Protecting Real-time Video Chat against Fake Facial Videos Generated by Face Reenactment

Jiacheng Shang
Dept. of Computer Science, Montclair State University

Jie Wu
Dept. of Computer and Information Science, Temple University

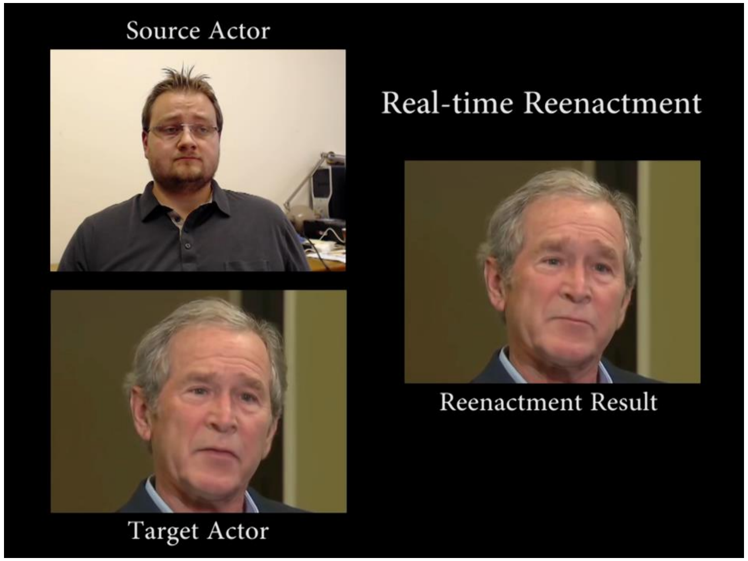
MONTCLAIR STATE UNIVERSITY

Power of Video

- Deliver much more information
- Various applications
    - E.g. Video calling and video conference

Threats of DeepFakes

- Videos are usually assumed to be true
- High-quality fake facial videos using deep learning (even in real time)

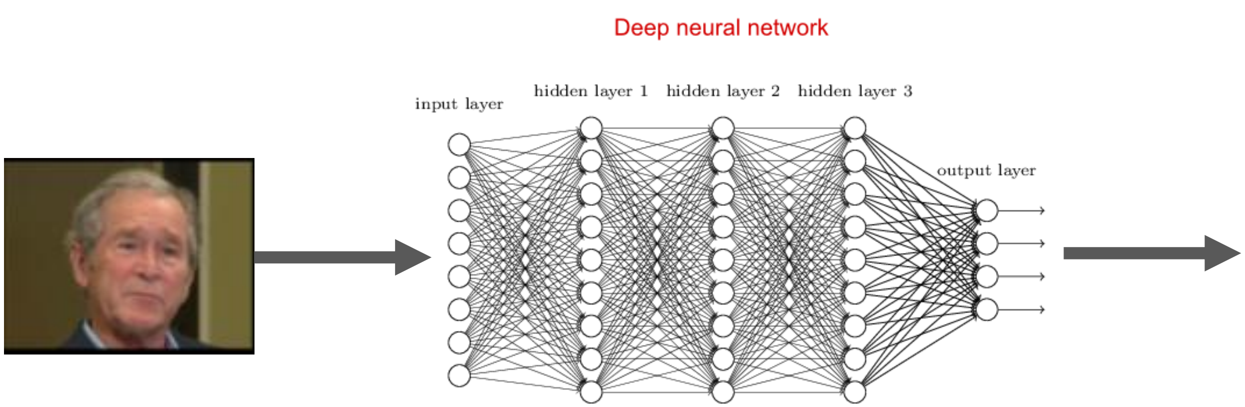


Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral)

