. This capability can be used to spread misinformation and disinformation, but it can also enable other criminal activities such as fraud and defamation. Already, real-time audio deepfakes have been reportedly used to commit financial crimes [11]. This trend is expected to become more common if videoconferencing systems are unable to protect against puppeteered videos. These videos are likely to be even more convincing than audio-only deepfakes, making them a potent tool for malicious actors. It is crucial to develop effective security measures to prevent puppeteering attacks and ensure the authenticity of input signals in talking head videoconferencing systems.

Currently, there are no defenses against against puppeteering attacks in low-bandwidth talking head videoconferencing systems. Initially, this problem may seem identical to detecting deepfake videos. However, even when these videoconferencing systems are operating as intended, they create a synthetic version of the speaker at the receiver, i.e. the system deepfakes an authentic speaker in order to save bandwidth. As a result, deepfake detectors are ill-suited to protecting against puppeteering attacks.

In this paper, we propose a new system to defend against puppeteering attacks in low bandwidth talking head videoconferencing systems. Our defensive system exploits the fact that the facial expression and pose information sent to the receiver inherently contains biometric information about the driving speaker. We leverage this information to obtain measurements of the biometric distance between the driving and reconstructed speaker. If the biometric distance becomes large, this indicates that the driving speaker is a different person than the reconstructed speaker. Our system then flags the video transmission as a puppeteering attack.

Our proposed system has several desireable properties: It requies no modifications to the video encoding and transmission system, nor does it require the additional transmission of side information to detect puppeteering. Instead, it only utilizes information already available at the receiver. Biometric features describing the driving and reconstructed speaker are obtained using compenents already present in the system. Furthermore, our system operates with a very low computational cost, making it well suited to real-time puppeteering detection.

The main contributions of this work are as follows:

- We present the problem of puppeteering attacks in low-bandwidth talking head videoconferencing systems.
- We demonstrate that facial expression and pose information transmitted by the sender inherently contains biometric information about the driving speaker. We show that this information can be used to identify discrepencies between the driving and reconstructed speaker.
- We propose a new defense against puppeteering in low-bandwidth talking head videoconferencing systems. To the best of our knowledge, this is the first defense against puppeteering attacks in these systems. Our defense is requires no modifications to the video encoding and transmission system, and can operate in real time.
- We present a series of experiments to verify the performance of our proposed defense. To do this, we develop a new dataset that can be used for benchmarking defenses against puppeteering attacks in low bandwidth talking head videoconferencing systems. We present experimental evidence that our defense does not exhibit bias in terms of race/ethnicity and sex .

2. Background Work

Talking Head Video Systems. Talking head video systems are a type of artificial intelligence-driven technology used to generate highly realistic and dynamic facial animations or video sequences. These systems create virtual characters or "talking heads" that can mimic human-like speech, facial expressions, and emotions. The primary goal of talking head video systems is to provide more engaging and interactive experiences in various applications such as virtual assistants, video games, film, and telecommunication. Non-AI-based talking head video systems typically rely on traditional computer graphics and animation techniques, such as: keyframe animation [26, 36, 41], blendshapes [12, 25, 27], morph target [8, 15, 39], facial motion capture [5, 33, 37], to generate facial animations and movements. These approaches often require more manual intervention, time, expensive equipments and large amount of expertise compared to AI-based systems. Recent developments in AI-based "talking head" systems involve extracting facial features combined with facial expression or emotion features from both source and target videos. These systems then learn a transfer function that adapts the source's features to fit the target's features, resulting in a more natural and accurate representation. Notable work using this paradigm includes Face2Face [42], DaGAN [20], Reenact-GAN [47], SAFA [43], and X2Face [46].

