Turkish Journal of Science & Technology
17(2), 435-450, 2022

Research Paper
https://doi.org/10.55525/tjst.1142626

Active Face Spoof Detection Using Image Distortion Analysis

Betul AY[1*], Peter ANTHONY[2]
[1,2] Department of Computer Engineering, Faculty of Engineering, Firat University, Elazig, Turkiye
[*1] betulay@firat.edu.tr, [2] pettony1@gmail.com

(Geliş/Received: 08/07/2022; Kabul/Accepted: 15/08/2022)

Abstract: With the rising use of facial recognition systems in a range of real-world scenarios and applications, attackers are also increasing their efforts, with a number of spoofing techniques emerging. As a result, developing a reliable spoof detection mechanism is critical. Active-based techniques have been shown to be good at finding spoofs, but they have a number of problems, such as being intrusive, expensive, hard to compute, not being able to be used in many situations, and usually needing extra hardware. This research presented an active-based robust spoof detection technique capable of detecting a wide range of media or 2D attacks while being less intrusive, less expensive, low in complexity, and more generalizable than other active-based techniques. It doesn't require any additional hardware, so it can easily be integrated into current systems. The distortion variations of video frames of the user's face collected at varying distances from the camera are analyzed to detect spoofing. Both the legitimate and spoof attack datasets were created using real-world facial photo and video data. The proposed approach achieved a spoof detection accuracy of 98.18% using both machine learning classifiers and a deep learning model, with an equal error rate and a half total error rate as low as 0.023 and 0.021, respectively.

Key words: Face Anti-spoofing, Face Spoof Detection, Face Recognition, Biometrics, Image Distortion Analysis.

Yüz Tanıma Sistemleri İçin Kararlı Aktif Tabanlı Yüz Sahteciliği Tespiti

Öz: Bir dizi gerçek dünya senaryosu ve uygulamasında yüz tanıma sistemlerinin artan kullanımıyla birlikte, yeni ortaya çıkan farklı sahtecilik teknikleri kullanaraks saldırganlar da çabalarını artırmaktadır. Bunun sonucu olarak, güvenilir bir sahtecilik tespit mekanizması geliştirmek kritik öneme sahiptir. Aktif tabanlı tekniklerin sahtekarlıkları bulmada iyi olduğu gösterilmiş olsa da bunların, müdahaleci, pahalı, hesaplanması zor olmaları, birçok durumda kullanılamamaları ve genellikle ekstra donanıma ihtiyaç duymaları gibi bir takım sorunları vardır. Bu çalışma, daha az müdahaleci, düşük maliyetli, karmaşıklığı düşük ve diğer aktif tabanlı tekniklerden daha genelleştirilebilen, çok çeşitli medya veya 2D saldırıları tespit edebilen aktif tabanlı sağlam bir sahtekarlık algılama tekniği sunmaktadır. Önerilen yöntem herhangi bir ek donanım gerektirmez, bu nedenle mevcut sistemlere kolayca entegre edilebilir. Kullanıcının yüzünün kamera yardımıyla farklı mesafelerden elde edilen video karelerindeki bozulma değişimleri, sahtekarlığı tespit etmek için analiz edilmiştir. Hem gerçek hem de sahte saldırı veri kümeleri, gerçek dünyadan yüz fotoğrafı ve video verileri kullanılarak oluşturulmuştur. Önerilen yaklaşım, hem makine öğrenimi sınıflandırıcıları hem de derin öğrenme modeli kullanılarak, sırasıyla 0,023 ve 0,021 kadar düşük bir hata oranı ve yarı toplam hata oranıyla %98,18'lik bir sahtekarlık algılama doğruluğu elde etmiştir.

Anahtar kelimeler: Yüz Sahteciliğine Karşı Koruma, Yüz Sahteciliği Algılama, Yüz Tanıma, Biyometri, Görüntü Bozulma Analizi.

1. Introduction

Face recognition is one of the most widely used biometric systems today, and its use has increased dramatically and rapidly over the last decade. Facial recognition, after fingerprints, is the second most frequently used biometric technology in the world, according to the International Biometric Group (IBG) [1]**.** This rising popularity is attributed to the fact that it is more convenient, contactless, and non-invasive than other biometric systems like fingerprint and iris [2,3]. Face recognition is also a biometric system that has a wider functioning range and application potential than others [4]. Unfortunately, the majority of existing facial recognition systems are vulnerable to spoof attacks.

A spoofing attack occurs when a phony piece of evidence is provided to a biometric system for authentication, resulting in false approval [5–8]. Face spoofing is the act of masking a person's face in order to get around a biometric system and acquire unauthorized access or privileges. Face photographs or videos of a person can now be easily obtained through social media or shot from afar without the individual's consent [9]. This is accomplished by simply presenting or replaying illegally obtained photographs or videos. As efforts to combat spoofing operations have increased, spoofing artifacts have morphed into a variety of methods.

[*] Corresponding author: betulay@firat.edu.tr. ORCID Number of authors: [1] 0000-0002-3060-0432, [2] 0000-0002-9010-3075

Face anti-spoofing, also known as face spoof detection, is the challenge of preventing fraudulent face verification. Despite the fact that the problem of face spoof detection has been around for over a decade [10], it did not undergo a genuine revolution until around the year 2010 under the European project TABULA RASA, which was focused on biometric spoof threats [11]. Another element that has contributed to significant progress in creating approaches for identifying face spoofing attacks is the compilation and sharing of publicly available face spoof datasets. This has relieved researchers of the burden of data collection, allowing them to focus on developing viable ways to counter various types of attacks [12–15].

These important factors have resulted in the release of numerous strategies for preventing both 2D and 3D face spoofing attacks. For spoof detection, researchers have proposed a number of approaches, ranging from the use of motion cues [16–18] to the current deep learning approaches [19,20] with competitive performance. Face spoof detection techniques can be divided into two categories: active and passive. The mode of enrollment, or the approach utilized to acquire the user's data, differentiates between the two types. Active techniques have proven to be quite reliable, with good anti-spoofing capabilities. Active techniques, on the other hand, involve user participation, which makes them intrusive and limits their applicability. Some active techniques, such as multi-modal [21,22] and sensor-level approaches [18], might be computationally challenging and costly due to the additional hardware required. A semi-active strategy is proposed here, with less intrusiveness and interaction, a wider application range, no additional hardware requirements, and hence a lower cost.

The possibility of image distortion caused by the proximity of the face to the camera in combating face spoof attempts is investigated in this paper. The decision to use image distortion analysis (IDA) for spoof detection is based on three findings:

- The typical photographic phenomena of face distortion in video when the camera is very close to the face, which is caused by the uneven 3D-surface of the human face, as contrast to the flat surface of 2D materials such as prints and screens.
- Image distortion has had a lot of success in the field of biometrics and security systems for finger print spoof detection [23], steganalysis [24,25], and facial liveness detection [2,26].
- When the camera gets closer to an object, that object grows in size faster than things in the background. A scatter plot of measured pixel intensity versus distance to the camera revealed an exponential falloff in intensity with distance from the camera in a study by Bryson et al [27]. This is due to the fact that the item and the background image create an uneven 3D surface. 2D attacks have flat surfaces, so when filmed by the camera, a different observation is predicted.

This study aims at employing image distortion for developing a robust active face spoof detection technique that is less intrusive, less expensive, and more generic than current active-based approaches and requires no additional hardware.

