Ryan Collins Prof A. Krokhin MEng Computer Science

Facial Liveness Testing: For The Web

April 3, 2019

Abstract

Context TODO

- Verify the results of the Image Quality Assessment test.
- Assess the outcome Convolutional Neural Networks on classifying real/spoofed images.
- Design and implement a new 3D based liveness test, aimed to prevent mask attacks.
- Determine the outcome of fusing the three above methods together, and how successful this is.

Aims

Method TODO

Results TODO

Conclusions TODO

Facial liveness, convolutional neural networks, image quality metrics

1 Introduction

Currently, username and password authentication is commonplace throughout the web. However, username and password based authentication systems have a number of problems. Some common passwords can be broken using dictionary attacks, especially if they consist partially or entirely of a word in a standard dictionary. Furthermore, the process of shoulder surfing is possible (watching out for someone's password, and how they type it).

While there are different measures of detecting liveness, each method is specialised towards defending against a given attack. The aim of this project is to understand the existing liveness detection methods, which type of attack they aim to prevent, and how effective they are. Once this has been achieved, the aim shall be to bring each of these methods together, hopefully improving the effectiveness of such a system by encorporating multiple methods.

In this context, we propose a novel new 3D-based liveness test, based on a two part approach: (i) VRN based 3D reconstruction (ii) VoxNet based 3D classification. We also confirm the success of the Image Quality Assessment method for Facial Liveness, and provide an improve

2 Related Work

As defined in **?**, the types of face spoofing attacks can be described under three sections: Photo Attack, Video Attack and Mask Attack.

2.1 2D Spoofing Attacks

Photo and Video Attacks are both 2D spoofing attacks, which involve using a previously retrieved photo/video, and holding it in front of a camera. In the case of photo attacks, a single photo is used, where in video, some video would be played back on a screen. **?**.

With video-based facial recognition systems, motions of some form can be used to determine whether the person is real or spoofed, such as blinking, head movement and others. In the method defined in **?**, structure from motion was used on the video to produce a 3D model of a user, with the depth channel being used to determine whether a person is real, or whether it's simply an image. They also extended this by fusing this method with audio verification. The fusion of multiple methods provides greater reliablity. However, while SFM works with video, it doesn't work with a single image, and it also doesn't work if a video with little motion is provided. This fusion was completed using a Bayesian Network

While motion based methods are video-only, quality based methods are useful for both videos and images (either by extracting key video frames or using all video frames and combining the results).

While there are various quality metrics that have been used, combining a large number of them can yield some increased accuracy. By combining 25 different metrics, , yielding the resulting metric values into a large vector, and using that as input to a classifier (an LDA), this yields fairly high accuracy. **?**. This is an example of combining many items to yield better results. While each metric on its own isn't that great, using them all together yields better results.

Recently, deep learning based approaches have been applied to facial liveness (both video and image based).

In particular, Convolutional Neural Networks are a key approach to this to learn features (e.g. texture based methods). Due to the existing datasets available, training CNNs has been difficult due to lack of data where overfitting has been common. The method proposed in **?** uses CaffeNet, inputting both the full image along with the isolated face. The output yielded general texture differences, as well as specific facial texture differences. Another interesting idea proposed in this paper is the fusion of two algorithms together to produce an outcome, therefore reducing the false reject rate.

Outside of the facial liveness field, image classification on the imagenet dataset has proven popular and yielded some fairly good results.

AlexNet One of the initial models was AlexNet, which has 5 convolutional layers (with some max pooling layers), and two globally connected layers. This model was used to classify 1.3 million high resolution images into 1000 classes. **?** Since the AlexNet paper was written, newer methods have been built that performed better.

VGG16 Network The VGG16 model improved AlexNet by replacing larger filters by more smaller filters one after another. However, VGG requires a high amount of computational power, something that's not easily deployable on a real-life system due

Cite VGG lack of deployability

to high time and space requirements.

GoogLeNet Inception GoogLeNet Inception is an improved module that approximates a space Convolutional Network with a normal dense construction. Since a small number of neurons is effective, computational requirements are kept small.

Cite GoogLeNet Inception

Residual Networks Another key problem with deep convolutional networks on Image Classification is the vanishing gradient problem, where early layers have very small gradients during the training process and are therefore much more difficult to train. Residual Networks avoid this problem by allowing a direct path in links between the input and output of a building block. . The overall outcome is far better accuracy than VGG and GoogLeNet while being more efficent than VGG in terms of computational power needed. While these aren't directly associated with facial liveness, the nature of image classification is fairly similar to facial liveness (since the image is simply a classifier with two outputs instead of 1000).

Explain Resnets better

Cite Residual Networks

While models exist, in order to test these models data is needed. One of the most common and earliest dataset for facial liveness is the NUAA dataset, which consists of photos of 15 subjects, with faked photos (both flat and warped) being placed in front of the camera. **?**

In 2012, the Replay-Attack dataset was first released , which consists of 1,300 video clips of both photo and video attacks. Each set of videos/images are taken under different illumination conditions, and various different attack methods were collected: printed photo, low resolution and high resolution screens with both photos and videos being displayed. **?**

2.2 3D Spoofing Attacks

Mask Attacks are a 3D spoofing attack, which involve creating a 3D mask of someone and wearing it. **?** These are much less prevalent, but with 3D printing becoming more mainstream, this could potentially get more prevalent in the future.

In 2013, the Mask Attack Dataset (MAD dataset) was released. **?**

Improve this section, explaining 3D based measures, more about the MAD dataset and attacks, etc.

3 Solution

3.1 Image Quality Assessment based liveness test

For 2D spoofing attacks, spoofed images are typically lower quality than the real images, and thus by measuring the image quality one can train a classifier to detect real and spoofed images respectively.

