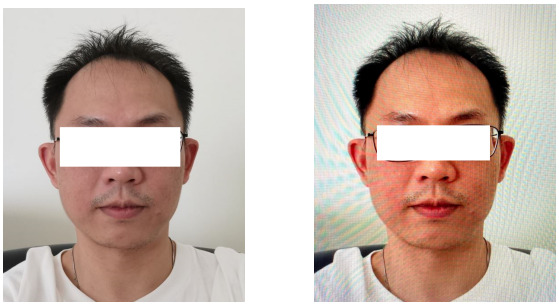
Liveness Detection for Images and Audio

Wei Yuen Teh
Wau Labs Sdn Bhd
Subang Jaya, Malaysia
tehweiyuen@waulabs.com

Ian K. T. Tan
School of Mathematical and Computer Sciences
Heriot-Watt University Malaysia
Putrajaya, Malaysia
i.tan@hw.ac.uk

Heng Kiat Tan
Wau Labs Sdn Bhd
Subang Jaya, Malaysia
kit@waulabs.com



(a) Live (b) Spoof

Fig. 1: Sample of a live selfie and a selfie spoofed using the screen replay method

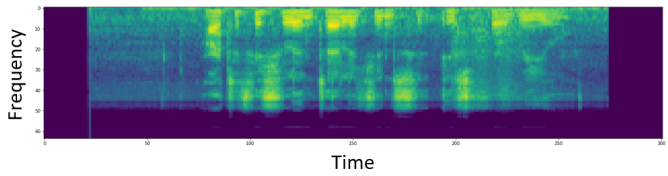


Fig. 2: Mel Spectrogram of an audio recording used for biometric verification.

***Abstract*—In this demo paper, we provide a brief overview of liveness detection in the visual and audio modalities. We provide links to our online demos, and discuss some of the technical details behind their implementations, including the use of Convolutional Neural Networks and Mel Spectrograms for the audio modality.**

I. Introduction

Liveness detection is the task of determining if a given media source contains a direct or "live" representation of a person. The most common modalities for this task are images and audio, where the former involves liveness detection on selfies, and the latter on voice recordings. For biometric authentication systems, liveness detection plays a crucial role alongside facial/voice recognition systems to ensure that the user is not attempting to replicate the biometric features of another person.

In general, liveness detection systems aim to classify images into two classes: live and spoof, where the live class indicates a live representation of a person, and the spoof class indicates a recreation. There are various spoofing methods depending on the input modality. For images, this can come in the form of an image of a screen depicting a person's selfie (a.k.a. screen replay method), or an image of printed material. For audio, spoofing can involve the playback of a person's voice from an electronic device. Figure 1 depicts an example of a live selfie and one spoofed using the screen replay method.

II. Methodology

Liveness detection was already studied prior to the deep learning era [1]. However, deep learning is particularly suited to the task of liveness detection. This is because the features required to identify spoofed media can largely be learned through induction and do not require any reasoning skills (an ability that current deep learning systems still lack). For example, spoof images taken from a screen often include the Moire effect. While careful control of lighting and camera angles can significantly reduce the visual impact of this effect, it is challenging to completely eliminate, especially at the individual pixel level. As such, kernel-based architectures like convolutional neural networks (CNNs) are well suited to picking up these repetitive frequency-based patterns. Likewise, audio that is spoofed by being re-recorded after being played back from another device contains a slightly different sound signature due to compression from both the playback device's speakers and the recording device's microphone.

We have found that CNNs are well suited for both image and audio liveness detection. Vision Transformers (ViTs) [2] have proven to be competitive with CNNs on image classification tasks over the past few years, especially when trained on large datasets such as JFT-300M. However, as mentioned above, the frequency-based features present in spoofed images are more easily learned by kernel-based methods like CNNs, especially when only limited training data is available, as is the case for this domain.

For audio, the sound signature differences described above can be visualized through the use of Mel Spectrograms (Figure 2), which can then be classified by a CNN. We are currently also working on the addition of voice and phrase matching as features to complement voice liveness detection. These features determine if two given voice samples are recorded by the same person, and contain the same uttered phrases. We are exploring the use of Siamese Networks [3] for this task, where each network embeds a voice sample, and the distance between the embeddings is used to determine feature

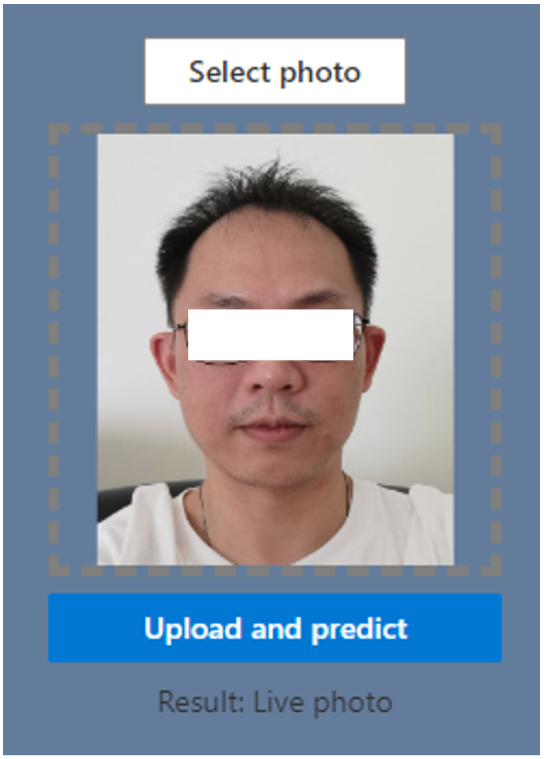


Fig. 3: Screenshot of output from our web demo

similarity.

III. Applications

Various sectors require the use of secure user authentication. Since liveness detection is a critical component in ensuring a user's physical presence, this, by extension, means that liveness detection is widely applicable across industries. Perhaps one of the most salient examples is the banking industry, where multinational banks have recently started implementing various forms of biometric identification, such as through the use of facial or voice recognition. If these recognition systems do not incorporate some form of liveness detection, malicious actors would easily be able to bypass these systems by recreating likenesses of the target user through methods described above. Complementary features such as the voice and phrase matching mentioned in Section II also provide added value. For example, the combination of liveness detection, voice matching, and phrase matching would allow for the use of voice-signature-based passwords in applications such as phone banking.

IV. Demo

Wau Labs is a startup that aims to establish a more research-oriented approach towards the development of AI models within the region. We operate primarily as a Contract Research Organization (CRO), with a long-term focus on the field of Explainable AI (XAI). Figure 3 shows a screenshot from our web demo for selfie liveness detection. This demo, as well as a demo for voice liveness detection, is publicly available on our website at https://www.waulabs.com/demos. Users will be able to upload their own images/audio to see if it's classified as live/spoof by the model. A mobile demo application that runs on-device will be available at the conference.

References

[1] G. Pan, Z. Wu, and L. Sun, "Liveness detection for face recognition," in *Recent Advances in Face Recognition*, K. Delac, M. Grgic, and M. S. Bartlett, Eds. Rijeka: IntechOpen, 2008, ch. 9. [Online]. Available: https://doi.org/10.5772/6397

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[3] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015, p. 0.