Brief of thesis:
Fake videos detection using Remote Photoplethysmography method

Le Huu Hoan
Student, Ho Chi Minh University of Technology, Vietnam
Email: lhhoan.ai@gmail.com

***Abstract*—Deepfake videos were created for entertainment purposes, but recently, these techniques have been used for fake identifications of humans. Therefore, Deepfake detection is an important technique for preventing that. The type of video considered is a photo with animation or a masked person video because these video are easy to create. This project uses Remote Photoplethysmography (rPPG) and Long short time memory (LSTM) to classify fake videos. Besides that, this model also used a Convolutional neural network (CNN) to recognize the appearance of the face. The achieved result is an accuracy of 92.33% on the video of the test dataset. The accuracy of the segment is 98.85% on test and 93.19% on validation. Link to github project: https://github.com/Maxlee2704/Fake-video-detection.**

1. Introduction

Fake videos were initially created to make fun, entertaining video. Recently, in Vietnam and around the world, fake videos have been used for bad purposes, so fake videos detection is an important task for identifying faces. Many techniques were applied to detect the frequency of blink [5], movement of the lip, etc., or detect the error of eyes, hair, mouth [7], etc. These methods have certain effectiveness but the fake videos are becoming more complex and harder to differentiate. The mentioned methods which use vital signs, and biological signals may deal with these problems. Intel released a paper: "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals" that proved 96% of accuracy in exposing fake videos [2]. They analyzed Remote Photoplethysmography (rPPG) and Convolutional neural networks as detecting methods in their research. The other paper: "DeepFakes Have No Heart: A Simple rPPG-Based Method to Reveal fake videos" used rPPG and Support Vector Machine (SVM) as a feature to classify, and gained the accuracy result of 96.68% accuracy [1]. rPPG is a sequence of data, and it can be appropriate with Recurrent neural networks, but recent researches only focus on CNN and SVM. Therefore, in this research, we will study about Recurrent neural network and compare it with other models.

2. Related work

Vital signs such as heart rate are important information to identify. There are numerous rPPG methods available for estimating heart rate. In this research, we call the signal which is extracted from the rPPG method is rPPG signal. This signal is used as input for deep learning models such as Convolutional neural networks, Recurrent neural networks, etc.

2.1. Fake videos

There are many types of fake videos and they can be divided into 2 groups:

- Face swapping video: a video that uses the face of person B and connects with the face of person A. Many models or applications is used this technique such as Faceswap, ROOP, etc.
- Optical motion transformation: video uses the motions of person B and converts it into person A. Some preventative models include Thin-plate motion, Revive, MyHeritage, etc. Looking at 1, a right picture is a video that is created by using the motion of a middle video and converts it into a left picture.

This project will target videos using transformation of motion because recently, videos of this type have been used for bad purposes.

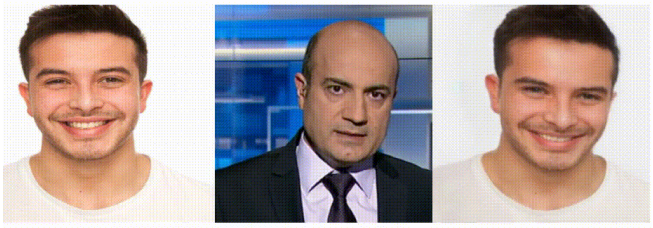

Figure 1. Optical motion transformation video

2.2. Remote Photoplethysmography

Heart rate estimation is a task in health care. When a heart pumps blood into the vessels, it causes changes in blood volume at this time. Therefore, the amount of absorbed and reflected light is different according to the temporal domain. These changes can present as a signal and we can count the number of peaks of the signal or use

the frequency domain to estimate heart rate. With video, Remote Photoplethysmography (rPPG) and Ballistocardiography (BCG) are two popular methods to extract these signals. However, Remote Photoplethysmography is usually used more than Ballistocardiography because BCG is so noisy with head movement. In general, rPPG method measures subtle changes in the intensity of pixels in low frequency (from 0.5 Hz to 3.3 Hz). The rPPG method offers various approaches for handling it, including: GREEN [10] , CHROM [3], POS [11], LGI [8],etc. Each method has different strengths and weaknesses. According to experiments of [4], LGI has the best result in measurement heart rate in the different conditions. In this brief summary, the effectiveness of each method is evaluated.