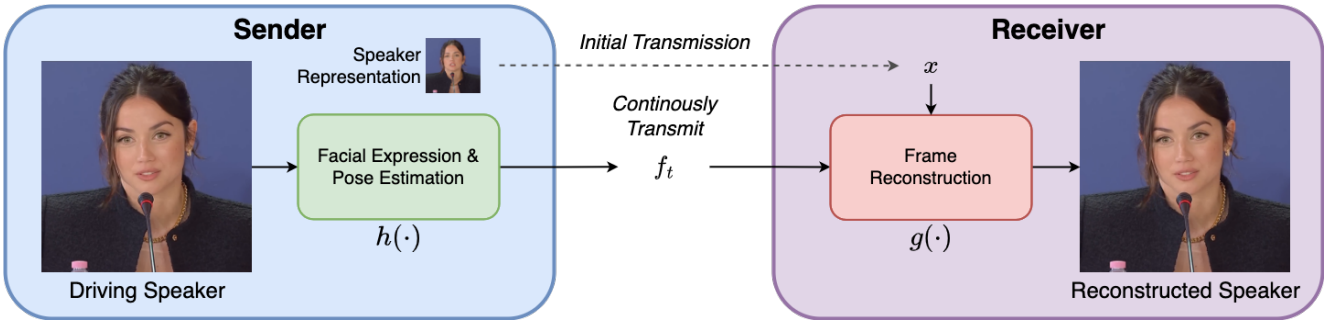

Figure 2. Overview of a low-bandwidth talking head videoconferencing system.

A possible application for talking head video Systems lies in enhancing low-bandwidth video transmission systems. In these systems, low utilization of bandwidth can be achieve in two ways: 1) By transmitting highly compressed facial embeddings from the sender, which the receiver then uses to reconstruct the face in the video stream [14, 32]; or 2) The sender initially sends low-level representations of the face and background, followed by facial landmarks for subsequent frames, allowing the receiver to reconstruct the corresponding face and background using this information. [1]

Deepfakes And Synthetic Image Detectors. In order to combat these concerns, researchers have developed many techniques to detect synthetic media. Some of these work has focused on detecting deepfakes. Deepfake detectors work by leveraging priors about the human face's anatomy structure to identify subtle inconsistencies or artifacts in the generated video. State-of-the-art approaches [2, 6, 10, 21, 23, 29, 45, 49] use deep learning to do this and they have achieved very strong results in multiple public datasets [13, 34, 50]. Other research has been done to detect synthetic images, as well as identify video editing and origin. These systems work by looking for either specific forensic traces left by the image generation process, or anomalies in the locally edited media. Notable approaches includes [3, 4, 9, 16–19, 22, 24, 28, 30, 31, 35, 38, 40]. However, since these approaches will likely false alarm authentic, self-reenacted videos as being deepfaked, or synthesized, they are not effective in identifying misuse of this technology.

3. Problem Formulation

In this section, we describe the problem of puppeteering in low bandwidth talking head video systems. We begin by describing how these systems operate, including system components relevant to this paper. Next, we provide details of how puppeteering attacks are launched.

3.1. Low-Bandwidth Talking Head Videoconferencing Systems

Low-bandwidth talking head videoconferencing systems are designed reduce the amount of information that must be transmitted to a receiver in video conferencing and similar applications. They do this by encoding a talking head video of a speaker at the sender side, then transmitting the encoded information to a receiver. The receiver decodes the video by using the transmitted information as input to a generator, which creates a synthetic video of the speaker.

These systems operate by first, sending a representation x of the speaker's face to the receiver. This representation is learned from the initial portion of the video, often the first video frame. Typically, x is a single video frame of the speaker containing the speaker's neutral face or a representation of the speaker's face in a latent space learned by a generative adversarial network (GAN).

At the sender's side, facial expression and pose information f_t at time t is extracted from the current frame I_t using a system using a system $h(\cdot)$ such that

$$f_t = h(I_t). \tag{1}$$

The resulting expression and pose feature vector f_t is then transmitted to the receiver. Additional information, such as features that capture any motion present in the background may also be captured and transmitted. These additional features are not relevant to this work, and for simplicity we will omit them further discussion without loss of generality.

The receiver decodes each video frame by using a system $g(\cdot)$ which takes as input the current expression and pose information from the sender, along with the representation of the speakers face sent at the beginning of the transmission. This produces a reconstructed frame I'_t containing a synthesized version of the speaker's face with the desired pose and expression such that

$$I'_t = g(f_t, x). \tag{2}$$

In many systems, g corresponds to a generator pre-trained as part of a GAN to synthesize a realistic human face.