2. Related Literature

The utilization of image distortion features as a descriptor for face spoof detection has been broadly adopted. Authors have devised a number of methodologies based on image distortion analysis (IDA) and found encouraging results.

Image distortion analysis feature vector was constructed in [2] by combining four (4) separate distortion features, including chromatic moment, color diversity, specular reflectance features, and blurriness features, to create a strategy against printed photo and replay attacks. For differentiating real and fake faces, an ensemble classifier made up of multiple support vector machines (SVM) was used to learn these characteristics. The same features were retrieved by the authors in [26] using the four IDA features from the MSU-MFSD dataset, which include printed photo and replay attacks. Training using SVM and an artificial neural network yielded 94.4% and 88.9% accuracy, respectively. Even though they used a non-intrusive approach, their spoof detection performance is not reliable enough for real-world use.

There have also been other active-based techniques proposed. The authors of [28] generated NIR differential (NIRD) images using an active near-infrared (NIR) light source. To achieve spoof detection, they used two separate features based on NIRD images. The pixel consistency between facial and non-facial regions was studied, and context signals were used to distinguish faked faces from real ones. Although their method proved effective in detecting spoofs, it is limited in terms of applicability because it involves the use of additional hardware and it is also quite intrusive. The authors of [29] proposed a multi-spectral imaging approach that used both visible (VIS) and near-infrared (NIR) spectra for imaging. The authors in [30] integrate face and voice. A user is required to present his/her voice and also speak. A specialized piece of hardware was utilized in [16] to capture pupil orientation as a clue for liveness assessment. In an approach in [1], to ascertain liveness, the user is asked to open and close his mouth or blink his eyelids. All of these active-based approaches suffer from some or all of the issues that active approaches encounter, such as the need for additional hardware, high intrusiveness, high cost, computational complexity, and reduced ability to fit into existing systems, among others.

Distinctively, the approach proposed in this research work computes the distortion changes between a user's face captured at two different distances to deliver a robust active spoof detection technique, yet less intrusive and less expensive. The authors in [31] propose a method that is similar to the one given in this paper, but their method requires eight (8) face frames at varied distances from the camera for a single sample, and the feature is represented by a 7 by 2147 matrix. The method provided in this paper, on the other hand, requires just two frames to be represented in a vector space. Furthermore, the technique does not require additional hardware, making it easy to integrate into existing systems.

3. Methodology

3.1 System Model Design

The system architecture in this work is organized into three core modules, as shown in Figure 1: the Frames Selection Module, the Distortion Feature Extraction Module, and the Classification Module. Face videos are fed into the frame selection module, and two frames are chosen from each of two facial videos captured at varying distances from the camera. The Feature extraction module then extracts a number of facial landmarks from each frame and computes features about how the distortion of the face evolves over time. Finally, the classification module uses the obtained features to perform binary classification to discriminate between fake and real faces.

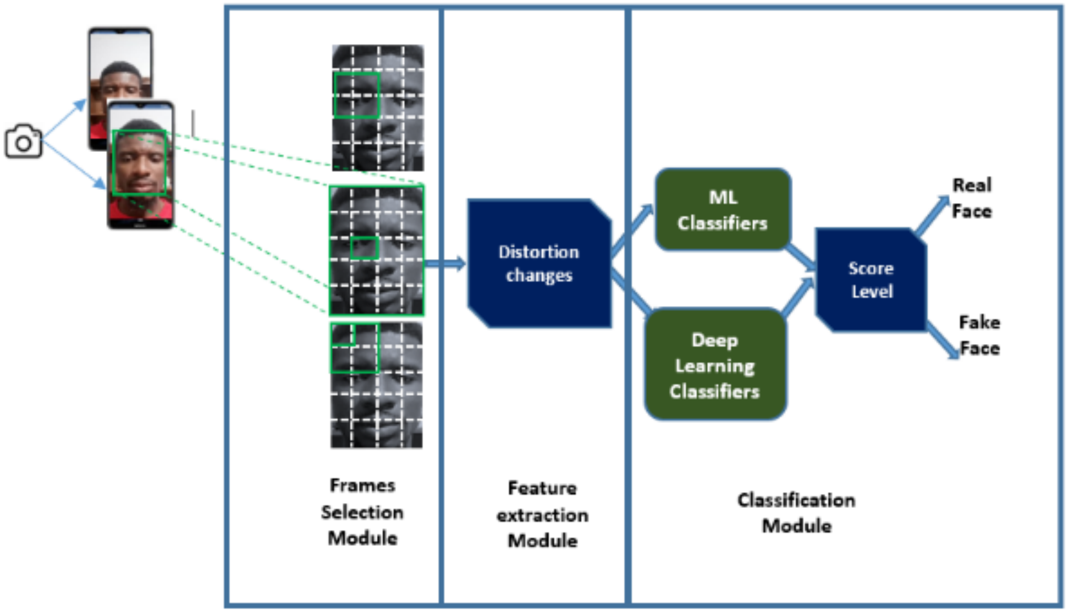


Figure 0. Workflow design for the proposed approach.

3.1.1 Frame selection module

At first, as the user's face approaches or moves away from the mobile device, the device's camera captures two video clips at two different specified distances between the camera and the user's face, referred to as D1 and D2. A number of frames of the user's face are contained in each video. Using Dlib face detector [32], the frame selection module then detects and extracts a K-sequence of frames (D1f1, D1f2, …, D1fk) and (D2f1, D2f2, …, D2fk) from the video frames captured at distances D1 and D2, respectively. The frames are then paired D1fi and D2fi, for i=(1,2,…, k.).

3.1.2 Feature extraction module

Using the extracted frames D1fi and D2fi as input, the feature extraction module detects a number of facial landmarks and calculates the geometric distances (d) between the different facial landmarks in each frame, as well as the relative distance (r) between the two frames (D1fi and D2fi) which form the feature vector for detecting distortion changes in the face frames. From each video frame, 68 facial landmarks are detected using the Dlib 68 facial landmark predictor [33]. As shown in Figure 2, the 68 facial landmarks are spread out over the face and include:

- The chin region with 17 points
- The eyebrows region with 10 points, 5 points for each eyebrow
- The nose stem region with 4 points
- Below nose with 5 points
- The eyes region with 12 point, 6 point for each eye
- And the lips or mouth region with 20 points,

The 68 face landmarks are denoted as ($p1$, $p2$, ..., $p68$) where pi = (xi, yi) is the coordinate.

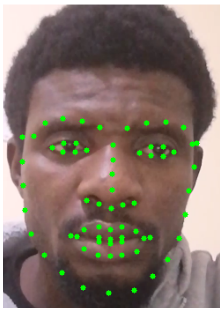


Figure 2. A face with detected landmarks using Dlib 68 point facial landmarks detector.

The facial distortion is captured for each paired frame D1fi and D2fi, i=(1,2,…, k.) by first computing the geometric distance between any two facial landmarks ps and pt for each frame using the formula below:

$$d = \sqrt{(xs - xt)^2\ (ys - yt)^2} \tag{1}$$

Where $s, t \in \{1, 2, ..., 68\}$ and $s \neq t$.

