**Machine Learning Classification and Feature Selection for Efficient Fake Face**

**Synthesized Video Identification**

A project report submitted to **Indian Institute of Engineering Science and Technology**

in partial fulfillment for the award of the degree of

### Bachelor of Technology in

**Information Technology**

### By

**Arijit Dalui (Enrollment Number- 510819100)**

**Soumik Mukhopadhyay (Enrollment Number- 510819102)**

**Ritaban Bhattacharya (Enrollment Number- 510819101)**

**Under the supervision of Prof. Ruchira Naskar**



### Department of Information Technology

**Indian Institute of Engineering Science and Technology, Shibpur May, 2023**

### DEPARTMENT OF INFORMATION TECHNOLOGY

**INDIAN INSTITUTE OF ENGINEERING SCIENCE AND TECHNOLOGY, SHIBPUR, HOWRAH - 711103, INDIA.**

**CERTIFICATE**

This is to certify that the project report entitled **“Machine Learning Classification and Feature Selection for Efficient Fake Face Synthesized Video Identification”** submitted by Arijit Dalui (Roll No. 510819100), Soumik Mukhopadhyay (Roll No. 510819102), Ritaban Bhattacharya (Roll No. 510819101) to Indian Institute of Engineering Science and Technology towards partial fulfillment of requirements for the award of the degree of Bachelor of Technology in Information Technology is a record of Bonafede work carried out by them under my supervision. This dissertation, in my opinion, is worthy of consideration for the purpose for which it is submitted and it fulfills the requirements of the regulations of this Institute. The results incorporated in this dissertation are original and have not been submitted to any University or Institute for the award of any Degree or Diploma.

Dr. Ruchira Nasar Professor

Department of Information Technology

Indian Institute of Engineering Science and Technology, Shibpur

Dr. Prasun Ghosal

Head of Department

Department of Information Technology

Indian Institute of Engineering Science and Technology, Shibpur

**ACKNOWLEDGEMENT**

We are grateful that we had the opportunity to experience an academic journey at IIEST, Shibpur that lasted four years. These four years have been an eye-opening experience for us and we consider ourselves fortunate to have learnt life-changing lessons which shall always be useful for us.

We are indebted to Professor Ruchira Naskar of the Department of Information Technology, IIEST, Shibpur for guiding us through the minute aspects of research in the domain of surface generation and plane detection. We were able to explore some interesting problems and implement certain niche concepts under his supervision.

We are also grateful to all the faculties of the Department of Information Technology, IIEST, Shibpur. The knowledge imparted through the courses as part of the B. Tech programme will always be useful in preparing us for future technological challenges.

We are also grateful to Arnab Kumar Das, PhD scholar of the Department of Information Technology, IIEST, Shibpur, for his valuable guidance and precious inputs that were crucial to the development of the project in its entirety.

Finally, we would like to thank all our friends and fellow students for enriching our experience in IIEST, Shibpur.

**Arijit Dalui (Examination Roll No.): 510819100**

**Ritaban Bhattacharya (Examination Roll No.): 510819101**

**Soumik Mukhopadhyay (Examination Roll No.): 510819102**

# Table of contents

# *Abstract*

# *Introduction*

# *Related Works*

# *Dataset*

# *Approach*

# *DCT-FADE Approach*

# *Multi-stage Multi-modal video identification*

# *Feature Extraction*

# *Classification*

# *Conclusion*

# *References*

# Abstract

Recent developments in deep learning have enabled media synthesis and manipulation to reach previously unheard-of degrees of realism. The widespread use of deepfake technology to produce fake media has the potential to have a large negative influence on the reliability of multimedia data, including videos, photos, and audio recordings. The frequency domain spectrum of deepfake videos is examined in this paper, and a unique method is suggested that uses high-frequency Discrete Cosine Transform (DCT) coefficients that are retrieved from videos as a recognizable fingerprint. We create a robust model that can identify between genuine videos and their deepfakes by examining the variance of differences between subsequent video frames' DCT coefficients. We extend our work by not only using a binary classification model to classify real and fake videos but also, we go beyond classification in our research and look into the sources of bogus videos. We offer a thorough study of deepfake media by integrating our feature-based classification model with a look into the origin of the fake videos. Our comprehension of deepfake techniques and their possible impact on multimedia integrity is improved by this all-encompassing approach. We ran tests on the publicly accessible dataset Face Forensics++ to determine how well our suggested model worked. Amazingly, our model produced excellent results, with a multilevel classification accuracy of 98.56% and a binary classification accuracy of 98.91%, respectively.