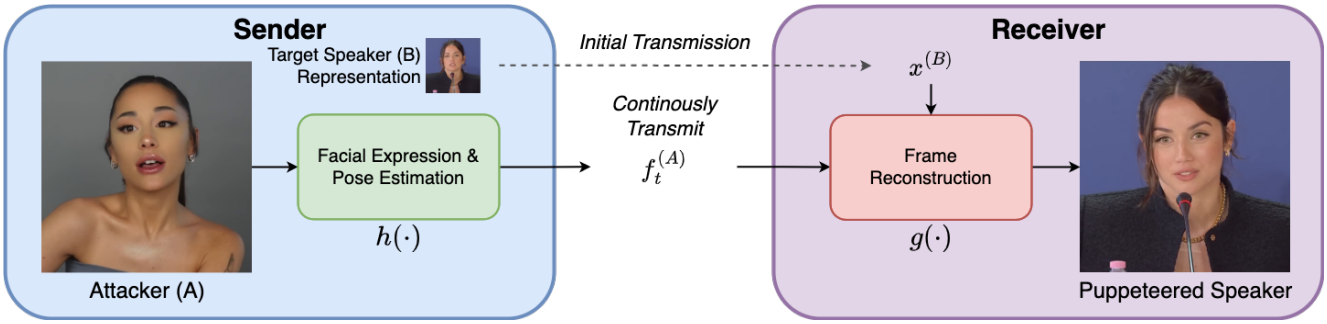

Figure 3. Overview of a puppeteering attack in a low-bandwidth talking head videoconferencing system.

An overview of the complete encoding, transmission, and decoding process at time t can be seen in Fig. 2 .

3.2. Puppeteering Attacks

The low-bandwidth talking head videoconferencing systems described above are vulnerable to puppeteering attacks, in which the reeconstructed speaker at the receiver side is actually controlled in real-time by a different person at the sender side. An overview of a puppeteering attack can be seen in Fig. 3.

In a puppeteering attack, an attacker (Speaker A) on the sender side first obtains a representation $x^{(B)}$ of a target speaker's face (Speaker B). When the video transmission is initiated, the attacker sends $x^{(B)}$ to the receiver instead of a representation of their own face. After this, they allow the video system at the sender side to observe their face, and produce a facial expression and pose vector $f_t^{(A)}$ which they send to the receiver. The receiver uses $f_t^{(A)}$ along with $x^{(B)}$ to construct a video frame $\hat{I}_t = g(f_t^{(A)}, x^{(B)})$ with the face of Speaker B, but with the facial expression and pose of Speaker A. As a result, the viewer at the receiver side sees a video of Speaker B that is actually controlled by the actions of the attacker.

4. Proposed Approach

4.1. Exploiting Biometric Side-Information

In a puppeteered video, the biometric identity of the driving speaker is different from that of the reconstructed speaker. Our proposed system leverages this fact to detect puppeteered videos. While the identity of the driving speaker is not directly observable to the receiver, the receiver does have access the series of facial expression and pose vectors f_t sent by the driving speaker. These vectors inherently capture biometric information about the driving speaker. By analyzing the reconstructed video and comparing it to the corresponding f_t's, our system is able to identify biometric differences between the driving and reconstructed speaker present in puppeteering attacks.

To gain further intuition how this is possible, let us first

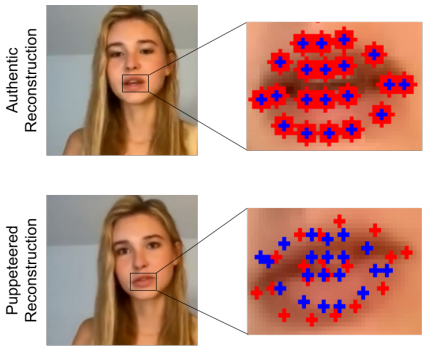

Figure 4. Example showing the effect of puppeteering on facial landmark postions

examine talking head video systems such as X2Face [46], in which f_t directly corresponds to facial landmark positions of the driving speaker as part of the driving features. In systems such as this, if the driving speaker is the same as the reconstructed speaker, then facial landmark positions extracted from the reconstructed speaker should closely match the facial landmark positions of sent by the driving speaker. This can be seen in the top row of Fig. 4, which shows the difference between the landmark positions extracted from a video frame synthesized by X2Face in red and from the driving speaker in blue. Here, the facial landmarks from the driving and reconstructed speaker closely align. In general, there may be small differences between the landmark positions from the driving and reconstructed speaker due to reconstruction error.

