Detection of Deep-Morphed Deepfake Images to Make Robust Automatic Facial Recognition Systems

Presenter: Alakananda Mitra

A. Mitra¹, S. P. Mohanty², P. Corcoran³, and E. Kougianos⁴
University of North Texas, Denton, TX, USA.¹,²,⁴ and
National University of Ireland, Galway, Ireland³.
Email: alakanandamitra@my.unt.edu¹, saraju.mohanty@unt.edu²,
peter.corcoran@nuigalway.ie³, and elias.kougianos@unt.edu⁴
Outline

- Facial Recognition System
- Attacks on Facial Recognition System
- Deep-Morphed Deepfake Attack
- Proposed Solution
- Results
- Conclusions & Future Work
Identification of Individual in Smart City
Facial Recognition System (FRS)

- **Facial Recognition System**
  - Biometric based Identification System - Unique to the User

- **Non-invasive Identification System** - No Touching required

- **Process of Identifying or Verifying the Identity of a Person using his/her Face**

- **Steps for FRS:**
  - **Face Detection**: Detecting and Locating Human Faces in Images/ Videos
  - **Face Capture**: Changes Information (Features) of a face into a Set of Vectors
  - **Face Match**: Verifies if Two Faces are of the Same Person
Attacks on FRS

- Susceptible to Attacks
  - Presentation Attack: A Biometric Spoof Detected when Presented to a Biometric Sensor
  - Indirect/Channel Attack: When Data Moves in the Network without Encryption
  - Face Morphing Attack (FMA): Morphed Image
    - Traditional – Landmark Points Based
    - Deep-Morphed Deepfake – GAN Generated
      - (MorGAN, StyleGAN, FSGAN)
Deep-Morphed Deepfake

- **Deepfake** = **Deep Learning** + **Fake**

- Created by Deep Learning Networks
  - **Generative Adversarial Networks** (GANs)

- Sophisticated Images

- Make Face Morphing Easy and Realistic

- Rampant in Social Media and Websites

- Change the Perception of TRUTH

- Threat to Biometrics Based Facial Recognition Systems
Face Morphing Attack on FRS of Smart City

- Citizens Submit their Existing Photo ID to the City Council Office
- Hostile – Person I1; Victim – Person I2
- ID of Hostile Person I1 Tampered with Morphed Photo from Victim Person I2
- Photo of the ID Matched with the Hostile Person I1
- Registered in the FRS
Hostile Person I\textsubscript{1} comes to a Smart City Facility

I\textsubscript{1}’s Face matches with the Data Stored

Gains Access to that Facility
Proposed Solution For The Problem

Problems

- Misuse of FRS
- Innocent People Victims
- Hostile People take Advantages

Solution Proposed

- CNN based network detects Deep-Morphed Deepfake Images
- Is used to detect images submitted for registration in FRS of smart cities
- Light Weight - IoT friendly model, makes the Registration Process easy and not localized to Council Office
# Related Works

<table>
<thead>
<tr>
<th>Papers</th>
<th>Dataset</th>
<th>Methods</th>
<th>AUC/ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matern et al. [2019]</td>
<td>DeepfakeTIMIT</td>
<td>Visual Aspects + Logistic Regression + MLP</td>
<td>Low</td>
</tr>
<tr>
<td>Yang et al. [2019]</td>
<td>DeepfakeTIMIT</td>
<td>Head pose + Facial expression + dlib + SVM</td>
<td>Low</td>
</tr>
<tr>
<td>Afchar et al. [2018]</td>
<td>DeepfakeTIMIT</td>
<td>Mesoscopic Features</td>
<td>High</td>
</tr>
<tr>
<td>Zhou et al. [2018]</td>
<td>DeepfakeTIMIT</td>
<td>Steganalysis + Deep learning feature</td>
<td>Low</td>
</tr>
<tr>
<td>Nguyen et al. [2019]</td>
<td>DeepfakeTIMIT</td>
<td>Capsule network</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Proposed Method [2021]</strong></td>
<td>DeepfakeTIMIT</td>
<td><strong>CNN based</strong></td>
<td><strong>Highest</strong></td>
</tr>
</tbody>
</table>
CNN Based Detection Method

