Passive Video liveness detection using Vision Transformer

Harish Krishnamoorthy Murali[1]
Pranav Kumar Ayee Goundar Venkatesan[1], Vicknesh Balabaskaran[1], Faizan Ahmad[1]
[1]System Design Engineering, University of Waterloo
{hkrishna, pkayeego, vbalabas, faizan.ahmad1}@uwaterloo.ca

Abstract

This paper presents a passive video liveness detection system designed to enhance security in online authentication and authorization services. As traditional authentication methods are increasingly susceptible to spoofing attacks using static images, videos, or deepfake technology, there is a growing need for advanced biometric solutions. The proposed system applies computer vision and machine learning techniques to accurately distinguish between live users and fraudulent attempts in real-time, without requiring active user interaction.

1 Introduction

Liveness detection is a biometric security measure that ensures the individual attempting to log in or perform an online transaction is a real, live person, rather than a fraudulent representation such as a photograph, video, or mask. It differentiates between genuine users and impostors attempting to deceive the system with false biometric data [1, 2].

As the reliance on online services continues to grow, the necessity for advanced security measures becomes increasingly critical. Liveness detection not only fortifies defenses against fraud but also enhances user confidence by providing a secure environment for online transactions [3].

Liveness detection provides a significant layer of security to prevent unauthorized access and fraudulent activities using spoofing methods such as videos, images, or deepfake technology. There are two main approaches: passive and active. Passive video liveness detection operates unobtrusively, analyzing the video feed for subtle cues such as eye blinking, facial micro-expressions, and texture patterns without requiring any user interaction. In contrast, active video liveness detection involves user participation, asking individuals to perform specific actions like nodding, smiling, or turning their head. In this paper, we focus on the passive method, leveraging advanced deep learning algorithms to detect liveness based on naturally occurring facial movements and characteristics.

2 Deep Learning Models

Two models were tested and compared for the task of liveness detection: Vision Transformers (ViT) and Convolutional Neural Networks (CNN). The following sections provide a brief description of each model.

2.1 Vision Transformer (ViT)

The Vision Transformer (ViT) [4] operates by breaking images into patches and processing them through a transformer architecture. This model captures long-range dependencies within the images, which makes it effective in distinguishing between live and spoof attempts. Transfer learning was applied to further fine-tune a pre-trained ViT model for improved performance on the dataset. The ViT model was trained for binary classification, differentiating between live and spoofed images.

A ViT model [5], with 86.6 M parameters pretrained on ImageNet-21k [6] was fine tuned on the custom dataset we created after adding a final layer of 2 neurons to classify the image as either real or fake. Performance dip was noticed after 5 epochs hence the model training was stopped early to achieve better accuracy.

2.2 Convolutional Neural Network (CNN)

The CNN model utilizes convolutional layers to extract spatial features from the images. It processes the images through a series of layers including Conv2D, MaxPooling2D, and fully connected layers for classification. The CNN architecture was optimized for binary classification tasks, using techniques like dropout to prevent overfitting and improve generalization. The CNN was also trained on the same dataset, focusing on detecting liveness in real-time scenarios.

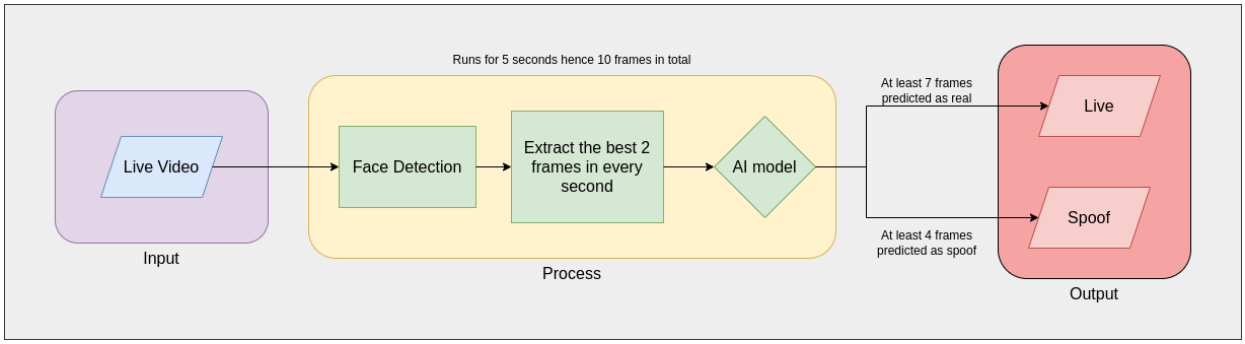

Figure 1: System Architecture Diagram

3 Dataset Specification

Two datasets were used to train and test the models:

- **NUAA Photograph Imposter Dataset [7]:** This dataset contained over 12,000 images captured under different lighting conditions, but it lacked sufficient racial diversity for generalization.

- **Custom-Made Dataset:** To address the lack of diversity, a more comprehensive dataset was created, combining images from different sources [8, 9, 10, 11, 12, 13]. We carefully curated this dataset by gathering data from multiple diverse sources to ensure it accurately represents a wide range of human characteristics, including races and genders. To enhance inclusivity, we ensured that the dataset captures features from individuals across different ethnic groups and gender identities. Additionally, we accounted for environmental factors such as lighting conditions by including images and videos captured under various scenarios, from natural sunlight to artificial and low-light settings. This meticulous process was aimed at reducing bias and promoting fairness in model training and evaluation, making the dataset robust and adaptable for real-world applications.