If the driving speaker is different than the reconstructed speaker, then they will not share the same facial geometry. This will cause facial landmark positions extracted from the reconstructed video to differ significantly from those sent in f_t. This can be seen in the bottom row of Fig. 4, which shows the difference between the landmark positions extracted from a puppeteered video frame synthesized by X2Face and those from the driving speaker.

We note that some systems, such as SAFA, do not transmit explicit facial landmark locations as f_t. Instead,

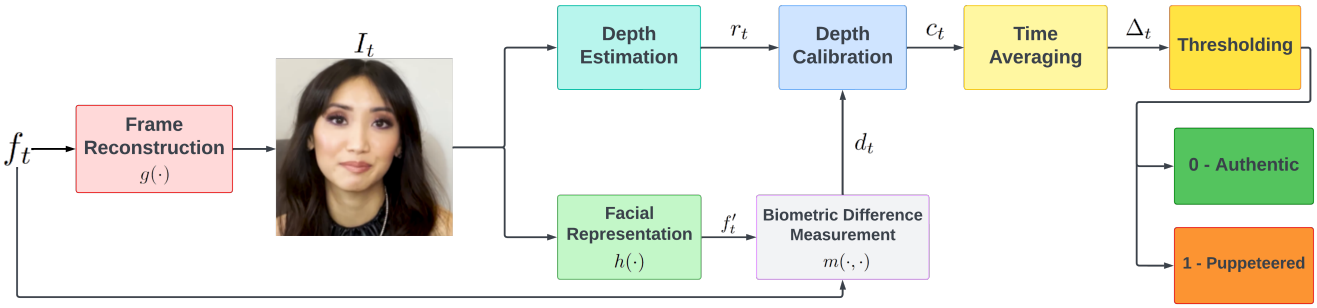

Figure 5. An overview of our propsed defensive system.

these systems encode facial expression and pose information through other means, such as a learned embedding. These embeddings, however, implicitly capture facial landmark position and other biometric information about the speaker. As a result, we are able still able to use these features to expose biometric differences between the driving and reconstructed speaker.

4.2. Detecting Puppeteering

Our proposed system detects puppeteering attacks by exploiting the biometric information that f_t captures about the driving speaker as described above. A diagram providing an overview of our system can be seen in Fig. 5.

Baseline Biometric Distance Measurement: First, our system obtains a baseline measurement of the biometric distance the reconstructed speaker and the driving speaker. To do this, we estimate the facial expression and pose features f'_t from the reconstructed video frame $\hat{I}$ such that

$$f'_t = h(\hat{I}_t). \tag{3}$$

We note that h is already available to the receiver because it is required to encode and transmit their face back to the sender. Next, our system captures the difference between the driving and reconstructed speaker's biometric information as

$$d_t = m(f_t, f'_t) \tag{4}$$

where $m(\cdot,\cdot)$ is an appropriate metric that measures the difference between f_t and f'_t. In practice, we have found that using $m(f_t, f'_t) = (\sum_k |f_t - f_p|^2)^{1/2}$ is sufficient to achieve strong system performance.

Controlling For Depth Variation: When a speaker moves farther from the camera, their face becomes smaller. As a result, the differences between f_t and f'_t caused by puppeteering also become relatively smaller. The opposite of this is true when the speaker moves closer to the camera. Our system must account for this when differentiating between values of d_t cased by puppeteering and those that naturally occur due to imperfect reconstruction of an authentic speaker.

To do this, our system makes an initial reference estimate r_0 of the speaker's distance from the camera in the first video frame. At each subsequent frame, we estimate the speaker's depth r_t and calcualte a depth-calibrated biometric distance c_t between the driving and reconstructed speaker according to

$$c_t = d_t\left(\frac{r_t}{r_0}\right). \tag{5}$$

Controlling For Natural Reconstruction Errors: As previously noted, a low-bandwidth talking head video system will not perfectly reconstruct an authentic driving speaker at the receiver. As a result, there will be natural variation between f_t and f'_t. This variation will be larger at some times due to temporally isolated conditions that make it difficult for the video system to accurately synthesize the driving speaker. This could be due to sudden motion, irregular facial expressions or poses, or a number of other factors. If only the instantaneous biometric difference c_t is used to detect puppeteering, then our system will false alarm when this occurs.