Each frame's 68 facial landmarks resulted in 2278 pairwise distances $d1$, $d2$, ..., $d2278$. The geometric vector (geo) = ($d1$, $d2$, ..., $d2278$, w, h) is formed by concatenating the 2278 pairwise distances with the width (w) and height (h) of the detected face. The relative distance between the paired frames D1fi and D2fi is then calculated using the formula: r = ($r1$, $r2$, ..., $r2278$, rw, rh), where:

$$r_j = \frac{d_{D1,i}}{d_{D2,i}} \tag{2}$$

for i = 1, 2, ..., 2278.

$$\mathrm{rw} = \frac{\mathrm{w_{D1,}}}{\mathrm{w_{D2,}}} \tag{3}$$

$$\mathrm{rh} = \frac{\mathrm{h_{D1,}}}{\mathrm{h_{D2,}}} \tag{4}$$

Therefore, each pairwise selected frame D$_1$fi and D$_2$fi is represented by a 1 × 2280 feature vector (FV).

3.1.3 Classification module

Finally, in the classification module uses classification algorithms to determine whether the feature vector (FV) was taken from a genuine face or a spoof attack. For comparison, four different models, including machine learning models (Linear Discriminant Analysis, Support Vector Machine, and K-Nearest Neighbor) and deep learning models (Convolutional Neural Network) were employed. Details of the models are discussed below:

Support Vector Machines (SVM): An SVM is a supervised learning-based classifier with a generalized model for conducting binary classifications. Nonlinear classification tasks can also be completed using the kernel approach. SVM is frequently used to solve face recognition problems due to its exceptional sparsity and robustness properties [34]. The maximum margin hyperplane of the learning samples' solution is the Support Vector Machine's decision boundary. SVM also uses hinge loss functions to calculate empirical risks and add regularization terms to the solution system in order to maximize structural risks. Face spoof detection is typically thought of as a binary classification task, and SVMs are classifiers with a lot of promise for handling it. It is important to note that this study used a hand-crafted feature-based approach to get a very large feature size, and Support Vector Machines are very good at learning from datasets with high dimensions.

Linear Discriminant Analysis (LDA): Linear discriminant analysis (LDA) has been widely used in face recognition as a supervised approach [35], as well as in face presentation attack detection [36]. LDA can be used to reduce dimensionality and classify data. The goal of LDA is to identify an appropriate projection in the projective feature space that maximizes the between-class scatter matrix while minimizing the within-class scatter matrix. In the past, image data was frequently used directly as a classification input, but when dealing with high-dimensional face data, linear discriminant analysis frequently suffers from a limited sample size problem. It is assumed that LDA is a good fit because the hand-crafted feature-based technique used in this work produced large, high-dimensional features.

K-Nearest Neighbor (KNN): Among machine learning classifiers, the KNN classifier is thought to have the simplest approach. The authors of [37] used Local Binary Pattern descriptors to approach the spoof detection problem; analysis of their results suggests that KNN fared better in terms of accuracy and execution time. In a similar spirit, the authors in [38] used the Eigen vector approach to extract characteristics. Their findings also reveal that KNN outperforms SVM in terms of accuracy and execution time. For comparison, KNN is also used in this study.

Convolutional Neural Network (CNN): Deep learning can be used to generate promising results in the field of computer vision, and it has been shown to be useful in solving the challenge of face spoof detection [39].

Table 1. Summary of the CNN architecture employed

Layer No	Layer type	Activation Function	Value
1	Convolution (Conv1D)	Relu	Unit=32, Kernel size=3
2	MaxPooling1D	None	Stride= 2
3	Dropout	None	0.4
4	Flatten	None	None
5	Dense	Relu	Unit=1024
6	Dropout	None	0.4
7	Dense	Sigmoid	Unit=2

The CNN architecture employed in this research is as follows: To extract various features from input images, a 1-dimensional convolutional layer (Conv1D) with 32 units as the input layer and an input dimension of (2280 x1) is used. To learn and approximate the relationship between network variables, the kernel of the input layer is set to 3 using a Relu activation function. Then, in order to reduce computational costs, a Maxpooling stride 2 is used to reduce the size of the feature map obtained from the Convolutional layer. The third layer has 1024 units and is fully connected. Finally, the output layer is a fully-connected layer with two units and a sigmoid function (due to the binary classification). To combat overfitting, two Dropout layers with a 0.4 value were added. The first fully-connected layer comes after the pooling layer, and the second comes after the pooling layer. Just before the fully connected layer, a flatten layer is added. The model is tuned at a learning rate of 0.001 and decay of 1e-6 using the Adam optimizer. In Table 1, the model's summary is provided.

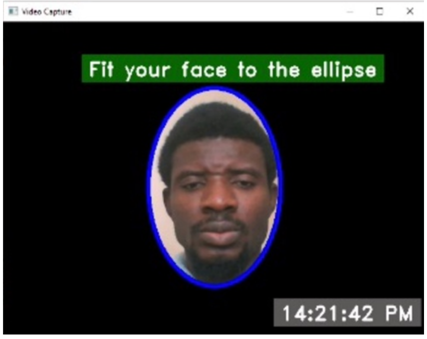
3.2 Data Collection and Dataset Generation

This section explains the complete data gathering procedure as well as the methodologies used to generate the datasets. The genuine photo dataset, the printed-photo attack dataset, the screen-photo attack dataset, and the replay attack dataset are the four datasets that will be discussed.

3.2.1 Data collection

Eighty (80) volunteers were used to collect data for this study, with 56 men and 22 women making up the group. Volunteers range in age from 15 to 40 years of age.

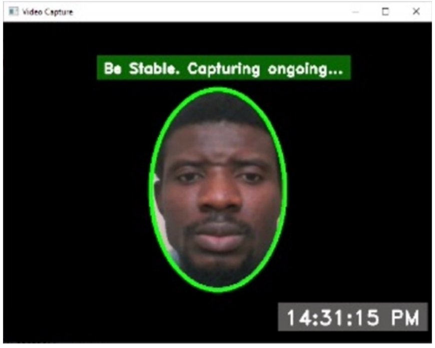
We collect selfie facial recordings from participants in two distinct device positions, with the face covering the middle of the image. Each participant is asked to capture two 1080HD frontal face video clips at 24fps (frames per second) selfies with his or her phone held at controlled distances D1 and D2 from his or her face. The video clips at each distance are 3 seconds long, with each frame measuring 1920 x 1080 pixels and the face filling the middle of the frame. The capturing at two predefined distances (20 cm and 50 cm) from the device camera are aided by a simple interface program built in python (see Figure 3, 4 and 5). Initially, the participant holds the phone and tries to fit his face into the elliptical shape on the interface (see Figure 3). Once the face is fitted to the elliptical shape (implying a 50 cm distance between the face and the camera), the program automatically takes 3 seconds of facial video frames and adjusts the settings for the next distance, as illustrated in Figure 4. and Figure 5. The elliptical shape increases automatically to capture the face at a closer range (20 cm). Each participant provides a set of two facial video data (at D1 = 20 cm and D2 = 50 cm).



Fit your face to the ellipse

14:21:42 PM

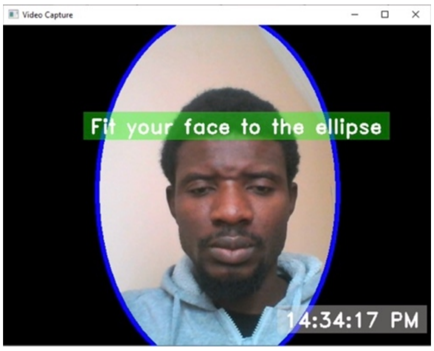
Figure 3. Interface with ellipse set to capture face at 50cm away from the camera.



