Panoptic Defenses for Secure Computer Vision

by

Akshay Agarwal

Under the supervision of

Dr. Richa Singh

Dr. Mayank Vatsa

Indraprastha Institute of Information Technology Delhi
October, 2020

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Panoptic Defenses for Secure Computer Vision

by

Akshay Agarwal

Submitted
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

to the
Indraprastha Institute of Information Technology Delhi
October, 2020

Certificate

This is to certify that the thesis titled "**Panoptic Defenses for Secure Computer Vision**" being submitted by **Akshay Agarwal** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

October, 2020
Dr. Richa Singh RSingh
Professor

October, 2020
Dr. Mayank Vatsa
Professor

Indraprastha Institute of Information Technology Delhi
New Delhi

Acknowledgment

First and foremost, I would like to thank my supervisors, Prof. Richa Singh and Prof. Mayank Vatsa, for their continuous support and motivation/deadlines. I can not even imagine this journey is coming to an end, which I think would have been impossible without the guidance they have provided. I would also like to thank them for the parties at the end of every semester or sometimes even in the middle of the semesters, which use to refresh the mood and keep the momentum going.

I want to thank Dr. Noore for hosting me at West Virginia University and enriching my professional and personal experience. I would also like to thank Dr. Nalini Ratha for his knowledge-dense guidance during my research in sparse learning. I acknowledge my seniors in the IAB Lab who have been gracious in sharing their knowledge and experience with me: Dr. Anush, Dr. Samarth, Dr. Tejas, and Dr. Gaurav. I must also express my gratitude to all of my co-authors who have made conducting research with them fun and exciting: Rohit, Soumyadeep, Naman, Daksha, Akhil, Anannya, Suril, Puspita, Shruti, Maneet, and Saheb. I want to acknowledge my labmates with whom at times, I might have disagreements, but in the end, without them, this roller coaster journey would have been challenging.

I want to thank Priti ma'am for accommodating and helping with all of my administrative requests. I want to thank IIIT-Delhi and the Government of India for Visvesvaraya, Ph.D. Fellowship by providing financial support during my Ph.D. and funding my research. I would also like to thank my parents, who supported me with confidence, patience, and encouragement at every step of the way. To conclude this section, I would like to dedicate this dissertation to everyone mentioned above, especially my parents, who are waiting for a long time to see *'doctor'* in front of my name. I want to thank everyone mentioned above for making this journey a memorable one.

Panoptic Defenses for Secure Computer Vision

by

Akshay Agarwal

Abstract

As the deployment and usage of computer vision systems increase, protecting these systems from malicious data has also become a critical task. The primary source of information in any computer vision system is the input data, and authenticity of the data is integral to the reliability of a system. With advancements in electronic equipment, especially communication mediums such as mobile phones and laptops, digital data acquisition has become an easy task. Such huge enablements of the cameras and digital contents have raised severe concerns such as capturing unauthorized biometrics data, video voyeurism, and sexting. Apart from that, in the case of person recognition, it is generally seen that when the testing image is captured using the different sensor/camera, the performance significantly drops. In other vital scenarios, digital images are used as evidence in the court of law and criminal investigation. While the image source might be authentic, the image itself might be a spoof or corrupted in a way to fool the machine learning algorithms. The attacks on computer vision algorithms have become advanced enough to trick the machine learning systems and deceive human visual systems. Therefore, proper authentication of digital images and videos is necessary. While many of these challenges of computer vision systems are dealt with individually, this dissertation provides a *'panoptic'* view to address the challenges ranging from image source identification to the classification of anomalies, using machine learning algorithms. This dissertation focuses on detecting and mitigating the spectrum of attacks on the data level. The four major contributions are (i) sensor identification to ascertain that the image is captured from an authenticated device, (ii) detecting digital attacks, (iii) detecting physical attacks, and (iv) detecting adversarial attacks.

In the case of large human identification projects such as India's Aadhaar project and Integrated Automatic Fingerprint Identification System (IAFIS) of the FBI, a variety of acquisitions devices are used. While it is important to ensure that the images are captured from authenticated devices only, the images captured from these different devices vary significantly in terms of the quality, texture, and illumination, which makes the matching of these images also a challenging task. As the first contribution, we have proposed a camera source identification algorithm and a novel feature selection algorithm to identify the biometric image sensor used for acquisition. The proposed algorithm yields more than 99% classification accuracy on several databases with images captured using multiple cameras. We have also prepared and released two multi-sensor iris databases to promote research on this problem.

The next two problems we have addressed in this dissertation are presentation attacks on face recognition systems, through physical presentation attacks and digital attacks such as morphing. A variety of presentation attack instruments have been used, starting from the simple print and replay attacks, to more sophisticated mediums such as silicone masks, latex masks, or wax faces. The

proposed presentation attack detection algorithm utilizes a combination of wavelet decomposition and texture feature extraction with support vector machine classifier to distinguish between real and attacked faces. The proposed algorithm outperforms state-of-the-art algorithms, including classifiers based on hand-crafted image features and deep CNN features under several generalized settings, including multiple spectrum. We have also prepared a multi-spectral (i.e., visible, near-infrared, and thermal) face presentation attack database. It is one of the largest publicly available databases in the physical presentation attack domain.

The second contribution focuses on detecting digital manipulations such as morphing and swapping. Morphing is the technique to blend two or more faces to create one morphed image, which can be used to create a duplicate identity, and two individuals can get authorized access using the same identity. We first prepared a large scale database using multiple images collected from multiple mediums such as mobile applications and internet websites. We propose a novel feature extraction algorithm to detect the digital alterations that can encode the artifacts developed due to morphing or swapping. The proposed feature extraction algorithm first filter the image patches and encodes the irregularities as a difference in those local regions. We have observed that because of the sophisticated digital alteration tools, these differences are minute. Therefore, to highlight the irregularities, we assign the weights to the difference value based on its magnitude. Once the features are extracted, a machine learning classifier is trained for binary classification (i.e., real or altered).

The massive success of deep convolutional neural networks has significantly increased their usage in machine learning inspired solutions. However, it has been observed that deep learning algorithms are susceptible to intelligently crafted minute noises, and are popularly known as adversarial examples. The adversarial attacks can be both targeted and untargeted. The impact of these adversarial attacks can be seen in the physical world where the simple misclassification of '*stop sign*' to '*increase speed*' can cause harm to pedestrian and the autonomous vehicle. Therefore, the detection of adversarial examples is essential for rightful and confident usage of deep learning-based solutions in the real world. As the final contribution, novel detection algorithms are developed to detect different kinds of adversarial attacks. The proposed solutions are the first in the community which can detect such vast and challenging scenarios, and yield the *panoptic* defense against adversarial examples being *agnostic* to the databases, adversarial attack algorithms, and CNN architectures.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Publications

JOURNALS

Published

1. **A. Agarwal**, R. Singh, M. Vatsa, and N. Ratha, **"Image Transformation based Defense Against Adversarial Perturbation on Deep Learning Models"**, IEEE Transactions on Dependable and Secure Computing (TDSC) Special Issue on AI/ML for Secure Computing, 2020, (DOI: 10.1109/TDSC.2020.3027183) **Impact Factor: 6.864**

2. **A. Agarwal**, R. Keshari, M. Wadhwa, M. Vijh, C. Parmar, R. Singh, and M. Vatsa, **"Iris sensor identification in multi-camera environment"**, Information Fusion, 2019, volume 45, Pages 333-345 **Impact Factor: 13.669**

3. G. Goswami, **A. Agarwal**, N. Ratha, R. Singh, and M. Vatsa, **"Detecting and Mitigating Adversarial Perturbations for Robust Face Recognition"**, International Journal of Computer Vision (IJCV), 2019, volume 127, pages 719-742 **Impact Factor: 5.698**

4. **A. Agarwal**, G. Goswami, M. Vatsa, R. Singh, and N. Ratha, **"DAMAD: Database, Attack, and Model Agnostic Adversarial Perturbation Detector"**, IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2020 **Impact Factor: 8.793** (Minor Revision)

Under Submission

1. **A. Agarwal**, R. Singh, M. Vatsa, and A. Noore, **"MagNet: Detecting Digital Presentation Attacks on Face Recognition"**, 2020

2. **A. Agarwal**, R. Singh, and M. Vatsa, **"Generalized Face Presentation Attacks Detection in NIR + VIS Spectrum"**, 2020

3. **A. Agarwal**, R. Singh, M. Vatsa, and N. Ratha, **“Cognitive Data Augmentation for Adversarial Defense via Pixel Masking”**, 2020

4. **A. Agarwal**, R. Singh, M. Vatsa, and N. Ratha, **“Parameter Agnostic Architecture for Secure Computer Vision”**, 2020

5. **A. Agarwal**, R. Singh, and M. Vatsa, **“On Adversarial Robustness of Face Presentation Attack Detection Algorithms”**, 2020

PUBLISHED PEER REVIEWED CONFERENCE ARTICLES

Top-Tier Publications:

1. **A. Agarwal**, M. Vatsa, and R. Singh, **“Role of Optimizer on Network Fine-tuning for Adversarial Robustness”**, AAAI-21 Student Abstract and Poster Program (SA-21), 2021

2. A. Mehra, **A. Agarwal**, M. Vatsa, and R. Singh, **“Detection of Digital Manipulation in Facial Images”**, AAAI-21 Student Abstract and Poster Program (SA-21), 2021

3. R. Singh, **A. Agarwal**, M. Singh, S. Nagpal, and M. Vatsa, **“On the Robustness of Face Recognition Algorithms Against Attacks and Bias”**, AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 13583–13589

4. G. Goswami, N. Ratha, **A. Agarwal**, R. Singh, and M. Vatsa, **“Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks”**, AAAI Conference on Artificial Intelligence (AAAI), 2018, pp. 6829-6836

First Author Publications:

1. **A. Agarwal**, R. Singh, M. Vatsa, and N. Ratha, **“Noise is Inside Me! Generating Adversarial Perturbations with Noise Derived from Natural Filters”**, IEEE CVPR Workshop on Adversarial Machine Learning in Computer Vision (CVPRW), 2020, pp. 774–775

2. **A. Agarwal**, R. Singh, and M. Vatsa, **“The Role of ‘Sign’ and ‘Direction’ of Gradient on the Performance of CNN”**, IEEE CVPR Workshop on Media Forensics (CVPRW), 2020, pp. 646–647

3. **A. Agarwal**, M. Vatsa, R. Singh, **“CHIF: Convolized Histogram Image Features for Detecting Silicone Mask based Face Presentation Attack”**, IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), 2019, pp. 1–5

4. **A. Agarwal**, A. Sehwag, R. Singh, and M. Vatsa, **“Deceiving Face Presentation Attack Detection via Image Transforms”**, IEEE International Conference on Multimedia Big Data (BigMM), 2019, pp. 373–382

5. **A. Agarwal**, A. Sehwag, M. Vatsa, R. Singh, **“Deceiving the Protector: Fooling Face Presentation Attack Detection Algorithm”**, IAPR International Conference on Biometrics (ICB), 2019, pp, 1–6

6. **A. Agarwal**, M. Vatsa, R. Singh, N. Ratha, **“Are Image-Agnostic Universal Adversarial Perturbations for Face Recognition Difficult to Detect?”**, IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2018, pp. 1–7

7. **A. Agarwal**, R. Singh, M. Vatsa, and A. Noore, **“SWAPPED! Digital Face Presentation Attack Detection via Weighted Local Magnitude Pattern”**, IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 659-665

8. **A. Agarwal**, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore,**“Face Presentation Attack with Latex Masks in Multispectral Videos”**, IEEE CVPR Workshop on Perception Beyond the Visible Spectrum (CVPRW), 2017, pp. 275-283

9. **A. Agarwal**, R. Singh, and M. Vatsa, **“Fingerprint Sensor Classification via Mélange of Handcrafted Features”**, International Conference on Pattern Recognition (ICPR), 2016, pp. 3001-3006.

10. **A. Agarwal**, R. Singh, and M. Vatsa, **“Face Anti-spoofing using Haralick Features”**, IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), 2016, pp. 1-6.

Other Publications:

1. S. Chhabra, **A. Agarwal**, R. Singh, and M. Vatsa, **“Revisiting Adversarial Attacks via Visual Imperceptible Bound”**, International Conference on Pattern Recognition (ICPR), 2020

2. N. Sanghvi, S. K. Singh, **A. Agarwal**, M. Vatsa, and R. Singh, **“MixNet for Generalized Face Presentation Attack Detection”**, International Conference on Pattern Recognition (ICPR), 2020

3. D. Anshumaan, **A. Agarwal**, M. Vatsa, and R. Singh, **“WaveTransform: Crafting Adversarial Examples via Input Decomposition”**, ECCV Workshop on Adversarial Robustness in the Real World (ECCVW), 2020

4. M. Gupta, V. Singh, **A. Agarwal**, M. Vatsa, and R. Singh, **“Generalized Iris Presentation Attack Detection Algorithm under Cross-Database Settings”**, International Conference on Pattern Recognition (ICPR), 2020

5. A. Goel, **A. Agarwal**, M. Vatsa, R. Singh, and N. Ratha, **“DNDNet: Reconfigured CNN for Adversarial Robustness”**, IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision (CVPRW) , 2020, pp. 22–23

6. A. Goel, **A. Agarwal**, M. Vatsa, R. Singh, and N. Ratha, **“Securing CNN Model and Face Template using Blockchain”**, IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2019, pp. 1–6

7. A. Goel, **A. Agarwal**, R. Singh, M. Vatsa, and N. Ratha, **“DeepRing: Protecting Deep Neural Network with Blockchain”**, IEEE CVPR Workshop on When Blockchain Meets Computer Vision and Artificial Intelligence (CVPRW), 2019.

8. P. Majumdar, **A. Agarwal**, R. Singh, and M. Vatsa, **“Evading Face Recognition via Partial Adversarial Tampering”**, IEEE CVPR Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security (CVPRW), 2019.

9. S. Mehta, A. Uberoi, **A. Agarwal**, M. Vatsa, R. Singh, **“Crafting A Panoptic Face Presentation Attack Detector”**, IAPR International Conference on Biometrics (ICB), 2019, pp, 1–6

10. A. Goel, A. Singh, **A. Agarwal**, M. Vatsa, R. Singh, **“SmartBox: Benchmarking Adversarial Detection and Mitigation Algorithms with Application to Face Recognition”**, IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2018, , pp. 1–7

11. D. Yadav, N. Kohli, **A. Agarwal**, M. Vatsa, R. Singh, and A. Noore, **“Fusion of Handcrafted and Deep Learning Features for Large-scale Multiple Iris Presentation Attack Detection”**, IEEE CVPR Workshop on Biometrics (CVPRW), 2018, pp. 572-579

12. T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, **A. Agarwal**, M. Vatsa, and R. Singh, **“Face Anti-spoofing with Multifeature Videolet Aggregation”**, International Conference on Pattern Recognition (ICPR), 2016, pp. 1035-1040.

13. R. Keshari, S. Ghosh, **A. Agarwal**, R. Singh, and M. Vatsa, **“Mobile Periocular Matching With PRE-POST Cataract Surgery”**, In IEEE International Conference on Image Processing (ICIP), 2016, pp. 3116-3120.

14. A. Sankaran, **A. Agarwal**, R. Keshari, S. Ghosh, A. Sharma, M. Vatsa, and R. Singh, **“Latent Fingerprint from Multiple Surfaces: Database and Quality Analysis”**, In Seventh IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), 2015, pp. 1-6.

UNDER SUBMISSION PEER REVIEWED CONFERENCE ARTICLES

1. **A. Agarwal**, M. Vatsa, and R. Singh, **“Supervised Mixup: Protecting the Likely Classes for Adversarial Robustness”**, 2020

2. **A. Agarwal**, R. Singh, and M. Vatsa, **“Adversarial Robustness through Cost Effective Network Fine-Tuning”**, 2020

3. A. Mehra, **A. Agarwal**, M. Vatsa, and R. Singh, **“DeepFake Detection: Impact on Face Recognition and Presence of Ethnicity Bias”**, 2020

Chapter 1

Introduction

The current era is the era of computer vision backed by machine learning and deep learning. The computer vision systems are now pervasive and have influenced every domain of life, ranging from security from various malware and attacks, health care, finance. Computer vision's massive success can be attributed to the advancements in the algorithms, the availability of high computing resources, and large databases for a variety of tasks. While the use and success of vision networks is colossal, several threats also exist, raising concern for its effective deployment. A typical computer vision algorithm comprises of the acquisition of data such as images, text, and speech. The data is processed to extract meaningful information, followed by the decision required to be performed on the data such as object classification and segmentation. The steps of a typical machine learning algorithm are shown in Figure 1-1. However, each stage of the machine learning pipeline is vulnerable to attacks [276]. At each stage, an attack can be performed to compromise the intended functioning of an algorithm. For example, at the sensor level, fake data can be presented, at the feature extraction level, correct features can be replaced with a spoof feature set. Similarly, the decision of the machine learning algorithm can be overridden so that the applications yield incorrect output. Imagine a scenario of an autonomous car, where the wrong decision to *'increase speed'* in place of the *'stop'* can harm the pedestrians or the vehicle itself. Therefore, the security of any machine learning algorithm is of paramount importance for their successful deployment and usage.

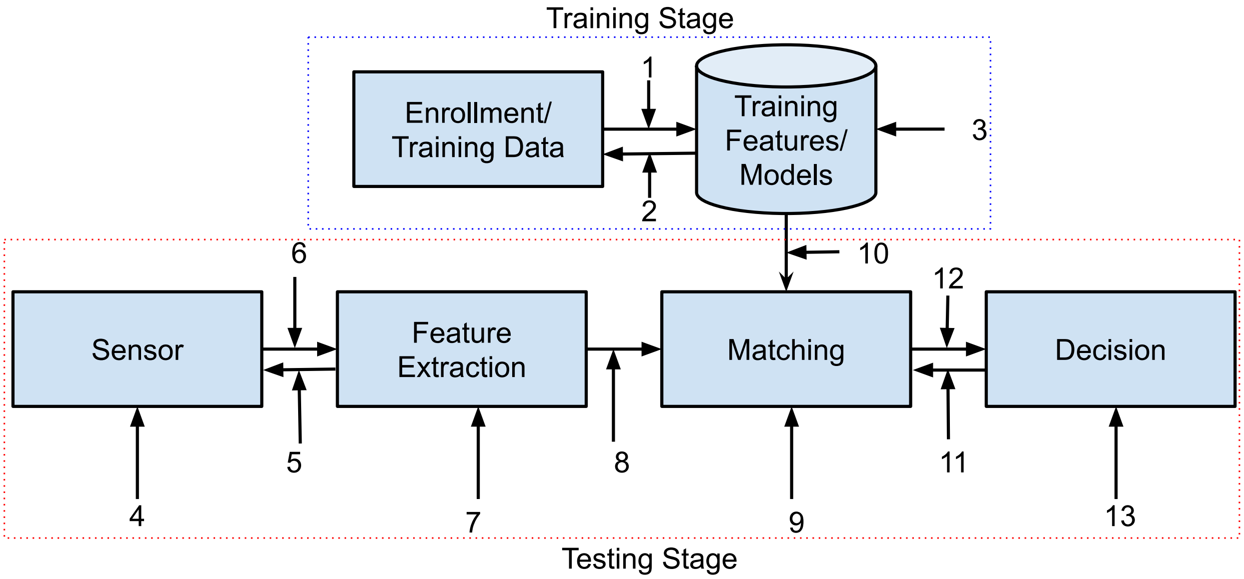


Figure 1-1: Illustrating the different stages of a computer vision system and its vulnerability points. The number at each stage represents that the attack can hinder its intended working. For example, point 1 represents a non-repudiation attack, rendering the denial of acquisition of data by the system going to be used in enrollment.

1.1 Attacks on Computer Vision Pipeline

The security of any machine learning algorithm can also be seen from the following aspects: confidentiality, integrity, availability, and privacy. Confidentiality and privacy cover the protection of essential information corresponding to the training model and data. The threats on model information can be termed as a model inversion attack where the aim is to extract the model's information such as parameters [101, 331]. It is also seen that the machine learning algorithms memorize the training data, and through inversion attack, the data can be extracted successfully. Therefore, the extraction of the financial or medical data from their trained networks can lead to severe privacy breach [43, 102, 323]. On the other hand, integrity and availability can be seen from the algorithm's intended behavior. The attacker might incorporate an excessive amount of false positives in the system to raise the question about the integrity of the system [303]. The system's availability can be affected in the following terms, such as the speed of processing, consistency in the decision, and accessibility. Figure 1-1 shows the typical computer vision pipeline and the associated threats at each point. The attacks can be broadly divided into three categories: (i) data acquisition and preparation, (ii) learning of feature representation and classifier, and (iii) decision making.

1.1.1 Data Acquisition and Preparation Attacks

The attacks on data acquisition and preparation level can be divided based on the timing of computer vision algorithms, i.e., training or inference. The attacks at the time of training are described using points *one* to *three* in Figure 1-1. Level *one* can be referred to as a non-repudiation attack where the system refuses to accept that it has received the acquired data. Even if the data is obtained, it might be of a low quality, which is not useful to properly represent the class of the data; therefore, the systems asks for recapturing the data. The attack on point *two* can stop passing this information where the system will never know that the earlier data was of low quality and its need of re-acquisition of data. The attack on point *two* can also be viewed as jamming of the acquisition stage where an attacker keeps on sending the message for an indefinite time that the acquired data was of low quality, and please send the new data. The attack on point *three* can be understood from the end of the biometric recognition system, which usually keeps the high-resolution gallery images and stores them as a template for matching. The attack can lead to the corruption of these templates either through noise insertion or replacing the template with the desired one. These attacks can be viewed as backdoor attacks where an attacker modifies the training template in such a way so that when that pattern comes in the testing images, the image will be classified in the desired class. The backdoor attacks are not limited to biometric systems but can also be applied on general purpose machine learning algorithms such as deep neural network can itself be learned on corrupted training data [130, 199, 337]. Similarly, at the time of testing, system needs the data to make any decision. Point *six* can be seen as a non-repudiation attack as performed on point *one*. On the other hand, the attack on point *five* can be the same as the attack on point *two*, but now it is performed at the testing phase. The attack on point *one* and *six* can also be complicated by bypassing the acquisition module and passing the previously modified digitized version. Other than that, the attack can be performed in real-time where the acquired data is digitally tampered, such as face morphing and adversarial perturbations can be used to fool facial and object recognition systems. The attack on point *four* can be seen as the presentation of fake data to the sensor itself. Examples include the presentation of a photo of the person for face recognition, gummy fingerprint made of artificial materials, and acquisition of contact lens iris images. In other words, the acquired data itself is fake [218].

1.1.2 Attack on Feature Extractor and Classifier

The attacks on the feature extraction and classification levels can also be performed in the white-box settings. The white-box setting refers to the condition where an attacker has access to the module to be attacked. The attack point *seven* refers to the overriding scenario of the feature extraction process. The attack is performed through the Trojan horse to make sure the desired feature are extracted for matching. While it is difficult to replace the feature extractor, the attack on point *eight* refers to replacing the target features from the pre-computed set of features. The attack setting is based on the assumption that the representation technique of the feature extractor in the target system is known. In other words, it should not happen that the replaced fraudulent features have a different distribution from the features computed from the actual feature extractor.

Apart from manipulating the feature extraction process or the features itself, the matcher or the classifier can itself be corrupted to produce the desired confidence score. In the case of a cloud machine learning system or when the machine learning system is stored at the cloud location, the communication is happened using wireless mediums. This communication channel, referred to as point *ten*, can be intercepted to modify either the transfer of template or decision of the system.

In the traditional machine learning classifier, domain knowledge is used to extract the features separately; later, the classifier is learned to make the decision. In contrast, in the era of deep learning, the feature extractor and classifiers are combined and treated as a single stage of the system. The deep learning classifiers extract the discriminating features automatically from the data while learning the decision function. In deep learning classifier, the attack on the features extraction stage is the modification or corruption of the individual component such as filter maps in case of a convolutional neural network (CNN) or the entire layer. Similarly, the classification softmax probabilities can be modified to increase the confidence of the incorrect class (in the case of targeted attack) or decrease the confidence of the correct type (in case of untargeted attack). The targeted attack represents the misclassification of an image in the desired class the attacker wants. The untargeted attack represents the scenario where the predicted level is just not the actual class.