4 Results & Analysis

4.1 Model Comparison

The system's performance was compared using Vision Transformer and Convolutional Neural Network models.

As shown in Table 1, the ViT model outperformed the CNN model in terms of accuracy and robustness, with significantly lower false acceptance and rejection rates. Due to its superior performance, the ViT model was chosen for real-time testing.

Table 1: Comparison of ViT and CNN based models by performance on test dataset.

Metric	ViT (Fine-tuned)	CNN
Accuracy	82.43%	78.04%
Recall	87.16%	44.59%
FAR	22.29%	45.27%
FRR	12.83%	55.41%

4.2 Real-Time Processing

A real-time processing pipeline was built using computer vision techniques, enabling the system to detect liveness from live video feeds. The system leverages the accuracy metrics obtained from the performance evaluation of the Vision Transformer (ViT) model on fake detection for single images. By mathematically extrapolating these metrics, we evaluated how the model would perform on video data.

We capture 10 frames of the user, pass them to the model, and get predictions for their liveness. The video is accepted for authorization only if at least 7 out of these 10 frames are predicted as real. The threshold of 7 out of 10 was selected because it allows for up to 3 false positives, while still providing a high fake detection accuracy. By performing predictions on multiple frames and basing the liveness decision on this majority rule, we significantly increase the model's overall performance.

To calculate the accuracy of the system for predicting the liveness of a video, we need to consider the binomial distribution. We need to find the probability that 7, 8, 9, or 10 frames out of 10 are predicted as live.

The probability mass function for the binomial distribution is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where:

- n is the number of trials (frames), which is 10.

- k is the number of successful trials (correctly predicted live frames).
- p is the probability of success on a single trial (0.8243).

We need to calculate the sum of probabilities for $k = 7, 8, 9, 10$.

$$P(X \geq 7) = \sum_{k=7}^{10} P(X = k)$$

For ViT model, which gave us a test accuracy of 82.43%, this value comes up to be 91.64% (using $p = 0.8243$ in the above equation).

Similarly, to find how well it detects fake videos, we find $P(Y \geq 4)$, where Y is a random variable describing the number of times the model correctly predicts a frame as fake. This comes up to be 99.67%. This means that the system can predict live video as live with an accuracy of 91.64% and fake videos as fraudulent with an accuracy of 99.67%.

By making a liveness detection decision based on multiple frames, the system's performance is boosted from an accuracy of 82.43% to 91.64% for real cases and to 99.67% for fake cases. These accuracies align with the product expectations as we cannot tolerate any fake users logging into the system, which would cause a security breach.

5 Conclusion & Future Work

This paper successfully developed a robust video liveness detection system, integrating Vision Transformers for accurate liveness prediction. The system achieved over 90% accuracy in real-time tests, making it highly suitable for secure banking applications. The Vision Transformer model proved to be more effective than traditional CNN approaches, particularly in reducing false rejection and false acceptance rates.

Future work will focus on expanding the dataset to further improve the performance of the system across different demographics and to evaluate the performance of real-time processing in a wider range of test data. Additionally, more advanced hardware could enhance the processing speed, making the system even more responsive for real-time applications.

References

[1] K. A. Omotoye, S. Misra, M. Kaushik, R. Ogundokun, and L. Garg, "Facial liveness detection in biometrics: A multivocal literature review," in *International Conference on Information Systems and Management Science*. Springer, 2021, pp. 195–209.

[2] E. A. Raheem, S. M. S. Ahmad, and W. A. W. Adnan, "Insight on face liveness detection: A systematic literature review," *International Journal of Electrical & Computer Engineering*, vol. 9, no. 6, 2019.

[3] S. Chakraborty and D. Das, "An overview of face liveness detection," *arXiv preprint arXiv:1405.2227*, 2014.

[4] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[5] R. Wightman, "timm/vit_base_patch16_224. augreg_in21k_ft_in1k," https://huggingface.co/timm/vit_base_patch16_224.augreg_in21k_ft_in1k, 2021, [Pretrained vision transformer model].

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[7] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*. Springer, 2010, pp. 504–517.

[8] TrainingDataPro, "Caucasian people liveness detection dataset," 2024, accessed: 2024-06. [Online]. Available: https://www.kaggle.com/datasets/trainingdatapro/caucasian-people-liveness-detection-dataset

[9] TrainingDataPro, "On-device face liveness detection," 2023, accessed: 2024-06. [Online]. Available: https://www.kaggle.com/datasets/trainingdatapro/on-device-face-liveness-detection

[10] TrainingDataPro, "Real vs fake: Anti-spoofing video classification," 2023, accessed: 2024-06. [Online]. Available: https://www.kaggle.com/datasets/trainingdatapro/real-vs-fake-anti-spoofing-video-classification

[11] TrainingDataPro, "Web camera face liveness detection," 2023, accessed: 2024-06. [Online]. Available: https://www.kaggle.com/datasets/trainingdatapro/web-camera-face-liveness-detection

[12] AxonData, “Liveness detection: Real and display attacks (5k),” 2024, accessed: 2024-06. [Online]. Available: https://www.kaggle.com/datasets/axondata/liveness-detection-real-and-display-attacks-5k

[13] Tapakah68, “Printed 2d masks with holes for eyes attacks,” 2023, accessed: 2024-06. [Online]. Available: https://www.kaggle.com/datasets/tapakah68/printed-2d-masks-with-holes-for-eyes-attacks