Be Stable. Capturing ongoing...

14:31:15 PM

Figure 4. Interface capturing detected face at 50 cm away from the camera.



Fit your face to the ellipse

14:34:17 PM

Figure 5. Interface with ellipse set to capture face at 20 cm away from the camera

Image samples from the data obtained are shown in Figure 6. Face photos on the left were obtained at a distance of 20 cm between the face and the camera, whereas the photos on the right were acquired at a distance of 50 cm between the face and the camera.

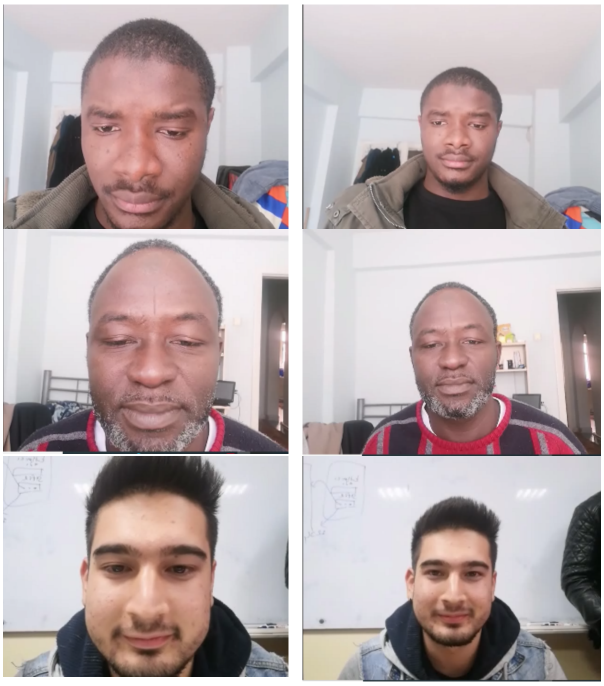


Figure 6.Sample face images from the dataset

3.2.2 Dataset generation

In this study, four different dataset classes were generated, including real and fake cases: The legitimate dataset, the printed-photo-attack dataset, the screen-photo-attack dataset, and the replay-attack dataset were created from the facial videos and photos collected.

Legitimate Dataset: The genuine dataset comprises of facial videos collected while using a mobile phone with the help of the capture control interface program from a distance of D1 = 20 cm to a distance of D2 = 50 cm. As a result, the genuine dataset contains eighty (80) videos for D1 = 20 cm and eighty (80) videos for D2 = 50 cm. That's a total of 80 paired videos.

Printed-Photo Attack Dataset: For the printed photo attack, one facial frame is manually extracted from each participant's frontal facial video clips taken at distance of 50cm from the camera. The images captured at 50 cm were chosen because of the obvious distortion that would be visible on the face image at 20 cm, and the majority of people will not disclose their facial photos/videos shot at such a close range as 20 cm. The extracted facial frames are printed on photo quality paper using a high-quality industrial Direct Impression printer. This makes the size of the face similar to that of a real face. Finally, using the mobile phone and the capture control interface program, each of the printed images is recaptured for three seconds at 24 frames per second, at distances of D1 = 20 and D2 = 50 cm. There are 80 paired videos in total in the printed-photo attack dataset (80 for D1 = 20 cm and 80 for D2 = 50 for each photo).

Screen photo-attack dataset: For the screen-photo attack dataset, the manually extracted facial frames from frontal facial video clips captured at distance D= 50cm between the camera and face, are displayed to the controlled capture system as described in section 5.2.1 using a mobile device. the system records 2 videos for each image at controlled distance D1=20 and D2=50cm, each for 3 seconds at 24fps. In total, the screen-photo attack dataset consists of 80 pairwise videos (that is 80 for D1=20cm and 80 for D2=50 for each photo).

Replay-Attack Dataset. We used frontal facial video clips collected from each participant at a fixed distance of 50 cm between the camera and their face for the replay attack dataset. Each video clip is played in loop mode using a mobile device before being presented to the controlled capture system, which consists of a mobile phone and the interface program. The replay device is moved closer or farther away from the camera until the facial region fits within the ellipse, and then the video is recorded. For each replayed video, the system records two facial video clips at a controlled distance of D1 = 20 cm and D2 = 50 cm, each for 3 seconds at 24 frames per second. Hence, a total of 80 pairwise videos (that is, 80 for D1 = 20 cm and 80 for D2 = 50 for each photo) are generated for the replay-attack dataset.

3.3 Evaluation metrics

Kanika and Jaspreet [40] outlined five characteristics that are typically utilized in face spoof detection evaluation: FRR (false rejection rate), FAR (false acceptance rate), EER (equal error rate), HTER (half total error rate), and accuracy. These parameters aid in identifying serious threats posed by a spoofing database to any face recognition system. These metrics are calculated using following results:

- TP (True Positive): The number of attacks that have been identified as attacks.
- TN (True Negative): The number of authentic samples that have been identified correctly as genuine.
- FP (False Positive): The number of legitimate samples that are incorrectly identified as fake.
- FN (False Negative): The number of attacks that have been identified as genuine samples.

3.3.1 False Acceptance Rate (FAR)

The false acceptance rate, or FAR, is a measure of the tendency that a biometric security system may accept an unauthorized user's access attempt wrongly. It's simply the percentage of times an unauthorized person's face is accepted erroneously. The FAR of a system is commonly calculated by dividing the number of false acceptances by the number of identification tries. The following is the FAR formula:

$FAR = FP/(FP + TN).$

3.3.2 False Rejection Rate (FRR)

The false rejection rate is a measure of a biometric system's likelihood of wrongfully rejecting or refusing access to a legitimate user. It's just the percentage of authorized users whose identification is wrongly rejected. The FRR of a system is commonly calculated by dividing the number of erroneous rejections by the number of identification tries. The formula is as follows: $FRR = FN/(TP + FN)$

3.3.3 Equal Error Rate (EER)

A lower false acceptance rate (FAR) will lead to a higher false rejection rate (FRR), and vice versa. The Equal Error Rate refers to the point where the lines meet (EER). The EER is the point of equality in FAR and FRR, as shown in Figure 3.8. Equal Error Rate refers to an algorithmic method of error margin in which false rejections and false acceptances are equalized. This rate is the result's common value, or common ground. The smaller the error margin, the more precise the biometric system is. In terms of misleading data, it's just a mathematical means of scoring out errors and error margins.

EER is very important for figuring out how accurate data is and comparing the results of two systems

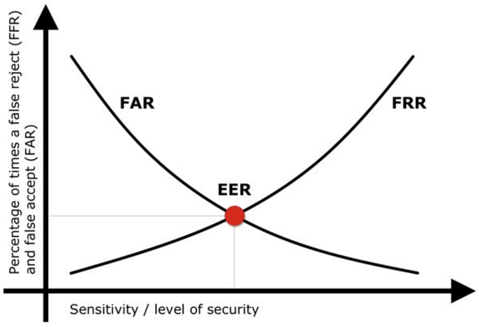


Figure 7. Illustration of EER derived from the plot of FAR vs FRR

If the false acceptance rate is reduced to the bare minimum, the false rejection rate is likely to increase. In other words, the more secure the access control system is, the less convenient it will be because users will be wrongly rejected by it. As a result, choosing between FAR and FRR is a personal decision that will result in either a more secure (but less user-friendly) or less secure system (but more user-friendly). There are few systems on the market that can provide a high level of security while still providing user-friendly access management. If a company doesn't have a system like this, user comfort usually comes before security [41].