1.1.3 Decision Attacks

While the entire pipeline of the computer vision systems, including data acquisition, features extraction, and classifier, is secured, the final decision taken by the system can be overridden. The attack on point *eleven* to *thirteen* represents the decision level attack. A hacker can alter the final decision of the system to be passed to the application process. The points *eleven* and *twelve* refers to the communication channel, which can be tampered as a non-repudiation attack where the decision is not passed to the application, and the application keeps on waiting. The overriding of the final decision, i.e., *thirteen*, can result in the availability of the application even if the machine learning system is perfect. The attack on a decision can significantly increase the false positive or false negative in the system. The false positive represents the classification of a negative class image into a positive class; whereas, the false negative is the classification of the positive class image into a negative category.

1.2 Defense Against Computer Vision Attacks

The earlier section shows the vulnerability of each stage of any computer vision system and demands the defense for its correct usage. In this section, we briefly describe the type of defenses that can be built to protect the integrity of each stage of the computer vision algorithms.

The attacks on data acquisition and preparation levels can be detected by building the binary classifier to identify whether an image is clean or modified. For example, the attack at the sensor level is widespread in the biometrics domain where a spoof image is acquired from the sensor, a field termed as presentation attack detection [218]. The area aims to classify whether the acquired image is the real image or the fake one. Similarly, another famous attack on data level is modifying data through noise and blending some parts of two or more images. Again, the popular defense against this kind of manipulation is the development of a binary classifier where the network uses both real and modified versions of the images [296, 363]. While the attack on data can be of different type and can vary significantly, it is believed that the real data follow a specific pattern or distribution. Inspired from this, anomaly detection based techniques can also be developed where the classifier is tried to learn the real class. Based on the representation of real and fake classes, the attack images can be filtered out [60, 340].

The attacks on other computer vision systems are hard to perform compared to the attacks on the data level. The probable reason for such difficulty is the system's knowledge, such as if the attacker wants to modify the feature extractor or classifier, the access to these modules is required.

The defense on these levels can be provided through cryptography [226], where the feature extractor itself or the extracted features can be encrypted to protect them from any manipulation. A similar defense can be applied to the decision module where the decision is encrypted and passed through a secure medium to the final application [112, 113].

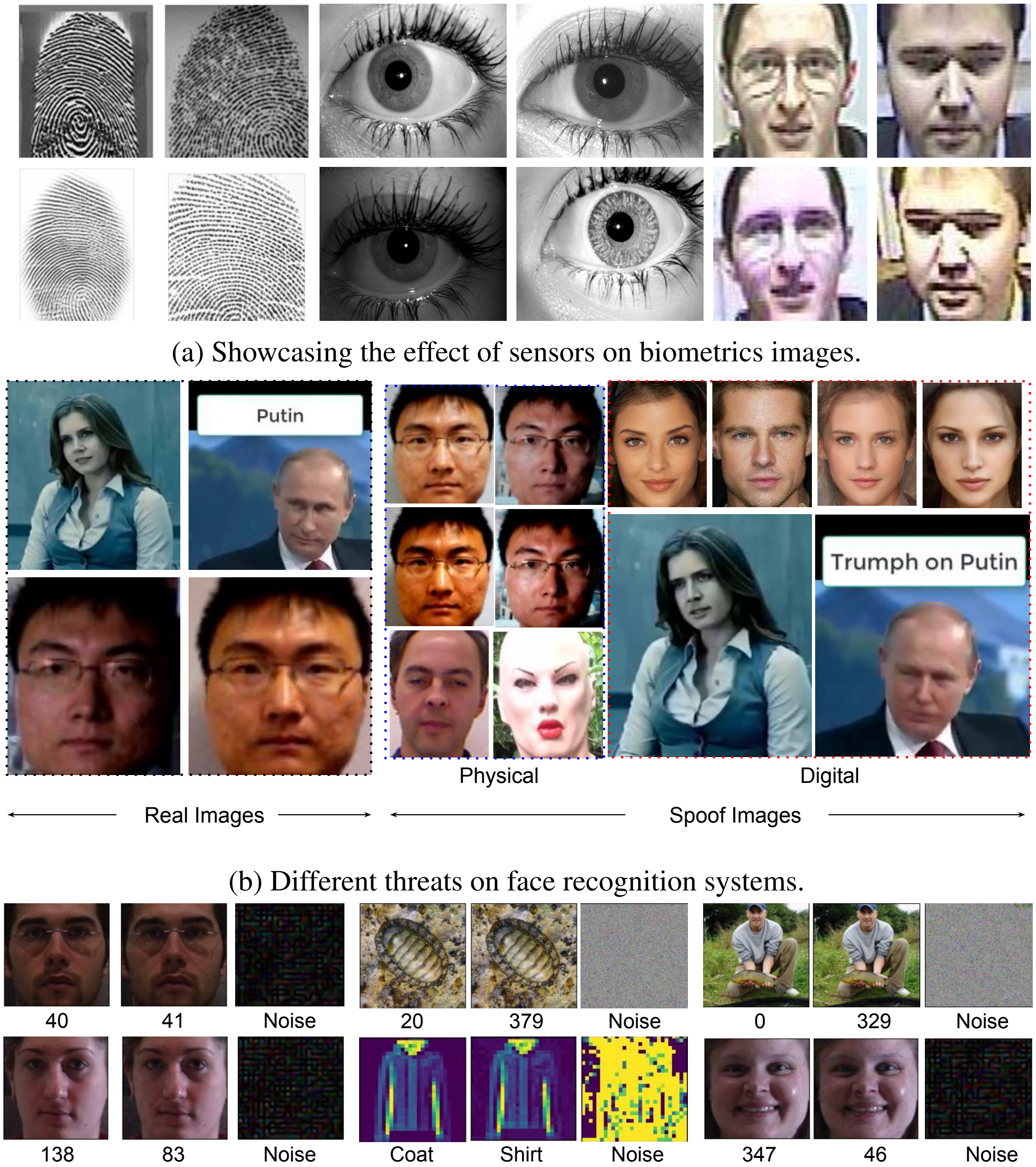
1.3 Research Contributions

Based on the urgent need for security mechanisms to protect the integrity of computer vision systems, this thesis provides *panoptic* solutions against various threats. In this dissertation, we focused on the attacks at data level and the source cameras used to capture these databases. The first and foremost component of any machine learning algorithm is the input data; therefore, the integrity of the input source, and input data itself is of utmost importance. The importance of authenticity can also be analyzed from the point of digital forensics as the images are considered as a critical source of evidence in the court of law. Therefore, it might be of potential interest to law enforcement agencies to know the source of digital content. The rise of digital contents and its importance in computer vision algorithms raises the following two serious questions:

- which source camera has been used in the acquisition of digital content, and
- is the digital content authentic, or has it been tampered with?

On the other hand, in biometrics recognition, the problem of sensor interoperability is a well known concept [73, 256]. The interoperability is referred to as matching images captured from different sensors/cameras. When the images collected from different sensors are used for matching, a significant drop in accuracy is observed. For example, Pillai et al. [252] have shown that the performance of iris recognition degrades significantly when cross sensor images are used for matching. To overcome such limitations, one simple and effective strategy was proposed by Arora et al. [22] and Marra et al. [220] that if we can first identify the sensor, then we can apply the camera-specific image enhancement to improve the matching accuracy. However, very few efforts have been made so far to identify sensors used in the acquisition of biometrics images.

The data is not only useful for very speific computer vision algorithm; however it has became a part of our day-to-day lives because of the explosion of various social media platforms. Therefore,



(a) Showcasing the effect of sensors on biometrics images.

(b) Different threats on face recognition systems.

(c) Vulnerability of deep CNN architecture against minute adversarial perturbations.

Figure 1-2: Illustrating the depth of the dissertation ranging from the identification of sensors used in the data acquisition which is the primary component of any computer vision algorithm. Apart from that, the defense against several kinds of anomalies data effective in fooling computer vision algorithms including face recognition and object recognition built on both traditional and deep neural networks. We proposed several panoptic solutions to develop the generalized solutions for each sub-problem which varies in terms of imaging spectrums, attacking instruments, attacking algorithms, databases.

the presence of fake information can create disharmony between the group of people or spread of misinformation. For instance, face is one of the quickest and first medium of communication among the humans, the presence of high quality fake face images such as DeepFake[1] or synthetic images [373] are not at all desirable for a healthy system. For example, the recent deepfake video showing the USA president launch of Apollo 11 is 'disaster'[2].

Therefore, as discussed above, the authentication of images and sources of images is an essential and challenging task. Much research has been done to detect the image source; however, those are mainly focused on natural images and are less useful for biometrics images. Similarly, the advancements can be seen towards securing the computer vision algorithms from anomalies. The problem of generalizability still makes them far behind reality. Such a bottleneck can also be attributed to the advancements in the acquisition devices, imaging sensors, and types of attacks. In this dissertation, we have proposed solutions for the problem of source camera detection and identifying whether the images are real or not at the sensor and image level. Figure 1-2[3] showcases the challenges of computer vision algorithms corresponding to variation in the input images because of sensors to several vulnerabilities against threats. This dissertation proposes solutions ranging from the identification of image source sensor to classification various anomalies presented in computer vision algorithms. The contributions of this dissertation include the following:

- **Sensor/Camera identification:** In literature, several algorithms have been proposed for the detection of cameras used to acquire natural images. However, minimal work has been performed to identify the sensors/cameras used for biometrics data acquisition. While the identification of cameras in natural images can be related to digital forensics, it leads to sensor interoperability, matching accuracy of the images captured using different sensors is lower than the images captured from the same sensor. As shown in Figure 1-2 (a), the images acquired from the different sensors might vary significantly, which increases the inter-class variability in the features. When such features are matched together, a significant drop in the recognition accuracy is observed. The benefit of identifying image sensors can help in preprocessing the image related to sensor characteristics to boost the matching performance.

[1] https://github.com/Qingcsai/awesome-Deepfakes

[2] https://nypost.com/2020/07/20/mits-deepfake-video-of-nixon-announcing-apollo-11-disaster-surfaces/

[3] The images are taken from multiple databases such as ImageNet [87], Presentation attacks [66, 93], Multi-PIE [127], and SCFace [126].

In this regard, we have proposed a novel algorithm for identifying sensors used to acquire biometrics images, including fingerprint and iris. The proposed algorithms utilize the amalgamation of features based on texture, intensity, and image quality. A new feature selection algorithm is proposed based on bacteria foraging using a support vector machine (SVM) fitness function. To the best of our knowledge, this is the first work that shows the effectiveness of sensor identification in multiple spectrums, including visible and near infra-red (NIR).

- **Physical presentation attack detection:** These days, face recognition algorithms are extensively used for authentication in security-related areas such as mobile unlocking, digital payments, and border access. However, it is seen that these systems are vulnerable to presentation attacks. Several defense-related research efforts have been undertaken to counter these limitations for the detection of physical presentation attacks. However, most of the algorithms developed so far have a restriction of generalizability against attack instruments such as 2D photo, 2D video, and silicone masks. Besides that, limited work has been undertaken to protect the face recognition systems working on different spectra, such as near infra-red (NIR) and thermal. In this dissertation, we aim to provide a panoptic solution by handling the variations that exist in the cross-spectrum images. By looking at the limited work in the field, we have first prepared one of the largest multi-spectral face presentation attack databases to make a more substantial impact. Simultaneously, a unified face presentation attack algorithm is developed, which is generalizable against various kinds of presentation attack instruments, including silicone masks, wax faces, and 2D attacks. The proposed algorithm computes the texture features at the global and local levels to highlight the attack images' inconsistencies. Extensive experimentation using multiple existing and proposed multi-spectral database has been performed.

- **Digital presentation attack detection:** Face morphing/swapping is another form of attack which can be used to gain the identity of someone else. It is found that face morphing is effective in generating duplicate identity, which means two different individuals can share the same identity. Face morphed images are useful in fooling the commercial systems, the human observers, and CNN-based face recognition algorithms. Therefore, to increase the robustness of face recognition systems against morphing, we have performed the following tasks. First, we have prepared a large scale database of morphing faces using multiple applications, including messaging apps, namely Snapchat. Second, we propose a Weighted Local Magnitude Pattern (WLMP) feature extraction algorithm. The WLMP descriptor computes the weighted difference value within the neighborhood of face regions. The proposed WLMP is combined with SVM for classifying the images into real and morphed class.

- **Adversarial examples detection:** The massive success of deep learning algorithms has made it popular across many fields, including face recognition, object recognition, and autonomous driving. However, deep learning algorithms are vulnerable to adversarial noise examples. It is found that images with subtle adversarial noise can be misclassified into the wrong category, which the human can correctly classify into the correct class. An image with a completely random structure/noise can be classified into the desired target class in extreme scenarios. Hence, for proper deployment of deep neural networks in various real-world applications, its robustness against adversarial examples needs improvement. To protect the integrity of CNN models, two defense algorithms are proposed. In the first solution, a computationally efficient detection algorithm is proposed utilizing several image transformations. The algorithm's effectiveness is presented by reducing the impact of adversarial noise on the recognition system. We propose a detection algorithm by score fusion of features computed over the CNN filter maps and non-linear embedding learned using auto-encoder in the second contribution. The proposed solution aims to provide a panoptic solution agnostic across various domains, including databases, architectures, and attack generation algorithms. The evaluations are performed using multiple databases used for object recognition and face identification, CNN architectures, and adversarial attack algorithms under several challenging *intra* and *cross* domain experimental settings.

Chapter 2

Biometric Sensor Identification in Multi-Camera Environment

2.1 Introduction

Biometric systems are utilized in a variety of applications including national identification projects such as India's Aadhaar project[1], as well as law enforcement applications such as the Integrated Automatic Fingerprint Identification System (IAFIS) of the Federal Bureau of Investigation (FBI)[2]. Among different kinds of biometric modalities, iris, face, and fingerprint are the most popular and accurate modalities [147]. With large scale projects, multiple kinds of devices are used for acquisition. Since the wavelength, LEDs for illumination, and sensor properties varies across devices, as shown in Figure 2-1, iris images of the same person, captured using different sensors in the same lighting condition, can show significant variations. Several researchers [22, 23, 74, 108, 283] have shown that interoperability leads to reduction in the performance of biometric recognition algorithms.

For any biometric system, it is important to protect the integrity of the system and to detect the attacks, if the system is compromised. For example, at the image level, an attacker may be sending biometric images from a sensor which is not certified or not in the system. In such a scenario, a mechanism is required to identify if the biometric image is captured from an authentic

[1] UIDAI https://uidai.gov.in/

[2] https://www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis

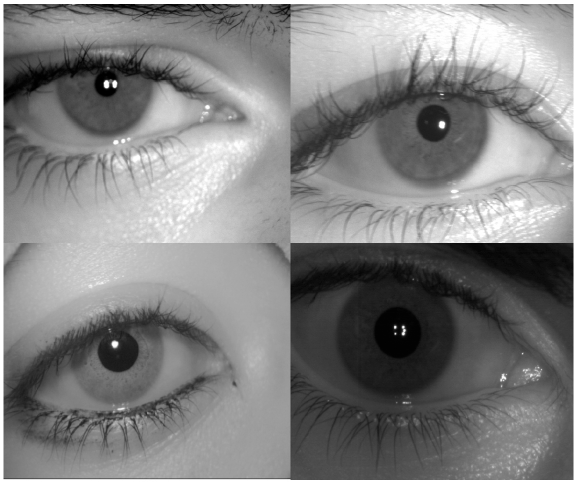


Figure 2-1: Illustrating the effect of sensor variation with iris images captured using different iris scanners in same (indoor) environment.

source. Therefore, for large scale applications where biometric sensors from several manufacturers are being used, it is important to ascertain the authenticity of images provided to the system for processing at both de-duplication and authentication stages. Further, in applications pertaining to law enforcement and legal-court cases, establishing that the biometric images used for recognition are captured using authentic devices are also of interest. While there are several sensor/camera identification approaches for natural images, very limited research is performed in iris sensor classification. In literature, researchers have generally used a particular type of features for biometric sensor classification. However, with technological advancements, the camera characteristics are now getting broadened. For example, iris images are captured in near infrared spectrum but researchers are also exploring the use of visible imagery for iris recognition. In such cases, it is important that the sensor classification algorithms are invariant to spectrum and other such variations. Therefore, in this research, we propose an efficient algorithm for iris sensor classification. The contributions of this chapter are four-fold:

1. Design a novel algorithm for biometric sensor identification using feature selection and fusion of intensity, wavelet, entropy, Haralick features, and image quality measures.

2. Design a novel feature selection algorithm that incorporates Support Vector Machine (SVM) fitness function into a bacteria foraging framework.

3. Present two multi-sensor iris databases that can be used for sensor classification as well as

interoperability research.

4. Experimental evaluation with multiple large databases containing images captured in near infrared and visible spectrums.

Section 2 presents the literature review of camera classification algorithms. Section 3 describes the details of the proposed algorithm. Section 4 explains the proposed bacteria foraging feature selection algorithm. Section 5 presents the details of the multi-sensor biometric databases used for performance evaluation, and Section 6 presents the experimental results pertaining to biometric sensor classification.

2.2 Literature Review

As mentioned previously, there are very limited research directions that attempt to identify the source sensors in a biometric system pipeline. However, the literature in camera classification is very thorough. Table 2.1 presents a brief summary of the algorithms that have been proposed for camera identification. Generally, sensor level noise pattern or image level features are modeled for device identification. For natural images, Kharrazi et al. [160] propose statistical features such as average pixel value over the RGB channels individually, RGB pairs correlation between each sub-band of color images, neighbor distribution center of mass correlation, RGB pairs energy ratio, and wavelet domain statistics to identify the digital camera from the images. Bayram et al. [35] observe that Color Filter Array (CFA) plays an important role in sensor identification. They propose to classify the source camera based on traces of the proprietary interpolation algorithm provided by a digital camera. Lukas et al. [203] suggest that every digital camera has a uniquely identifiable fingerprint as its reference pattern which can be considered as a high-frequency spread spectrum watermark. They propose to extract this pattern from images using wavelet denoising filter. They further [204] propose to identify camera based on sensor's pattern noise such as Fixed Pattern Noise (FPN) and Photo-Response Non-Uniformity Noise (PRNU). Before camera identification, they record reference pattern noise for each camera. The reference pattern noise serves as spread spectrum watermark. Khanna et al. [158] considered two processes of image acquisition: an image captured by digital camera and the image generated by the scanner. They use pattern

Table 2.1: Brief review of existing sensor classification algorithms in literature.

Author	Technique	Image contains
Kharrazi et al. [160]	Texture features with SVM	Natural scenes
Bayram et al. [35]	CFA features with SVM	Natural scenes
Lukas et al. [203]	Reference pattern with correlation matching	Natural scenes
Lukas et al. [204]	PRNU with correlation matching	Natural scenes
Chen et al. [63]	PRNU with maximum likelihood and correlation matching	Natural scenes
Khanna et al. [158]	Pattern noise with SVM	Natural scenes
Çeliktutan et al. [59]	Quality features with SVM	Natural scenes
Chen et al. [62]	PRNU with correlation	Natural scenes
Goljan et al. [116]	PRNU with correlation matching	Natural scenes
Khanna et al. [159]	Reference pattern	Natural scenes
Xu and Shi [354]	Uniform local binary pattern with wavelet features	Natural scenes
Arora et al. [22]	Image statistics and entropy with SVM	Iris images
Jin et al. [153]	Statistical correlation with neural network	Natural scenes
Tomioka et al. [329]	PRNU with correlation	Natural scenes
Thai et al. [324]	Generalized likelihood ratio	Natural scenes
Lawgaly et al. [173]	Sensor pattern noise	Natural scenes
Li et al. [180]	Wavelet decomposition with normalized cross-correlation	Natural scenes
Satta [292]	Sensor pattern noise with reliability map	Natural scenes
Cozzolino et al. [79]	camera-related features	Natural scenes
Ding et al. [358]	Hand-crafted and Deep learning	Natural scene
Yang et al. [358]	Residual CNN	Natural scene
Samaras et al. [288]	PRNU	Cell phones
Bartlow et al. [32]	Wavelet decomposition based PRNU with correlation matching	Fingerprint
Kalka et al. [156]	Wavelet decomposition based reference pattern with normalized cross-correlation	Iris
El-Naggar and Ross [92]	Gabor features and statistical features with neural network	Iris
Agarwal et al. [13]	Multiple hand-crafted features with SVM	Fingerprint
Banerjee and Ross [31]	Multiple sensor pattern noise	Iris
Freire-Obregón et al. [103]	CNN	Periocular
Chowdhury et al. [70]	Energy features from wavelet decomposed images	Ear

noise correlation along with SVM to distinguish between the scan and non-scanned images and classify the digital image source. Chen et al. [63] consider photo-response non-uniformity noise for sensor identification by deriving the maximum likelihood estimator of PRNU. Chen et al. [62] propose to use PRNU as a feature for digital images to classify source digital camera and check the integrity verification of their proposed framework after applying selected image processing operations. Goljan and Fridrich [116] propose to identify the source camera from cropped and scale images. They suggest brute force search to determine scaling factor and normalized cross-correlation for determining the cropping parameter. Çeliktutan et al. [59] use binary similarity between the bit planes. Along with the Binary Similarity Measure (BSM) features, higher order wavelet statistics and image quality measures are calculated for camera identification. Each image is decomposed using the quadrature mirror filter up to four levels and skewness, kurtosis, mean and variance of the wavelet coefficients are computed in the three orientations. For image quality measure, the reference image is obtained by smoothing the original image with the Gaussian filter.

Bartlow et al. [32] propose fingerprint sensor identification using wavelet based Wiener filtering approach to generate the PRNU reference pattern. Quadratic mirror filters are used for the decomposition of images into sub-bands. For each sub-band, four sub-local windows of size $3, 5, 7$, and 9 are used for calculating the variance and the minimum variance is selected from each window. Finally, the noise residual is calculated after taking the average of all the noise patterns of the camera images. Correlation coefficients are calculated between the test and noise residual images for classification. Khanna et al. [159] verify the method presented by Lukas et al. [203] on the basis of unprocessed and processed images. Processed images are generated using JPEG compression, resampling and adding the effect of malicious processing. Jin et al. [153] have created a model to distinguish between original and fake images. A camera color filter array is used to achieve separability. The main idea is to reverse engineer the interpolation function by learning the Demosaicking inter-pixel correlation. The pixels are selected using the mean square error and PCA is applied followed by the neural network. Xu and Shi [354] have used uniform local binary pattern and diagonal sub-band of the wavelet decomposed image of the original image. Tomioka et al. [329] use PRNU noise pattern to cluster the pixels of images for each camera model. For camera identification, they create cluster pairs and experimentally observe thresholds for camera classification. Lawgaly et al. [173] observe that noise pattern is different when acquisition conditions are changed such as bright images and dark images provide different noise pattern. Using this motivation, authors propose weighted averaging-based Sensor Pattern Noise (SPN) estimation. Thai et al. [324] propose a likelihood ratio based statistical test for the camera identification problem. They suggest that a strong statistical test can help in enhancing the confidence over false alarm rate. Likelihood ratio test can be modeled for known and unknown parameters of images and cameras. In order to handle unknown parameters, they suggest that Generalized Likelihood Ratio Test (GLRT) can be applied. Since images come from different sources, it is more practical to consider it as an unknown parameter of the image and known parameter of cameras. To make it more robust, they also consider the case where both image and camera parameter are unknown. Li et al. [180] use sensor pattern noise for camera identification. Each image is decomposed into wavelet sub-bands and the Wiener filtered images are subtracted from it. Sub-bands of the wavelet decomposed image are converted into the spatial domain and PCA is applied on this filtered image. A reference pattern for each sensor is calculated using the mean of the reduced dimension images. Normalized cross-correlation is calculated to identify the sensor of the test images from each of the reference sensor feature vector. Satta [292] propose reliability map as a parameter while estimating the noise at the pixel value. The reliability map for each pixel is considered based on the high frequency information around the neighboring pixels. This reliability map is used to provide weight for SPN thereby making it more robust. For the camera model identification, Cozzolino et al. [79] have proposed the noiseprint intending to enhance the camera-related features and suppress the image level content. Ding et al. [358] have used the domain knowledge to combine the hand-crafted and deep learning features for source camera identification. Yang et al. [358] have proposed the residual fusion network trained on three subcategories of input images for source camera identification. Kalka et al. [156] use Pixel Non-Uniformity (PNU) pattern of each camera sensor. The reference pattern for each sensor is generated by subtracting the original images of each sensor from its wavelet denoised images. The local variance of the different size of windows is calculated and Wiener filter is applied for each sub-band of wavelet decomposition. A reference pattern for each camera is computed after taking the mean of the reference pattern of all the images of that sensor. Pearson's product moment is calculated between the test reference pattern and sensor reference pattern. El-Naggar and Ross [92] have proposed the iris sensor classification using 50 Gabor and 68 statistical features. The features are extracted from the inner half portion of normalized iris image. Agarwal et al. [13] have proposed the fingerprint sensor classification using the combination of multiple hand-crafted features. The comparison of the proposed algorithm with several existing algorithms shows the effectiveness in identifying the sensor. They have also performed the feature selection using various algorithms and indicates that the Random Subset Feature Selection (RSFS) improves the identification performance by using only 19% of the total features. Banerjee and Ross [31] have done the comparison study of iris sensor classification using multiple Sensor Pattern Noise (SPN) pattern. Four different SPN methods: basic, enhanced, maximum likelihood, and phase based are used for evaluation. They have found that basic and enhanced SPN provides better classification performance in comparison to maximum likelihood and phase based algorithm. Freire-Obregón et al. [103] have proposed deep learning architecture for mobile camera identification used in the acquisition of periocular images. Chowdhury et al. [70] extracted the energy features from wavelet decomposed images for camera identification utilized for ear biometric acquisition.