To control for these effects, our system calculates a time averaged value of the biometric distance Δ_t between the driving and the reconstructed speaker as

$$\Delta_t = \tfrac{1}{W}\sum_{\ell=0}^{W-1} c_{t-\ell}, \tag{6}$$

where W is the width of a sliding window over which c_t values are averaged.

Puppeteering Detection: Finally, our system uses the time averaged biometric distance Δ_t to detect puppeteering by comparing it to a detection threshold τ. Because puppeteering induces large biometric distances, values of Δ_t greater than τ indicate that the video is puppeteered.

5. Experiments

Below, we present the details and results of a series of experiments conducted to evaluate the performance of our proposed defensive system.

Figure 6. Example of authentic self reenacted videos as well as puppeteered videos in our experimental dataset.

	Proposed	CNN Ensemble	Efficient ViT	Cross-Efficient ViT
DaGAN	99.31%	66.80%	76.26%	69.81%
Reenact GAN	94.83%	69.73%	76.96%	68.58%
X2Face	99.80%	68.24%	79.00%	78.15%
SAFA	98.92%	67.35%	74.86%	67.81%
Average	98.03%	68.03%	76.77%	71.09%

Table 1. Puppeteering detection accuracies achieved by our proposed defensive system as well as several leading deepfake detectors.

5.1. Dataset

To conduct our experiments, we created a dataset of talking head videos reconstructed by the receiver, along with the facial expression and pose vectors used to reconstruct them. To do this, we first collected a set of pristine videos of multiple speakers, which we used to drive a talking head video system. We gathered these pristine videos by excerpting segments from celebrity interviews publicly distributed on Youtube. Each pristine video corresponds to a 20 to 30 second clip of a single, front facing speaker. Three pristine videos with different backgrounds and settings were collected from each of 24 different celebrities, resulting in a total of 72 pristine videos. To ensure diversity in our dataset and to help identify any biases that may be inherent in our

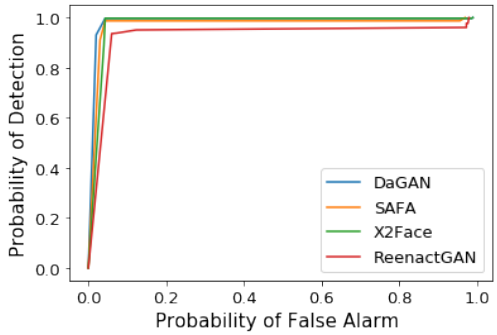

Figure 7. ROC curves showing the performance of our defensive system for different talking head video systems.

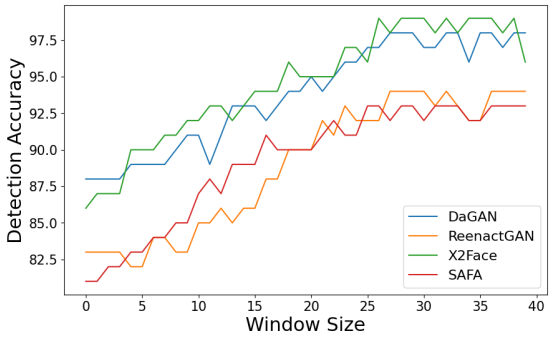

Figure 8. Plot showing the puppeteering detection accuracy vs. temporal averaging window size W.

system, celebrity speakers were chosen to be equally split across sex (i.e. 12 male and 12 female speakers) as well as across four racial/ethnic groups: Black, White, Hispanic, and Asian (i.e. 6 speakers from each group).

The set of pristine videos was then used to create both authentic and puppeteered talking head videos, as would be reconstructed by the receiver in a low-bandwidth talking head video system. Reconstructed talking head videos were created using four different networks: DA-GAN [20], SAFA [43], X2Face [46], and ReenactGAN [47]. The set of facial expression and pose features used to create each video were also retained.