3.3.4 Half Total Error Rate (HTER)

The half total error rate is an error rate that is equal to the sum of the FAR and FRR error rates. The HTER formula is as follows: HTER = (FAR + FRR)/2.

HTERs or other variations of HTERs are used in most biometric systems for measurement and comparison. It should be noted that in most benchmark databases used in biometric system literature, there is a significant imbalance between the number of genuine and fake accesses. It is most likely due to the fact that obtaining the former is more expensive than obtaining the latter. This imbalance explains why HTER, rather than the usual classification error used in machine learning literatures, is used for model comparison [42].

3.3.5 Accuracy

The ratio of correctly predicted samples to total predictions is used to determine the accuracy of a biometric system. It is calculated as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN).

It's worth noting that the lower the EER or HTER, the better the detection result.

4. Results And Discussion

4.1 Experimental Setup

To ascertain the authenticity of a face, the frame selection module takes K-frames from each paired facial video collected at distances D1 = 20 cm and D2 = 50 cm as mentioned in section 3.2. The feature extraction module then extracts the features, generating a vector of 1 x 2280 for each paired frame D1fi and D2fi, for i= (1,2,..., K), K is set to 30 for attack samples in this investigation, while K is set to 60 for genuine samples. This is to avoid a situation where the classes are unbalanced. As a result, 30 samples are taken from each pairwise video of attack instances, resulting in 30x80 = 2400 samples for each spoof case, for a total of 2400x3 = 7200 (printed-photo, screen-photo, and replay attack) samples for the attack dataset. The legitimate dataset also generates 60x80 = 4800 samples for real videos. There are a total of 12,000 samples, which makes a matrix of 12000x2280 for real and fake data samples.

Three experiments were carried out to verify the effectiveness of our approach. First, the full dataset of 12000 samples was trained using four distinct models (SVM, LDA, K-NN, and CNN models) in a 3:1 ratio for training and testing (9000 and 3000, respectively). In section 4.2, the findings are provided.

Secondly, the datasets for each spoof case are supplied to the trained models for classification in order to gain insight into their performance based on different spoof examples. The figures in section 4.3 indicate the results.

Finally, each model is saved and a simple program is created to evaluate all of the models in real time using the saved models and the video capture control system described in section 2.2. (see Figure 5,6 and 7). The models were presented with live faces as well as all of the spoof cases.

4.2 Results

We trained four models with the dataset. SVM, LDA, K-NN, and CNN. Figure 8 provides accuracy results for these models for three types of spoof attacks: printed photo, screen photo and replay video attacks.

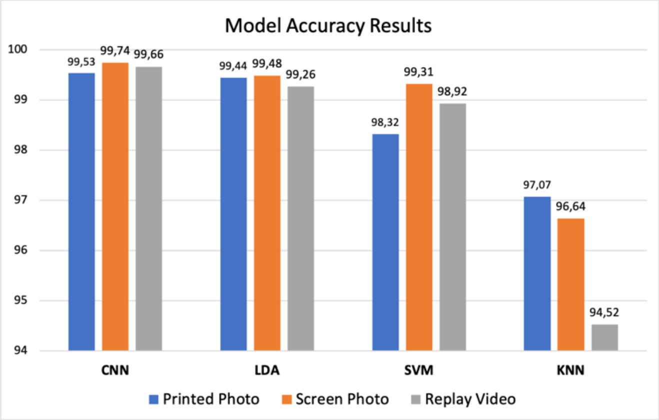


Figure 8. Model Accuracy Results

4.2.1 SVM Classifier Results

The SVM classifier is trained on the main dataset, which contains 12000 samples, 4800 of which are valid and 7200 of which are spoof cases. The SVM classifier had an accuracy of 96.07 percent, as shown in Table 4.1. This demonstrates the model's overall ability to correctly detect both real and fake faces. It also has 0.053 and 0.031 FAR and FRR, respectively. With a false acceptance rate of 0.053, the model is expected to wrongly accept or provide access to 53 unauthorized users in a sample of 1000. Similarly, a false rejection rate of 0.031 means that the model can correctly accept 969 authentic faces out of 1000, while failing in only 31 situations. This also demonstrates that the SVM model's capacity to accurately recognize real faces is slightly superior to its ability to correctly identify spoof cases. The model also achieved an EER and HTER of 0.034 and 0.042, respectively, which are good indicators of performance because of the lower the EER and HTER, the better.

In the second experiment, the SVM model is put to the test on individual fake cases to see how well it does. As shown in Figure 8, the results show that it is better at finding screen-photo threats than printed-photo and replay attacks.

4.2.2 Results from LDA Classifier

The accuracy of 98.03%, FAR and FRR of 0.022 and 0.018, respectively, revealed by linear discriminant analysis, demonstrates an excellent performance across all parameters. This indicates that just 22 illegal individuals are likely to be mistakenly accepted or granted access in a sample of 1000. Similarly, a false rejection rate of 0.018 means that the model is likely to reject or refuse access to as few as 18 authorized users in a sample of 1000 authentic faces. EER and HTER values of 0.014 and 0.020, respectively, show how robust the model is (See Table 2).

Figure 8 reveals that Linear Discriminant Analysis performs somewhat better in detecting screen photo attacks, followed by printed-photo attacks, and least in replay attacks, with an accuracy of 99.44 percent, 99.48 percent, and 99.26 percent, respectively, in the unit testing.

4.2.4 Results from KNN Classifier

With an accuracy of 92.90%, FAR and FRR of 0.082 and 0.062, respectively, the K-NN classifier has the worst performance (See Table 2). When given a sample of 1000 real and spoof faces, a false prediction of 82 and 63 is expected for each. The low performance could be due to the dataset's high dimensionality. This correlates to the fact that, while K nearest neighbors (KNN) is one of the simplest nonparametric classifiers, its accuracy is impacted by nuisance features in high-dimensional settings [43].

In the unit testing phase, the K-NN classifier, in contrast to other classifiers that were more robust in detecting screen-photo attacks, had its best performance in printed-photo attacks with 97.07%, which is slightly better than its performance in screen-photo and replay attacks, with 96.64 % and 94.52 % accuracy, respectively (See Figure 8).

4.2.4 Results from CNN Model

The CNN model is trained with the architecture described in Table 3.1, with the learning rate set to 0.001, weight decay set to 1e-6, and the maximum iteration set to 100. The CNN model performed the best, as shown in Table 2, with an accuracy of 98.18% and FAR and FRR of 0.036 and 0.007, respectively. An FRR of 0.007 indicates that the model has strong recall for real samples. The CNN model also got EER and HTER values of 0.023 and 0.021, respectively, which shows very good performance.

In the unit testing phase, the CNN model performed better in the screen-photo attack than in the other two attacks, as shown in Figure 8, with an accuracy of 99.74%, 99.66%, and 99.53% for screen-photo attack, replay attack, and printed-photo attacks, respectively.

