Facial Liveness Testing: For The Web

Student Name: Ryan Collins
Supervisor Name: Prof A. Krokhin

Submitted as part of the degree of MEng Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University
March 28, 2019

***Abstract* —**

Context

TODO

Aims

- 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.

Method

Results

TODO

Conclusions

TODO

***Keywords* —** Facial liveness, convolutional neural networks, image quality metrics

I 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

II RELATED WORK

As defined in [4], the types of face spoofing attacks can be described under three sections: Photo Attack, Video Attack and Mask Attack.

A 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. [4].

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 [1], 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. [2]. 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 taken to facial liveness (both video and image based).

B 3D Spoofing Attacks

Mask Attacks are a 3D spoofing attack, which involve creating a 3D mask of someone and wearing it. [4] These are much less prevalent, but with 3D printing becoming more mainstream, this could potentially get more prevalent in the future.

III SOLUTION

A 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 **(author?)** [2], 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

.

B 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.

Missing figure

TODO: insert diagram of architecture of our final resnet model

TODO: insert citation for trying pretrained imagenet

C 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.

C.1 Point Cloud Reconstruction

In order to classify an image/video, a 3D point cloud needs to be created, containing many 3D points (x, y, z) 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 (**(author?)** 3) which are more specific and designed for reconstructing faces based on images.

C.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.

For points, PointNet is a model that can be used

EXPLAIN POINTNET

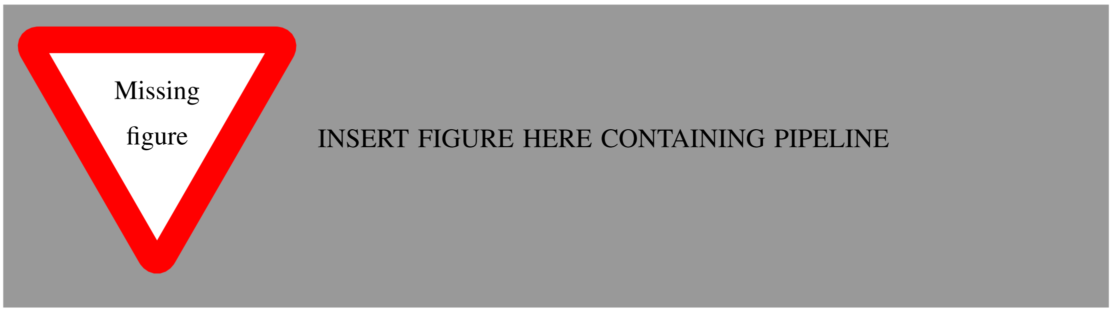
However, as we are using voxels, there is a more specialised architecutre called Voxnet that is designed for classifying 3d volume-based objects. VoxNet takes in a point cloud and converts this to an occupancy grid. This is then fed through two convolutional layers, pooled, and then goes through a dense layer before reaching the classifier output (a dense layer with the k outcomes).

Instead of training on our existing datasets, we first train a VoxNet model on the SUOD dataset, in order to learn the basic features surrounding 3D classification. This largely uses the tools provided in the original implementation of VoxNet, but with a custom Keras implementation. This pretrained model is then saved, for use in our final implementation.

Using this original VoxNet implementation, the last dense layer is then ignored, with it being extended to contain further dense layers, and an eventual output classifier (for fake/real respectively).

C.3 How this overall system functions

Each image is first preprocessed: the image is fed through the 3D reconstruction network, and the output is a set of Voxels. Using these voxels, they are then resized into a (24 x 24 x 24) shape to provide a basis for the network. From here, they are fed through several convolutional layers, before then being fed through several dense layers for classification.



D Visualisation and Demonstration

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

IV RESULTS

TODO results

V EVALUATION

TODO evaluation

VI CONCLUSIONS

References

[1] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland. Multimodal person recognition using unconstrained audio and video. In *Proceedings, International Conference on Audio-and Video-Based Person Authentication*, pages 176–181. Citeseer, 1999.

[2] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, Feb 2014.

[3] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *International Conference on Computer Vision*, 2017.

[4] Sandeep Kumar, Sukhwinder Singh, and Jagdish Kumar. A comparative study on face spoofing attacks. 05 2017.