Exposing Deep-Fake Videos by Detecting Face Warping Artifacts

Paper about detecting deep-fake images and videos without deep learning model using normal face image datasets. Deep-fake images can only generate images of limited resolutions and when these images warped with normal images, it leads to distinctive artifacts. Model has two advantages. First, easier way to make samples that works similar with real deep-fake image which is hard to make. Second, artifacts are generally existing in deep-fake, model is more robust compared to others.

1. Introduction

AI-based fake video named Deep-Fake takes specific individual target video input and makes output video that target’s face changed. The main issue is to train neural network to automatically mapping facial expressions of the source to target.

**Deep-Fake algorithm can only synthesize face images of fixed size since limitation of computation resources and production time. To match the configuration of the source’s face, they must undergo an affine warping and this leaves distinct artifacts due to resolution inconsistency between warped face area and surrounding context**. These artifacts can be used to detect Deep-Fake videos.

Model detects artifacts by comparing generated face areas and their surrounding regions with CNN. Simulating the resolution inconsistency in affine face warping: 1. Detect faces and then extract landmarks to compute the transform matrices to align the faces to a standard configuration(원본 이미지에서 얼굴 요소만 뽑아내어 정면을 보는 모양으로 만드는 transform matrices를 구함). 2. Apply Gaussian blurring to aligned face and this will affine warped back to original image using inverse of the estimated transform matrix (1에서 구한 transform matrices의 역 행렬을 사용해서 Gaussian blurring을 거친 얼굴 이미지를 다시 원본 이미지에 삽입). To make diverse dataset, align faces into multiple scales. This method for creating negative data without training deep-fake model saves plenty of time and resources.

2. Related Works

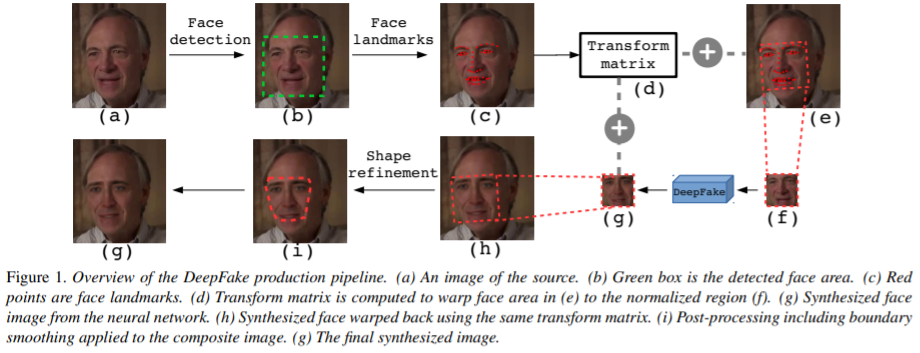
GAN Generated Image/Video Detection

Since there’s no many training images with eye blinking, deep-fake images lack realistic eye blinking and this is detected with CNN/RNN model to expose deep-fake videos. There’s many other methods that considering the head poses or color disparity of images.

3. Methods

**Deep-Fake algorithm can only synthesize face images of fixed size since limitation of computation resources and production time (lack of time for producing perfect image). To match the configuration of the source’s face, they must undergo an affine warping and this leaves distinct artifacts due to resolution inconsistency between warped face area and surrounding context**.

Follow is method of making deep-fake image.



a. Source image에서 얼굴 부분 detection

b. 얼굴에서의 landmark를 따 놓음

c, d, e, f. landmark를 기준으로 정방형 사진이 되게 하는 transform matrix를 구함

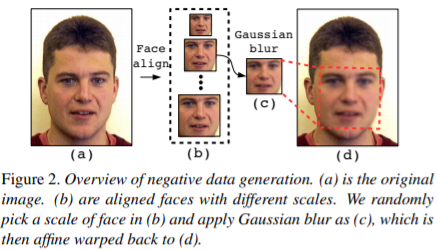
g. neural network를 거쳐서 만들어진 새로운 얼굴 landmark 사진

h. 새로운 landmark 사진을 transform matrix의 역 행렬을 통하여 source image에 fitting시킴

i. fitting된 사진을 보정함 (shape refinement)

For training dataset, used 24,442 JPEG face images as positive examples, simplify the negative example generation procedure by simulating the affine face warping step directly.