4.2.5 Validation on Real-Life Test Scenario

Based on each test case, the results of the real-time test cases using the saved trained models and the capture control system is presented. Figures 9 is the output of live test on a live face. Whereas, Figure 10, 11 and 12 are the output of the live tests using the spoof cases.

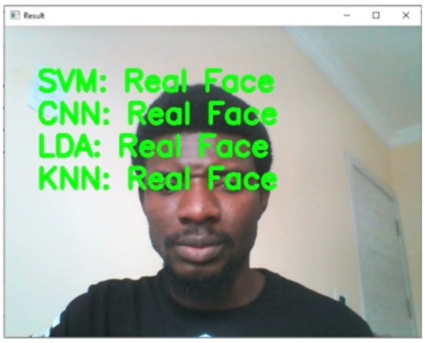


Figure 9. Result of the models on live face.

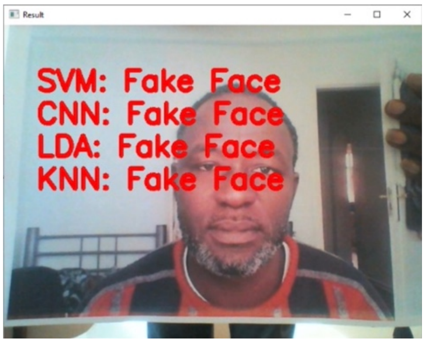


Figure 10. Result of the models on printed-photo attack

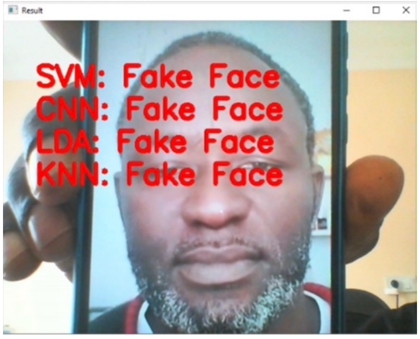


Figure 11. Result of the models on screen-photo attack

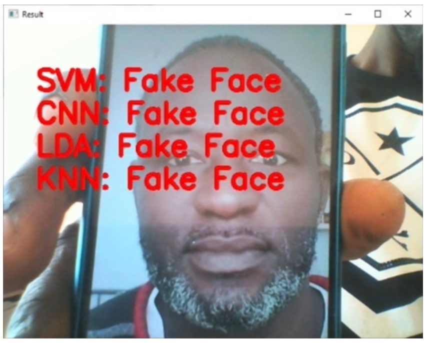


Figure 12. Result of the models on replay attack

4.3 Comparison of The Model Performances

Table 2 provides a comparison of the performances of all models across all five metrics for evaluation. Values in bold indicate the best performance on each metric.

Table 2. Comparison of the model performances across five metrics

Model	Accuracy	FAR	FRR	EER	HTER
SVM	96.07%	0.053	0.031	0.034	0.042
LDA	98.03%	**0.022**	0.018	**0.014**	**0.020**
KNN	92.90%	0.082	0.063	0.053	0.073
CNN	**98.18%**	0.036	**0.007**	0.023	0.021
Best Performance	**98.18%**	**0.022**	**0.007**	**0.014**	**0.020**

The saved models were also evaluated against time. For prediction, each model is provided an input of 2400 data samples. Figure 13 depicts their results. The LDA model is incredibly fast, with the best performance in terms of time spent by predicting the 2400 samples in about 490 milliseconds. The CNN model has the longest execution time of 10.901 seconds. SVM and KNN performance in terms of execution time are nearly the same, with 3.712 seconds and 3.869 seconds for SVM and LDA, respectively.

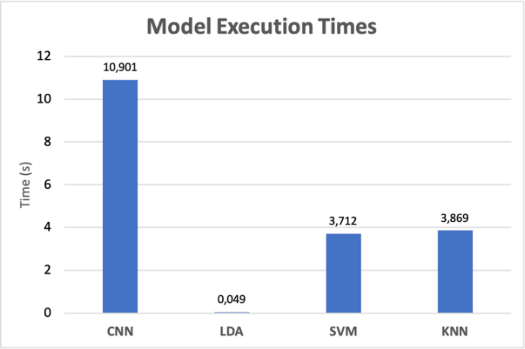


Figure 13. Results of the models performances against time taken for prediction

4.4 Discussion

The results reported in this work demonstrated that the strategy proposed in this study is resilient and thus practical, with all models above 92 percent accuracy. The CNN model and the LDA classifier are at the top of the list. Based on the performance evaluation of the four models over the five metrics, the CNN model outperformed the others in two of the measures, with accuracy and false rejection rate (FRR) of 98.18 percent and 0.007, respectively. That is, given a set of data samples, the CNN model will yield a greater accuracy rate, and the likelihood of blocking access or wrongly rejecting an authorized user is extremely low, as demonstrated by an FRR of 0.007. demonstrating that only 7 erroneous rejections are expected in a sample of 1000. The linear discriminant analysis, on the other hand, surpassed the others in terms of three other measures, FAR, EER, and HTER, with values of 0.022, 0.014, and 0.020, respectively. Although the CNN has a very low FRR, it has a much larger FAR, which is why it is defeated by the LDA in EER and HTER values. A lower EER suggests a healthy FAR/FRR margin.

In contrast to the results provided in [37]and [38], where KNN performed better than other classifiers, KNN had the lowest performance across all measures in this study. This is consistent with the authors' belief in [43] that in the presence of a high dimensional feature space, KNN performance is hampered by nuisance features.

CNN performed the worst in the execution time test, with 10.901 seconds of execution time over 2400 samples predicted. LDA performed the best, with an execution time of 0.049 seconds (490 milliseconds) for the

2400 samples. The rapid and good performance of LDA is attributable to the reduction in dimensionality, which is not the case with other models.

Table 3. Comparison of other approaches with the proposed approach

Technique	Category	Strength	Limitation	Performance

Fusion of face and Voice [30]	Active	• Very robust
		• Can detect 3D attacks
			• Highly Intrusive
			• Extra Hardware requirement
			• Low application range
			• Expensive
			• Computational complexity
				FAR=0.017
				FRR= 0.011
				FTC= 0.031
				EER= 0.01

Pupil tracking using led sensor[16]	Active	• Robust against print attacks	• Intrusive • Helpless in cut-photo and replay attacks • Extra Hardware requirement • Cannot easily fit into existing system •Time consuming	-
DoG[13]	Passive	• Fast response time • Non-Intrusive	• Poor generalization (Vulnerable to variation in acquisition)	EER=0.17

Ensemble of Texture descriptors (LBP+GDP+GLTP+ LDiP+LGBPHS+ LPQ)	Passive	• Non-Intrusive s	• Computational complexity
			• Poor generalization
			• Limited to 2D attacks
				RR= 0.98.39
				FRR= 0.49