2.3 Proposed Biometric Sensor Classification Algorithm

In this research, we propose a novel sensor classification algorithm that is based on image and texture properties. Figure 2-2 illustrates the steps involved in the proposed algorithm. The proposed algorithm has three components: feature extraction, feature selection, and classification.

2.3.1 Feature Extraction

Since the images from different sensors may vary in multiple characteristics such as image properties, texture, and brightness, we propose a combination of four different image-based and texture-based features for sensor classification.

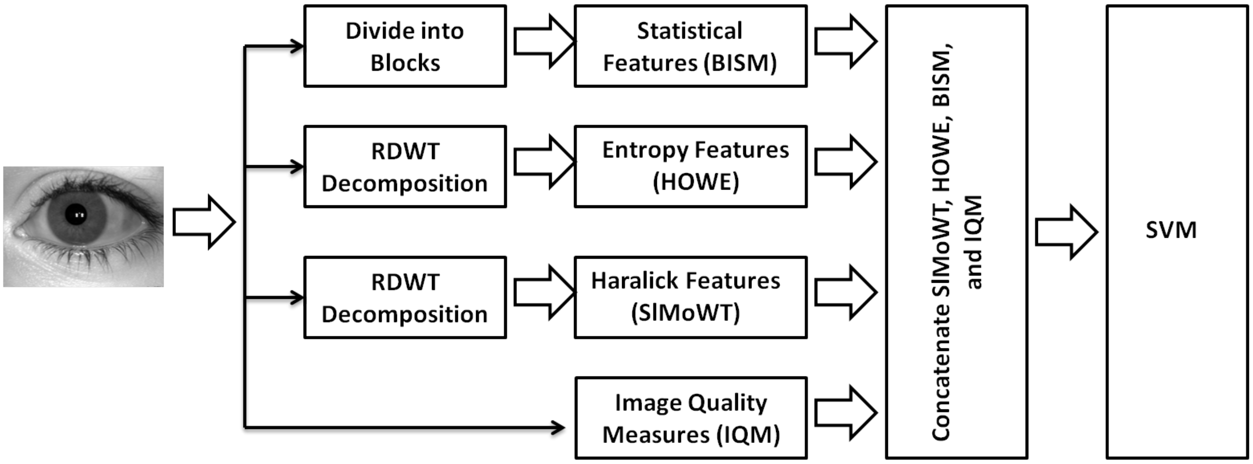

Divide into Blocks
Statistical Features (BISM)
RDWT Decomposition
Entropy Features (HOWE)
RDWT Decomposition
Haralick Features (SlMoWT)
Image Quality Measures (IQM)
Concatenate SlMoWT, HOWE, BISM, and IQM
SVM

Figure 2-2: Block diagram of the proposed algorithm for biometric sensor identification.

Block Image Statistical Measure

It is generally observed that with a higher number of LEDs or change in wavelength, the brightness and contrast of the images change, leading to variations in intensity values. It is to be noted that local variations can also be observed due to specular reflection. Therefore, to model and learn these variations across sensors, we use a first-order image statistics via combination of both local and global features. The global characteristics are encoded by dividing the image into four equal blocks and computing the mean and standard deviation of each block.

$$f_{1-1} = \{\mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3, \mu_4, \sigma_4\} \tag{2.1}$$

Similarly, the local characteristics are extracted by tessellating the image into 16 small patches and computing the mean and standard deviation of each block. The global features are of length 8 and the local features are of length 32. As shown in Figure 2-3, combining both the features yield the first feature vector of length 40 and we called it as Block Image Statistical Measure (BISM).

$$\begin{aligned} f_{1-2} = \{&\mu_{11}, \sigma_{11}, \ldots, \mu_{14}, \sigma_{14}, \mu_{21}, \sigma_{21}, \ldots \\ &\mu_{24}, \sigma_{24}, \sigma_{31}, \ldots \mu_{31}, \sigma_{31}, \ldots \mu_{34}, \sigma_{34}, \\ &\mu_{41}, \sigma_{41}, \ldots \mu_{44}, \sigma_{44}\} \end{aligned} \tag{2.2}$$

$$F_1 = \{\ f_{1-1}, f_{1-2}\} \tag{2.3}$$

High Order Wavelet Entropy

Discrete Wavelet Transform (DWT) [305] is the process of decomposing an image into four different sub-bands: approximation (H_a), horizontal (H_h), vertical (H_v), and diagonal (H_d). The size of each sub-band is one-fourth the size of the input image. Redundant discrete wavelet transform (RDWT) [100] is a variant of DWT which decomposes an image into four sub-bands but does not downsample the image. Thus, each sub-band is of the same size as the input image. RDWT provides the texture properties of an image and encodes the low and high frequency information

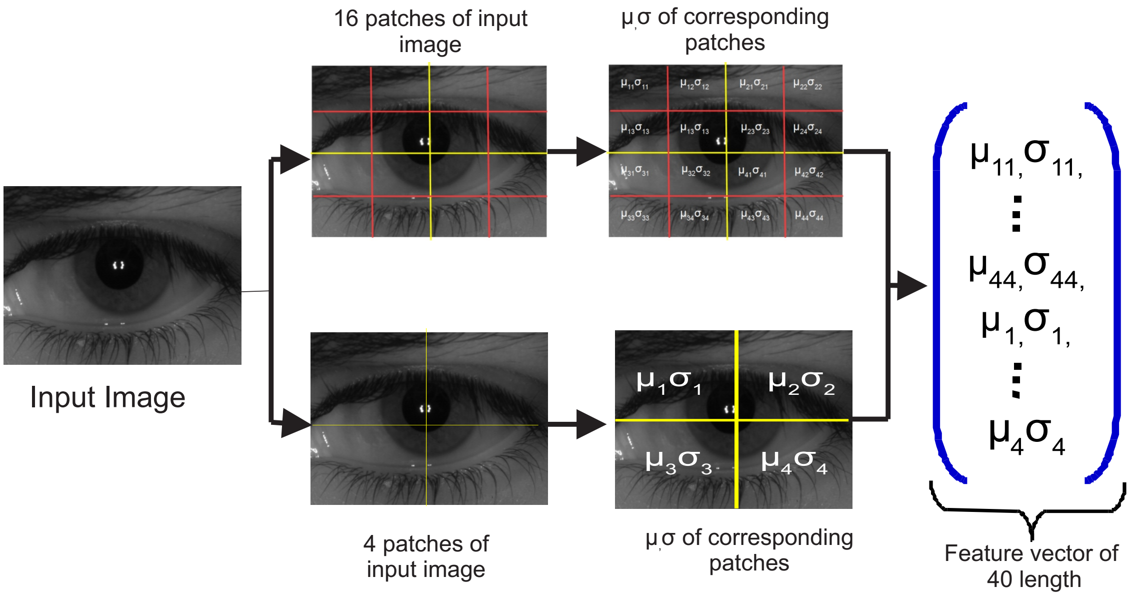

16 patches of input image
μ,σ of corresponding patches
Input Image
4 patches of input image
μ,σ of corresponding patches
Feature vector of 40 length
μ11,σ11,
μ44,σ44,
μ1,σ1,
μ4σ4
μ1σ1
μ2σ2
μ3σ3
μ4σ4

Figure 2-3: Extracting the first set of features: mean and standard deviation (BISM) from global and local patches of iris image.

in different sub-bands.

Biometric image is decomposed to the first level of RDWT and entropy of each sub-band is computed. Figure 2-4 shows RDWT decomposition of an iris image.

$$E(H_i) = \frac{\mu_i - \sum H_i(x,y)}{\sigma_i} \tag{2.4}$$

where $E(\cdot)$ is the entropy of sub-band H_i, and μ_i and σ_i are the mean and standard deviation of the i^{th} sub-band respectively. The entropy of each sub-band up to two levels is concatenated to generate the second feature vector F_2 and called as High Order Wavelet Entropy (HOWE).

$$F_{2-1} = \{E(H_{a1}), E(H_{h1}), E(H_{v1}), E(H_{d1})\}$$

$$F_{2-2} = \{E(H_{a2}), E(H_{h2}), E(H_{v2}), E(H_{d2})\}$$

$$F_2 = \{F_{2-1}, F_{2-2}\} \tag{2.5}$$

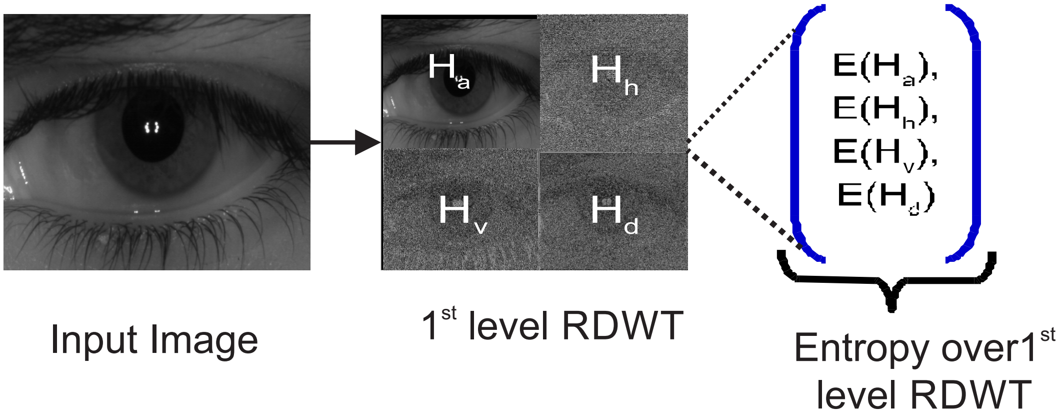


Figure 2-4: Steps involved in extracting entropy based HOWE features from RDWT decomposed iris images.

Texture Measure and Single-level Multi-orientation Wavelet Texture

Haralick et al. [136] proposed Haralick features that describe an image based on the texture properties computed by gray-tone spatial-dependence matrices (GSDM). Gray-tone spatial dependency matrix is computed for each grayscale intensity image by computing how many times a pixel with one value occurs adjacent to a pixel with another value in the horizontal direction. These matrices have four orientations ($\theta = 0^0, 45^0, 90^0, 135^0$). The square GSDM matrix is of size $N_g \times N_g$, where N_g is the number of gray levels in the image.

Haralick feature vector can encode the texture and global statistical properties of an image which can help in sensor classification. Out of the 14 features proposed by [136], we have used 13 features as explained below:

- Angular Second Moment:

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)^2 \tag{2.6}$$

- Contrast:

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\}, |i-j| = n \tag{2.7}$$

- Correlation:

$$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{2.8}$$

- Sum of Squares: Variance

$$f_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2 p(i,j) \tag{2.9}$$

- Inverse Difference Moment:

$$f_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2} p(i,j) \tag{2.10}$$

- Sum Average:

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{2.11}$$

- Sum Variance:

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i) \tag{2.12}$$

- Sum Entropy:

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) log \left\{ p_{x+y}(i) \right\} \tag{2.13}$$

- Entropy:

$$f_9 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) log(p(i,j)) \tag{2.14}$$

- Difference Variance:

$$f_{10} = \sum_{i=0}^{N_g - 1} i^2 p_{x-y}(i) \tag{2.15}$$

- Difference Entropy:

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) log \left\{ p_{x-y}(i) \right\} \tag{2.16}$$

- Information Measures of Correlation:

$$f_{12} = \frac{HXY - HXY1}{max\{HX, HY\}} \tag{2.17}$$

$$f_{13} = (1 - exp\{-2(HXY2 - HXY)\})^{\frac{1}{2}} \tag{2.18}$$

where,

$$HXY = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) log(p(i,j)),$$

$$HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) log\{p_x(i)p_y(j)\},$$

$$HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j) log\{p_x(i)p_y(j)\}$$

$$F_{3i} = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}\} \tag{2.19}$$

$$where,\ i \in \{H_a, H_h, H_v, H_d\}$$

$$F_3 = \{\ f_{3H_a}, f_{3H_h}, f_{3H_v}, f_{3H_d}\} \tag{2.20}$$

Here, μ_x, μ_y, σ_x, and σ_y are the mean and standard deviations of p_x and p_y - the marginal probability density functions, x and y are the coordinates (row and column) of an entry in the co-occurrence matrix, $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$, and HX, HY are the entropies of p_x and p_y.

In this research, we have computed Haralick features of RDWT decomposed images for $\theta = 0^0$. For an image, after RDWT decomposition, 13 Haralick features are extracted for each sub-band.

Table 2.2: Image quality features.

Image Quality Measures	
Mean Square Error	Median Spectral Phase Distortion
Average Difference	Normalized Absolute Error
Structural Content	Normalized Cross-Correlation
Peak Signal to Noise Ratio	Laplacian Mean Squared Error
Spectral Phase Distortion	Spectral Magnitude Distortion
Weighted Spectral distortion	Median Spectral Magnitude Distortion
Maximum Difference	Weighted Median Spectral distortion
Czekanowski distance	Mean Absolute Error
Mean of image	Mean of Wavelet sub-bands (4)

Combining the Haralick features from all four sub-bands yields a feature vector of size 52 and called as Single-level Multi-orientation Wavelet Texture (SlMoWT). Along with this, 13 Haralick texture features are also computed over the original image without decomposing it using wavelet filters and termed as Texture Measure (TM).

Image Quality Measures (IQM)

Across different biometrics recognition applications, images are captured using different kinds of sensors in different environments, and therefore they are of different qualities. Figure 2-6 shows sample images from different sensors. From the figure, it is clear that the imaging condition and the sensor used for acquiring the images have a large effect on the output images. To encode this information, we extract 21 image quality features F_4 (listed in Table 2.2) and combine this information with the other proposed features. Image quality measure is well explored for steganalysis [27], compression, and imaging artifacts [28, 94]. To calculate the IQM features, desired reference image is obtained by filtering the original image using a Gaussian filter. in this research, we have filtered the image using a Gaussian filter of size 3×3 with the sigma value of 0.5. Some of the image quality measures computed are explained below:

The commonly used terms in the equations below such as k=1...3 represent color channels and C, $\hat{C}$ represents the original and reference image respectively. $N \times N$ is the size of the input image.

- Minkowski measures compute the difference between the pixel intensities of the original image and the reference image. $\gamma = 1$ corresponds to mean absolute error and $\gamma = 2$

corresponds to mean square error.

$$M_\gamma = \tfrac{1}{K}\sum_{k=1}^{K}\left\{\tfrac{1}{N^2}\sum_{i,j=1}^{N}|C_k(i,j)-\hat{C}_k(i,j)|^\gamma\right\}^{\frac{1}{\gamma}} \tag{2.21}$$

- Czekanowski Distance

$$C = \frac{1}{N^2}\sum_{i,j=0}^{N-1}\left\{1-\frac{2\sum_{k=1}^{K} min(C_k(i,j),\hat{C}(i,j))}{\sum_{k=1}^{K}(C_k(i,j)+\hat{C}(i,j))}\right\} \tag{2.22}$$

- Spectral Measure

$$\Gamma_k(u,v)=\sum_{m,n=0}^{N-1} C_k(m,n)exp\left(-2\pi im\frac{u}{N}\right)exp\left(-2\pi in\frac{v}{N}\right) \tag{2.23}$$

where $\Gamma_k(u,v)$ denotes the Fourier transform of the original image and $\hat{\Gamma_k}(u,v)$ denotes the Fourier transform of the corrupted image.

- Magnitude spectra and phase spectra of the image with block and without block are used as features. Magnitude and phase spectra are defined as $\varphi(u,v) = arctan(\Gamma(u,v))$ and $M(u,v) = |\Gamma(u,v)|$.

$$M=\frac{1}{KN^2}\sum_{k=1}^{3}\sum_{u,v=0}^{N-1}||\Gamma_k(u,v)|-|\hat{\Gamma_k}(u,v)||^2 \tag{2.24}$$

- The weighted sum of the block based magnitude and phase distortion is also computed.

$$J^l = \lambda J^l_M + (1-\lambda)J^l_\phi \tag{2.25}$$

where

$$J^l_M = \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{u,v=0}^{N-1}\left(|\Gamma_k(u,v))|-\left|\hat{\Gamma_k}(u,v)\right|\right)^2\right)^{1/2}$$

$$J^l_\varphi = \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{u,v=0}^{N-1}\left(|\phi_k(u,v))|-\left|\hat{\phi_k}(u,v)\right|\right)^2\right)^{1/2}$$

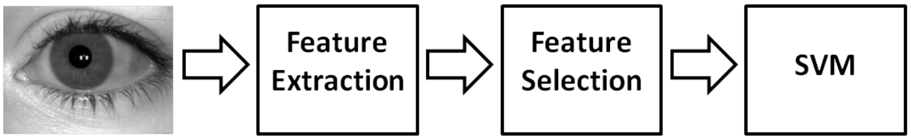


Figure 2-5: Block diagram of the proposed algorithm with feature selection.

where, l represents the block number and λ is the weighted factor between the magnitude and phase distortion. J_M and J_ϕ are the magnitude and phase distortions respectively. We have set the value of $\lambda = 2.5 \times 10^-5$.

2.3.2 Sensor Classification

Since the data has been collected via multiple sensors, multi-class SVM [334] is used for classifying the features. The multi-class SVM is trained as a one-vs-all model in which an SVM model is learned to differentiate one class from all other classes. For the c-class classification problem, c SVMs are learned with radial basis function (RBF) kernel that discriminate the i^{th} class from the remaining classes.

Given a training data set (x_i, y_i) where, x_i is the input data such that $x_i \in \mathbb{R}^m$, y_i represents the class label, $i = 1, \cdots, n$, n is the number of data points, and m is the dimension of feature vector x_i. The four feature sets are extracted (feature set I (F_1), feature set II (F_2), feature set III (F_3)), and feature set IV (F_4)) concatenated, and provided as input to SVM. During testing, the concatenated feature vector $[F_1, F_2, F_3, F_4]$ is computed for the test image and the trained multi-class SVM is used for classifying it to one of the sensor classes.

2.4 Proposed Feature Selection for Sensor Classification

It is our assertion that some features have higher discrimination power as compared to others and therefore they should be given higher importance. Further, redundant features or features not contributing towards improving the classification performance should be discarded. *Feature selection* is an important approach that provides mechanism to select important features that can enhance the overall classification performance [290]. Popular feature selection approaches such as random subset feature selection, sequential forward selection, sequential floating forward selection

[257, 275, 345] are statistical approaches. On the other hand, there exist evolutionary optimization based feature selection algorithms that select features inspired by *genetic selection* mechanisms observed in nature. For example, Ludwig and Nunes [202] proposed genetic learning based feature selection [202] which uses mutation and crossover based optimization. In this research, we propose bacteria foraging based feature selection for sensor classification.

Bacteria is the simplest form of life in the universe which evolves based on the foraging approach. Bacteria foraging works on the assumption that the bacteria (e.g. E.Coli) searches for the nutrients, in a conducive environment, and evolves (i.e. synthesize and replicate). This natural evolution is utilized in developing bacteria foraging algorithm [246]. In literature, Yadav et al. [355] proposed bacterial foraging optimization for biometric feature fusion. Inspired by their results, we propose bacteria foraging based feature selection algorithm (BFFS) and apply it for sensor classification. As shown in Figure 2-5, input to the proposed feature selection are features extracted from the images and output is the selected features which are used in SVM classification.

Generally, the loss function used in bacteria foraging based feature selection is some form of *distance* (or health), in this research, we have used SVM loss function in different stages of bacteria foraging, chemotaxis, reproduction, elimination, and dispersal. In the feature selection approach, we propose to utilize SVM loss function as the optimization function for bacteria foraging. In other words, the optimization of bacteria foraging based feature selection is controlled with minimizing the SVM loss function. Mathematically, for the input feature vector x_i and associated class labels y_i, selected features from bacteria foraging are used as input to the SVM and loss function is optimized as,

$$min \left[\frac{1}{2}||w||^2 + C \sum_i \psi_i \right]$$

$$\text{subject to } y_i \left\{ w\varphi(\mathcal{F}(x_i)) + b \right\} \leq (1 - \psi_i) \tag{2.26}$$

Here $\mathcal{F}(\cdot)$ is the output of feature selection from bacteria foraging based optimization. C is the cost function, ψ is the slack variable, and φ is the kernel function. w and b are parameters of the classifier representing the normal vector to separating hyper-plane and bias, respectively. The proposed feature fusion and classification approach is divided into two parts: training and testing.

Training: Using the labeled training data, BFFS model is trained as follows:

- **Step-1:** Features are extracted from the training images, using the feature extraction algorithms. The training feature vector is defined as (x_i, y_i), where x_i is the feature and y_i is the sensor class label.

- **Step-2:** Initialize bacteria foraging parameters, e.g. amount of bacteria, reproduction steps, chemotactic steps, elimination-dispersion steps, and probability of dispersion.

- **Step-3:** In bacteria foraging based optimization, there are four main stages. In chemotaxis stage, the movement (swim and tumble) of E.Coli bacterium is simulated which attempts to find the optimal solution (i.e. feature combination). In the next stage, i.e. swarming, the fitness function is calculated. In this research, probabilistic output of SVM classifier is calculated and used as the fitness function. The third step is reproduction which simulates the concept that the least healthy bacterium (which has the lowest fitness value) dies and healthier ones are used for reproduction. Based on the SVM probabilities, bacteria are sorted, top 50% of the population is used for reproduction and the remaining are discarded. In reproduction, the surviving bacteria are split into two, along with replacement based regularization, to generate the bacteria and ensure that the size of total population remains the same. In the elimination and dispersal process, it is ensured that each bacterium with less than a specified fitness value (less than a threshold) are removed and replaced with a healthy bacterium.

- **Step-4:** Previous step is iterated till convergence or maximum iteration (defined as 500 in our experiments) and features associated with the best performing bacterium (i.e. with maximum sensor classification accuracy) are used as the selected features with SVM classification.

Testing: For a test sample x_{test}, the selected features, learned from previous step, are extracted. The learned SVM classifier with the selected features are used to predict the class label $y_{predict}$.

2.5 Biometric Databases, Protocol, and Existing Algorithms for Sensor Identification

The results of the proposed sensor classification algorithm can be evaluated with databases containing images captured from multiple sensors. In this research, we have evaluated the effectiveness of the algorithm on three biometric modalities, i.e., iris, fingerprint, and face. First, we described the iris databases used for experimentations followed by the description of fingerprint and face databases.

2.5.1 Iris Databases

We have used four different data sets: all four sets contain images from multiple sensors with varying characteristics. The first two sets, multi-sensor database A and B, are created by the authors: database A contains images from four different sensors and database B contains images from three different sensors. These two sets are created with a time difference of more than two years and will be made publicly available to the research community[3]. Typically, the iris databases prepared in literature are captured using different state-of-the-art iris sensors including hand-held devices, in both controlled and uncontrolled environment. The iris sensors such as OKI IRISPASS-h, Vista, Crossmatch, and Delta are hand-held devices. Due to the variation in environment and sensor, the information content of the iris images also varies. The specification of databases and sensors used in this research are given in Tables 2.3 and 2.4.