MONTCLAIR STATE UNIVERSITY

Fake Facial Video Detection

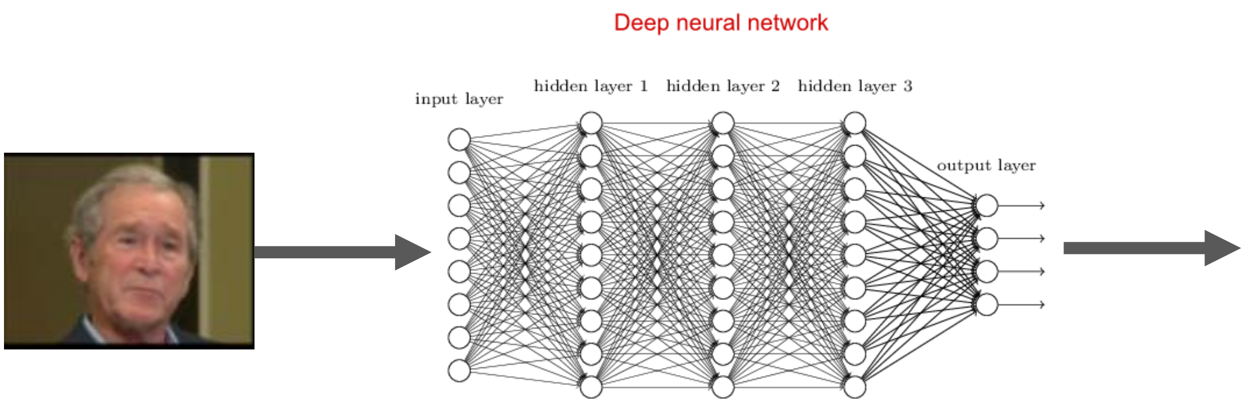
- Many fake facial video detection systems have been proposed based on deep learning



Real or Fake

Fake Facial Video Detection

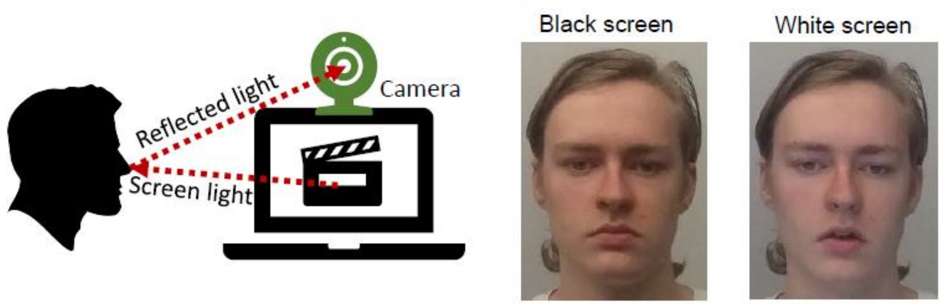
- However, they fail to answer two questions
  - **Generality**: Can their detection systems be generally used to detect all types of fake facial videos?
  - **Cost**: Is there any low-cost detection scheme?



Real or Fake

System Overview

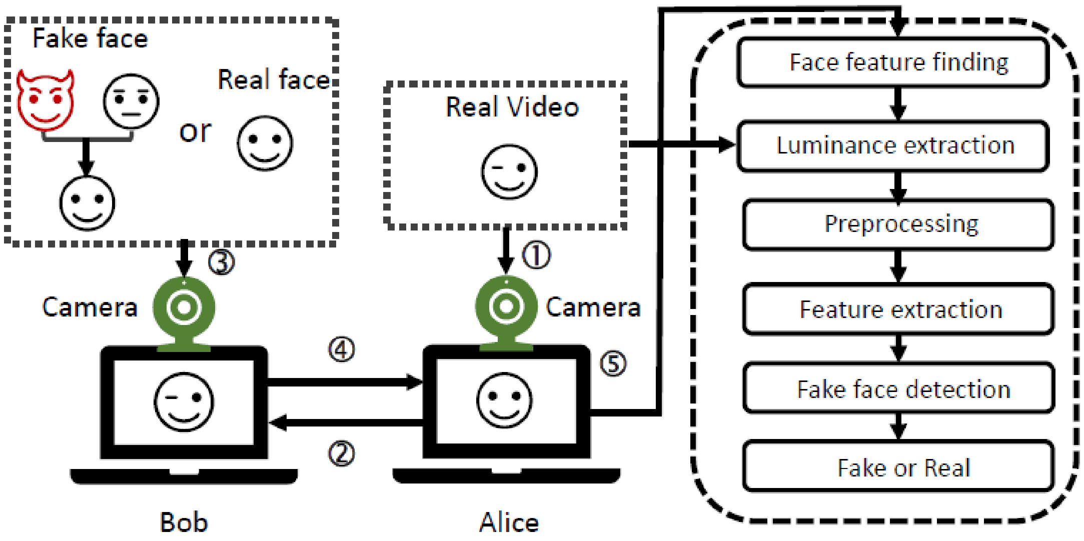
- Utilizing the face reflected light
  - The screen light can be reflected by the face
  - The reflected light can be captured by the webcam
  - The normal user can change the luminance of the screen light by changing the area of light metering