3D sensor [18]	Passive	• Robust • Non-intrusive	• Time complexity • Require extra hardware • Helpless in 3D attacks • Cannot fit into existing systems easily	Accuracy=100%
Eye-blinking and Mouth movement[1]	Acvtive	• Robust against • No extra hardware required • Less expensive	• Instrusive • Weak in detecting cut-photo and replay attacks • Slow response time	-
Face and Finger print[21]	Active	• Very robust • Captures 3D attacks	• Highly Intrusive • Extra Hardware requirement • Low generalization ability • Expensive	EER=0.012
Facial distortion changes	Active	• Robust • No extra hardware requirement • Less expensive	• Intrusive • Computational complexity	Accuracy= 99.48%
Temporal and depth information	Passive	• Robust • Non-intrusive	• Computational Complexity • Require extra hardware • Helpless in 3D attacks • Cannot fit into existing systems easily	Intra DB: ACER= 0.73 on SiW; ACER= 1.3 on OULU-NPU Cross DB: HTER = 17.5 on Replay Attack DB; HTER = 24.0 on CASIA-MFSD

Proposed approach	Active	• Can easily fit into existing system
		• Less intrusive
		• No extra hardware requirement
		• Less expensive
		• Robust
		• Less complexity
			• Intrusive
			• Limited to 2D attacks
				Accuracy=98.18%
				FAR=0.036
				FRR=0.007
				EER=0.023
				HTER=0.021

A useful question to pose is, "In this scenario, which model should be picked between CNN and LDA?" The truth is that it is determined by the aim and priority of the system to be built or deployed. Some system objectives may emphasize FAR while others may prioritize FRR. As a result, the choice of FAR or FRR is a question of preference, resulting in either a more secure (but less user-friendly) or less secure (but more user-friendly) system [41]. Consider the following scenario: you want to avoid people queuing at the entry since the system is not working properly (false rejection). In such instances, users may accept that convenience takes precedence. However, when users assume a high level of security, the situation is reversed. However, when it comes to biometric systems, a low equal error rate should be taken into account.

In comparison to the efforts in [31], which are closer to the one proposed in this study, while they achieved slightly higher accuracy, a single sample is represented by a 7 × 2147 matrix as a feature set taken from 8 frames at different distances in their technique. In contrast, the approach suggested in this work requires only two frames recorded at 20 cm and 50 cm from the camera to create the feature vector for a single sample. As a result, their approach becomes more computationally complex and time-intensive.

Table 4.6 further clearly shows the benefits of the proposed approach over existing active-based approaches. In [30] and [21], where face recognition is paired with voice or fingerprint recognition, a user is additionally asked to speak or present his or her fingerprint, which is collected using an audio device or fingerprint reader. Although they are relatively resistant to all types of face spoof attacks, they are very intrusive and require additional hardware such as a fingerprint scanner, which is also costly. Whereas with the suggested solution, no additional hardware is required, and it is less intrusive because a user will only need to move his face closer or farther, which is usual for users before any facial recognition system to do even without instruction. A user is asked to blink his/her eye and open and close his/her mouth in the effort presented in [1], which does not require extra hardware and tends to fit into existing systems. Unfortunately, it is quite intrusive and extremely vulnerable to cut-photo or replay attacks, whereas the approach provided here has achieved 99.66 percent accuracy against replay-attacks in the unit testing scenario. Furthermore, every user confronted with such a system is aware that the liveness test is based on blinking of the eye and movement of the mouth, and as such, it suggests the next point of action to the impostor, as opposed to the method proposed in this study, where the liveness cue is not obvious to the user.

Table 3 provides a comparison of the proposed approach to other approaches while highlighting their strength and limitations.

5. Conclusion

This study is inspired by the common phenomenon in photography of distortion of faces in video where the camera and the face are in close proximity, which is caused by the uneven 3D surface of the human face, and proposed a face spoof detection approach that uses the distortion changes of video frames of user's faces captured at different distances from the camera to deliver a robust active-based spoof detection while being less intrusive, less expensive, and with more application range.

An analytical experiment was carried out and assessed across five metrics to verify the effectiveness of the approach: accuracy, FAR, FRR, EER, and HTER. The results demonstrate excellent and competitive performance, with the best performances in CNN and LDA models, with accuracy = 98.18%, EER = 0.023, HTER = 0.021 and accuracy = 98.03%, EER = 0.022, HTER = 0.020, respectively. Hence, it is obvious that the approach proposed in this study is practical, and the handcrafted features are discriminative enough against 2D attacks and can be investigated further against other types of attacks, such as 3D attacks. It is also worth noting that the proposed approach has various advantages over alternative active-based approaches, such as reduced intrusiveness, no additional hardware required, fast response time, lower computing complexity, and ease of integration into existing systems.

Finally, despite the variety of threats, spoof detection tasks are typically viewed and approached as binary classification problems. As such, it is recommended that future research handle spoof detection in terms of multi-class classification, taking into account the various types of attacks. Feedback on the most prevalent types of attacks faced by the facial recognition system can thus be collected, analyzed, and used for further research and development. Cases of previously unknown or emerging threats can also be tracked.

References

[1] A. Singh, P. Joshi, G.C. Nandi, Face recognition with liveness detection using eye and mouth movement, 2014 Int. Conf. Signal Propag. Comput. Technol. (ICSPCT 2014). (2014) 592–597.

[2] D. Wen, H. Han, A.K. Jain, Face Spoof Detection With Image Distortion Analysis, IEEE Trans. Inf. Forensics Secur. 10 (2015) 746–761. https://doi.org/10.1109/TIFS.2015.2400395.

[3] L. Sun, G. Pan, Z. Wu, S. Lao, Blinking-Based Live Face Detection Using Conditional Random Fields, in: S.-W. Lee, S.Z. Li (Eds.), Adv. Biometrics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007: pp. 252–260.

[4] C.N. Karson, Spontaneous eye-blink rates and dopaminergic systems., Brain. 106 (Pt 3) (1983) 643–653.

https://wwww.unboundmedicine.com/medline/citation/6640274/Spontaneous_eye_blink_rates_and_dopaminergic_systems_.

[5] S. Kumar, S. Singh, J. Kumar, A comparative study on face spoofing attacks, in: 2017 Int. Conf. Comput. Commun. Autom., 2017: pp. 1104–1108. https://doi.org/10.1109/CCAA.2017.8229961.

[6] J. Galbally, S. Marcel, J. Fierrez, Biometric Antispoofing Methods: A Survey in Face Recognition, IEEE Access. 2 (2014) 1530–1552. https://doi.org/10.1109/ACCESS.2014.2381273.

[7] D. Menotti, G. Chiachia, A. Pinto, W.R. Schwartz, H. Pedrini, A.X. Falcão, A. Rocha, Deep Representations for Iris, Face, and Fingerprint Spoofing Detection, IEEE Trans. Inf. Forensics Secur. 10 (2015) 864–879. https://doi.org/10.1109/TIFS.2015.2398817.

[8] A. Pinto, W. Schwartz, H. Pedrini, A. Rocha, Using Visual Rhythms for Detecting Video-Based Facial Spoof Attacks, Inf. Forensics Secur. IEEE Trans. 10 (2015) 1025–1038. https://doi.org/10.1109/TIFS.2015.2395139.

[9] J. Yang, Z. Lei, D. Yi, S.Z. Li, Person-Specific Face Antispoofing With Subject Domain Adaptation, IEEE Trans. Inf. Forensics Secur. 10 (2015) 797–809. https://doi.org/10.1109/TIFS.2015.2403306.

[10] I. Pavlidis, P. Symosek, The imaging issue in an automatic face/disguise detection system, in: Proc. IEEE Work. Comput. Vis. Beyond Visible Spectr. Methods Appl. (Cat. No.PR00640), 2000: pp. 15–24. https://doi.org/10.1109/CVBVS.2000.855246.