Multi-sensor NIR Iris Database A

This database contains iris images captured from four different sensors: 1) CG4 iris, 2) CM Iris, 3) Delta Capture[4], and 4) Vista[5]. These four devices are named as Device I, Device II, Device III, and Device IV respectively. For all four sensors, the resolution of images is 640×480. There are 2731 images pertaining to 136 unique irises of 68 subjects and at least five samples of each subject are captured with each sensor. Figure 2-6 shows sample images of three subjects. All the images

[3] http://iab-rubric.org/resources/msirDB.html

[4] http://www.deltaid.com/technology.php

[5] http://www.vistaimaging.com/EY2H_product.html

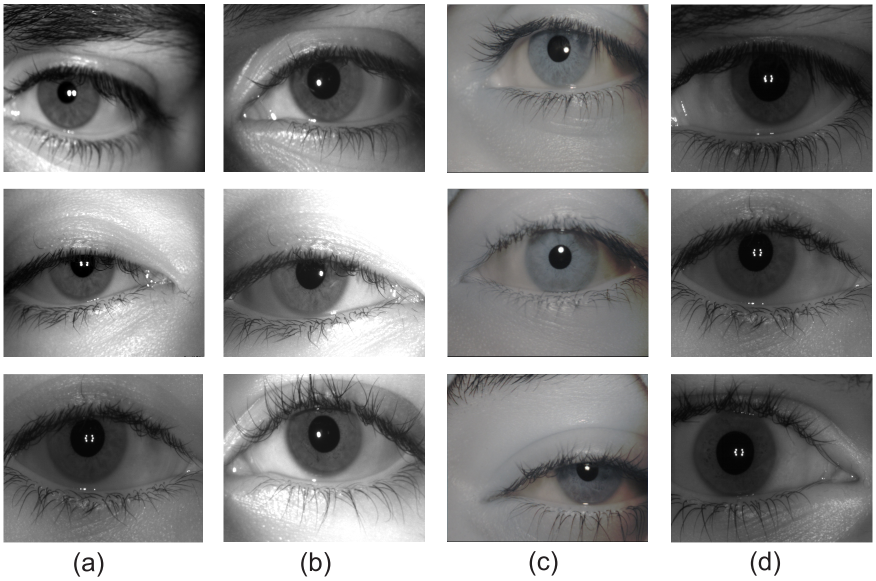


Figure 2-6: Sample iris images from database A captured by (a) CG4, (b) CM, (c) Delta, and (d) Vista. Every column represents different sensor and every row represents different subject.

are captured in room under the same environment.

Multi-sensor NIR Iris Database B

This database consists of 3390 images pertaining to 104 subjects, with a distance of 2-8 inches between the subject and sensor. The images are captured in the daytime lab environment. A minimum of five samples per eye per subject are collected from every sensor and an additional sample of the subjects wearing eyeglasses or contact lenses is also captured. The data is collected using three different sensors having a resolution of 640×480. The three sensors are: CrossMatch I Scan 2[6], Cogent Systems CIS-202[7], and IriShield USB MK 2120u[8]. These three devices are named as Device I, Device II, and Device III, respectively. Figure 2-7 shows sample images acquired by the three sensors. The sensors used for NIR databases are invariant to ambient illumination and operate in the range of 700-900 nm only. The CrossMatch I scan 2 used in the proposed databases covers the eye portion with the help of the shield provided and hence inturn reduces environmental factors.

[6] http://www.crossmatch.com/i-scan-2/

[7] http://solutions.3m.com.sg/wps/portal/3M/en_SG/Cogent/Border-Control/Government/Biometric-Capture-Devices/Dual-Iris-Scanner/

[8] http://www.iritech.com/products/hardware/irishield

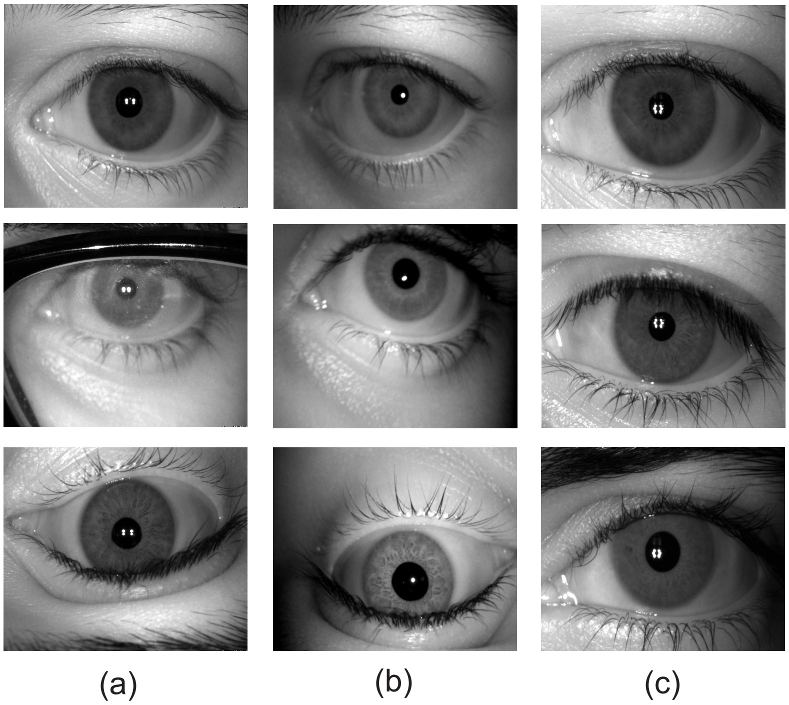


Figure 2-7: Sample iris images from database B captured by (a) Cogent Systems CIS-202, (b) IriShield USB MK 2120u, and (c) CrossMatch I Scan 2. Every column represents different sensor and every row represents different subject.

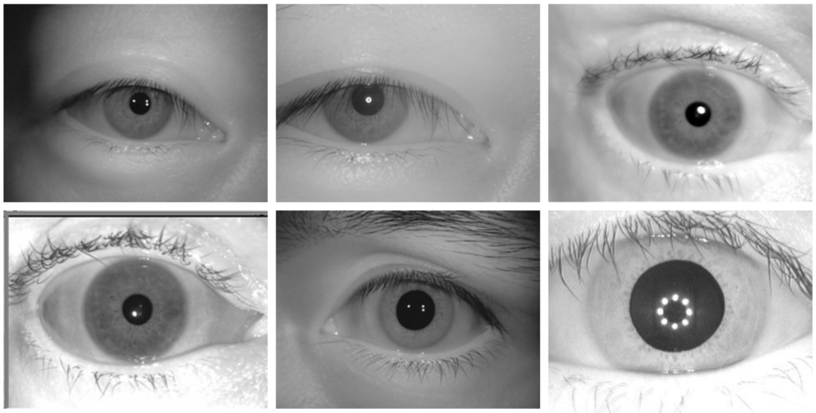


Figure 2-8: Sample images from database C. Every iris image is captured from a different sensor.

Multi-sensor NIR Iris Database C

Along with the above mentioned two databases, we have also used publicly available databases to perform the experiment. The dataset C contains images from three public databases:

- CASIA v2 [2] is collected using two different sensors (A) OKI-IRISPASS-h and (B) CASIA-

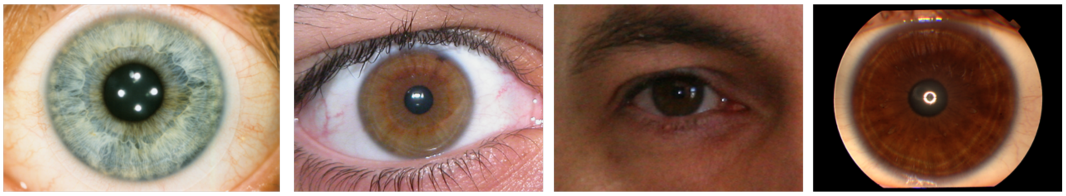


Figure 2-9: Sample images from database D: the iris images are captured from four different sensors.

Table 2.3: Summarizing the characteristics of all four iris biometric data sets.

Dataset	Constituent Databases	Spectrum	Sensors	Sensor Models	Data	Train Data	Test Data
A	In-house (IIIT-D)	NIR	4	CG4, CM, Delta, and Vista	2,731	800	1,931
B	In-house (IIIT-D)	NIR	3	Cogent Systems CIS-202, IriShield USB MK 2120u, and CrossMatch I Scan 2	3,090	600	2,490
C	CASIA v2 [2]	NIR	2	OKI-IRISPASS-h and CASIA-IrisCamV2	2,400	400	2,000
	CASIA v3 (Lamp and Interval) [3]	NIR	2	CASIA close-up iris camera and OKI IRISPASS-h	18,851	400	18,451
	ND cross-sensor [5]	NIR	2	LG2200 and LG4000	45,140	400	44,740
D	UPOL [6, 89]	VIS	1	SONY DXC-950P 3CCD	384	200	184
	UBIRIS.v1 [261]	VIS	1	Nikon E5700	1,876	200	1,676
	UBIRIS.v2 [262]	VIS	1	Canon EOS 5D	11,101	200	10,901
	Miles Research [4]	VIS	1	Self developed	816	200	616

IrisCamV2. It contains 1,200 images captured from 60 different subjects. It is an indoor environment database collected in a single session.

- CASIA v3 [3] has three different subsets: CASIA-Iris-Interval, CASIA-Iris-Lamp, and CASIA-Iris-Twins. in this research, we have used only the first two subsets. CASIA v3 Interval is collected in two different sessions from 249 subjects and CASIA iris Lamp is collected while the lamp is on or off, of 411 subjects. These are collected using two sensors (a) CASIA close-up iris camera and (b) OKI IRISPASS-h. A total of 2,639 and 16,212 images are available from each sensor respectively.

- ND cross-sensor [5] is prepared for cross-sensor iris recognition challenge. It contains 117,503 and 29,939 iris images captured using two different sensors LG2200 and LG4000 respectively. In this research, we have used a subset of the database with 23,250 and 21,890 images from each sensor respectively.

Table 2.4: Characteristics of some of the iris sensors used for capturing images used in this research.

Characteristics	Crossmatch I SCAN 2	Iris Shield MK 2120 u	Cogent CIS 202	OKI IrisPass-h
Spectrum	NIR	NIR	NIR	NIR
Iris Diameter	>200 Pixels	>200 Pixels	>210 Pixels	–
Interface	USB 2.0	USB 2.0	USB 2.0	USB 1.1
Operating Temperature	0^0 to 49^0 C	0^0 to 45^0 C	0^0 to 40^0 C	5^0 to 35^0 C
Resolution	640×480	640×480	640×480	640×480
Operating System	Windows	Windows	Windows	Windows
Certificate	ISO/IEC 19794-6	ISO/IEC 19794-6	ISO/IEC 19794-6	–
Eyes	Dual	Single	Dual	Single

In total, database C contains $66,391$ images out of which more than 98% images are used for testing. Figure 2-8 shows sample images from the dataset.

Multi-sensor VIS Iris Database D

All the images used in the previous datasets are captured using near infrared sensors. However, researchers have also explored the usability of visible spectrum images for iris recognition. Therefore, we have performed the experiments on visible spectrum databases as well to show the effectiveness of the algorithm. The databases used for this purpose are:

- UBiris v1 [261] database is collected in two different sessions and contains total 1876 images. The main importance of this database is the presence of different kinds of noise in the images. It is collected using a Nikon E5700 camera model.
- UBiris v2 [262] is captured in unconstrained environment. It contains 11,101 images captured from 522 different classes.
- Miles research database [4] is collected using a self developed eye camera and a total of 816 images are available.
- UPOL database [6, 89] contains RGB images captured using a SONY DXC-950P 3CCD camera. For each eye, three images are captured and the total number of classes is 128.

In total, database D contains $14,177$ images out of which 800 images are used for training and the remaining (approximately 94.3%) of the images are used for testing. Sample images from this set are shown in Figure 2-9.

Table 2.5: Summarizing the fingerprint database characteristics.

Database	Sensors	Sensor Models	Data	Train Data	Test Data
FVC 2002 [214]	4	(a) Identix (Optical) (b) Biometrika (Optical) (c) Precise Biometrics (Capacitive) and (d) SFinGe v2.51 (Synthetic)	3,200	800	2,400
FVC 2006 [54]	4	(a) CrossMatch (Optical) (b) Digital Persona (Optical) (c) Atmel (Thermal-sweeping) and (D) SFinGe v3.0 (Synthetic)	6,720	800	5,920
IIIT-D MOLF [291]	4	(a) Lumidigm Venus IP65 Shell (Optical) (b) Secugen Hamster-IV (Optical) (c) CrossMatch L-Scan Patrol (Optical) and (d) Latent (CMOS sensor)	16,400	800	15,600
CASIA Cross Sensor [1]	3	(a) UrU 4000B (Optical) (b) Authentec AFS-II (Capacitive) and (c) Symwave sw6800 (Capacitive)	3,000	600	2,400

From the multi-sensor iris database A, 800 images pertaining to the first 20 subjects are used for training while 1931 images corresponding to the remaining 48 subjects are used for testing. From database B, 900 images pertaining to the first 30 subjects are used for training and 2490 images of the remaining 74 subjects are used for testing. Both training and testing databases contain the images from all the sensors. Similarly, from databases C and D, first 200 images corresponding to each sensor are used for training and remaining images are used for testing.

2.5.2 Fingerprint Databases

The results of the proposed algorithm are shown by combining 4 publicly available databases containing fingerprint images from 15 different sensors. These databases are Fingerprint verification competition (FVC) 2002 [214] and 2006 [54], IIIT-D MOLF [291], and CASIA multi-sensor [1] database. Table 2.5 summarizes the characteristics of all four databases and Figure 2-10 shows sample images from these databases.

- *FVC 2002* [214] is collected using four different sensors. We are using set A of this database for our experiments.

- *FVC 2006* [54] is also acquired from four different fingerprint sensors and we are using Set A of this database.

- *IIIT-D MOLF* database [291] is captured using five different capturing methods. in this research, we are using data corresponding to four sensors. It contains 16,400 images collected from 100 different subjects.

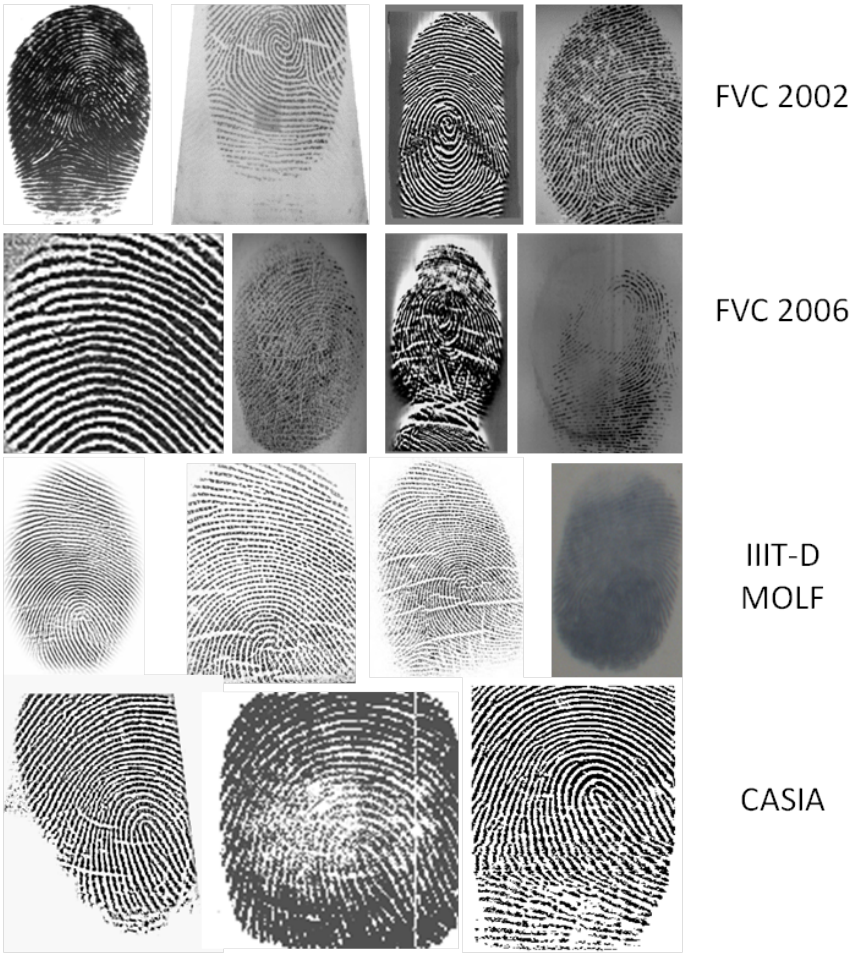


Figure 2-10: Sample fingerprint images from the four databases - every image is collected from a different sensor.

- *CASIA Cross Sensor* database [1] is collected using three different sensors. 10 fingerprint impressions per person from 100 different fingers are captured for this database.

Combining all four databases, the experiments are performed with $17,960$ images collected from the optical scanner, $2,800$ images from capacitive, $4,400$ images from CMOS, $1,680$ images from thermal sweeping, and $2,480$ synthetically generated images. From each database, 200 images corresponds to particular sensor are used for training the SVM model and the remaining images are used for testing. As shown in Table 2.5, out of the total $29,320$ images, $3,000$ images are used for training and the remaining (around 90%) images are used for testing.

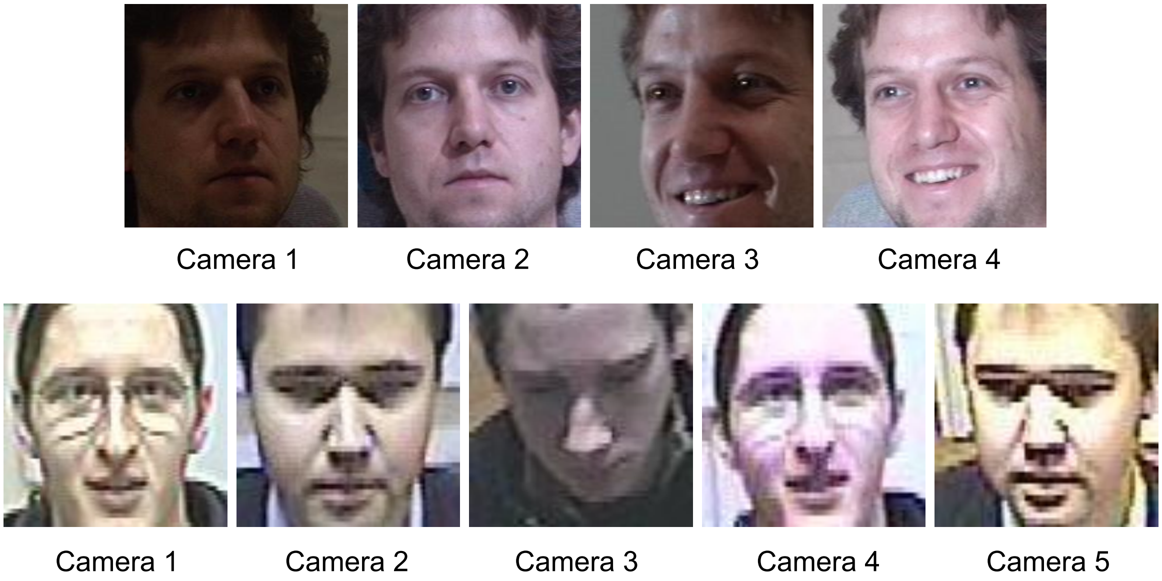


Figure 2-11: Face images of Multi-PIE and SCFace database. The images varies in terms of cameras, quality, illumination, expression, and pose.

2.5.3 Face Databases

The proposed algorithm, which is an amalgamation of features, have also been utilized for the camera classification of face images. For this purpose we have utilized two databases namely SCFace [126] and Multi-PIE [127]. The Multi-PIE database is captured using 15 sensors, and we have used the images of 4 sensors containing frontal and semi-frontal faces. The images also vary in terms of illumination and expression. The SCFace database contains low-quality images captured using surveillance cameras. The database in total contains images from 8 cameras: five out of which are visible spectrum cameras, whereas, remaining three are infra-red cameras. For camera classification, we have used the images acquired using visible spectrum cameras. The face images of both databases are shown in Figure 2-11.

The images of both the databases are divided into independent subject training and testing sets. In total, the Multi-PIE database consists of $7,611$ images, out of which 60% images are used for training and 40% images for testing. The SCFace database used for camera classification has $1,939$ images. Similar to Multi-PIE, the database is divided into 60% and 40% training and testing subject independent subset, respectively. Both the databases are also combined, where the training set of both is used for training the sensor classifier, and the testing set is used for reporting the detection accuracy.

2.5.4 Existing Algorithms for Comparison

We have compared the performance with following sensor classification algorithms.

1. **Blind identification of source cell-phone model** [59]: Binary Similarity Measure (BSM) is extracted between the bit planes. Along with the BSM features, higher order wavelet statistics and image quality measures are also calculated for identification of the camera. Each image is decomposed using the quadrature mirror filter up to four levels and mean, variance, skewness, and kurtosis of wavelet coefficients in the three orientations are calculated. For image quality measure, the reference image is obtained by smoothing the original image with a Gaussian filter of size 3×3 and sigma 0.5. After combining these three feature vectors, SFFS feature selection is applied to reduce the dimensionality of the feature vector. Finally, probabilistic SVM is used both with feature and decision level fusion.

2. **PCA-based denoising of sensor pattern noise for source camera model identification** [180]: This paper proposes to use Sensor Pattern Noise (SPN) as features for camera identification. Each image is decomposed into wavelet sub-bands and subtracted from its corresponding Wiener filtered images. Sub-bands of the wavelet decomposed image is then converted into the spatial domain and PCA is applied on this filtered image. The reference pattern for each sensor is calculated using the mean of the sensor images after getting the reduced dimension images. Normalized cross correlation is calculated to identify the sensor of the test images from each of the reference sensor feature vector.

3. **Identifying sensors from iris images using pixel non-uniformity** [156]: In this research, the sensor identification techniques are applied based on the Pixel Non-Uniformity (PNU) pattern of each camera sensor. The reference pattern for each sensor is generated by subtracting the original images of each sensor from its wavelet denoised images. One reference pattern for each camera is computed by taking the mean of the reference pattern of all the images of that sensor. Pearson's product moment is calculated between the test reference pattern and sensor reference pattern.

Table 2.6: Performance of individual components of the proposed sensor classification algorithm on multisensor iris databases A and B.

Algorithm	Feature Dimension	Classification Accuracy %	
		Database A	Database B
Block Image Statistical Measure (BISM)	40	97.2	85.2
Texture Measure (TM)	13	98.8	89.2
Single level Multi Orientation Wavelet Texture (SlMoWT)	52	99.6	93.3
Entropy on RDWT level 1	4	62.1	74.5
Entropy on RDWT level 2	4	67.2	58.1
High Order Wavelet Entropy (HOWE)	8	74.2	86.2
Image Quality Measures (IQM)	21	99.6	99.7
Proposed - 1	134	**100**	**100**

2.6 Experimental Results for Sensor Classification

The results of the proposed sensor classification algorithm are evaluated on all biometric modalities described above. The experiments are performed with and without the proposed feature selection: the results and analysis without feature selection (referred to as Proposed - 1) and then with proposed feature selection (termed as Proposed - 2). First, The detailed experimental analysis about individual databases, components of the algorithm, comparison with existing algorithms is performed on iris modality. Later, the proposed algorithm's performance is also evaluated on the fingerprint and face databases and discussed in brief.

2.6.1 Iris Sensor Classification Results

Since the proposed algorithm is a combination of four features, we have analyzed the performance of each feature set individually and also in combination.

Analysis of Proposed Algorithm without Feature Selection (Proposed - 1)

Table 2.6, Table 2.7 and Table 2.8 report the sensor classification accuracy and confusion matrix on both the in-house databases, A and B, for individual features and combination of multiple features. We first evaluate the performance of individual features involved in the proposed algorithm. The results are computed by optimizing SVM with grid search over different types of kernels and their parameters. We have evaluated the results of linear kernel, RBF kernel with sigma parameter varying from 1 to 10, and polynomial kernel of degrees 1 to 10.