For all methods explained below, with the input of video, [R, G, B] data can be extracted. In this case, R, G, and B is the red channel, green channel, and blue channel which are calculated mean of all pixel in the region of interest (ROI) and normalized. ROI contains n pixels and is calculated as:

$$R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{1}$$

$$G = \frac{1}{n}\sum_{i=1}^{n} G_i \tag{2}$$

$$B = \frac{1}{n}\sum_{i=1}^{n} B_i \tag{3}$$

where R_i, G_i, and B_i is the intensity of pixel i in red channel, green channel, and blue channel.

2.2.1. CHROM. Considering this model of light to skin: According to Fig. 2, skin receives light from the environment, one part is absorbed and the other part, including specular reflection and diffuse reflection, is reflected into camera. Intensity changes are caused by changes in blood pulse volume. This has relevant to diffuse reflection, so our mission is to record changes in diffuse reflection. Based on color theory, CHROM gave this formula:

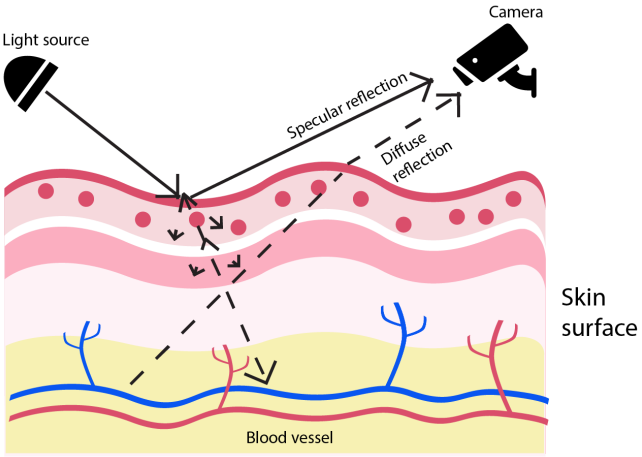

Figure 2. Bioglogical theory of rPPG

$$S = X_f - \alpha Y_f \tag{4}$$

where S is the rPPG signal, X_f, Y_f is a new color space that eliminates the specular reflection component. f stands for bandpass filtered signal. α is standard variation of X_f divided by standard variation of Y_f.

To convert from RGB to a new color space, we can use this formula:

$$X = R - G \tag{5}$$

$$Y = 0.5R + 0.5G - B \tag{6}$$

2.2.2. POS. According POS method, a Plane Orthogonal-to-Skin is used to project RGB into it. This method will remove effect of different skin and illuminance. Formulas of POS method are presented below:

$$S_1(t) = G_n(t) - B_n(t) \tag{7}$$

$$S_2(t) = G_n(t) + B_n(t) - 2R_n(t) \tag{8}$$

$$h(t) = S_1(t) + \alpha.S_2(t) \tag{9}$$

where h(t) is rPPG signal and α is calculated in the same way as in CHROM approach.

2.2.3. LGI. This method aims to eliminate changes caused by movement. To deal with this, they project signal into $P = I - V.V^T$ and V is an eigenvector standing for variation of signal.

$$\overrightarrow{S'} = P.\overrightarrow{S} \tag{10}$$

where S and S' are raw signal and processed signal after projection.

3. Method

This section is about creating a dataset for training and testing model as well as building a model based on theories of rPPG methods and Deep learning approaches.

3.1. Dataset generation

In Vietnam, fake videos have recently been used to swindle. These videos are usually a type of optical motion transformation, so this project surveyed this type of video. Thin-plate motion model [13] is chosen to create videos for training and testing. Real videos are collected from CelebDF [12] and Faceforensics++ [9]. Data on human faces comes from the internet and CelebFaces [6]. The dataset is divided into train-validation-test with the ratio of 60:15:25.

TABLE 1. STRUCTURE OF DATASET

	Train	Validation	Test
Real video	325	139	198
fake videos	427	183	193
Sum	1208	322	391

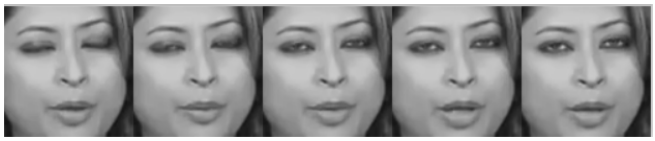

Figure 3. Series of frames containing a face

3.2. Feature selection

Based on the rPPG method, this project hypothesizes that subtle changes in intensity that are created by the rPPG method follow rules with real video. With fake videos, subtle changes are random. Therefore, this can be used to classify real or fake videos. According to [2], signals in the temporal and frequency domain have important information that is effective in classifying fake or real videos. To obtain a signal in the frequency domain, the Discrete Fourier transform (DFT) is applied by a formula:

$$X(k) = \sum_{n=0}^{L-1} x(n) e^{-j\frac{2\pi kn}{N}}, k = 0, ..., N-1 \qquad (11)$$

Although rPPG features are good ones, in experiment, if features of appearance such as the face, the model's accuracy will be improved. Therefore, frames containing faces are also used to feed into CNN for classification.