MONTCLAIR STATE UNIVERSITY

System Overview

- Goal: detect the liveness of the face in the video by measuring the correlation between luminance signals of the screen light and face-reflected light



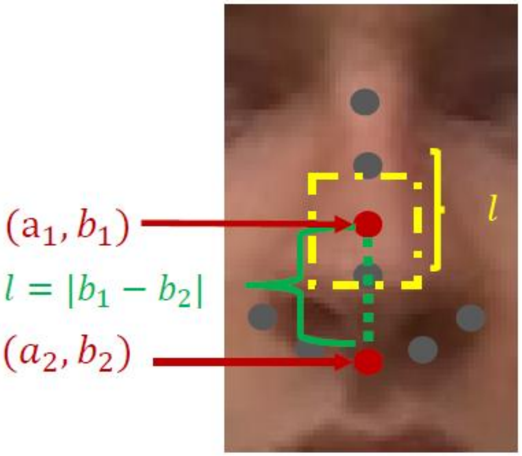
Luminance Extraction

- Extract relative luminance information of the screen light
  - Compress each frame of the screen into a single pixel
  - Use the luminance value of the compressed pixel to represent the overall luminance of the transmitted video
  - The luminance of a pixel is defined as

$$C = 0.2126R + 0.7152G + 0.722B,$$

Luminance Extraction

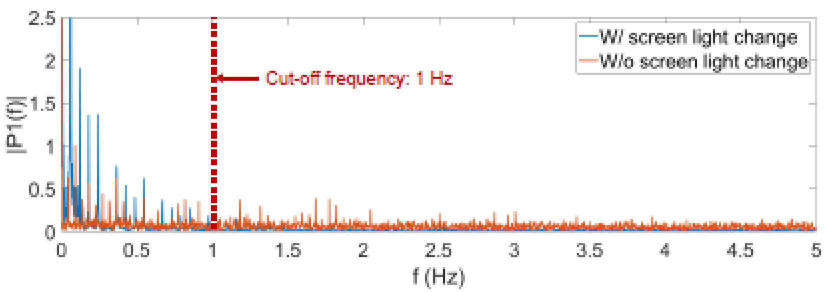
- Not all facial parts can be used to measure luminance changes.
- We find that the lower part of the nasal bridge has the most stable images and hard to be occluded in most cases

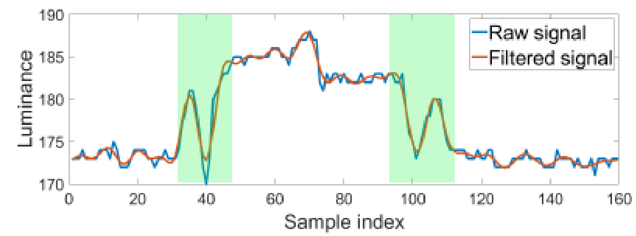


MONTCLAIR STATE UNIVERSITY

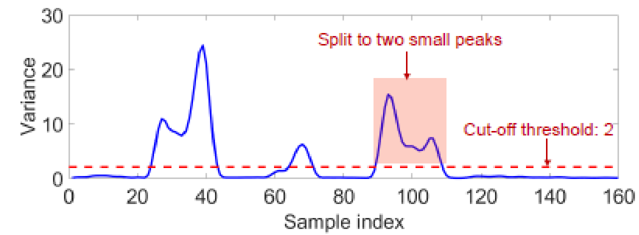
Preprocessing

- Raw luminance signal contains noise
  - Object movement in the scene
  - Inaccurate face localization can lead to jittering in the interested area,