Table 2.7: Confusion matrix on iris database A using the features individually and the combined features.

Features	Device	CG4	CM	Delta	Vista	Total
BISM	CG4	471	12	0	1	484
	CM	19	459	1	7	486
	Delta	0	5	474	1	480
	Vista	0	0	0	481	481
	Total	490	476	475	490	1931
SlMoWT	CG4	484	0	0	0	484
	CM	3	481	0	2	486
	Delta	0	1	479	0	480
	Vista	0	2	0	479	481
	Total	487	484	479	481	1931
HOWE	CG4	333	94	0	57	484
	CM	184	277	0	25	486
	Delta	27	0	453	0	480
	Vista	121	160	0	200	481
	Total	665	531	453	282	1931
IQM	CG4	480	4	0	0	484
	CM	4	482	0	0	486
	Delta	0	0	480	0	480
	Vista	0	0	0	481	481
	Total	484	486	480	481	1931
Proposed - 1	CG4	484	0	0	0	484
	CM	0	486	0	0	486
	Delta	0	0	480	0	480
	Vista	0	0	0	481	481
	Total	484	486	480	482	1931

It can be observed from Table 2.6 that the classification accuracy on set A is higher compared to set B. Among all three individual features, Single level Multi-orientation Wavelet Texture (SlMoWT) features provide the best classification accuracy of 99.6% and 93.4% on databases A and B respectively. Block Image Statistical Measure (BISM) features also provide high accuracy on database A, however, on database B, the accuracy reduces to 85.2%. The entropy of RDWT for both level-1 and level-2 provides very low accuracies in the range of 58.1% to 74.5%. The image quality measure provides 99.6% and 99.7% accuracy on databases A and B respectively. However, on combining the four sets of features, the accuracy improves to 100% on both database A and database B. The results show that even though a combination of features increases the dimension-

Table 2.8: Confusion matrix on iris database B using the three features individually and the combined features.

Features	Device	Cogent	IriShield	CrossMatch	Total
BISM	Cogent	727	115	44	886
	IriShield	210	645	21	876
	CrossMatch	157	86	485	728
	Total	1094	846	550	2490
SlMoWT	Cogent	734	62	90	886
	IriShield	98	748	30	876
	CrossMatch	79	9	640	728
	Total	911	819	760	2490
HOWE	Cogent	815	20	51	886
	IriShield	167	645	64	876
	CrossMatch	256	97	375	728
	Total	1238	762	490	2490
IQM	Cogent	886	0	0	886
	IriShield	2	868	6	876
	CrossMatch	0	0	728	728
	Total	888	868	734	2490
Proposed - 1	Cogent	886	0	0	886
	IrisShield	0	876	0	876
	CrossMatch	0	0	728	728
	Total	886	876	728	2490

ality to 134, it gives perfect classification accuracy on both the databases.

Analyzing the confusion matrix in Table 2.7 and Table 2.8 shows the number of images for which the algorithms incorrectly predict the acquisition device. It can be observed that individual features primarily misclassified the images captured by CG4 and CM interchangeably whereas the images captured from the other sensors are generally correctly classified. Since sensor classification is a pre-processing step in the entire recognition pipeline, it is important that the algorithm is accurate and requires less time. On using the combined features with the proposed algorithm, out of 1931 images in database A, no image is misclassified. Similarly, in database B, all images are correctly classified to their respective sensor.

As mentioned previously, we have also compared the results with several existing sensor classification algorithms. Using the same experimental protocol, Table 2.9 summarizes the classification accuracies on both the databases along with time required on a 2.2 GHz $i7$ desktop with

Table 2.9: Comparing the performance of the proposed algorithm on iris databases with some existing algorithms. Top value in each column is highlighted.

Algorithm	**Classification Accuracy** (%)			**Average Time (Seconds)**
	Database A	**Database B**	**Database C**	
Çeliktutan et al. [59]	68.4	99.6	85.2	13.0
Li et al. [180] Reference pattern on 128 × 128	63.1	58.2	52.4	0.7
Li et al. [180] Reference pattern on 256 × 256	66.6	57.9	58.0	0.8
Kalka et al. [156]	43.8	52.3	37.6	6.9
HOG+SVM [155]	93.1	89.3	86.2	0.10
Proposed - 1	**100**	**100**	**98.6**	**0.35**

Table 2.10: Sensor classification results using SVM for VIS multi-sensor iris database D.

Algorithm	**Color Channel**				
	RGB	**R**	**G**	**B**	**Gray Scale**
Çeliktutan et al. [59]	57.0	–	–	–	20.0
Li et al. [180] - Reference pattern on 128 × 128	–	–	–	–	31.2
Li et al. [180] - Reference pattern on 256 × 256	–	–	–	–	52.3
Kalka et al. [156]	–	–	–	–	53.4
HOG+SVM [155]	86.8	84.5	85.7	86.0	82.2
Proposed - 1	**99.8**	**99.2**	**99.0**	**99.4**	**98.8**

8GB RAM in Matlab programming environment. We have also computed iris sensor classification performance when HOG features [155] are extracted and SVM is used for classification. The results show that the proposed algorithm is not only highly accurate but also computationally very fast. On the other hand, existing algorithms are either accurate but time consuming or fast but not accurate. For instance, the algorithm of Çeliktutan et al. [59] yields similar accuracies on database B but requires significantly large amount of time.

We next show the results on database C, which contains iris images captured in the NIR spectrum. Database C contains more than 66,000 images out of which more than 98% of the images are used for testing and the remaining for training. To the best of our knowledge, in the iris sensor classification literature, no one has presented experimental results with such a large database. Table 2.9 summarizes the classification results of the proposed and existing algorithms for sensor classification. The table shows that due to lot of variations in large database, the accuracy of both proposed algorithm and Çeliktutan et al. [59] reduces. However, with combination of texture and image quality features, the sensor classification accuracy improves significantly. This shows that image quality measures efficiently encode the quality of images captured from different sensors and therefore are good features to be used for sensor classification.

For the unconstrained environment, iris image acquisition in visible spectrum is gaining significant importance these days. Mobile devices are also being equipped with cameras which work in the visible spectrum and hence can be used for iris data acquisition. Therefore, the last experiment is performed on the visible spectrum iris database - D. To the best of our knowledge, this is the first research presenting results on visible spectrum iris sensor classification. Table 2.10 shows the classification results and comparison with some existing feature extraction technique used for general sensor classification. Since the sensors capture color images, we present the results with individual color channels and grayscale images. The results show that computing the features from RGB images yields better accuracy than individual channels including gray scale. We observed that blue channel is better compared to the other two color channels for iris sensor classification.

Analysis of Proposed Algorithm with Feature Selection (Proposed - 2)

The performance of the proposed SVM fitness function based bacteria foraging feature selection algorithm is compared with six existing approaches:

- **Mutual information (MI)** [257] is based on the selection of the most relevant features using mutual information between the features and class labels. This feature selection algorithm computes the score of the features to show the usefulness of features with respect to the class.

- **Statistical dependency (SD)** [257] measures the correlation between the features and its class label for feature selection. The main goal of statistical dependency is to check whether the feature values are random or have a correlation with the corresponding class label.

- **Random subset feature selection (RSFS)** [257, 275] is based on finding the subset of features which is more useful than the average number of features available. This technique randomly selects a subset of features and calculates the classification performance using k-nearest neighbor classifier. RSFS provides the relevance value to each feature based on the classification performed using a subset of features in each iteration.

- **Sequential forward selection (SFS)** [345] selects the most important feature in each iteration and adds it in the selector. Once the feature is selected, it cannot be removed. The feature set that yields highest accuracy is selected.

- **Sequential floating forward selection (SFFS)** [263] uses the basic strategy of sequential forward selection and after some n number of desired features are added to the selector, features with the least significance are removed. The selection algorithm continues iteratively until the best feature selection criteria is not satisfied.

- **Feature selector based on genetic algorithm and information theory (GF)** [202] performs feature selection based on the maximization of the mutual information and genetic learning approach. This is one of the closest existing approach compared to the proposed feature selection algorithm.

Using the same experimental protocols and databases, Table 2.11 shows the results of different feature selection algorithms on the proposed features. For the proposed algorithm, we observe that 50 features are sufficient for iris sensor identification on databases A, B, and C. Comparison with the combined set of features (134) shows that on database C, proposed feature selection yields maximum accuracy of 100% while reducing the dimension as well. Comparison with existing algorithms shows that the genetic learning based approach yields the second best performing results with 100 selected features. Overall, Table 2.11 shows the superior performance of the proposed algorithm compared to existing algorithms. On the multi-sensor VIS iris database D, the proposed feature selection approach yields an accuracy of 99.85% with 50 features and 99.9% with 100 features. Compared to without feature selection approach (Proposed - 1), it is a slight improvement (from 99.8% to 99.85% for RGB input) but the dimensionality is significantly reduced, i.e. from 402 (134×3 features for RGB channels) to 50.

Table 2.12 shows the number of features selected from individual feature sets using different feature selection techniques on the database A. The table shows that for most of the techniques, BISM and SlMoWT features are more discriminative than others, followed by IQM features. This shows that both intensity based and wavelet based features are important in iris sensor classification.

Table 2.11: Sensor classification results obtained by applying different feature selection techniques and SVM classifier on the proposed feature set.

Algorithm	**No. of Top Selected Feature Dimension**	**Classification Accuracy %**		
		Database A	**Database B**	**Database C**
Sequential floating forward selection†	3, 4, and 5	99.0	99.5	58.3
Sequential forward selection†	3, 6, and 5	99.0	99.8	58.3
Mutual Information	50	**100**	**99.9**	92.0
Mutual Information	100	**100**	99.9	98.7
Statistical dependency	50	**100**	**99.9**	92.9
Statistical dependency	100	99.9	99.9	98.7
Random subset feature selection†	45, 59, and 23	100	99.6	98.7
Genetic based	50	**100**	99.6	98.9
Genetic based	100	**100**	99.9	99.0
Proposed - 2 (Bacteria Foraging)	50	**100**	**100**	**100**
Proposed - 2 (Bacteria Foraging)	100	**100**	**100**	**100**

†The number of selected features are mentioned for each database A, B, and C respectively.

Table 2.12: Number of features selected from the proposed feature sets by the features selection techniques on multi-sensor database A.

Algorithm	**Total features selected**	**BISM (40)**	**HOWE (8)**	**SlMoWT (52)**	**TM (13)**	**IQM (21)**
SFFS	3	0	0	1	0	2
SFS	3	0	0	1	0	2
Mutual Information	50	3	0	29	9	9
Mutual Information	100	31	6	37	11	15
Statistical dependency	50	3	1	28	8	10
Statistical dependency	100	31	6	37	11	15
RSFS	45	15	2	19	5	4
Genetic based	50	27	0	12	2	9
Genetic based	100	40	8	40	3	9
Proposed - 2 (Bacteria Foraging)	50	25	1	12	2	10
Proposed - 2 (Bacteria Foraging)	100	38	6	40	4	12

2.6.2 Fingerprint Sensor Classification Results

Table 2.13 shows the classification results of the proposed algorithm and comparison with existing sensor classification techniques [32, 155, 180]. Among these existing algorithms, only Bartlow et al. [32] is related to fingerprint sensor identification. The results show that the proposed algorithm yields the classification accuracy of above 96% and is almost 29% higher than the next best performing algorithms. Further, Table 2.13 also shows that individual features are at least 4.4% less accurate than the mélange of features. We also observe that if we remove any of the feature set, then the accuracy is reduced. Further, in the proposed algorithm, selection of wavelet decomposition and mother wavelet can have a significant impact on the performance. The proposed algorithm uses RDWT with DB1 mother wavelet. We also performed comparison with RDWT and DWT

Table 2.13: Fingerprint sensor classification results of the proposed and existing algorithms on the multisensor database.

Algorithm	**Classification Accuracy (%)**
HOG [155]	67.18
Li et al. [180] (128 × 128)	34.33
Li et al. [180] (256 × 256)	50.68
Bartlow et al. [32]	51.84
Statistical features	90.44
Entropy	52.85
Haralick over image	76.32
Haralick (images + RDWT)	92.08
IQM	90.79
Proposed	**96.52**

with two other mother wavelets: DB9 and biorthogonal 9/7. We observed that the classification accuracy with all the combinations varies within 1.5% of each other and RDWT + DB1 yields the best results.

2.6.3 Face Sensor Identification Results

The proposed algorithm yields $\mathbf{87.45\%}$ 'five-class' sensor classification accuracy on SCFace database using radial basis function (RBF) kernel. On the other hand, the accuracy of the proposed algorithm for 'four-class' camera identification on the Multi-PIE database is $\mathbf{94.51\%}$. On the combined database (Multi-PIE and SCFace) which contains images from 'nine-cameras' is used for sensor classification, the proposed algorithm achieves $\mathbf{91.81\%}$ detection accuracy. We have also implemented texture features such as LBP and fine-tuned ResNet-18 for sensor identification. The proposed algorithm outperforms the ResNet-18 classifier by 3% and LBP + SVM combination by 6%.

2.7 Summary

Sensor classification or source camera identification is an important research challenge that has been addressed in the context of cameras, scanners, and fingerprint sensors. However, to the best of our knowledge, classification of biometric sensors is still an open research problem. In this research, we propose a novel biometric sensor classification algorithm, consisting of three main steps: (i) multiple features are extracted which are based on image statistical properties, along with image quality, entropy, and Haralick features from RDWT sub-bands, (ii) SVM fitness function based bacteria foraging feature selection is used to select the optimal feature set, and (iii) SVM is used for sensor classification. The results show that the proposed algorithm yields more than 99% accuracy on the multi-sensor iris databases. Apart from that, the proposed feature extraction algorithm yields an accuracy of over 96% and 91% on fingerprint and face images, respectively. The proposed algorithm computationally efficient, which makes it effective for real-time applications. We have also prepared a multi-sensor iris database that contains more than 6000 images corresponding to multiple sensors. It is our assertion that availability of such databases will promote research on sensor classification and interoperability.

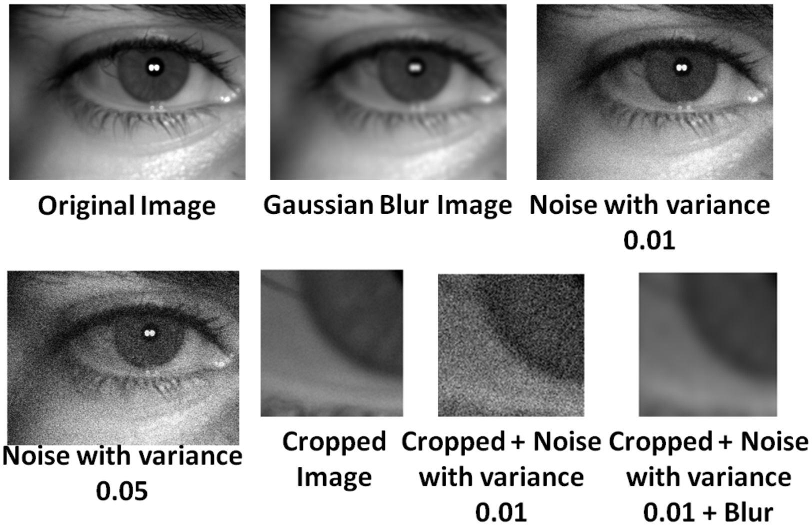
Appendix

Other than iris sensor identification, we have studied the effect or sensor interoperability on iris images using the collected database. Further, the resiliency of the proposed sensor identification algorithm is evaluated under the manipulation of images from different image artifacts such as noise and missing information.

Sensor Classification with Manipulated Images

For mobile applications that involve image transfer, the output image can be noisy due to the communication channel. Therefore, it is important that the sensor classification algorithm should be able to identify the sensor even in presence of such artifacts in the image. Figure 2-12 shows sample images where the Gaussian noise and blur are added to the image with varying mean and variance parameters of Gaussian noise. In this research, we have tested the performance with some specific attacks:

- **Gaussian Blur** is a commonly used smoothing method to minimize the noise in images. We have applied a Gaussian blur with significantly large $\sigma = 5$ which can reduce detailed information needed for sensor classification.



Original Image

Gaussian Blur Image

Noise with variance 0.01

Noise with variance 0.05

Cropped Image

Cropped + Noise with variance 0.01

Cropped + Noise with variance 0.01 + Blur

Figure 2-12: Sample images from database A after manipulation with artifacts.

- **Gaussian Noise**: We have used it as an attack and performed experiments with two parameter settings. Figure 2-12 shows the outputs with different noise variations.
  - $\mu = 0$ and $\sigma^2 = 0.01$ (low level of noise)
  - $\mu = 0$ and $\sigma^2 = 0.05$ (high level of noise)
- **Cropped and Gaussian Noise**: There can be instances when the complete image is not available. To emulate such a scenario, we have cropped a part of the original image from the center of size, 120×120, and added Gaussian noise with $\mu = 0$ and $\sigma^2 = 0.01$.
- **Cropped, Gaussian Noise, and Blur**: The last experiment is performed with a combination of individual artifacts. The images can be affected by hybrid effects and in that scenario, the proposed algorithm should also be robust enough to work effectively. To perform this attack first the original sample is cropped from the center of size 120×120 and then Gaussian noise with $\mu = 0$ and $\sigma^2 = 0.01$ is added followed by the Gaussian blur with $\sigma = 5$.

Figure 2-13 shows the result of iris sensor classification with image manipulations on database A. The results show the reliability of the proposed algorithm on attack images where some manipulations are performed to deceive the system. The proposed algorithm achieves perfect iris sensor classification on original images but under these attacks, the accuracy drops down. Gaussian noise and blur have the similar effect on the sensor classification performance. Gaussian blur and noise

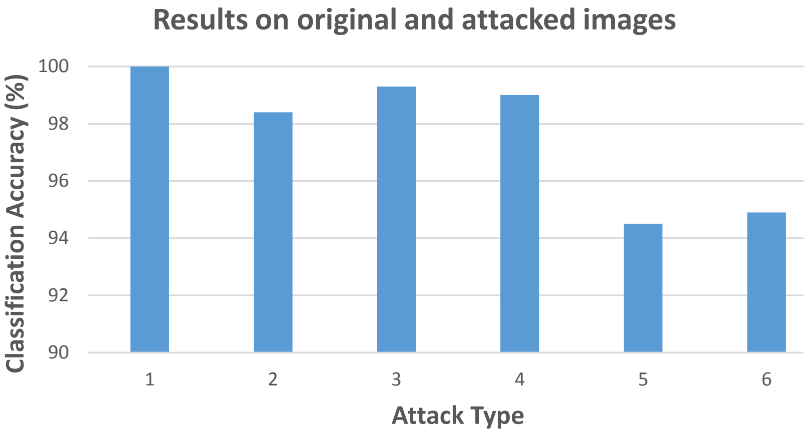


Figure 2-13: Sensor classification accuracy of the proposed algorithm (Proposed-2) for database A after image manipulation: (1.) Original image, (2.) Blur $\sigma = 5$, (3.) Noise $\mu = 0$ and $\sigma^2 = 0.01$, (4.) Noise $\mu = 0$ and $\sigma^2 = 0.05$, (5.) Crop and Noise $\mu = 0$ and $\sigma^2 = 0.01$, (6.) Crop and Blur $\sigma = 5$ and Noise $\mu = 0$ and $\sigma^2 = 0.01$

Table 2.14: Sensor classification accuracy (%) using the proposed and existing state-of-the-art algorithms for multi-sensor database A after artifacts addition.

Artifacts	**Proposed - 2**	**HOG +SVM [155]**	**Çeliktutan et al. [59]**
Original images	**100**	93.1	68.4
Gaussian blur with sigma 5	**98.4**	92.7	68.8
Gaussian noise with mean 0 & variance 0.01	**99.3**	95.5	82.4
Gaussian noise with mean 0 & variance 0.05	**99.0**	93.4	92.3
Gaussian noise with mean 0 & variance 0.01 on 120 × 120 size image cropped from center	**94.5**	85.1	71.2
Gaussian noise with mean 0 & variance 0.01 and Gaussian blur of sigma 5 on 120 × 120 image cropped from center	**94.9**	83.1	85.4

have the maximum performance difference of 0.9% in the proposed framework. When the attack with the noise of low and high level is performed the proposed algorithm achieves an accuracy of 98.9% and 98.6% respectively. The combinations of attacks degrade the sensor classification performance more than the individual attacks. On combining all three attacks: crop, Gaussian noise, and blur, the sensor classification performance reduces by 5.1% as compared to original image performance.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Boosting Face Presentation Attacks Detection in Multi-Spectral Videos Through Score Fusion of Wavelet Partition Images

3.1 Introduction

Face recognition algorithms have been gaining more interest than ever, both for their increasing usage [132, 299] and their limitations [58, 111, 218, 234, 278, 350]. While researchers are attempting to make the face recognition algorithms generalizable to unseen scenarios, it is important that the face presentation attack detection (PAD) algorithms are also generalizable and inclusive. Existing research in face presentation attack detection has primarily focused on detecting different kinds of attacks captured in visible spectrum. While the algorithms have received nearly perfect classification performance on individual attacks, the focus is towards designing unified algorithms for different kinds of presentation attacks in visible images [33, 194, 225].

The usage of face recognition algorithms in near infrared spectrum is also increasing. With surveillance cameras operating in both visible and NIR images, it is important to that the PAD algorithms are developed for NIR spectrum face images as well. Fig. 3-1 shows sample real and print attack images in visible (VIS) and near-infrared (NIR) spectrum. It can be observed that the illustration of attacks in both the spectra are quite different, and hence attack detection in

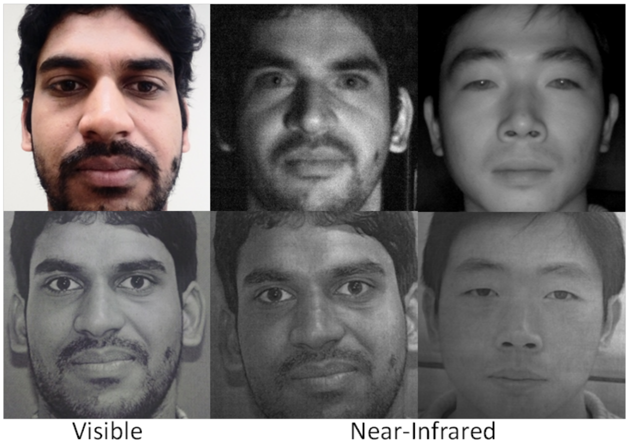


Figure 3-1: Real and attack samples in VIS and NIR spectrum. First row is the real samples and second row is the attack samples.

images of different spectrums require specialized algorithms. Realizing this, recently researchers have started working towards PAD in NIR spectrum images as well [16, 192, 369]. This research focuses on extending the usability and efficiency of presentation attack detection algorithms in multiple spectrums and ethnicity variations.

3.1.1 Related Work

In this section, we present a brief overview of the existing algorithms in multi-spectrum (visible + near infrared) presentation attack detection. Face presentation attack detection in other than visible spectrum is still in nascent stages. Pavlidis and Symosek [248] presented an algorithm to detect disguised faces system using images captured in NIR spectrum. In 2014, Yi et al. [361] performed experiments with print attack by printing the photo of 100 clients on coarse paper in both VIS and NIR spectrum. However, the database was not made publicly available. Later, Chingovska et al. [67] and Raghavendra et al. [267] also analyzed the vulnerability of face recognition systems in NIR spectrum. They showcased that face recognition systems working in NIR spectrum are also susceptible to presentation attacks such as replay and print. Therefore, effective presentation attack detection algorithms are required to protect the surveillance system operating in NIR spectrum. Chingovska et al. [67] also released the first publicly available VIS and NIR presentation attack database (MSSPOOF). The database comprises images from 21 subjects. Three photos of each client in VIS and NIR spectrum are selected and printed using the black and white printer. The database contains 70 real access images and 144 attack images for each client. They reported more than 88% Spoof False Accept Rate (SFAR) using NIR attack on NIR face recognition system. However, no counter presentation attack algorithm is presented in this paper.