Follow is method of negative example in this paper.

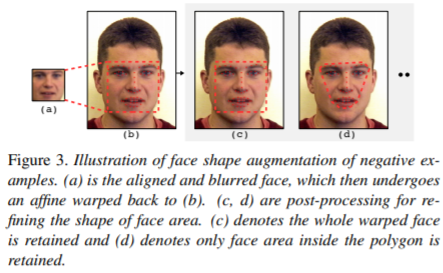


1. 원본 이미지에서 얼굴을 찾고 그 부분만 dlib라는 software를 사용하여 추출함.

2. 추출한 이미지를 여러 가지 사이즈로 정렬시킨 후 random하게 특정 사이즈의 이미지만 Gaussian blur with kernel size 5 X 5로 진행함. 이는 실제 deep-fake image에서 affine face warping을 할 경우 다른 resolution inconsistency가 생기는 것을 고려하기 위함임.

3. smoothed face는 원래 이미지 사이즈와 같은 사이즈로 affine warp back되어 deep-fake production pipeline에서의 artifacts와 같은 역할을 수행함.

To enlarge the training diversity, changed color information such as brightness, contrast, distortion, sharpness for all training examples. Also, changed the shape of affine warped face area to simulate different post-processing procedure in deep-fake pipeline. Shape of affine warped face area can be processed based on face landmarks. For example, convex polygon shape in **Figure 3(d)** is created based on face landmarks of eye browns and the bottom of mouth.



**Final inputs of our model network is region of interest(RoI) of positive, negative examples.** RoI chosen as rectangle areas that contain both face and surrounding area. RoI represented as [y0–y0’, x0-x0’, y1+y1’, x1+x1’] y0, x0, y1, x1 are minimum bounding box that contains all face landmarks, variables y0’, x0’ y1’, x1’ are random value between [0, h/5] and [0, w/8]. **RoIs resized to 224 X 224 to feed to the CNN models for training.** VGG16, ResNet50, ResNet101, ResNet152 are used for training model. **Crop RoI of each training example by 10 times and average predictions of all RoIs as the final fake probability.**

4. Experiments

**For each training batch, randomly select half positive examples and convert them into negative examples following the pipeline in Figure 2. Batch size 64, learning rate started from 0.001, decay 0.95 after each 1000steps, in SGD optimizer.** VGG16 epoch 100, ResNets load imageNet pretrained model fine tune them using data for epoch 20.

4.1 Evaluations on UADFV

Validation, deep-fake video dataset UADFV, 98 videos, 32752 frames in total. Used Area Under Curve(AUC) metric on two setting, image based evaluation, video based evaluation. ResNet50 is best performance and this implies classification-relevant information diminishes when layers go deeper. Each environment in frames are different, if certain number of frames in video are detected as fake then total video is deep-fake generated. Feed all frames of the video to CNN model and return average the top third of the output score as the overall output of the video.

4.2 Evaluations on Deep-fake TIMIT

Deep-fake TIMIT contains two fake video sets, lower quality, 64 X 64 input/output size model, higher quality 128 X 128 size model. Each set has 32 subjects, each subject has 10 videos with face swapped and 512 X 384 sized video and lasts for 4 seconds. Original dataset Vid TIMIT and fake videos from deep-fake TMIT for validation 10537 original images and 34023 fake images for each quality set. Overall result is CNN model is effective in detecting the existence of such artifacts.

4.3 Comparing with State-of-the-arts

This method focuses on more intuitive aspect in deep-fake video generation: resolution inconsistency in face warping, which is thereby more robust to deep-fake videos of different source