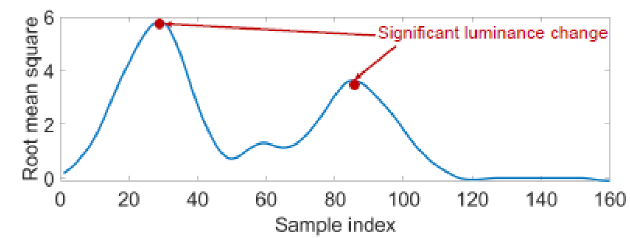




(a) The raw and filtered luminance signal



(b) Variance signal



(c) Smoothed variance signal

MONTCLAIR STATE UNIVERSITY

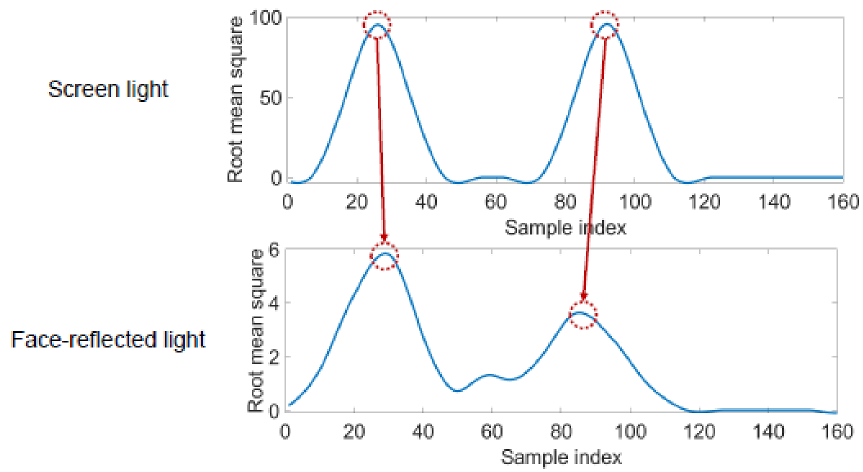
Feature extraction

- Luminance change behavior
  - For any significant luminance change in one signal, we can always find a matched luminance change in another one.
  - We define two behavior similarity metrics z_1 and z_2

$$z_1 = \frac{1}{N} \times F(T, R),$$

Num. of luminance changes in the screen light

Num. of matched luminance changes

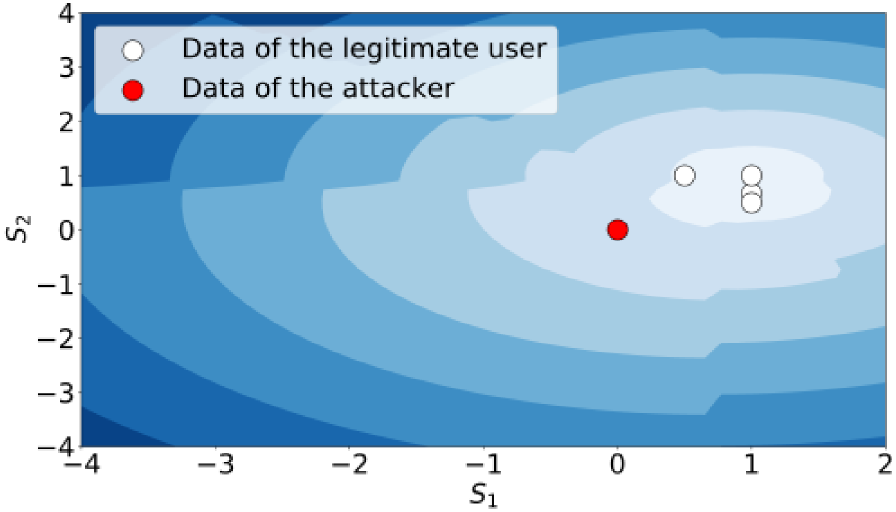


Feature extraction

- Luminance change trend
  - Evaluate the correlation of their trends
  - Reduce the impact of network delay
    - Average time difference between each pair of matched luminance change
  - Each signal is cut into two segments with equal length
  - Measure correlation using Pearson correlation coefficient for each pair of segments
    - Use the smaller one of them as the third feature
  - Use the maximum dynamic time warping (DTW) distance (expressed with z4) between each pair of segments as the fourth feature

Fake Facial Video Detection

- Detection for a single video clip
  - Build with good classification performance using only the data of a limited number of legitimate users.
  - Local outlier factor (LOF) model