**Introduction**

In 2017, the Deepfake phenomenon emerged notably as a consequence of artificial intelligence (AI) methods for creating synthetic media and their distribution on the internet. These methods included tweaking (or generating) audiovisual content using ad-hoc machine learning generative models, such as the Generative Adversarial Network (GAN) [1]. On television as well as internet, we found videos and images of high-profile individuals that could initially seem real, but could really be the output of an AI process that produces incredibly accurate fake media. Such fraudulent tapes are prone to be misused in personal defamation lawsuits, crimes involving child pornography, and deceiving court proceedings and the general public. The horrific consequences of deepfakes were witnessed when a Reddit user of the same name posted doctored porn clips on the site [2]. In the clips, renowned individuals like Gal Gadot, Taylor Swift, Scarlett Johansson, and others had their visages switched with those of porn performers. Recent reports claim that deepfake films were employed as a political tool during elections. In order to stop the dissemination of false information in the run-up to the 2020 US elections, Facebook banned Deepfake and its synthetic videos [3]. The usage of deepfake video in a news story as opposed to actual footage demonstrated how this affects the behavior of the general public, according to Shin et al. [4]. The abusive, misleading, and fraudulent use of this technology has increased hazards rather than possibilities for stakeholders. We can invariably see the necessity for an immediate and trustworthy solution to confront the Deepfake technology, considering that anybody could be its next target.

In this paper, an explainable method of Deepfake detection, a DCT-FADE (DCT Frequency-based Authentication for Deepfake Evaluation) technique is proposed based on the analysis of Discrete Cosine Transform (DCT) [5] video coefficients. It was demonstrated through experiments using Deepfake frames of human faces that the spectral frequencies

contain an effective generative process signature. The main contributions of this research are the followings:

* A novel approach for detecting fake faces is developed using the Discrete Cosine Transform, which produces excellent generalization results.
* This “explainable” technique allows us to identify deepfake-generated anomalous frequencies that significantly deviate from the frequencies of an authentic image.

In addition to that we present a machine-learning classification model for the blind

detection of Deepfake videos. For this purpose, we propose a multimodal detection

technique by combining an efficient set of prefabricated Histogram of

Oriented Gradients (HOG) based features, and a set of features automatically

learned by Convolutional Neural Networks (CNNs). Our contributions in this

paper may be summarized as follows:

* Development of a multimodal Deep-Fake detection scheme combining both deep learning and traditional computer vision techniques, hence optimizing performance efficiency.
* We propose a multi-stage classification approach, for identifying a Deepfake video, followed by identifying its source network using the above set of features.

# Related Works

# Approach

# DCT-FADE Approach

The Discrete Cosine Transform (DCT) is applied to an N × N square matrix of pixel values to generate an N × N square matrix of frequency coefficients. In our case, N is set to 8 because using larger blocks would increase the time required to perform DCT calculations, resulting in an unreasonable tradeoff between compression and processing time. Therefore, the image is typically divided into more manageable 8 × 8 blocks before applying DCT.

A DCT filter was further applied to those both real and fake frames. The filter yielded fruitful results, as the DCT image matrix for a fake and real frame was easily distinguishable by the naked eye as shown in Fig. 1.



Fig. 1. DCT spectrum analysis. (A) DCT frequency spectrum for pristine test

video. (B) DCT frequency spectrum for deepfake video.

The DCT matrix, which is 128×128, and the high-frequency terms which reside in the top-left corner of the DCT matrix are considered and made into an 8×8 matrix. These high frequency

values are taken since they are the most determining factors. The flattened values of the high-frequency matrix were graphically plotted for the real frame and four corresponding fake frames as shown in Fig. 2.



Fig. 2. (A) Real vs. fake video fingerprints (B) Zoomed fingerprints.

The general trend noticed was that the deepfake curve deviated the most from the real curve, hence having the most difference.

The variance of real and fake data was plotted in Fig. 4. Fig. 4 (B) represents a zoomed portion of Fig. 4 (A). The variance of real and fake data with different thresholds shown in Table I are very promising. The threshold helps to normalize the variance Var(X) by not taking into account the fewer high-frequency outliers The variances as extracted from real and fake data respectively were multiplied by Real × 105 are tabulated, which clearly indicates the fake videos are having higher variance values than real videos.

**Table-I**

**Change in variance with different threshold values**

|  |  |  |
| --- | --- | --- |
| **Threshold** | **Real\*(10^5)** | **Fake\*(10^5)** |
| -0.001 to 0.1 | 1.210 | 1.240 |
| -0.001 to 0.101 | 1.220 | 1.250 |
| -0.001 to 0.103 | 1.210 | 1.251 |
| -0.001 to 0.110 | 1.221 | 1.252 |
| -0.003 to 0.110 | 1.550 | 1.580 |
| -0.009 to 0.105 | 2.170 | 2.130 |
| -0.005 to 0.1 | 1.810 | 1.850 |

The sample real and fake video frames from which the data was taken is shown in fig 3.