Using each of the four networks, we created a set of both authentic self-driven videos as well as a set of puppeteered videos. Authentic videos were created by using each of the 72 pristine videos to drive a self-driven reconstruction. Puppeteered videos of each speaker were created by using a pristine video from a different speaker to drive the system. For each speaker, a set of 18 puppeteered videos were made using two different driving speakers. To produced higher quality reconstructions, the driving speakers for each puppeteered video were selected to match the race/ethnicity and sex of the target speaker. This process was repeated for a each of the 24 speakers, resulting in a set of 432 pupeteered videos per network.

In total, our dataset consists of 2016 talking head videos corresponding to approximately 14 hours of total video footage. Examples of this dataset can be seen in Fig. 6. This dataset can be downloaded at
https://gitlab.com/MISLgit/talking-head-puppeteering-defense/

5.2. System Performance

To assess our system's overall accuracy, we used it to identify puppeteering in each of the videos in our dataset. When conducting these experiments, our system used a window size of $W = 30$ frames, corresponding to 1 second intervals of each video. Puppeteering detection decisions were assessed at a window level. To compare our system's performance to existing approaches, we also analyzed each video using three leading deepfake detectors: Efficient ViT [10], Cross-Efficient ViT [10], and CNN Ensemble [6].

Puppeteering detection accuracies obtained by our system are shown in Table 1, as well as accuracies obtained by the three deepfake detection networks used for comparison. We can see that our system acheives strong puppeteering detection performance across all four talking head video systems, with an average detection accuracy of 98.03%. Additionally, Fig. 7 shows ROC curves capturing the performance of our defensive system on all four talking head systems. These ROC curves demonstrate that we can achieve strong puppeteering detection performance at low false alarm rates. We note that we are still able to achieve strong performance for SAFA even though facial expression and pose vector f_t used by SAFA do not correspond to explicit facial landmark positions. Instead, these correspond to learned abstract landmark representations. Despite this, we are still able to use SAFA's f_t's to measure the biometric distance between the driving and reconstructed speaker.

Comparison With Deepfake Detectors: The results in Table 1 clearly show that our proposed system significantly outperforms deepfake detectors. Our system achieves approximately a 20 percentage point increase in accuracy over the highest performing deepfake detector (Efficient ViT). This is not a surprising result, as deepfake detectors are intentionally built to for a different application . We provide more interpretation of this result in Section 6.

5.3. Effect of Window Size

We conducted additional experiments to examine the effect of the window size W in (6) on our system's overall accuracy. To do this, we repeated the experiments described in Section 5.2, only we let W vary from 1 to 40, i.e. a single frame to ~ 1.33 seconds. The results of this experiment were used to create the plots in Fig. 8, which show our system's puppeteering detection accuracy vs window size.

	White male	White Female	Asian Male	Asian Female	Black Male	Black Female	Hispanic Male	Hispanic Female
DaGAN	98.37%	99.60%	97.04%	98.13%	98.26%	99.15%	99.71%	99.42%
Reenact GAN	94.10%	93.58%	95.27%	95.74%	93.54%	96.08%	94.37%	96.84%
X2Face	99.37%	98.26%	99.46%	97.35%	98.14%	99.31%	98.46%	99.02%
SAFA	99.74%	99.91%	97.10%	98.75%	98.48%	99.23%	98.20%	98.61%
Average	97.99%	97.84%	97.22%	97.49%	97.19%	98.44%	97.44%	98.47%

Table 2. Our system's puppeteering detection accuracies conditioned on the race/ethnicity and sex of the reconstructed speaker.

From this plot we can see that our system's accuracy increases with W for all talking head video systems until W lies between 25 and 30 frames. After this point, the accuracy holds roughly constant as W is further increased.

6. Discussion

6.1. Why Deepfake Detectors Perform Poorly

Deepfake detectors are intentionally built to detect deepfake videos where a speaker's face has been generated to match a target speaker. This is a similar, yet distinct problem from detecting puppeteering in low-bandwidth talking head systems. While it is clear that a deepfake detector should produce a detection when analyzing a puppeteered video, it is not as clear what these detectors should output when presented with a authentic talking head video.