Data Processing → Feature Extraction → Classification → Prediction

Input Image

Face Detection, Crop, and Resize

Convolutional Neural Network

Fully Connected Layer (2 nodes) with Softmax Activation

\[ y^i = \frac{e^{y^i}}{\sum_j e^{y^j}} \]

Real, Fake
CNN Based Detection Method (Contd..)

- MobileNet V2 as Feature Extractor
- Depthwise Separable Convolution
- Linear Bottleneck between layers
- Shortcuts connect the bottleneck layers
- Last FC layer with ImageNet classes changed to a FC layer with softmax activation and two nodes
**CNN Based Detection Method (Contd..)**

- MobileNet V2 as Feature Extractor
- Depthwise Separable Convolution
- Linear Bottleneck between layers
- Shortcuts connect the bottleneck layers
- Last FC layer with ImageNet classes changed to a FC layer with softmax activation and two nodes
## Datasets

<table>
<thead>
<tr>
<th></th>
<th>Real</th>
<th>Fake</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data source</strong></td>
<td><strong># of Image</strong></td>
<td><strong>Data source</strong></td>
</tr>
<tr>
<td>VidTIMIT</td>
<td>34,004</td>
<td>DeepfakeTIMIT (HQ)</td>
</tr>
<tr>
<td>VidTIMIT</td>
<td>34,004</td>
<td>DeepfakeTIMIT (LQ)</td>
</tr>
</tbody>
</table>

### DeepfakeTIMIT
- 32 subjects
- Total of 620 videos
- A lower quality (LQ) with 64x64 in/out size
- A higher quality (HQ) 128x128 in/out size
- Fake image frame rate 25 fps

### VidTIMIT
- Same subjects' videos as DeepfakeTIMIT

<table>
<thead>
<tr>
<th>Data</th>
<th>Real # of Images</th>
<th>Fake # of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeepfakeTIMIT (HQ)</td>
<td>DeepfakeTIMIT (LQ)</td>
</tr>
<tr>
<td>Train</td>
<td>23,873</td>
<td>23,939</td>
</tr>
<tr>
<td></td>
<td>23,965</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>6,135</td>
<td>6,000</td>
</tr>
<tr>
<td></td>
<td>6,010</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>3,996</td>
<td>4,049</td>
</tr>
<tr>
<td></td>
<td>4,050</td>
<td></td>
</tr>
</tbody>
</table>
Implementation & Training Protocols

- Data Augmentation
- Transfer Learning
  - Accuracy Higher
  - Training Time Lower
- Feature Extractor kept frozen. Last FC layer trained for 10 epochs
- End-to-end network trained for 15 epochs
- Best model chosen from validation accuracy
- Same and Cross dataset evaluation
- Adam Optimizer learning rate 0.0002
- GeForce RTX 2060 laptop with GPU and 6GB shared memory+ 16GB total memory
Feature Visualization of MobileNet V2

Output of 32 Filters at Layer 2

Output of 32 Filters of 1280 Filters at Layer 153
Class Activation Map Visualization (GRAD-CAM)

Predicted
Wrong

Real

Fake

Predicted
Correct

Real

Fake

Pretrained on
ImageNet

Trained on
DF-TIMIT HQ
## Accuracy & Inference Time

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Testing Dataset</th>
<th>Accuracy (%)</th>
<th>Inference Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepfakeTIMIT (HQ)</td>
<td>DeepfakeTIMIT (HQ)</td>
<td>94.83</td>
<td>3.67</td>
</tr>
<tr>
<td>DeepfakeTIMIT (LQ)</td>
<td>DeepfakeTIMIT (LQ)</td>
<td>100.00</td>
<td>3.76</td>
</tr>
<tr>
<td>DeepfakeTIMIT (HQ)</td>
<td>DeepfakeTIMIT (LQ)</td>
<td>96.91</td>
<td>3.81</td>
</tr>
<tr>
<td>DeepfakeTIMIT (LQ)</td>
<td>DeepfakeTIMIT (HQ)</td>
<td>57.38</td>
<td>4.45</td>
</tr>
</tbody>
</table>