3.3. Model

This research suggests model using CNN and LSTM models to predict video. The model's diagram is illustrated below.

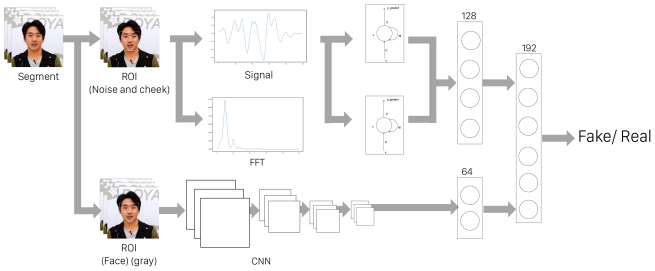

Figure 4. Structure of a model

Looking at Fig.4, the model is divided into 2 branches: one of the branches uses rPPG signal as input for LSTM model; the others one uses frames contraining face as input for CNN model. After that, they extract two vector and concatenate with each other. The model's output uses sigmoid function for probability.

3.4. Sliding window

A video is divided into many segment with length ω. Result from a model is for each segment. To predict the probability of classifying the video, sliding window method

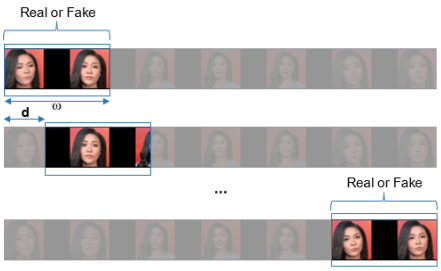

Figure 5. Sliding window for each segment

is used for predicting many segment. The segment preceding will be a segment after d frames of a video.

After having results from segments, a result of the video is indicated based on voting. If the number of fake segments is more than the real segments, this will be a fake videos, and the same vice versa. In case the number of fake segments is equal to the number of real segments, the decision is made based on the mean of probabilities of segments (mean score > 0.5: fake videos and mean score < 0.5: real video)

4. Evaluate

This model is evaluated by using the confusion matrix, F1_score, Precision, and Recall. Besides that,it is also compared with other models for better understanding and evaluation.

4.1. rPPG extraction method experiment

Although LGI is selected as one of the best method to extract rPPG by [4], in this research, brief comparison of efficiency between CHROM, POS and LGI method based on accuracy is also conducted. Looking at TABLE 2, the result

TABLE 2. COMPARISION BETWEEN EACH RPPG EXTRACTION METHOD

Method	CHROM	POS	LGI
Training Acc	73.81%	83.75%	**87.60%**
Validation Acc	78.23%	81.46%	**83.73%**

is the same as [4] and the LGI method has better accuracy than others. That is the reason why it is also chosen as a feature for the deep learning model.

4.2. Feature selection

To evaluate efficiency, the project conducted test with each feature. In this test, the research used the LGI method, and length of a segment is 150. The result is shown in TABLE 3.

As a result, both features (signal and discrete Fourier transform) has the best accuracy on training and validation tests. This is why projects also use these features.

TABLE 3. COMPARISION BETWEEN FEATURES

Feature	Train Acc	Valid Acc
Signal	82.43%	77.82%
DFT	88.34%	80.59%
Both	**87.60%**	**83.73%**

4.3. Size of segment

To select an appropriate segment, experiments on several sizes of segment were conducted.

TABLE 4. COMPARISION OF SIZES OF SEGMENT

ω	Train loss	Train acc	Valid loss	Valid acc
100	0.4420	78.56%	0.4767	78.55%
125	0.3752	82.14%	0.4650	81.47%
150	0.2863	87.60%	0.3932	**83.73%**
200	0.3702	85.03%	0.4420	81.13%

As a result, the best length of segment that produce the best accuracy for 83.73% is 150. This value will be used for sampling and feeding into model.