In our experiments, deepfake detectors' most frequent source of errors corresponded to them them flagging authentic self-driven videos as 'fake.' For example, this accounts for the vast majority of puppeteering detection errors produced CNN Ensemble. This is reasonable, since the face of the speaker in an authentic talking head video has still been synthesized using essentially the same means used to produce a deepfake. We note, however, that Efficient ViT is able to achieve puppeteering detection perfomances as high as 79.00%. This is only possible because Efficient ViT identifies a large portion of self-driven videos as 'real.'

6.2. Influence of Race/Ethnicity and Sex

To examine our system for implicit biases, we investigated the influence of race/ethnicity and sex on our system's performance. Table 2 shows our system's accuracy conditioned on the reconstructed speaker's race/ethnicity and sex. From this table, we can see that the average accuracies hold farirly consistent all groups. The standard deviation of our system's average accuracy for each group was 0.53 percentage points, with all group's average accuracies lying within two standard deviations from the mean. This indicates that our system is unlikely to produce incorrect decisions more frequently for a speakers of a particular race/ethnicity or sex.

We note that biases inherent in a low-bandwidth talking head videconferencing system are likely to propogate to our defensive system. This is because our system uses the f_t's produced by the video system to measure the biometric difference between the driving and reconstructed speaker. If the video system produces worse facial expression and pose representations for one sex or racial/ethnic group, then our system will likely perform worse for the same group.

6.3. System Limitations

Our system's performance depends on biometric information about the driving speaker contained in the transmitted facial pose and expression features. If a talking head system is developed that is able to completely disaggregate facial expression and pose information from a speaker's facial geometry, our system would not be able to defend this system. Additionally , our system only works if it has access to f_t's. Because of this, it is unable to identify puppeteered videos that have been fabricated offline then distributed over the internet such as deepfakes.

7. Conclusion

In this paper, we proposed a new system to defend against puppeteering attacks in low bandwidth talking head video systems. Our defensive system exploits the fact that the facial expression and pose information sent to the receiver inherently contains biometric information about the driving speaker, which can be used to identify discrepancies between the driving and reconstructed speaker. Our proposed system requires no modifications to the video encoding and transmission system and can operate in real-time with low computational cost. We presented a series of experiments to verify the performance of our proposed defense and developed a new dataset for benchmarking defenses against puppeteering attacks in low bandwidth talking head videoconferencing systems.

Acknowledgement

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number HR0011-20-C-0126.

References

[1] NVIDIA maxine. https://developer.nvidia.com/maxine, Oct. 2020. Accessed: 2023-3-18. 1, 3

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, page 38, 2019. 3

[3] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, pages 5–10, 2016. 3

[4] Belhassen Bayar and Matthew C Stamm. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Trans. Inform. Forensics and Security*, 13(11):2691–2706, 2018. 3

[5] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. Multi-scale capture of facial geometry and motion. *ACM transactions on graphics (TOG)*, 26(3):33–es, 2007. 2

[6] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)*, pages 5012–5019. IEEE, 2021. 3, 7

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, pages 35–51. Springer, 2020. 1

[8] Erika Chuang and Christoph Bregler. Mood swings: Expressive speech animation. *ACM Trans. Graph.*, 24(2):331–347, apr 2005. 2

[9] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE TPAMI*, PP:1–1, 2020. 3

[10] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*, pages 219–229. Springer, 2022. 3, 7

[11] Jesse Damiani. A voice deepfake was used to scam a ceo out of $243,000. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/, Sep. 3 2019. 2

[12] Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann. Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games*, I3D '06, page 43–48, New York, NY, USA, 2006. Association for Computing Machinery. 2

[13] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 3

[14] Dahu Feng, Yan Huang, Yiwei Zhang, Jun Ling, Anni Tang, and Li Song. A generative compression framework for low bandwidth video conference. In *2021 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2021. 1, 3

[15] Nico Galoppo, Miguel A. Otaduy, William Moss, Jason Sewall, Sean Curtis, and Ming C. Lin. Controlling deformable material with dynamic morph targets. In *Proceedings of the 2009 Symposium on Interactive 3D Graphics and Games*, I3D '09, page 39–47, New York, NY, USA, 2009. Association for Computing Machinery. 2

[16] Alessandra Gironi, Marco Fontani, Tiziano Bianchi, Alessandro Piva, and Mauro Barni. A video forensic technique for detecting frame deletion and insertion. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6226–6230. IEEE, 2014. 3