Later, Raghavendra et al. [266] proposed a presentation attack detection algorithm on the MSSPOOF (multispectral spoof) database using the combination of Laplacian pyramid and Fourier transform. They reported an Average Classification Error Rate (ACER) of 2.1% and 0.74% on VIS and NIR spectrum, respectively. Raghavendra et al. [267] prepared the extended multispectral presentation attack database (EMSPAD) from 50 subjects and attack samples were prepared by printing the images from two different printers. The database is captured in seven different spectral bands ranging from 425nm to 930nm. In total, the database contains $7,000$ print attack images and $3,500$ real access images from 50 subjects. A vulnerability of the face recognition system using printed attack samples is shown with perfect SFAR in 680nm spectra and more than 98% SFAR in 930nm spectra. Liu and Kumar [192] have proposed different CNN architectures to detect the presentation attack using multi-spectral face images. They have prepared the presentation attack database from 13 masked subjects and 9 real subjects. The limitation of the work is the unavailability of the database. Sun et al. [317] proposed consistency measure based presentation attack detection in multi-spectral imaging. Bhattacharjee et al. [42] have shown the vulnerability of VIS, NIR, and Thermal face recognition using deep CNN models under custom silicone mask attacks. George et al. [109] presented a new database along with multi-channel CNN for face PAD.

Heusch et al. [140] have presented a database in multiple spectrum to effectively present an study on the effect of SWIR imaging on presentation attack detection. Zhang et al. [369] have prepared the multi-spectral database comprising of 2D print attack. Along with that the squeeze and excitation network using ResNet-18 as a backbone network is also proposed to counter presentation attacks. Li et al. [178] extended the database by incorporating multiple attacks including 3D print and silica gel mask. The other popular face presentation attack databases are SiW [193], SiW-M [194], and OULU-NPU [47], however all are captured in visible spectrum. The details of the existing face presentation works can also be find in the comprehensive evaluation [150] and handbook [41, 219]. Other than multi-spectral presentation attack in face recognition, researchers have also explored fusion of multiple biometric modalities for effective attack resistant system. Wild et al. [346] perform the fusion of fingerprint and face. Similarly Tan and Schuckers [320] performed the fusion of fingerprint ridge and valley noise information. Bhardwaj et al. [39] have combined the physical and behavioral characteristics of fingerprint for better protection of multimodal system against presentation attacks. Gragnaniello [125] and Abhyankar and Schuckers [8] have analysed the input image in fourier and wavelet domain to extract discriminative features to check the liveness in fingerprint recognition system. Samartzidis et al. [289] have explored the ultra-violet spectrum to acquire face images as future biometrics. Recently, Spinoulas et al. [311] have presented a multi-spectral PAD biometric system.

3.1.2 Research Contributions

The literature review shows that there are few publicly available databases in the NIR spectrum and a significant lack of research in automated algorithms for presentation attack detection in the NIR spectrum images and algorithms that can detect PAD in multiple spectrum. Inspired by these, there are two primary contributions of this research. The first contribution is that we have prepared a large video-based attack database in the NIR spectrum from more than 340 subjects of two different ethnicities: Indian and Chinese. The second contribution of this paper is a state-of-the-art face presentation attack detection algorithm for different kinds of attacks in the NIR + VIS spectrum. The strength of the proposed algorithm can be seen through extensive experiments. The proposed algorithm not only yields effective performance with spectrum variations but also generalized against attack mediums, ethnicities, and databases. Experimental evaluation on multiple databases show that the proposed algorithm surpasses several deep learning, non-deep learning, and state-of-the-art algorithms by a significant margin.

3.2 Proposed Spoof-in-NIR Database

As discussed in the previous section, there are only two small databases in NIR spectrum. Therefore, to promote the research in this important problem, we first prepared a large video PAD database in NIR spectrum, termed as the Spoof-in-NIR database. In this section, we present the details of the proposed NIR presentation attack database. The NIR database contains face images from two completely different ethnicities: Indian and Chinese, which makes the database first of its type to cover multiple annotated ethnicities. The database contains the attack videos captured

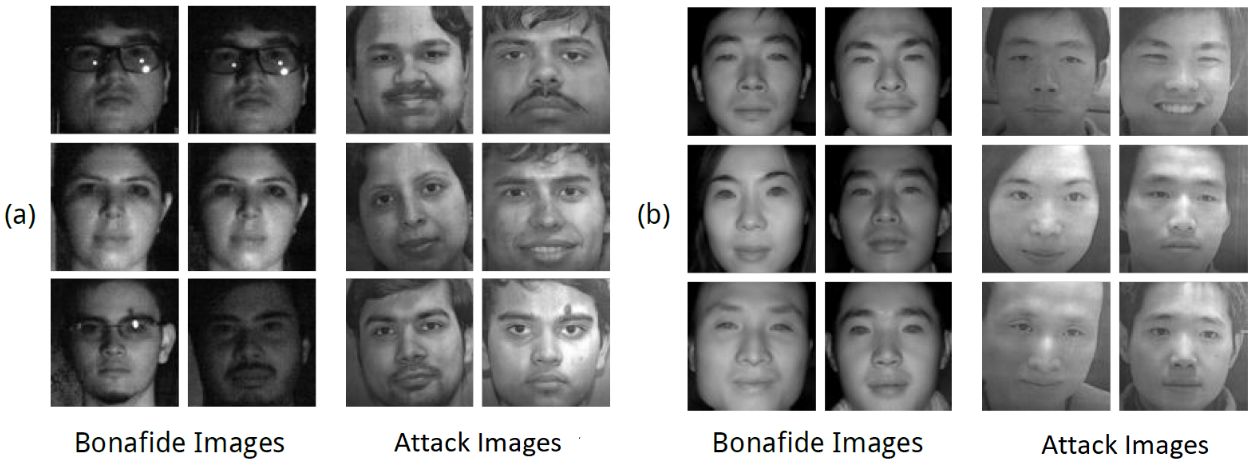


Figure 3-2: Sample face images from the proposed Spoof-in-NIR database. Images are shown from (a) Indian ethnicity and (b) Chinese ethnicity.

using the printed photograph of 400 genuine users. The database will be made publicly available to the research community[1].

3.2.1 Camera Setup

To collect the database in NIR spectrum, a camera is mounted on the tripod and subjects are asked to stand at a distance of approximately one meter and look into the camera. To ensure that the videos are captured in a relatively uncontrolled scenario, no other special instructions are given to the subjects. The database is captured in two different environments and background conditions: one inside and other is outside the building in the night time. To ensure that the videos are only captured in NIR spectrum, visible cut filter is placed in front of the camera. GO-5000-USB camera[2] is used to capture the videos at the frame resolution of $2,048 \times 2,560$. Frames are captured at the rate of 20 fps and stored as raw pixel so that the quality of images is not degraded because of compression.

3.2.2 Indian Face Presentation Attack Database

Each ethnicity subset comprises two sets: bonafide and attacked. The bonafide/real videos of the Indian database are captured from 152 subjects at night time using an NIR camera. To provide

[1] http://iab-rubric.org/resources.html

[2] http://www.jai.com/en/products/go-5000-usb

Table 3.1: Characteristics of proposed NIR face presentation attack database. *From NIR-VIS 2.0 database [181].

Ethnicity	Type	Sessions	Subjects	Videos	Images	Faces
Indian	Real	1	152	152	36,480	32,629
	Attack	2	150	300	72,000	70,221
Chinese	Real*	4	725	–	12,469	12,469
	Attack	1	98	98	7,840	7,799

variability in the data, videos are captured at two different locations and comprise variations in background and illumination conditions. The subjects also perform natural motions such as eye blinking and head movement.

To capture the attacked videos, frontal images of all the subjects are first captured using a DSLR camera in day time constrained environment. A black and white printout of frontal images of 150 subjects is taken using an HP color printer and videos of these images are captured to prepare the attacked video subset. Attack videos of Indian subset are captured in two different sessions with illumination variations. Attack and real videos are captured for a duration of 12 seconds to exhibit a real world surveillance scenario.

In total, 300 attack and 152 real videos are collected as part of this Indian database. The resolution of the real and attack frame is $2,048 \times 2,560$. Face detection is performed using Viola-Jones face detector [335]. Characteristics of the database are given in Table 3.1 and samples are shown in Figure 4-6(a).

3.2.3 Chinese Face Presentation Attack Database

To prepare the attack database with Chinese ethnicities, real visible spectrum frontal images of 98 subjects are randomly selected from the NIR-VIS 2.0 database [181]. Similar to the Indian database, these images are also printed on normal A4 paper using HP color printer. These prints are then placed on a fixed medium to capture the attack videos. 98 videos comprising $7,840$ images form the NIR attacked subset while real NIR samples from 725 individuals acquired from the CASIA NIR-VIS2.0 database [182] comprise the bonafide Chinese subset. In total, Chinese database contains $12,469$ real face images and $7,799$ attack face images. Characteristics of the database are given in Table 3.1 and sample images are shown in Figure 4-6(b).

Table 3.2: Protocol for intra database experiments.

Ethnicity	**Session**		**Folds**	**Results Reported**
	Real	**Attack**		
Indian	1	1	15	Video and Frame
	1	2	15	Video and Frame
Chinese	1	1	5	Frame

3.2.4 Experimental Protocol

To facilitate benchmarking the performance of different algorithms on this database, we propose two protocols: (1) intra-database and (2) cross-database.

Intra Database Protocol: Indian NIR dataset is divided into 15 folds, where at a time one fold is used for training the linear SVM classifier for attack detection, and remaining folds are used for evaluating the classifier. In the Indian subset, attack videos are captured in two different sessions, therefore the results for both the sessions are calculated separately. To report the results of a particular session, attack videos captured in that session are used. Similarly, Chinese NIR database is divided into five random folds. Due to the unavailability of the videos in the real subset of Chinese NIR dataset, only frame based results are calculated. For Indian NIR database both video based and frame based protocols are designed. In video-based, every video is classified as real or attack and in frame-based an individual entity of a video (frame) is classified as real or attack. The protocol is also summarized in Table 3.2.

Cross Database Protocol: The proposed NIR attack database is captured from two different ethnicities: Indian and Chinese, and hence can be utilized for cross database experiments. The proposed presentation attack database is the only available large database captured in NIR which can be utilized for cross database experiments. Cross database experiments are the true way to assess the robustness of the attack detection algorithm for better real world scenarios. The experiments are performed in two folds: 1) Presentation attack model is trained over the Indian NIR database and tested on Chinese NIR database and 2) Presentation attack model is trained over the Chinese NIR database and tested on Indian NIR database.

In the first experiment, training database (i.e. Indian NIR) is randomly divided into 15 folds, and each time one fold is used for training the classifier and testing is done on the complete Chinese NIR database. Average EER of 15 folds is reported for frame based detection. Since no real videos are there in Chinese database, video based results are not reported.

In the second experiment, training database (i.e. Chinese NIR) is randomly divided into five folds, and for learning the presentation attack detection model at a time, one fold is utilized. The learnt model is evaluated on the complete Indian NIR database. Average equal error rate (EER) across all folds in both video and frame based mode is reported.

3.2.5 Vulnerability of Face Recognition Against Attack

Researchers have illustrated that several publicly available commercial face recognition systems are prone to presentation attacks [16, 344]. To show the effect of the proposed NIR face database on the Commercial-Off-the-Shelf (COTS) systems, face identification experiment is performed. The gallery images are single frontal images of each subject. For instance, the gallery for Indian NIR attack database comprises 150 images, one of each subject. Each attack video is used as probe to perform the identification using FaceVACS[3]. Figure 3-3 shows the identification performance using COTS. It is important to observe that with all the attacked images as probe, COTS yields the rank-1 identification accuracy of 100% for the Chinese NIR subset while Indian NIR attack videos show more than 98% identification accuracy at rank-1. Face identification experiment thus shows the vulnerability of existing face recognition system against the attack in NIR spectrum.

3.3 Proposed Presentation Attack Detection Algorithm

Face presentation attack detection algorithms proposed in literature are primarily based on encoding texture measures and image artifacts such as moiŕe pattern [247], local binary patterns [45, 211, 273], and more recently, deep learning algorithms have been proposed [61, 209, 332]. A recent study [150] shows that hand-crafted image feature-based algorithms have the potential of handling multiple challenging presentation attacks at the time being computationally feasible. Therefore, in this chapter, we present a presentation attack detection algorithm based on combining the image features extracted from global as well as local facial regions. Raw images provide a global view of the real and attack data. This global information can help in extracting discriminative information such as foreground-background inconsistency and attacking medium boundary.

[3] http://www.cognitec.com

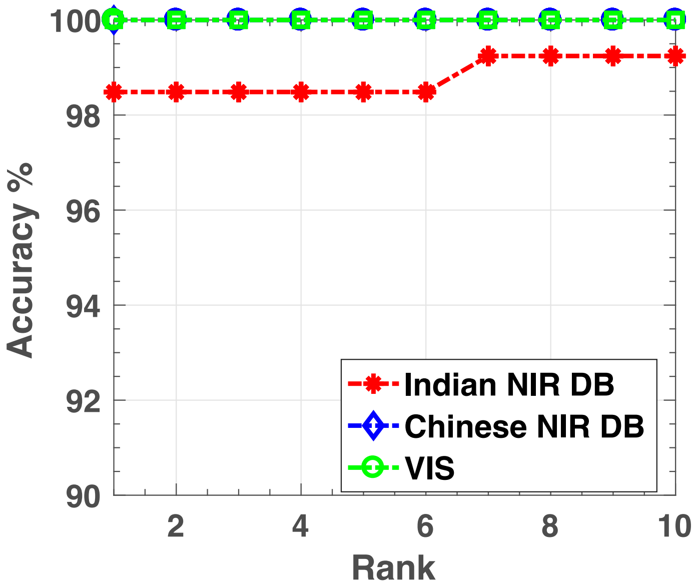


Figure 3-3: CMC curves of COTS for face identification on the proposed presentation attack database

While local facial regions may help in extracting textural information/artifact of the face. We hypothesize that since the attack images are mostly recaptured from the camera, they can be different in high-frequency content from its counterpart i.e., real face.

The proposed presentation attack detection algorithm encodes these assertions to differentiate between real and attacked images. Figure 3-4 illustrates the steps involved in the proposed algorithm. To highlight the low and high frequency for discrimination, the first step of the algorithm involves applying the wavelet decomposition on the input image/frame. Wavelet provides the multi-resolution decomposition of the frame and highlights the edges in multiple directions: horizontal, vertical, and diagonal [100]. Discrete wavelet transform (DWT) downsamples the subbands to $M/2 \times N/2$, where M and N are the height and width of the image respectively. Since downsampling leads to loss of information, we apply redundant discrete wavelet transform (RDWT). RDWT preserves the image size by creating subbands of the same size as the input image, thereby generating overcomplete representations.

As mentioned earlier, the effect of presentation attack may be visible in either facial regions or even as abnormalities in relation to foreground and background regions. Therefore, wavelet decomposition is applied in two steps. In the first step, RDWT decomposition is applied on the

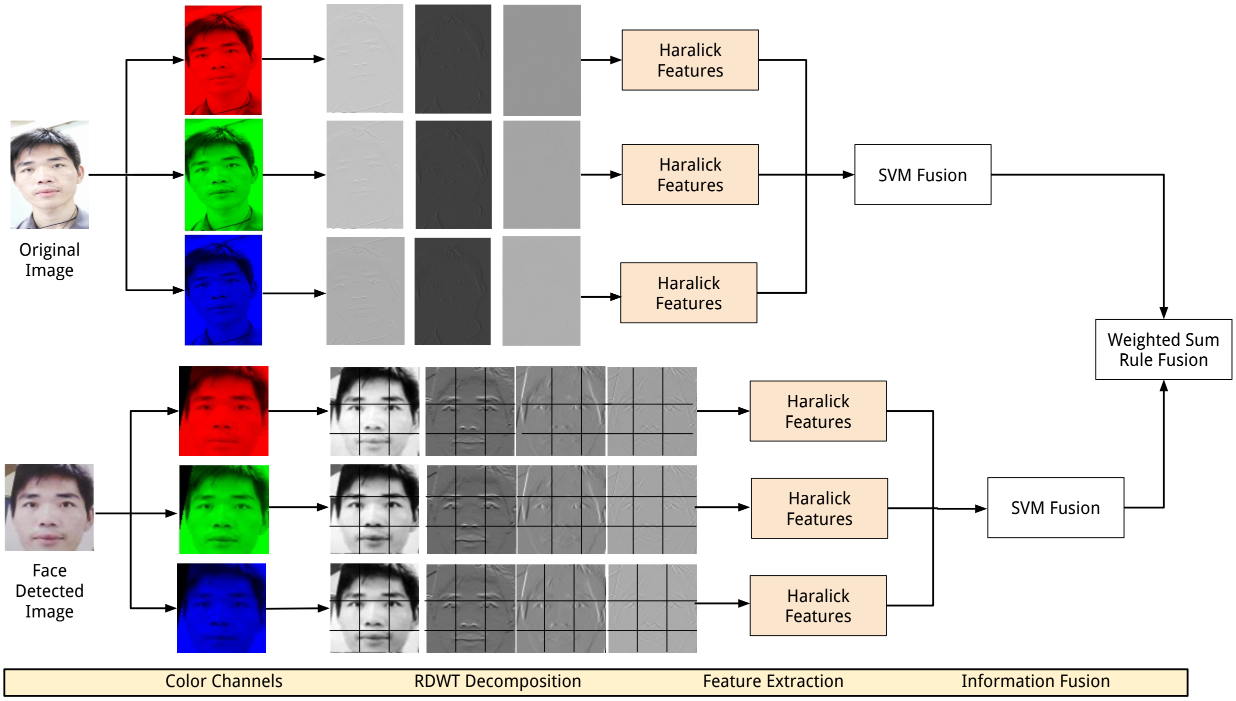


Figure 3-4: Illustrating the proposed presentation attack detection pipeline.

complete input image without face detection, while the second step involves first applying face detection followed by tessellating the facial region into nine non-overlapping facial patches.

Global information, which is directly computed from the raw images, is computed without tessellating the frames into multiple blocks. To compute the information from local texture, the face region is first divided into multiple blocks, and then each block is decomposed using wavelet filter. The patches generated are of the fixed size of 32×32. The textural features are then extracted by applying Haralick features [136] on the RDWT decomposed subbands, which can encode the image distortion based information present in the recaptured images such as intensity distribution and homogeneity.

The basic building block of Haralick features is the Co-occurrence Matrix, which contains the information of counts of how many times a certain pixel value is in the neighborhood of other possible pixel value. The Haralick features used in this research are listed in section 2.3.1. The Haralick features are extracted from each color channel in RGB image. While the Haralick features are used to encode the textural information, RDWT provides the high-frequency information presented in multiple orientations. For classification, linear Support Vector Machine (SVM) [334] classifier is used. Two different SVM classifiers are trained: one for the features extracted from the raw sensor images and second on the features extracted from the face region. The final score of a testing frame is calculated as the weighted sum of the scores obtained from two different trained SVM. To summarize, the steps involved in the proposed algorithm are discussed below:

1. Input image is decomposed into individual RGB channels

2. Each channel is then decomposed into four wavelet sub-bands using Redundant Discrete Wavelet Transform (RDWT) [100],

3. Haralick texture features are computed over each wavelet sub-bands and the original image without decomposition,

4. Haralick features obtained from each subband are concatenated and input to a linear Support Vector Machine (SVM) classifier for classification,

5. The face region is cropped using the eye-coordinates obtained using the Viola-Jones face detector,

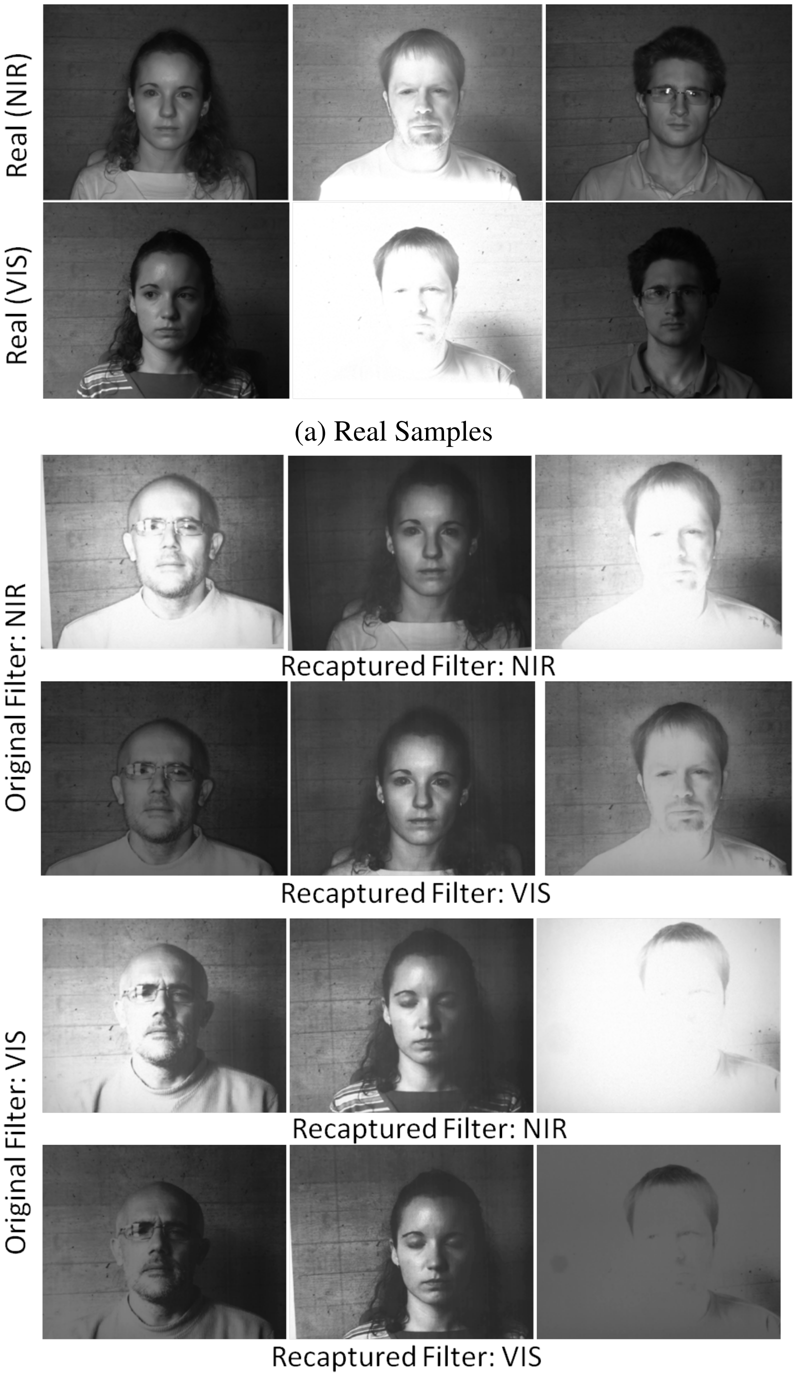
6. Face region is divided into nine non-overlapping blocks,

7. Each face block is decomposed into four wavelet sub-bands using Redundant Discrete Wavelet Transform,

8. Haralick texture features are computed over each wavelet sub-bands and the original face block without decomposition,

9. Haralick features obtained from each subband and original face are concatenated and input to a linear Support Vector Machine (SVM) classifier for classification,

10. The scores obtained from SVM in Steps 4 and 9 are fused using weighted sum rule fusion and then thresholded for classification.



(a) Real Samples

(b) Attack Samples. Original filter means the image used to perform attack is captured in that spectrum and Recaptured filter means the attack image is recaptured in that spectrum.

Figure 3-5: Real and Attack samples of MSSPOOF database.

3.4 Results on Multispectral PAD Databases

The performance of the proposed algorithm is computed on the proposed Spoof-in-NIR database along with the publicly available multispectral MSSPOOF database [67] and CASIA-SURF database [369]. The results are reported in terms of Bonafide (real) Presentation Classification Error Rate (BPCER), Attack Presentation Classification Error Rate (APCER), and Average Classification Error Rate (ACER) also known as Half Total Error Rate (HTER). BPCER is defined as the rate of the bonafide (real) samples classified as attack samples. APCER is the proportion of the attack samples classified as bonafide (real) samples. The results are also reported using Equal Error Rate (EER). EER is the error at which APCER and BPCER are equal, and HTER is defined as the average of APCER and BPCER.

Further, first, the results of the proposed algorithm on existing MSSPOOF and CASIA-SURF databases along with comparisons to state-of-the-art (SOTA) algorithms are reported. Later, the experiments performed on the proposed Spoof-in-NIR database are described. The comparison of the proposed algorithm with SOTA algorithms highlights the efficacy of PAD in multiple spectrum.

