Exposing DeepFake Videos By Detecting Face Warping Artifacts

Abstract

- 본 논문에서는 가짜 비디오를 구분해낼 수 있는 딥러닝 베이스의 method를 제한한다.
- 이전의 method 에서는 GAN 모델 등에서 생성된 딥페이크 이미지가 학습에 많이 필요하였지만, 본 논문에서 제안한 method에서는 가짜 이미지를 구분하기 위한 특징으로 affine 얼굴 뒤틀림의 아티펙트를 대상으로 하기에 생성 이미지가 필요 없다.
- 본 논문에서 제안된 method의 이점은 다음과 같다.
    - 새로운 딥페이크 이미지를 생성할 필요가 없으니, 시간과 비용을 많이 줄일 수 있다.
    - 본 논문의 method는 2개의 Deepfake video 데이터셋으로 evaluate 되어서 다른 method 보다 새로운 데이터에 대한 성능이 견고하다.

Introduction

인공지능을 베이스로 하는 가짜 이미지 생성인 딥페이크는 근래에 많은 관심을 모으고 있다. 그 기술이 발전하며 진짜 이미지와 구분이 어려운 가짜 딥페이크 이미지 생성이 가능해지고 있다.

본 논문에서는 딥페이크 비디오를 진짜 이미지에서 구분할 수 있는 딥러닝 베이스의 method를 제안하였다. 그 method는 딥페이크 알고리즘은 computation 자원과 production time의 제한 때문에 제한된 크기의 얼굴 이미지만 합성할 수 있으며 source face를 match 하기 위해 affine warping을 거치게 된다는 특징에 기반한다.

상기된 warping은 warped face와 다른 부분의 resolution inconsistency를 발생시키며, 이는 진짜 이미지와 딥페이크 이미지를 구분할 수 있는 여지를 제공하여 준다.

즉, 본 논문의 method는 인공지능에 의해 생성된 얼굴 부분과 근처 부분을 CNN을 사용해 비교한다는 것이다. CNN 모델을 학습시키기 위해서 affine face warping에 resolution

inconsistency를 simulate한다. 구체적으로 보자면, 먼저 이미지에서 얼굴을 detect 하고, 얼굴을 표준 구성으로 정렬하기 위해 쓰이는 transform metrics를 계산하기 위해 랜드마크를 추출한다.

Method

딥페이크 이미지를 제작할 때에 변화가 생기는 얼굴 부분에 아핀 변환이 일어난다. 이미지에 변환이 생긴 부분과 그 부분을 둘러싸고 있는 부분 사이에 resolution 차이가 발생할 것이다. 따라서, 본 논문에서는 CNN 모델을 제안해 딥페이크 영상을 구분해내었다.

CNN 모델의 학습은 인터넷에서 모은 이미지에 기반한다. 본 논문의 저자들은 여러 positive 이미지들을 모으고, 딥페이크 이미지를 생성하였다. 하지만 이러한 방법은 많은 시간과 비용이 필요하다. 따라서 본 논문의 저자들은 하기 이미지처럼 딥페이크 이미지들이 간단하게 affine warping step을 바로 진행하도록 하였다.

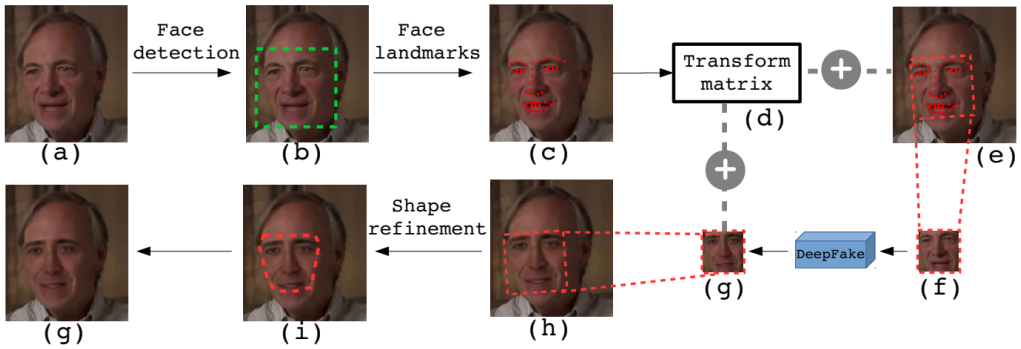

Figure 1. *Overview of the DeepFake production pipeline. (a) An image of the source. (b) Green box is the detected face area. (c) Red points are face landmarks. (d) Transform matrix is computed to warp face area in (e) to the normalized region (f). (g) Synthesized face image from the neural network. (h) Synthesized face warped back using the same transform matrix. (i) Post-processing including boundary smoothing applied to the composite image. (g) The final synthesized image.*

하기 이미지와 같이 본 논문의 저자들은 다음과 같은 순서로 CNN 모델을 학습시키기 위한 딥페이크 이미지를 생성하였다.