[11] R. TABULA, "Trusted biometrics under spoofing attacks," Http://Www.Tabularasa-Euproject.Org/. (n.d.).

[12] I. Chingovska, A. Anjos, S. Marcel, On the effectiveness of local binary patterns in face anti-spoofing, 2012 BIOSIG - Proc. Int. Conf. Biometrics Spec. Interes. Gr. (2012) 1–7.

[13] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, S. Li, A face antispoofing database with diverse attacks, 2012 5th IAPR Int. Conf. Biometrics. (2012) 26–31.

[14] X. Tan, Y. Li, J. Liu, L. Jiang, Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model, in: K. Daniilidis, P. Maragos, N. Paragios (Eds.), Comput. Vis. -- ECCV 2010, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010: pp. 504–517.

[15] N. Erdogmus, S. Marcel, Spoofing in 2D Face Recognition with 3D Masks and Anti-spoofing with Kinect, in: Biometrics Theory, Appl. Syst., 2013.

[16] M. Killioğlu, M. Taşkiran, N. Kahraman, Anti-spoofing in face recognition with liveness detection using pupil tracking, in: 2017 IEEE 15th Int. Symp. Appl. Mach. Intell. Informatics, 2017: pp. 87–92. https://doi.org/10.1109/SAMI.2017.7880281.

[17] T. Dhawanpatil, B. Joglekar, A Review Spoof Face Recognition Using LBP Descriptor, in: A.K. Somani, S. Srivastava, A. Mundra, S. Rawat (Eds.), Proc. First Int. Conf. Smart Syst. Innov. Comput., Springer Singapore, Singapore, 2018: pp. 661–668. https://doi.org/https://doi.org/10.1007/978-981-10-5828-8_63.

[18] G. Albakri, S. Alghowinem, The Effectiveness of Depth Data in Liveness Face Authentication Using 3D Sensor Cameras@, Sensors (Basel). 19 (2019).

[19] Y. Ma, L. Wu, Z. Li, F. liu, A novel face presentation attack detection scheme based on multi-regional convolutional neural networks, Pattern Recognit. Lett. 131 (2020) 261–267.

[20] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T.C.-H. Cheung, K.-W. Cheung, Integration of image quality and motion cues for face anti-spoofing: A neural network approach, J. Vis. Commun. Image Represent. 38 (2016) 451–460. https://doi.org/https://doi.org/10.1016/j.jvcir.2016.03.019.

[21] P. Wild, P. Radu, L. Chen, J. Ferryman, Robust multimodal face and fingerprint fusion in the presence of spoofing attacks, Pattern Recognit. 50 (2016) 17–25. https://doi.org/https://doi.org/10.1016/j.patcog.2015.08.007.

[22] B. Geng, C. Lang, J. Xing, S. Feng, W. Jun, MFAD: A Multi-modality Face Anti-spoofing Dataset, in: 2019: pp. 214–225. https://doi.org/10.1007/978-3-030-29911-8_17.

[23] T. Chugh, A.K. Jain, Fingerprint Spoof Detection: Temporal Analysis of Image Sequence, CoRR. abs/1912.0 (2019). http://arxiv.org/abs/1912.08240.

[24] V. Holub, J. Fridrich, Digital Image Steganography Using Universal Distortion, in: IH MMSec 2013 - Proc. 2013 ACM Inf. Hiding Multimed. Secur. Work., 2013. https://doi.org/10.1145/2482513.2482514.

[25] J. Cheng, A.C. Kot, S. Rahardja, Steganalysis of Binary Cartoon Image using Distortion Measure, in: 2007 IEEE Int. Conf. Acoust. Speech Signal Process. - ICASSP '07, 2007: pp. II-261-II–264. https://doi.org/10.1109/ICASSP.2007.366222.

[26] P.P.D. Raval, R.R. Sedamkar, S. Kulkarni, Face Spoofing Detection Using Image Distortion Features, in: 2017.

[27] M. Bryson, M. Johnson-Roberson, O. Pizarro, S. Williams, Colour-Consistent Structure-from-Motion Models using Underwater Imagery, in: 2012. https://doi.org/10.15607/RSS.2012.VIII.005.

[28] X. Sun, L. Huang, C. Liu, Context based face spoofing detection using active near-infrared images, in: 2016: pp. 4262–4267. https://doi.org/10.1109/ICPR.2016.7900303.

[29] Mohamed, Shaimaa, Ghoneim, Amr, Youssif, Aliaa, Visible/Infrared face spoofing detection using texture descriptors, MATEC Web Conf. 292 (2019) 4006. https://doi.org/10.1051/matecconf/201929204006.

[30] B.R. Naidu, P.V.G.D. Reddy, Fusion of face and voice for a multimodal biometric recognition system, Int. J. Eng. Adv. Technol. 8 (2019) 506–515.

[31] Y. Li, Z. Wang, Y. Li, R. Deng, B. Chen, W. Meng, H. Li, A Closer Look Tells More: A Facial Distortion Based Liveness Detection for Face Authentication, in: Proc. 2019 ACM Asia Conf. Comput. Commun. Secur., Association for Computing Machinery, New York, NY, USA, 2019: pp. 241–246. https://doi.org/10.1145/3321705.3329850.

[32] Dlib Python API Tutorials [Electronic resource] – Access mode: http://dlib.net/python/index.html, (n.d.). http://dlib.net/python/index.html.

[33] D.E. King, Dlib-ml: A Machine Learning Toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[34] E. Osuna, R. Freund, F. Girosit, Training support vector machines: an application to face detection, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 1997: pp. 130–136.

[35] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognit. 34 (2001) 2067–2070.

[36] S. Bharadwaj, T.I. Dhamecha, M. Vatsa, R. Singh, Computationally efficient face spoofing detection with motion magnification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work., 2013: pp. 105–110.

[37] C. Priyanka, Sharma; Neha, Spoofing Face Detection using LBP Descriptor and KNN Classifier in Image Processing, Int. J. Recent Technol. Eng. Volume-8 (n.d.).

[38] K. Samrity, Saini; Kiranpreet, KNN Classification for the Face Spoof Detection, Int. J. Sci. Eng. Res. Volume 10 (n.d.) 1101–1106.

[39] Y. Du, T. Qiao, M. Xu, N. Zheng, Towards Face Presentation Attack Detection Based on Residual Color Texture Representation, Secur. Commun. Networks. 2021 (2021) 6652727. https://doi.org/10.1155/2021/6652727.

[40] Kanika kalihal ; Jaspreet Kaur, A Review on Different Face Spoof Detection Techniques in Biometric Systems, Int. J. Sci. Res. Eng. Trends. Volume 5 (n.d.).

[41] RecogTech, FAR and FRR: security level versus user convenience, Https://Www.Recogtech.Com/En/Knowledge-Base/Security-Level-versus-User-Convenience, (Retrieved on 31st/01/2022). (n.d.) (Retrieved on 31st/01/2022).

[42] S. Bengio, J. Mariéthoz, A Statistical Significance Test for Person Authentication, Speak. Lang. Recognit. Work. (2004).

[43] H. Raeisi Shahraki, S. Pourahmad, N. Zare, K Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data, Biomed Res. Int. 2017 (2017) 7560807. https://doi.org/10.1155/2017/7560807.