3.4.1 Results on MSSPOOF Database

MSSPOOF database contains images with print attack performed in four different ways: 1) original image is captured in visible spectrum, while visible image is recaptured using visible and near infrared spectrum camera and 2) original image is captured in near-infrared spectrum while NIR image is recaptured back using visible and near infrared filter camera. Figure 3-5 shows sample images from the MSSPOOF database. The database contains a total of 630 real images in visible and 624 images in near infrared spectrums pertaining to 21 subjects, in seven different environmental conditions. The database is divided into three disjoint sets: train, dev, and test. Table 3.3 summarizes the characteristics of the MSSPOOF database and the number of images in train, dev, and test splits are given in Table 3.4. Raghavendra et al. [266] proposed two different protocols on the MSSPOOF database:

1. *Individual spectrum*: NIR and VIS, individual spectrum data in training and testing

2. *Combined spectrum*: both spectrum data in training and testing.

Table 3.3: Characteristics of MSSPOOF print attack samples.

Original Filter	**Recaptured Filter**	**Images**
VIS	VIS	756
	NIR	756
NIR	VIS	756
	NIR	756

Table 3.4: Images in each set of the MSSPOOF database.

Data		**Dev**	**Train**	**Test**
Bonafide (Real)	VIS	180	270	180
	NIR	179	265	180
Print Attack	VIS	218	324	216
	NIR	216	312	216

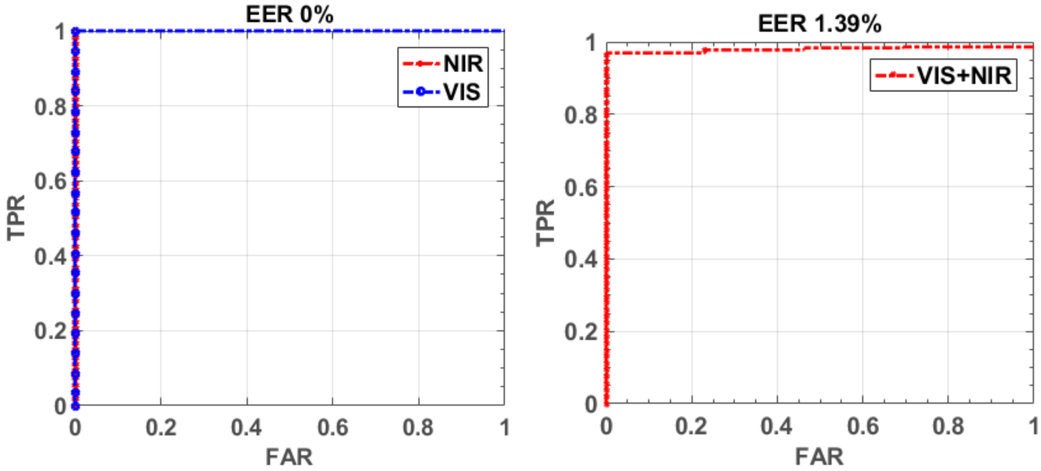
Table 3.5: Results (%) on individual and combined spectrum set of MSSPOOF database. *Results are taken from [266].

Spectrum	**VIS**			**NIR**			**VIS + NIR**		
Algorithm	**APCER**	**BPCER**	**ACER**	**APCER**	**BPCER**	**ACER**	**APCER**	**BPCER**	**ACER**
LBP-SVM* [83]	2.31	11.67	6.99	0.46	8.33	4.39	2.54	7.77	5.16
BSIF-SVM* [265]	5.55	4.44	5.00	4.16	2.22	3.19	3.47	3.33	3.40
LPQ-SVM* [211]	5.55	0.55	3.05	0.92	4.44	2.68	1.85	4.44	3.14
DoG-SVM* [163]	62.03	28.88	45.46	37.03	38.54	37.79	43.05	43.61	43.33
GLCM-SVM* [343]	**0**	97.22	48.61	**0**	96.08	48.04	0.00	98.05	49.02
L$_a$MT$_i$F [266]	4.16	**0**	2.08	0.92	0.55	0.74	3.00	2.50	2.75
Proposed	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

We have reported the results using these two protocols along with the experiments defined in this chapter to fully utilize the characteristics of the database. In the individual spectrum experiments: the presentation attack detection SVM model is trained using individual spectrum data such as for visible spectrum experiment, the partition belongs to visible spectrum is used for training and evaluating the classifier. In the combined spectrum experiment, the train set belonging to both the spectra is used to learn the classifier and similarly for evaluation, test sets belongs to both spectra are combined.

Results on Individual Spectrum Experiments

The results for individual spectrum experiments are summarized in Table 3.5. The proposed algorithm achieves 0% ACER and BPCER in both the spectrums which shows the consistency of the algorithm across spectrums. The second best performing algorithm is by Raghavendra et al. [266]


(a) Individual spectrum.

(b) Combined spectrum.

Figure 3-6: ROC curve of the proposed presentation attack detection algorithm on the MSSPOOF database.

(L$_a$MT$_i$F) and it achieves 0.74% and 2.08% ACER in NIR and VIS spectrum respectively. L$_a$MT$_i$F uses a combination of Laplacian pyramid based decomposition to extract the high-frequency information and Short Term Fourier Transform (STFT) for time and frequency features. The limitation of the L$_a$MT$_i$F algorithm is the high error rate for attack detection. Table 3.5 also shows results of several other texture based algorithms and it can be observed that the algorithms are generally ineffective in detecting either the bonafide presentation or attack presentation. The proposed algorithm is robust to both kinds of data, which is desired and required for real-world presentation attack detection algorithms integrated with face recognition systems. The misclassification of bonafide data as attack data can frustrate the genuine user because of the need to give face data again and again for recognition. At the same time allowing attack data as bonafide data can cause serious harm to the face recognition system. Figure 3-6(a) shows the perfect EER (i.e. 0%) of the proposed algorithm on MSSPOOF database.

Results on Combined Spectrum Experiments

In the combined spectrum experiment, the data belonging to both the spectrums are utilized to make a joint decision. To learn the presentation attack detection model, training set given in the database for both the spectrums is used and evaluation is performed on the test set of both the

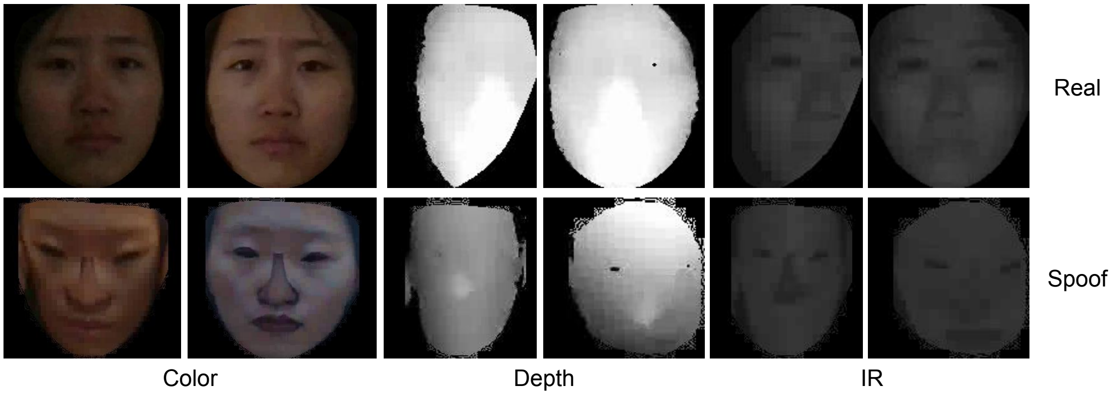


Figure 3-7: Real and spoof samples of CASIA-SURF database.

Table 3.6: Error rates (%) of the proposed and baseline algorithm on the CASIA-SURF database. (–) means the values are not provided in the baseline paper [370].

Type	Algorithm	Modality	EER	APCER	BPCER	ACER
Fused	Baseline	Color&IR	–	14.4	1.6	8.0
	Proposed		10.0	10.1	9.9	10.0
	Baseline	Color&Depth	–	4.3	5.6	5.0
	Proposed		1.8	1.7	1.9	1.8
	Baseline	Depth&IR	–	1.5	8.4	4.9
	Proposed		2.0	1.9	2.1	2.0
	Baseline	Color&Depth&IR	–	3.8	**1.0**	2.4
	Proposed		1.5	**1.6**	1.4	**1.5**

spectrums collectively. Table 3.5 shows the results of the proposed and existing algorithms on the combined set.

The proposed algorithm achieves 1.22% ACER on the combined spectrum database which is more than 55% better than the second-best algorithm L$_a$MT$_i$F. The results show that the proposed algorithm is robust towards different kinds of data. Similar to individual spectrum results, combined spectrum shows the ineffectiveness of the existing algorithm in detecting attack or bonafide data. The GLCM features yield the perfect error rate in detecting the presentation attack but rejecting the bonafide almost all the time. Similarly, the second best algorithm yields 3% APCER. ROC curve of the proposed algorithm for joint spectrum attack detection is shown in Figure 3-6(b). The proposed algorithm yields 1.39% EER on the joint spectrum dataset.

3.4.2 Experiments on CASIA-SURF Database

Recently proposed CASIA-SURF database [369] is one of the most extensive database for face presentation attack detection problems both in terms of modalities and subjects. The database consists of $21,000$ videos of 1000 subjects in RGB (color), IR, and depth modalities. Sample images of real and spoof class in each modality are shown in Figure 3-7. For each subject, one real video is captured while six fake videos are captured using eye, mouth, nose region cut of the flat and curved printed face. For example, in attacks 1 and 2, the eye region is cut from the flat and curved face photo. Similarly, in other attacks, either the eyes and nose or eyes, nose, and mouth, all portion is cut from flat and curved face photo. The color, depth, and IR videos are acquired using the Intel RealSense SR300 camera. The real faces are first printed out using an A4 color printer and later used by the attackers while exhibiting real life motions such as turn left or right, move up or down, walk-in or away from the camera. The database is divided into training, validation, and testing set and contains $148,089$, $48,789$, and $295,644$ cropped images, respectively.

In this research, we have used the pre-defined protocol for evaluation, and the results are reported on fused modalities. The baseline algorithm [370] consists of the ResNet-18 model as a backbone model where the first three ResNet blocks are used for feature extraction. The features from each modality, i.e., color, depth, and IR, are then fused using squeeze and excitation module. In the end, two blocks of ResNet are used for discriminative features learning, followed by global average pooling. The whole pipeline is trained using a softmax classifier. The comparison of the proposed algorithm with the baseline algorithm in terms of error rates are given in Table 3.6.

The scores computed over different modalities are fused using a weighted sum. The weight parameter across different modalities is learned, which yields the lowest EER on the validation set. The fusion of all three modalities, i.e., color, depth, and IR, outperforms the baseline algorithm by 37.5% (i.e., reduces from 2.4% to 1.5%) in terms of ACER. The APCER of the baseline algorithm (i.e., 3.8%) is more than two times than the proposed algorithm (i.e., 1.6%). The proposed fusion of color and IR with depth modality surpasses the baseline algorithm in identifying bonafide (i.e., real) images by at-least 66%. Other than the APCER of the proposed algorithm is significantly better than the recently proposed face PAD algorithm by Zhang et al. [369]. For example, when the fusion of color and IR data is performed, the APCER of the proposed and existing algorithm [369]

Table 3.7: Video and frame based EER ($\mu \pm \sigma$)% on the proposed NIR print attack database in Intra-database scenario.

Ethnicity	**Session**	Algorithm	**Video**	**Frame**
Indian	1	ResNet-18	12.8 ± 3.5	24.1 ± 5.8
		Proposed	4.3 ± 2.3	19.7 ± 1.9
	2	ResNet-18	10.2 ± 0.9	29.3 ± 2.3
		Proposed	0.8 ± 0.6	23.2 ± 1.6
Chinese	1	ResNet-18	-	27.4 ± 3.0
		Proposed	-	20.7 ± 0.8

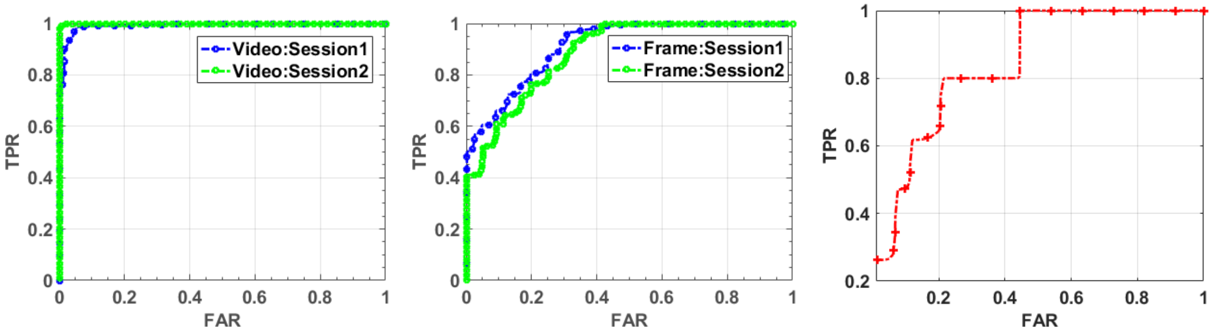
Table 3.8: Video and frame based EER ($\mu \pm \sigma$)% on the proposed NIR print attack database in cross-database scenario.

Train Set	**Test Set**	Algorithm	**Frame**	**Video**
Chinese	Indian(Session 1)	ResNet-18	51.4 ± 1.5	58.2 ± 0.95
		Proposed	48.4 ± 0.35	51.8 ± 0.24
	Indian(Session 2)	ResNet-18	51.2 ± 1.5	54.8 ± 1.35
		Proposed	49.2 ± 0.10	45.4 ± 0.89
Indian(Session 1)	Chinese	ResNet-18	52.0 ± 1.36	–
		Proposed	46.6 ± 1.20	–
Indian(Session 2)		ResNet-18	51.6 ± 1.05	–
		Proposed	46.9 ± 0.50	–

is 14.4% and 36.5%, respectively. The fusion of all modalities in the proposed algorithm yields 1.6% APCER, whereas, the existing algorithm yields 1.9% APCER. These superior performances of the proposed algorithm on one of the largest multi-modal presentation attack database depicts the strength of the proposed algorithm in identifying physical fake face data.

3.4.3 Experiments on Proposed Spoof-in-NIR Database

The proposed Spoof-in-NIR database contains images and videos of subjects from two different ethnicities: Indian and Chinese. Indian NIR database contains 152 real videos and 300 attack videos in two sessions. Chinese NIR database contains $12,469$ real frames taken from CASIA VIS-NIR 2.0 [181] and $7,799$ attack frames. According to the protocol described in Section II.D, the following experiments are performed on the proposed Spoof-in-NIR database: (i) intra-database and (ii) cross-database. For comparison, a ResNet-18 model [137] trained on ImageNet [87] is fine-tuned for PAD. The model is fine-tuned for 10 epochs using the Adam optimizer and learning rate set to 0.0001.



(a) Video based (b) Frame based (c) Frame based

Figure 3-8: ROC curve of presentation attack detection on proposed NIR spectrum database using proposed algorithm. (a) and (b) are on Indian ethnicity and (c) is Chinese ethnicity database.

Intra Database Experiments: Results of the proposed algorithm on the Spoof-in-NIR database are summarized in Table 3.7 and Figure 3-8. The proposed algorithm achieves 4.3% and 0.8% EER on session 1 and 2 of the Indian NIR spectrum dataset respectively for video based experiment. Similarly, for frame based experiment an EER of 19.7% and 23.2% is achieved for sessions 1 and 2 respectively. On the Chinese NIR dataset, the proposed algorithm yields 20.7% EER using frame based experiment. The reported results show that the classification of an individual frame is difficult as compared to video, where the information from multiple frames help in improving the results. Comparing the result with MSSPOOF database further shows the proposed database is more challenging. Availability of this database to the research community will further improve the state-of-the-art of presentation attack detection in multiple spectrums.

Cross Database Experiments: Table 3.8 shows the results of cross database which can also be seen as cross-ethnicity experiments on Spoof-in-NIR database. Session 1 and 2 of Indian NIR database represent which session's attack data is used for training/evaluation. The EER on the Indian NIR database, when trained with the Chinese set, is 51.8% and 45.4% for video based classification in session 1 and 2 respectively. Similarly, when the attack detection model is trained on the Chinese ethnicity subset, the proposed algorithm yields 48.4% and 49.2% EER for frame based classification. On the Chinese database, frame based EER are 46.6% and 46.9% when Session 1 and Session 2 data of Indian NIR is used for training the classification model, respectively. The high EER in cross-database experiments shows the challenging nature of presentation attack detection in NIR spectrum, which needs to be addressed properly using large scale attack databases.

Table 3.9: Characteristics of the existing VIS spectrum attack database used in this research

Database	**Attack**	**Unconstrained**
CASIA-FASD	Print and Replay	✓
Replay-Attack	Print and Replay	✓
MSU-MFSD	Print and Replay	✓
3DMAD	3D Hard Resin Mask	×
MSU USSA	Print and Replay	✓
SMAD	Silicone Mask	✓
WFFD	3D Wax Figure	✓

The proposed database is the first of its kind attempt to increase the research in this direction.

3.5 Experiments on Existing VIS Spectrum Databases

To further illustrate the effectiveness of the second contribution of this chapter i.e., development of robust algorithm across different attacks, additional experiments are performed on the existing benchmark databases in VIS spectrum. The performance of the proposed algorithm is evaluated on the CASIA-Face Anti-Spoofing Database (FASD) [372], Replay-Attack [66], MSU-MFSD [344], 3D Mask Attack Database (3DMAD) [93], MSU USSA databse [247], Silicone Mask Attack Database (SMAD) [217], and 3D wax figure face database (WFFD) [151]. These databases cover a wide spectrum of attacks such as print, photo, a replay of video, 3D hard resin mask, and the most challenging silicone mask. Table 3.9 summarizes the characteristics of these databases and a brief description of each is provided below.

CASIA-FASD database [372] contains three different kinds of attacks: cut photo (eye portions are cut to perform the eye blink), warped photo (to make it cylindrical as a real face), and replay of a video. It contains the videos in three different image qualities: low, normal, and high. Replay-Attack database [66] is captured in controlled and adverse environments. In the controlled environment, the background was kept fixed and the fluorescent lamp was used for illumination. In the adverse environment, the background is random and natural light is the source of illumination. MSU-MFSD database [344] is captured from 35 subjects and is one of the mobile face attack databases. Real videos are captured from two different devices: built-in camera of MacBook Air 13-inch laptop and a front facing camera of a Google Nexus 5 Android phone. To capture the attack, two different high-resolution cameras are used: Canon 550D single-lens reflex camera and an iPhone 5S back facing camera.

CASIA-FASD, Replay-Attack, and MSU-MFSD databases are challenging but contain 2D attacks only. To assess the effectiveness of the algorithm on the 3D attack, the 3DMAD database [93] is also used in this chapter. Advancement in the 3D reconstruction and 3D printer makes the availability of 3D mask easier. These masks can be worn and can effectively hide the identity of the person in day to day life. These masks are hard to detect in comparison to wearing the photo paper masks. 3DMAD database is captured from 17 subjects where each subject is wearing a different 3D mask. It is captured in three sessions, where the real access videos are captured in first two sessions and the third session covers the 3D mask attack. For each subject ten real and five 3D attack videos are captured, with total of 255 videos in the database. While the 3DMAD database is the challenging 3D mask attack database but it has some limitation. The masks used to prepare the database are hard resin masks which do not allow movements similar to natural face. In the real world, some challenging cases are found where the robbers have used the silicone masks to hide the identity from the surveillance cameras. These silicone masks are the soft mask which can properly fit the face and can move with the face. Manjani et al. [217] have prepared the Silicone Mask Attack Database (SMAD) which is currently the most challenging kind of attack to detect.

To tackle the limitations such as diversity in terms of background, illumination, and image quality, Patel et al. [247] have prepared one of the largest database. Such a database is essential to obtain generalizable and robust anti-spoofing methods, particularly in face unlock scenarios on smartphones. To create such a database we selected 1,000 live subject images of celebrities from the Weakly Labeled Face Database[4]. The public set of the MSU USSA database for face anti-spoofing consists of $9,360$ images (out of which $1,040$ are real images and $8,320$ spoof attack images) of $1,040$ subjects. To perform the experiments standard database protocol of 5 fold cross validation is performed.

Recently, a new era of 3D attack is highlighted in 3D attacks where wax figure faces are used as a possible adversary on face recognition systems [151]. The authors have shown that state-of-the-art face recognition algorithms such as OpenFace [19] and Face++[5] are vulnerable to wax

[4] http://wlfdb.stevenhoi.com

[5] https://www.faceplus plus.com/face-compare-sdk/

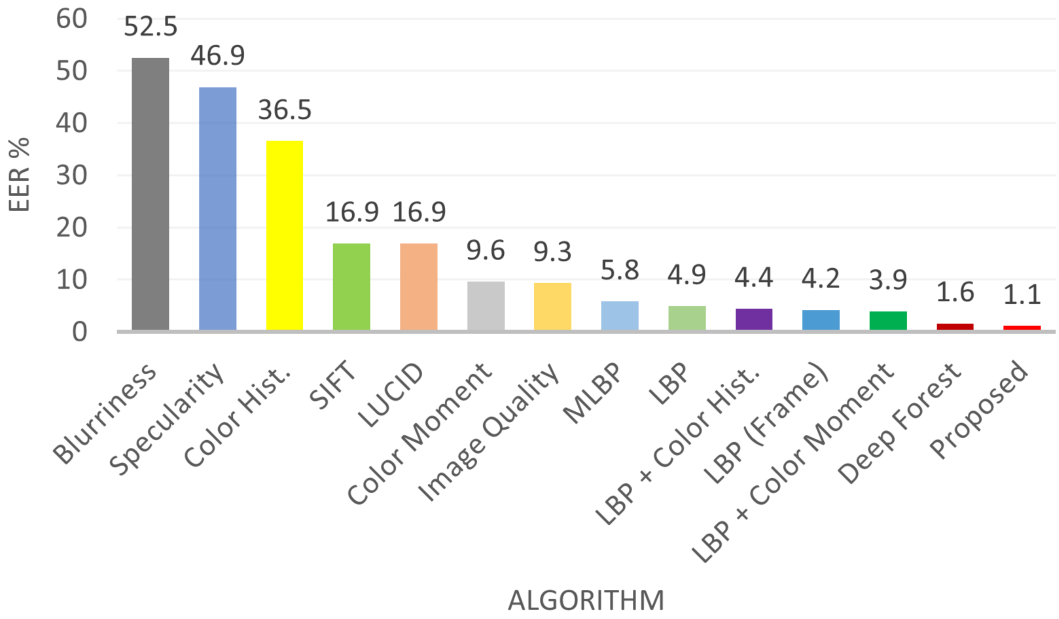


Figure 3-9: Comparison with state-of-the-art results on the MSU USSA database for presentation attack detection.

figure faces. These spoof faces have achieved at-least 92% Impostor Attack Presentation Match Rate (IAMPR) across multiple protocols. Therefore, the identification of wax faces from real faces is important and challenging because of properties similar to real faces. In this research, we have used two working conditions (protocols) provided by the authors. In the first protocol (Prot. 1), images captured under different recording devices and environments are used, whereas, in another protocol (Prot. 3), images captured in different and same recording devices and environments are combined. Protocol 1 consists 600 train, 200 development, and 440 test images; while, protocol 3 consists $1,320$ train, 440 development, and 440 test images.

To compare the results with existing state-of-the-art algorithms, original protocol of each database is followed and the results are reported both in terms of intra-database and cross database scenarios. Table 3.10 shows the comparison of the proposed algorithm with existing algorithms in video based attack detection. On one of the most challenging presentation attack i.e. silicone mask, the proposed algorithm outperforms the state-of-the-art performances [217, 301]. EER and HTER of the proposed algorithm on video-based detection are 7.7% and 6.9% which is more than 37% and 47% lower, respectively. Similarly, on CASIA-FASD database, the proposed algorithm gives the lowest EER value of 0.92%. Perfect EER on Replay-Attack, MSU-MFSD, and 3DMAD

Table 3.10: Comparison with state-of-the-art results on the video based presentation attack detection. Best result of existing algorithms is underlined.