The method used, based on the work of **?**, implements 24 different metrics with varying differences, and produces a vector for each image. Initially, classification was done using a Support Vector Machine (SVM), but after experimentation this proved to be fairly unreliable (yielding 70% accuracy on the test set). The classifier was later changed to use Linear Discriminant Analysis (LDA) which yielded a much improved accuracy (96% accuracy on the test set).

TODO: give more accuracy figures of accuracy here, I can't remember the exact numbers

.

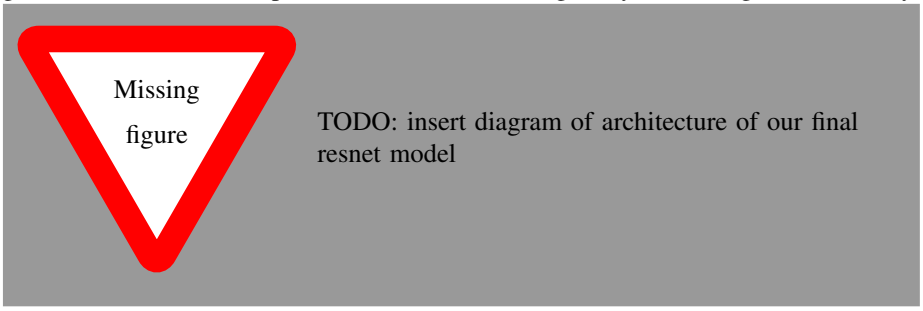
3.2 Residual Network based 2D liveness test

Recently, 2D convolutional neural networks have had great success in image classification tasks. Therefore, it might be possible to train a residual neural network (resnet) to classify for facial liveness tasks.

In order to simplify the process of training, an existing resnet model (ResNet50) was used, with only the final convolutional layer being set to trainable. This is because the initial convolutional layers contain the standard features contained within images, while the final one learns bundles of features. Internal feed forward activations use relu, while the external output uses the softmax activation function

Training was completed using the categorical cross-entropy loss function (as this is considered multiclass). We yield a 2-tuple output from this model, which is the probability of each possible case. We take the value with the highest probability as the true outcome.

The output of this ResNet model is then fed into a 2D Max Pooling layer, which then feeds into a feed forward neural network. Initially, the model was trained using the Adam optimiser, but this yielded poor accuracy (75% accuracy). Utilising the standard gradient descent (SGD) optimiser with a low learning rate yielded far greater accuracy.



TODO: insert citation for trying pretrained imagenet

3.3 A system for preventing 3D spoofing attacks

While the systems before might go partially towards preventing 3D spoofing attacks, though primarily considering the 2D image, we now propose a method that is designed for classifying facial liveness based on a 3D point cloud.

3.3.1 2D to 3D Conversion

In order to classify an image/video, the 2D image needs to be converted to a 3D representation of a user's face. While 3d reconstruction is easier with videos (using structure from motion or other multiview based methods), there also exist image-based reconstruction methods such as vrn (**?**) which are more specific and designed for reconstructing faces based on images. This also has benefits, as structure from motion is unable to reconstruct 3D from a single image, or from videos with very little motion.

The image was converted by first applying a facial detection algorithm on the image, and cropping the image down to provide only the face. This cropped image was then resized to be of size $(192, 192)$, still in colour. This cropped and resized image was then fed through the VRN network. After this, the network output was filtered and stacked to provide the voxel input required.

The code to operate this can be found under the *liveness.vox.reconstruction* namespace within the code.

3.3.2 3D point cloud classification

Once the 3D reconstruction is obtained, one can then classify this using some model to produce the fake/real metric.

VoxNet takes in a point cloud and converts this to an occupancy grid. This is then fed through two convolutional layers, pooled, and then goes through a dense layer before reaching the classifier output (a dense layer with the k outcomes).

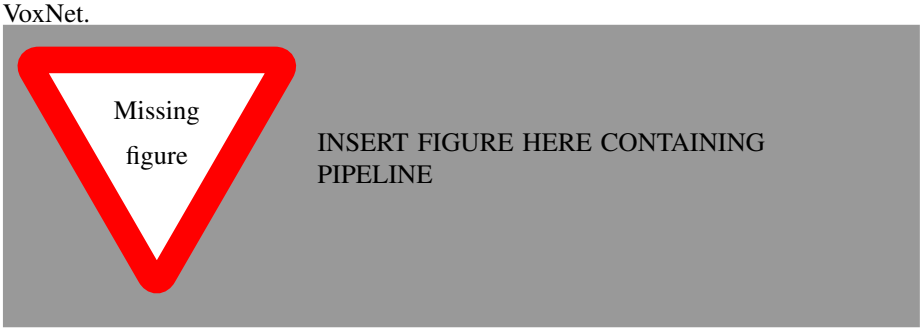
As a pretrained version of VoxNet wasn't readily available, the whole system was trained together from scratch.

3.3.3 Linking everything together

While each system is self-contained, linking them together took a little bit more work than expected. The models themselves couldn't be directly joined together, as VRN required extra postprocessing steps which couldn't be implemented using tensors within tensorflow. As such, the initial 2D to 3D conversion was required to be run as a preprocessing step.

To assist in the training phase, a generator was written in Python to conduct the postprocessing on the fly for each batch, which didn't require the entire preprocessing step to be done before training, thus reducing the peak memory usage problems. While an ImageDataGenerator was used previously, this isn't compatible with 3D, and therefore a custom module needed to be written.

Once the preprocessing had been completed, the preprocessed image was fed to the VoxNet.



3.4 Visualisation and Demonstration

In order to visualise the overall outcome of facial liveness, a generic model

4 Results

TODO results

5 Evaluation

TODO evaluation

6 Conclusions