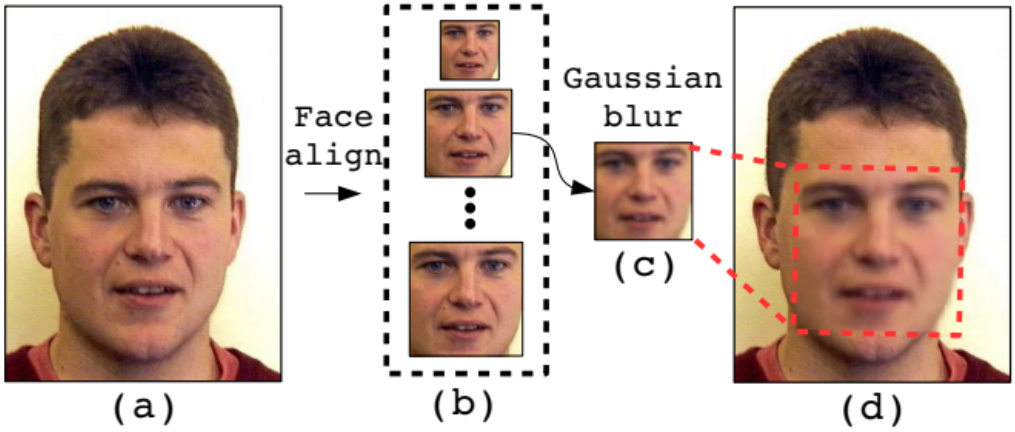

Figure 2. *Overview of negative data generation. (a) is the original image. (b) are aligned faces with different scales. We randomly pick a scale of face in (b) and apply Gaussian blur as (c), which is then affine warped back to (d).*

1. 본 논문의 저자들은 dlib을 통해 얼굴 부분을 추출해내었다.
2. 얼굴 부분을 여러 scale로 할당하고 랜덤하게 하나를 고른다. 그 후 5x5 커널의 gaussian blur을 통해 이미지를 부드럽게 한다.
3. 부드럽게 된 얼굴 이미지는 affine warp를 통해 원본 얼굴 이미지와 같은 크기로 변화한다.

학습의 다양성을 더하여주기 위해서 본 논문의 저자들은 color information을 변경하였다. 또한 affine warped face area를 변경하여 딥페이크 파이프라인의 후처리 절차를 다양화 하였다. 하기 이미지와 같이 affine warped face area가 얼굴의 landmark를 기반으로 생성된다(이미지 d와 같이).

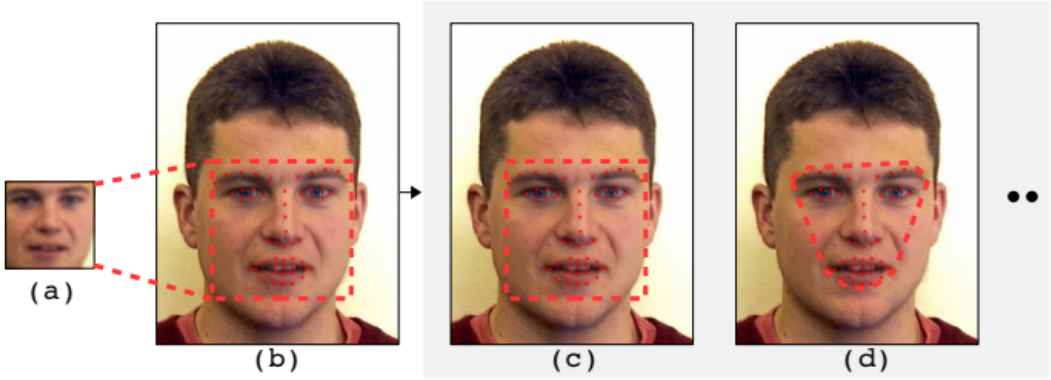

Figure 3. *Illustration of face shape augmentation of negative examples. (a) is the aligned and blurred face, which then undergoes an affine warped back to (b). (c, d) are post-processing for refining the shape of face area. (c) denotes the whole warped face is retained and (d) denotes only face area inside the polygon is retained.*

positive와 negative 이미지에서 region of interest(RoI)를 입력값으로 뽑아내었다. 본 논문의 저자들의 목표는 face area와 surrounding area의 차이를 보는 것이기에 RoI는 face area와 surrounding area를 모두 포함하는 부분으로 정의한다. 크기는 $[y_0 - \hat{y}_0,\ x_0 - \hat{x}_0,\ y_1 + \hat{y}_1,\ x_1 + \hat{x}_1]$ 이며, y_0, x_0, y_1, x_1 은 뺨의 외곽선을 제외한 얼굴을 모두 커버할 수 있는 minimum bounding box b 를 나타낸다. 그리고 $\hat{y}_0, \hat{x}_0, \hat{y}_1, \hat{x}_1$ 은 $[0, \frac{h}{5}]$ 와 $[0, \frac{w}{8}]$ 사이(h 는 얼굴의 height, w 는 얼굴의 width)의 랜덤한 값이다. 모든 RoI 이미들은 동일하게 224x224의 크기로 resize 되어 CNN 모델에 학습데이터로 들어간다.

개선 방안

- 간단한 CNN 모델 외에 더욱 정교하고 성능이 좋은 모델을 사용한다.
- 인공지능이 이미지를 생성한다면 그 규칙을 찾는 방법도 가능하지 않을까?