[17] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021. 3

[18] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018. 3

[19] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 3

[20] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. 2022. 1, 2, 7

[21] Brian Hosler, Davide Salvi, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1022, 2021. 3

[22] Brian C Hosler and Matthew C Stamm. Detecting video speed manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 670–671, 2020. 3

[23] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *ICASSP*. IEEE, 2021. 3

[24] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 3

[25] Pushkar Joshi, Wen C. Tien, Mathieu Desbrun, and Frederic Pighin. Learning controls for blend shape based realistic facial animation. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06, page 17–es, New York, NY, USA, 2006. Association for Computing Machinery. 2

[26] Midori Kitagawa and Brian Windsor. *MoCap for artists: workflow and techniques for motion capture*. CRC Press, 2020. 2

[27] John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)*, 1(8):2, 2014. 2

[28] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 384–389. IEEE, 2018. 3

[29] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. IEEE, 2019. 3

[30] Owen Mayer and Matthew C Stamm. Exposing fake images with forensic similarity graphs. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1049–1064, 2020. 3

[31] Tai D Nguyen, Shengbang Fang, and Matthew C Stamm. Videofact: Detecting video forgeries using attention, scene context, and forensic traces. *arXiv preprint arXiv:2211.15775*, 2022. 3

[32] Maxime Oquab, Pierre Stock, Daniel Haziza, Tao Xu, Peizhao Zhang, Onur Celebi, Yana Hasson, Patrick Labatut, Bobo Bose-Kolanu, Thibault Peyronel, and Camille Couprie. Low bandwidth video-chat compression using deep generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2388–2397, June 2021. 3

[33] Katherine Pullen and Christoph Bregler. Motion capture assisted animation: Texturing and synthesis. SIGGRAPH '02, page 501–508, New York, NY, USA, 2002. Association for Computing Machinery. 2

[34] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*. IEEE, 2019. 3

[35] Danial Samadi Vahdati and Matthew C Stamm. Detecting gan-generated synthetic images using semantic inconsistencies. *Electronic Imaging*, 2023. 3

[36] Yeongho Seol, Wan-Chun Ma, and J. P. Lewis. Creating an actor-specific facial rig from performance capture. In *Proceedings of the 2016 Symposium on Digital Production*, DigiPro '16, page 13–17, New York, NY, USA, 2016. Association for Computing Machinery. 2

[37] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM SIGGRAPH 2005 Papers*, pages 417–425. 2005. 2

[38] Milani Simone, Marco Fontani, Bestagini Paolo, Barni Mauro, Alessandro Piva, Tagliasacchi Marco, Tubaro Stefano, et al. An overview on video forensics. *APSIPA Transactions on Signal and Information Processing*, 1:0–0, 2012. 3

[39] Jesse Spencer-Smith, Heather Wild, Åse H Innes-Ker, James Townsend, Christy Duffy, Chad Edwards, Kristina Ervin, Nicole Merritt, and Jae Won Pair. Making faces: Creating three-dimensional parameterized models of facial expression. *Behavior Research Methods, Instruments, Computers*, 33:115–123, 2001. 2

[40] Matthew C Stamm, W Sabrina Lin, and KJ Ray Liu. Temporal forensics and anti-forensics for motion compensated video. *IEEE Transactions on Information Forensics and Security*, 7(4):1315–1329, 2012. 3

[41] L.M. Tanco and A. Hilton. Realistic synthesis of novel human movements from a database of motion capture examples. In *Proceedings Workshop on Human Motion*, pages 137–142, 2000. 2

[42] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2Face: Real-time face capture and reenactment of RGB videos. In *CVPR*. IEEE, 2016. 1, 2

[43] Qiulin Wang, Lu Zhang, and Bo Li. SAFA: Structure aware face animation. 2021. 1, 2, 7

[44] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1

[45] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu. GAN-generated faces detection: A survey and new perspectives. 2022. 3

[46] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2Face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, pages 690–706. Springer International Publishing, Cham, 2018. 1, 2, 4, 7

[47] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. ReenactGAN: Learning to reenact faces via boundary transfer. In *ECCV*, pages 622–638. Springer International Publishing, Cham, 2018. 1, 2, 7

[48] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 1

[49] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 3

[50] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390, 2020. 3