4.4. Result from training and testing

4.4.1. Training. This model is trained for 77 epochs with configuration:

- Optimizer: Adam
- Learning rate: 10^{-4}
- Batch size: 64
- Weight decay: 10^{-6}

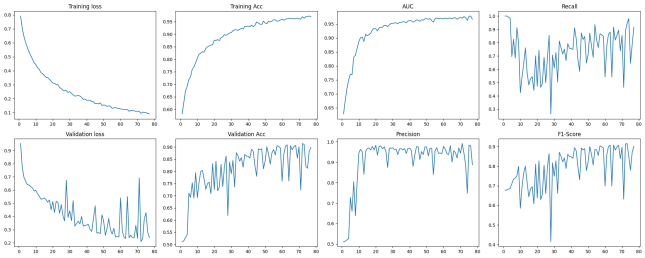

Figure 6. Training process

From Figure 6, it is realized that during the training process, loss function descreased but with epoch 70, loss started increased in validation test. This means overfitting problem is occured. Therefore, early stop method is used to end at the best epoch. The best epoch gains 98.85% on training data an 93.19% on validation data.

4.4.2. Testing. Result from test dataset gave confusion matrix, Recall, Recision and F1_score.

- Accuracy = 92.32%
- Recall = 0.90
- Precision = 0.94
- F1-score = 0.92

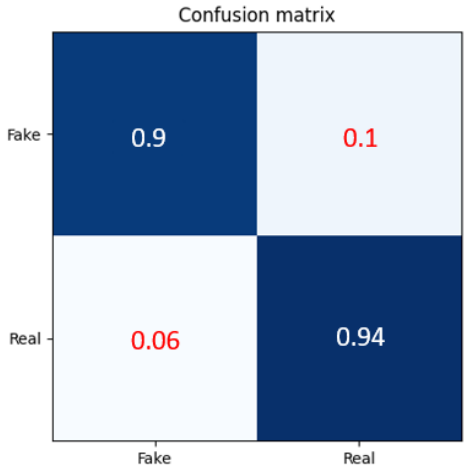

Figure 7. Confusion matrix

The results show that this model can be applied for fake videos detection. Based on the confusion matrix and scores, it is realized that True positive (90%) and True negative (94%) are almost equal. This means that this model has the ability to classify both classes (Real/fake), and an unbalanced data problem didn't occur. The accuracy is 92.32%, which is acceptable, but the model needs to be improved by the addition of a dataset or by giving more features from videos.

4.5. Comparison with CNN, SVM model

Our method is also compared with [2] and [1] are chosen for presentation in CNN and SVM methods.

TABLE 5. COMPARISION WITH OTHER MODELS

Model	Trai loss	Train Acc	Valid loss	Valid Acc
Fakecatcher [2]	0.5211	84.15%	0.5362	84.22%
SVM [1]	_	_	_	75.68%
LSTM+CNN	**0.0445**	**98.85%**	**0.2021**	**93.19%**

On the validation dataset, CNN and SVM methods' operating accuracy is 84.22% and 75.68%; while the proposed method's accuracy has been proven to be up to 93.19

5. Conclusion

The presented results demonstrate that the research's methodology is highly effective. The proposed method outperforms other previous models with the accuracy result of 93.19% on the validation set and 92.32% on the test dataset.. When applied to real-world fake videos, the model performs well compared to models using optical flow-based transformations. In contrast, the face-swapping models exhibit several limitations and require further research.

Acknowledgments

I would like to thank Assoc. Prof. Ha Hoang Kha, Head of the Department of Telecommunications Engineering, who

instructed the implementation of this project, and Assoc. Prof. Do Hong Tuan, Dean of the Faculty of Electrical and Electronics Engineering, who gave criticism and suggestions for the project.

References

[1] Giuseppe Boccignone, Sathya Bursic, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Sabrina Patania. *DeepFakes Have No Heart: A Simple rPPG-Based Method to Reveal Fake Videos*, pages 186–195. 05 2022.

[2] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[3] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[4] Fridolin Haugg, Mohamed Elgendi, and Carlo Menon. Effectiveness of remote ppg construction methods: A preliminary analysis. *Bioengineering*, 9(10), 2022.

[5] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[7] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.

[8] Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1254–1262, 2018.

[9] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 2019.

[10] Wim Verkruysse, Lars Svaasand, and J Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16:21434–45, 12 2008.

[11] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[12] Pu Sun Honggang Qi Yuezun Li, Xin Yang and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[13] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3647–3656, 2022.