ATTACH REAL FAKE VIDEO FRAMES

.

Fig. 4. (A) Differences per frame (B) Zoomed differences.

# Multi-stage Multi-modal Video Identification

# 

# Fig. 5: Workflow of the proposed method having 4 stages. (A) Preprocessing: Extracts frames from video and resize them into 64×128 pixels. (B) Feature Extraction: Extract features from the frames using HOG and multiple CNN models. (C) Classification-1: A binary classifier to distinguish between a real and a fake image. (D) Classification-2: A Multi-class classifier to identify Deepfakes [10], Face2Face [13], Face Swap [12], and Neural Textures [11] generated images.

# Feature Extraction

# As evident from Fig. 5, we use a feature extraction block to create a feature map from video. In this work, we combine both deep-learned features as well

# as conventional prefabricated image features for deepfake detection. We use the

# Histogram of Oriented Gradients (HOG) [14] feature descriptors, along with

# four different Convolutional Neural Network (CNN) based feature extractors for

# this purpose. The CNN architectures explored in this work are: ResNet-50[15],

# ResNet-101[16], VGG-16, and VGG-19[17]. Our aim in this work is to investigate

# the effectiveness of both prefabricated image features (as the HOG) and deep

# CNN-learned features in deepfake detection. HOG features have been prevalently

# used in the literature for determining the distribution of image gradient orientations

# in focused areas of an image [14]. HOG features are capable of identifying

# patterns and texture information in an image, by computing the magnitude and

# orientation of gradients of the pixels.

# The next mode of feature extraction in this work is the features learned in

# an automated way by deep CNNs. The ResNet architecture involves a number

# of residual blocks, each of which has a batch normalization layer, a Rectifier

# Linear Unit (ReLU) [18] activation function, and one or more convolutional layers.

# ResNet50 [15] and ResNet101 [16] are 50- and 101-layer deep convolutional neural network architectures respectively. Deep convolutional neural network

# architectures VGG-16 and VGG-19 [17], combine a deep stack of convolutional

# layers with tiny filters. These models can capture both local and global features

# in an image, including edges, textures, and object forms, making them useful

# for feature extraction in synthesized frames. The above CNNs are used here as

# feature extractors by removing the last classification layer and passing test video

# frames through the network, to obtain a fixed-length feature vector. This feature

# vector is later used in our classification step, to distinguish between fake

# and authentic videos. We select only 1.49%, 5.97% of ResNet and VGG generated features,

# and 100% of the HOG features. We generate a feature map containing

# 3780 features per video frame using HOG. The retrieved features from ResNet

# and VGG were substantially more numerous and may have had several outliers.

# In order to extract principle features in each of these extracted feature maps, we

# reduced the dimensionality using PCA [24] to 1500, while keeping computational

# power constraints and outliers in mind. PCA maximizes the variance along each

# axis, or principle component, by transforming the data into a new coordinate

# system. To ascertain which is best, the resulting feature maps were examined

# using both binary and multi-class classifiers.

# Classification

# The classification module is responsible for performing a binary classification

# between fake and real videos, as well as for detecting the source of a deepfake

# video by performing a four-way classification. The proposed classification module

# uses five different classifiers - Polynomial Kernel SVM [19], Random Forest [20],

# Decision Tree [21], K Nearest Neighbor (KNN) [22], Logistic Regression [23].

# For classification, we use the feature maps generated by the CNNs presented in

# Section 2.1 as well as the HOG features. Once the model determines a frame to be

# fake, it undergoes multi-way classification to determine whether it is Deepfake

# [10], Face2Face [13], FaceSwap [12] or NeuralTextures [11]. Our experimental

# results with related discussions are presented next.

# The classification results for both stages, binary and four-way, are presented in

# Table 1 and Table 2, respectively. All results presented are outcomes of a 15-

# fold cross-validation. The best results were seen obtained for binary fake vs. real

# video classification with the Polynomial SVM classifier on HOG and Resnet-101

# extracted features, with an accuracy of 98.91% and F1-score of 98.12 and 98.13

# respectively. For source identification, that is four-way classification the best

# results were obtained with the Polynomial SVM classifier on HOG extracted

# features, with an accuracy of 98.56% and F1-score of 97.18.

# Table 1 and 2