For Real images ➔ VidTIMIT dataset
Confusion Matrix

Predicted label

<table>
<thead>
<tr>
<th></th>
<th>3636</th>
<th>413</th>
</tr>
</thead>
<tbody>
<tr>
<td>fake</td>
<td></td>
<td></td>
</tr>
<tr>
<td>real</td>
<td>3</td>
<td>3993</td>
</tr>
</tbody>
</table>

True label

<table>
<thead>
<tr>
<th></th>
<th>Predicated Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive (TP): True</td>
<td>False Negative (FN): True</td>
</tr>
<tr>
<td>Model predicted: Fake</td>
<td>Model predicted: Real</td>
</tr>
<tr>
<td>False Positive (FP): True</td>
<td>True Negative (TN): True</td>
</tr>
<tr>
<td>Reality: Real</td>
<td>Reality: Real</td>
</tr>
<tr>
<td>Model predicted: Fake</td>
<td>Model predicted: Real</td>
</tr>
</tbody>
</table>

December 17, 2021

OCIT 2021 - Alakananda Mitra
## Detection Metrics

<table>
<thead>
<tr>
<th>Test Images</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,996 Real</td>
<td>100.0</td>
<td>90.0</td>
<td>95.0</td>
</tr>
<tr>
<td>4,048 Fake</td>
<td>91.0</td>
<td>100.0</td>
<td>95.0</td>
</tr>
<tr>
<td>Macro Average</td>
<td>95.0</td>
<td>95.0</td>
<td>95.0</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>95.0</td>
<td>95.0</td>
<td>95.0</td>
</tr>
<tr>
<td>Total 8,044</td>
<td>Accuracy (%)</td>
<td>95.0</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Accuracy} = \left( \frac{TP+TN}{TP+TN+FP+FN} \right) \times 100\\
\text{Precision} = \left( \frac{TP}{TP+FP} \right) \times 100\\
\text{Recall} = \left( \frac{TP}{TP+FN} \right) \times 100\\
\text{F1-score} = \left( \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \right) \times 100
\]
## Performance Comparison

<table>
<thead>
<tr>
<th>Papers</th>
<th>PERFORMANCE (%) For DF-TIMIT LQ</th>
<th>PERFORMANCE (%) For DF-TIMIT HQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matern et al. [2019]</td>
<td>AUC = 77.00</td>
<td>AUC = 77.30</td>
</tr>
<tr>
<td>Yang et al. [2019]</td>
<td>AUC = 55.10</td>
<td>AUC = 53.20</td>
</tr>
<tr>
<td>Afchar et al. [2018]</td>
<td>AUC = 87.80</td>
<td>AUC = 68.40</td>
</tr>
<tr>
<td>Zhou et al. [2018]</td>
<td>AUC = 83.50</td>
<td>AUC = 73.50</td>
</tr>
<tr>
<td>Nguyen et al. [2019]</td>
<td>AUC = 78.40</td>
<td>AUC = 74.40</td>
</tr>
<tr>
<td><strong>Proposed Method [2021]</strong></td>
<td><strong>ACC = 100.00</strong></td>
<td><strong>ACC = 94.83</strong></td>
</tr>
</tbody>
</table>
Conclusions & Future work

- Proposed a CNN based model for Detection of Deep-Morphed Deepfake images in context of Smart City facilities.
- Detected FSGAN generated images
- Light Weight model - makes the Registration Process easy and not localized to Council Office
- High Accuracy
- As future work, generalizability of the model can be obtained
Thank You!!