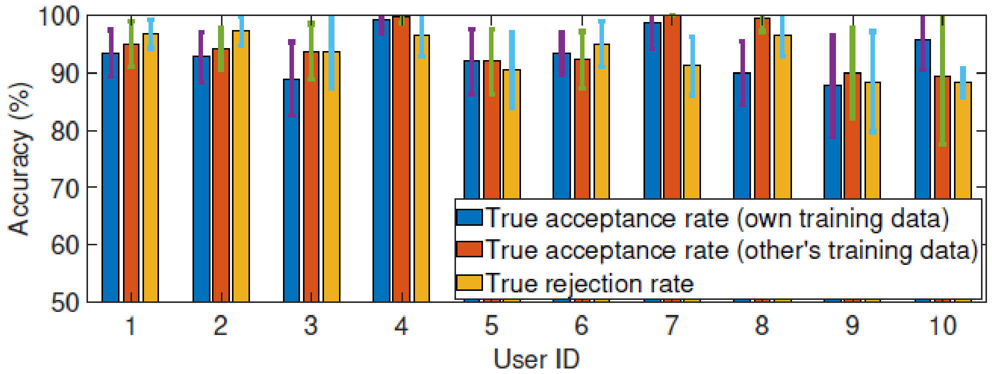


Evaluation

- Testbed
  - Screen: Dell 27-inch LED monitor with 85% brightness
  - Webcam: The front camera of Google Nexus 6 smartphone
  - Fake facial video: ICface
    - Generating the most visually convincing results of any open-source methods
  - 10 volunteers (four females and six males)
  - Each facial video is 15 seconds in length
  - Data processing: desktop computer with Intel(R) i7-8700 @ 3.2 GHz CPU and 32 GB of RAM

Overall Performance

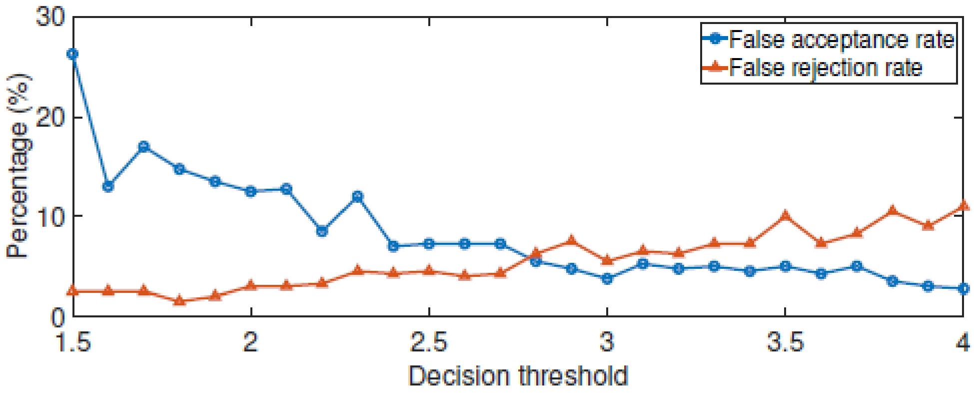
- An average true acceptance rate of 92.5% when the classifier is trained using own data.
- Achieve an average true acceptance rate of 92.8% with other's training data
- Reject attackers with average accuracy of 94.4%.



MONTCLAIR STATE UNIVERSITY

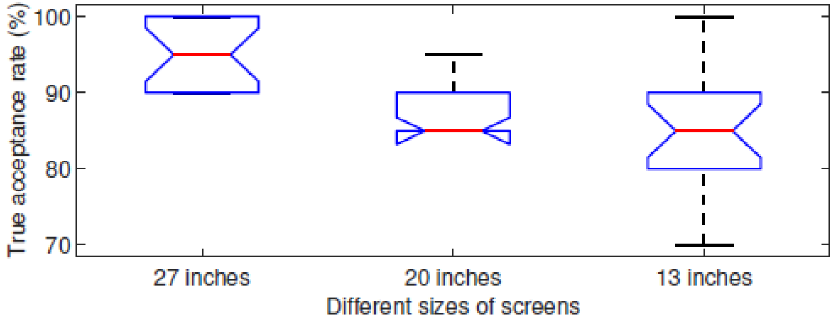
Impact of Decision Threshold

- When the decision threshold is between 2.8 and 3, our system can provide an equal error rate of about 5.5%.



Impact of Screen Size

- Screen size has a significant impact on the performance



Conclusion

- We show that the face reflected light can be leveraged to detect fake facial video with low cost and high generality.
- Our system only requires a limited number of training instances from the legitimate user and does not need to collect data from attackers.
- We develop a prototype and conduct comprehensive evaluations. Experimental results show that our system can provide an average true acceptance rate of at least 92.5% for legitimate users and reject face reenactment attackers with mean accuracy of at least 94.4% for each detection

MONTCLAIR STATE UNIVERSITY

Thank you