Algorithm	**CASIA-FASD**	**Replay-Attack**		**MSU-MFSD**	**3DMAD**	**SMAD**	
	EER	**EER**	**HTER**	**EER**	**EER**	**EER**	**HTER**
CNN [228] (2015)	–	–	0.75	–	–	–	–
IDA [344] (2015)	12.9	–	7.41	8.58	–	–	–
Spectral Cubes [254] (2015)	14.0	–	2.8	–	–	–	–
DMD + LBP + SVM [327] (2015)	21.8	5.3	3.8	–	–	–	–
Multicue Fusion [247] (2016)	5.88	–	14.6	8.41	–	–	–
Color Texture [45] (2016)	3.2	0.0	3.5	3.5	–	–	–
CNN (2017) [171]	–	–	0.8	–	–	–	–
C-SURF + Fisher Vector [46] (2017)	2.8	0.1	2.2	2.2	–	–	–
Deep Dictionary [217] (2017)	1.3	–	0.0	–	0.0	12.3	13.1
LGBP + GS-LBP [249] (2017)	2.53	–	3.13	8.54	–	–	–
Directional LBP [264] (2017)	4.44	–	4.88	3.33	–	–	–
Frame Diff + Fisher Score + LPQ (2017) [29]	4.62	5.60	4.80	2.50	–	–	–
Depth and patch CNNs [26] (2017)	2.67	0.79	0.72	–	–	–	–
Ultra-deep Neural Network [333] (2017)	1.22	1.03	1.18	–	–	–	–
Skin Blood Flow [341] (2017)	7.01	–	4.92	7.23	–	–	–
Multiscale quality [360] (2018)	12.7	–	5.38	–	–	–	–
Temporal Texture [242] (2018)	6.71	–	0.6	10.07	–	–	–
Motion CodeBook [91] (2018)	17.0	–	5.7	17.0	3.53	–	–
Texture Markov Feature [368] (2018)	8.0	4.0	4.4	7.5	–	–	–
3D CNN [179] (2018)	1.4	0.3	1.2	0.0	–	–	–
Locally Specialized CNN [134] (2018)	4.44	0.33	1.75	–	–	–	–
CNN + STN+ MIL [190] (2018)	–	–	1.8	–	–	–	–
Deep Dynamic Texture [301] (2019)	–	–	–	–	0.0	14.9	11.7
GFA-CNN [332] (2019)	–	–	–	7.5	–	–	–
Spoof Buster [49] (2019)	–	–	5.50	–	–	–	–
2-stream ResNet-18 + Attention [61] (2019)	3.15	0.21	0.39	–	–	–	–
Multi-Regional CNN [209] (2020)	–	–	1.6	–	–	–	–
CCoLBP+Ensemble Learning [250] (2020)	3.33	–	4.00	5.00	–	–	–
Color Texture Weighted Features [309] (2020)	7.34	2.32	7.39	–	–	–	–
SFDSF* [309] (2020)	15.38	5.15	6.06	–	–	–	–
FDCNN-AUTO** [309] (2020)	5.06	0.93	2.77	–	–	–	–
SfSNet [253] (2020)	3.3	–	3.1	–	–	–	–
Proposed	**0.92**	**0.0**	0.75	**0.0**	**0.0**	**7.7**	**6.9**

*Spatial-Frequency Domain Selection Feature **Features on Double Convolutional Neural Network and Autoencoder

shows the robustness of the algorithm across different attacks and acquisition/attack devices. The proposed algorithm outperforms various deep learning based [209, 217, 253, 309, 333], multi cue fusion based [247, 368], motion [91], and texture based [45, 250] algorithms. The ultra deep neural network proposed by Tu and Fang [333] combine pre-trained deep residual network with Long

Table 3.11: Comparison with state-of-the-art results on the frame based presentation attack detection. Best result of existing algorithms is underlined.

Algorithm	**CASIA-FASD**	**Replay-Attack**		**MSU-MFSD**	**SMAD**	
	EER	**EER**	**HTER**	**EER**	**EER**	**HTER**
Motion [21] (2011)	26.6	11.6	11.7	–	–	–
LBP [66] (2012)	18.2	13.9	13.8	–	–	–
CDD [357] (2013)	11.8	–	–	–	–	–
Motion + LBP [161] (2013)	–	4.5	5.1	–	–	–
LBP-TOP [85] (2014)	–	7.9	7.6	–	–	–
IQA [106] (2014)	32.4	–	15.2	–	–	–
CNN [356] (2014)	7.4	6.1	2.1	–	–	–
IDA [344] (2015)	–	–	7.4	8.5	–	–
Color Texture [45] (2016)	2.1	0.4	2.8	4.9	–	–
LGBP + GS-LBP [249] (2017)	2.5	–	3.13	8.54	–	–
Deep Dictionary [217] (2017)	–	–	–	–	14.7	15.0
Proposed	4.95	0.8	**2.1**	**0.0**	**10.9**	**10.7**

Table 3.12: Wax figure face detection error rates (%) using unconstrained (protocol 1) and real-world protocol (protocol 3) of WFFD database [151]. The proposed algorithm reduces the average classification error rate (ACER) and EER by 2.40% and 4.27%, respectively. Best results among existing algorithms are underlined.

Algorithm	**EER**			**APCER**			**BPCER**			**ACER**		
	Prot. 1	Prot. 3	Avg.	Prot. 1	Prot. 3	Avg.	Prot. 1	Prot. 3	Avg.	Prot. 1	Prot. 3	Avg.
M-Scale LBP	33.17	34.56	33.86	31.22	33.33	32.27	31.22	32.92	32.07	31.22	33.13	32.17
Color LBP	33.17	36.81	34.99	30.24	35.38	32.81	36.10	35.79	35.94	33.17	35.58	34.37
Reflectance	41.95	44.78	43.36	40.00	46.01	43.00	52.19	46.22	49.20	46.10	46.11	46.10
VGG-16	45.85	48.67	47.26	50.73	45.19	47.96	41.95	49.28	45.61	46.34	47.24	46.79
Proposed	23.50	35.68	**29.59**	25.50	35.68	**30.59**	22.00	35.91	**28.95**	23.75	35.79	**29.77**

Short Term Memory (LSTM) yields an EER of 1.22% and 1.03% on CASIA-FASD and Replay-Attack database respectively. The EER of the proposed algorithm is at-least 24% better than [333] on CASIA-FASD while 0% EER is achieved on Replay-Attack database. As shown in Figure 3-9, the proposed algorithm outperform several state-of-the-art presentation attack detection algorithm including recent deep forest [53] algorithm on one of the largest MSU-USSA database. The EER and standard deviation of the proposed, Deep Forest [53] and LBP + Color moment [247] algorithm are $1.1 \pm 0.3\%$, $1.6 \pm 0.6\%$, and $3.9 \pm 0.8\%$, respectively.

Table 3.11 shows that the proposed algorithm either achieves state-of-the-art or competitive results even for frame-based classification with all the challenging face spoofing databases. The EER value of 0.8%, 4.95%, and 0.0% is achieved on Replay-Attack, CASIA-FASD, and MSU-MFSD database respectively in frame based detection. HTER of 2.1% is achieved on Replay-

Attack database in the grand test attack scenario which is lower to various texture based algorithms [45, 249]. On SMAD database proposed algorithm shows an improvement of more than 25% to 47% from the baseline performance [217]. The detection error rate using proposed and existing algorithms on wax figure faces are reported in Table 3.12. Similar to other challenging attacks, the proposed algorithm outperforms several existing algorithms for wax faces detection including hand-crafted and deep learning algorithms. The proposed algorithm is 37.4$ better than VGG-16 based wax faces detection.

To further show the generalizability of the proposed PAD algorithm, cross database experiment are also performed and results are reported in the Table 3.13. For video based scenario when the anti-spoofing algorithm is trained on CASIA-FASD database average HTER value of 26.7% and 35.3% is reported on MSU-MFSD and Replay-Attack database respectively. The average HTER on CASIA-FASD and Replay-Attack is 23.7% and 32.2% respectively when the anti-spoofing model is trained using MSU-MFSD database. When the model is learnt using the Replay-Attack database, and tested on each subset of CASIA-FASD and MSU-MFSD the average HTER value reported is 33.3% and 23.9% respectively. The anti-spoof model trained using Replay-Attack database shows better generalizable and it is may be due to the fact that it is captured in different illumination, devices, and background. The comparison of the proposed countermeasure with the existing anti-spoofing algorithms is shown in Table 3.13.

The proposed algorithm outperform the state-of-the-results when the countermeasure is trained using Replay-Attack and MSU-MFSD database. The proposed algorithm improves the HTER of the 2^{nd} best performing algorithm [250] from 39.6% to 23.7% on CASIA-FASD database when the model is trained on MSU-MFSD. Similarly the proposed algorithm improves the performance on MSU-MFSD by 1.1% when model is trained using Replay-Attack. Recently proposed algorithms based on deep CNN by Chen et al. [61] and [316] yields an HTER value of 34.7% and 37.3% on CASIA database when the model is trained on Replay-Attack, whereas the HTER of the proposed algorithm is at-least 1.4% lower than them.

Boulkenafet et al. [44, 45] have studied the effect of texture features extraction in various color domain such as HSV and YCbCr for presentation attack detection. To further promote the research in this direction and to see the effective channels in RGB color domain, we have performed the presentation attack detection using individual color channel of RGB. To perform this experiments,

Table 3.13: Result in terms of avg. HTER of the cross database experiments and comparison with State-of-the-art results using video based countermeasure. Results of Spoof Buster [49] are reported when single database is used in training. (Top two results are highlighted)

Train Database	**Algorithm**	**Test Database**		
		CASIA-FASD	**MSU-MFSD**	**Replay-Attack**
CASIA-FASD	Motion [84] (2013)	–	–	50.2
	Spectral Cubes [255] (2015)	–	–	34.4
	LBP [44] (2015)	–	36.6	47.0
	Color Texture [45] (2016)	–	20.4	30.3
	LBP+ GS-LBP [249] (2017)	–	18.6	48.4
	Directional LBP [264] (2017)	–	26.3	21.6
	Frame Diff + Multi-Level + Fisher Score + LPQ [29] (2017)	–	50.4	50.3
	Multiscale quality [360] (2018)	–	–	38.1
	De-Spoofing [154] (2018)	–	–	28.5
	Texture Markov Feature [368] (2018)	–	32.4	32.3
	Motion CodeBook [91] (2018)	–	50.0	33.7
	Spoof Buster [49] (2019)	–	–	53.0
	Two stream ResNet-18 + Attention [61] (2019)	–	–	36.2
	CCoLBP+Ensemble Learning [250] (2020)	–	18.6	18.7
	SAPLC [315] (2020)	–	–	27.3
	FCN-LSA [316] (2020)	–	–	27.3
	Proposed	–	**26.7**	**35.3**
Replay-Attack	Motion [84] (2013)	47.9	–	–
	Spectral Cubes [255] (2015)	50.0	–	–
	LBP [44] (2015)	39.6	35.2	–
	Color Texture [45] (2016)	37.7	34.1	–
	LBP+ GS-LBP [249] (2017)	40.3	36.1	–
	Directional LBP [264] (2017)	46.6	31.1	–
	Frame Diff + Multi-Level + Fisher Score + LPQ [29] (2017)	42.6	38.0	–
	Multiscale quality [360] (2018)	39.0	–	–
	De-Spoofing [154] (2018)	41.1	–	–
	Texture Markov Feature [368] (2018)	45.9	37.7	–
	Motion CodeBook [91] (2018)	49.3	40.8	–
	Spoof Buster [49] (2019)	43.3	–	–
	Two stream ResNet-18 + Attention [61] (2019)	34.7	–	–
	CCoLBP+Ensemble Learning [250] (2020)	39.3	25.0	–
	SAPLC [315] (2020)	37.5	–	–
	FCN-LSA [316] (2020)	37.3	–	–
	Proposed	**33.3**	**23.9**	–
MSU-MFSD	Motion [84] (2013)	–	–	–
	Spectral Cubes [255] (2015)	–	–	–
	LBP [44] (2015)	49.6	–	42.0
	Color Texture [45] (2016)	46.0	–	33.9
	LBP+ GS-LBP [249] (2017)	40.6	–	45.3
	Directional LBP [264] (2017)	40.2	–	48.8
	Frame Diff + Multi-Level + Fisher Score + LPQ [29] (2017)	50.0	–	48.0
	Multiscale quality [360] (2018)	–	–	–
	De-Spoofing [154] (2018)	–	–	–
	Texture Markov Feature [368] (2018)	57.0	–	42.7
	Motion CodeBook [91] (2018)	47.7	–	30.6
	Spoof Buster [49] (2019)	–	–	–
	Two stream ResNet-18 + Attention [61] (2019)	–	–	–
	CCoLBP+Ensemble Learning [250] (2020)	39.6	–	27.2
	Proposed	**23.7**	–	**32.3**

random 30 frames from each video are first selected for features extraction. The performance is reported on three challenging benchmark databases: CASIA-FASD, Replay-Attack, and MSU-

Table 3.14: All and individual color channel classification results in terms of EER(%) using 30 frames for presentation attack detection

Database	**Color channel**			
	RGB	**R**	**G**	**B**
CASIA-FASD	4.4	3.3	5.5	4.4
Replay-Attack	0.0	1.1	0.2	0.1
MSU-MFSD	0.0	0.0	0.0	0.0

Table 3.15: Computational complexity of the proposed and existing PAD algorithms.

Algorithm	FPS
LBP+LDA [84]	5.2
CDD [357]	2.5
SPMT [308]	1.5
IDA [344]	3.8
SpoofNet [228]	69.0
SSD [308]	120.0
SPMT + SSD [308]	45.5
Proposed	0.1

MFSD.

Results with 30 frames of a train and test video on CASIA-FASD, Replay-Attack, and MSU-MFSD databases are reported in Table 3.14. R channel yields the lowest EER value of 3.3% on CASIA-FASD database which is lower than [247],[96]. On the Replay-Attack database, the lowest EER value of 0.1% is given by B channel which is equal to the EER reported by [46]. Boulkenafet et al. [46] have used all frames of a video while we have used only 30 frames for such comparable performance.

Strength of the Algorithm: In summary the strengths of the proposed generalized PAD algorithm are listed below:

- The proposed algorithm can be implemented in real time. The feature extraction time on core i7@ 3.4GHz CPU machine with Matlab environment is 0.1 frames per second (FPS) (Table 3.15). The huge deployment of face unlocking on mobile devices[6] needs the protection of them from presentation attack. The proposed algorithm with such low computational time and memory requirement can also be implemented on mobile devices;

[6] www.counterpointresearch.com/one-billion-smartphones-feature-face-recognition-2020/

- The proposed algorithm outperforms various state-of-the-algorithms including 3D CNN [179], SfSNet [253], Multi-Regional CNN [209], and deep dictionary [217], for a variety of presentation attacks including silicone mask attack [217] and 2D attacks;
- The proposed algorithm is also able handle new era of 3D attacks i.e., wax faces [151]. Average EER of the proposed algorithm is at-least 37.4% and 12.6% lower than VGG-16 deep learning and multi-scale (M-scale) LBP texture features;
- The proposed algorithm is generalizable across imaging spectrum (visible or NIR), attacks (2D or 3D), acquisition devices (mobile or high-def), and quality of images (low or high).

3.6 Summary

Similar to visible spectrum, face recognition in near-infrared (NIR) spectrum is also vulnerable to presentation attacks. In the literature, there is very limited research towards developing efficient and inclusive countermeasures for the attack in the NIR spectrum and designing a unified algorithm to design and evaluate the performance of PAD algorithms towards continuously evolving presentation attacks in multiple spectra. In this research, we attempt to contributes towards this space by creating a large NIR PAD face database that comprises videos with different kinds of attacks of Indian and Chinese ethnicities. We next present a presentation attack detection algorithm for efficiently differentiating between bonafide and attacked images in the NIR spectrum. The generalizability of the algorithm is established by evaluating the performance on 9 existing databases and comparing with state-of-the-art results reported in the literature. It is observed that the proposed algorithm yields best results on almost all the databases using all three metrics of APCER, BPCER, and EER. In cross-database evaluations, while the proposed algorithm yields the best results, the error rates are comparatively higher. As future work, we plan to improve the effectiveness of the algorithm so that the error rates can be further reduced without increasing the computational complexity.

Chapter 4

MagNet: Detecting Digital Presentation Attacks on Face Recognition

4.1 Introduction

With advancements in face recognition algorithms, the usage of face recognition is also increasing considerably [40, 88, 111, 224]. The high performance of algorithms and convenience of capturing face images have supported the applications to allow remote or unsupervised capture of face images for authentication, for instance, now online banking can be performed via face authentication. While this increases convenience and reduces fraudulent access, the security of these recognition systems is also an important task. It's importance can be observed from the launch of the Odin, IARPA project on biometric presentation attack detection[1].

Presentation Attacks are "defined as the attack on the system which in any way can affect the decision of a biometric system". They can be broadly classified into two categories: digital and physical. Physical attacks include physical methods of deceiving the system such as print and replay attack, 3D mask, and silicon mask attack. Digital presentation attacks include attacks such as morphing, swapping, and digital alterations. These attacks can be performed for multiple reasons, avoiding recognition, impersonating someone else's identity, or multiple people sharing an identity. Recently, researchers are also studying adversarial attacks which are digital in nature, however, they are targeted towards fooling specific deep learning architectures and are generally

[1] https://www.iarpa.gov/index.php/research-programs/odin

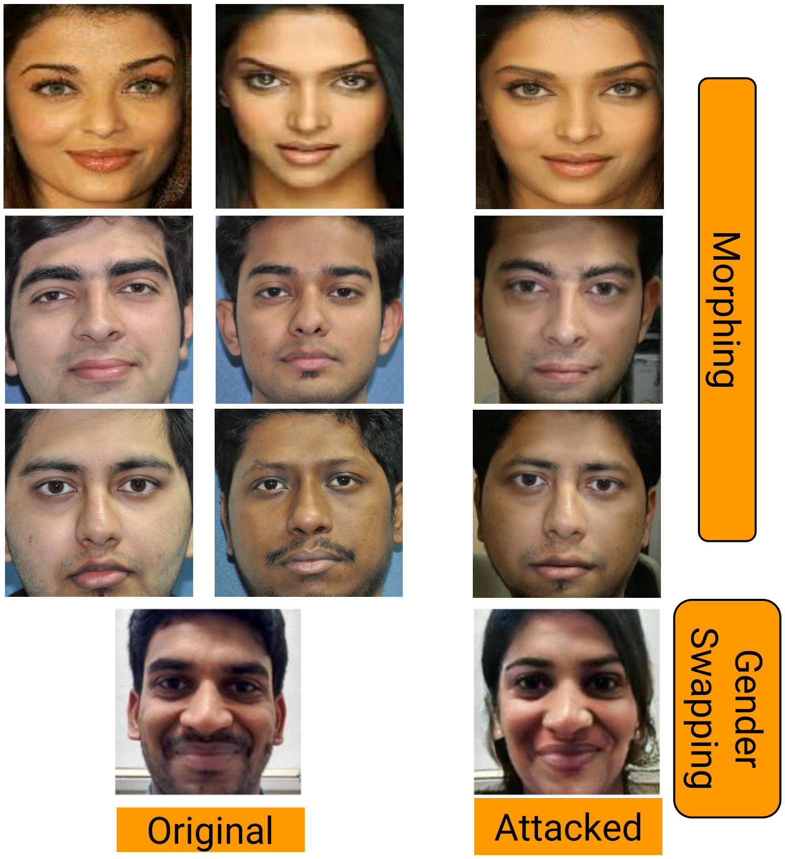


Figure 4-1: Illustrating the effect of perceptible digital alterations on face images.

imperceptible [18, 122–124, 319].

This chapter focuses on detecting digital alterations in face images. The effect of face morphing in enrollment was first introduced by International Organization for Standardization ISO 19792. In 2014, Ferrara et al. [97] showcased the vulnerabilities of commercial face recognition systems towards morphed images. They also showed that these morphed images are challenging to be detected by face recognition experts as well as automatic algorithms [98]. The popularity of face morphing applications around the world can be observed by the fact that *Facebook* has acquired one of the famous morph application called *MSQRD*. Figure 4-1 shows samples of digital alterations from multiple platforms. The first two columns show images of different subjects and the third column represents the morphed image which consists of equal facial features of both the identities. The morphed image in first row is generated using the Internet website called *morphthing.com*. The morphed images in the second and third rows are generated using

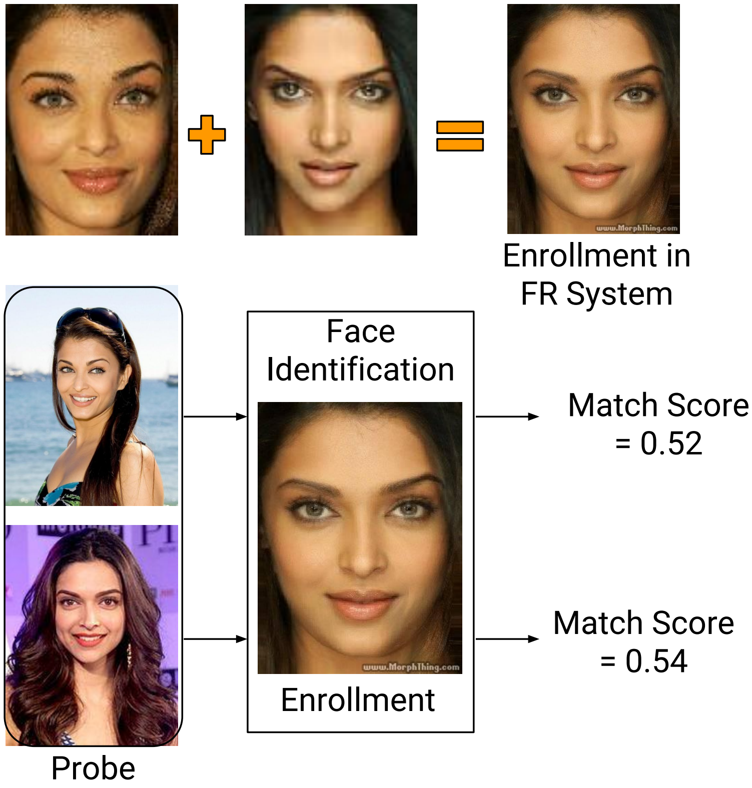


Figure 4-2: Illustrates the effect when enrolled templates are modified via morphing. The bottom row shows that the two different identities can claim the same identity through enrolled morphed image. Experiments are performed using COTS Face Recognition (FR).

the swap/morph feature of Snapchat application[2]. The gender swap image in last row is generated using another popular mobile application called FaceApp[3]. Similarly, recently *Instagram*[4], one of the most used social platform for story and information sharing with more than 1 billion users, has launched the face filters which can alter the facial properties in real time.

The similarity of source and target images in Figure 4-1 show that digital attacks like morphing can be used to both elude and create a duplicate identity. To experimentally visualize the effect of morphing on face recognition, Figure 4-2 shows the source and morphed images, and the recognition outcome of a face recognition system. Using a commercial-off-the-shelf (COTS) recognition system, it shows that the morphed image can successfully match with it's constituent source images. Inspired by the effectiveness of face morphing applications and limitation of face recognition algorithms, this research focuses on designing a novel algorithm to differentiate between digitally

[2] https://store.snapchat.com/
[3] https://faceapp.com/
[4] https://www.instagram.com/?hl=en

attacked images and original/ non-tampered images. As shown in Figure 4-1, different kinds of alterations introduce different effect on face images. In order to efficiently detect all the variations, we present a novel presentation attack detection algorithm that incorporates a new feature extraction algorithm. The contributions of this research can be summarized as follows:

- We propose a novel *MagNet* algorithm for effectively differentiating between digital presentation attacks and original non-tampered videos/frames, using the proposed WLMP feature descriptor;
- We present a new database termed as “***IDAge**nder*" - Digital Attack Face Database. It comprises three different databases generated using three separate techniques: (1) face swap/morph feature of Snapchat, (2) Internet website *http://www.morphthing.com/*, and (3) gender and age swap/morph feature of face transformation application called FaceApp;
- The effectiveness of the proposed algorithm is established using a series of experiments, including comparison with state-of-the-art presentation attack detection algorithms.

4.2 Related Work

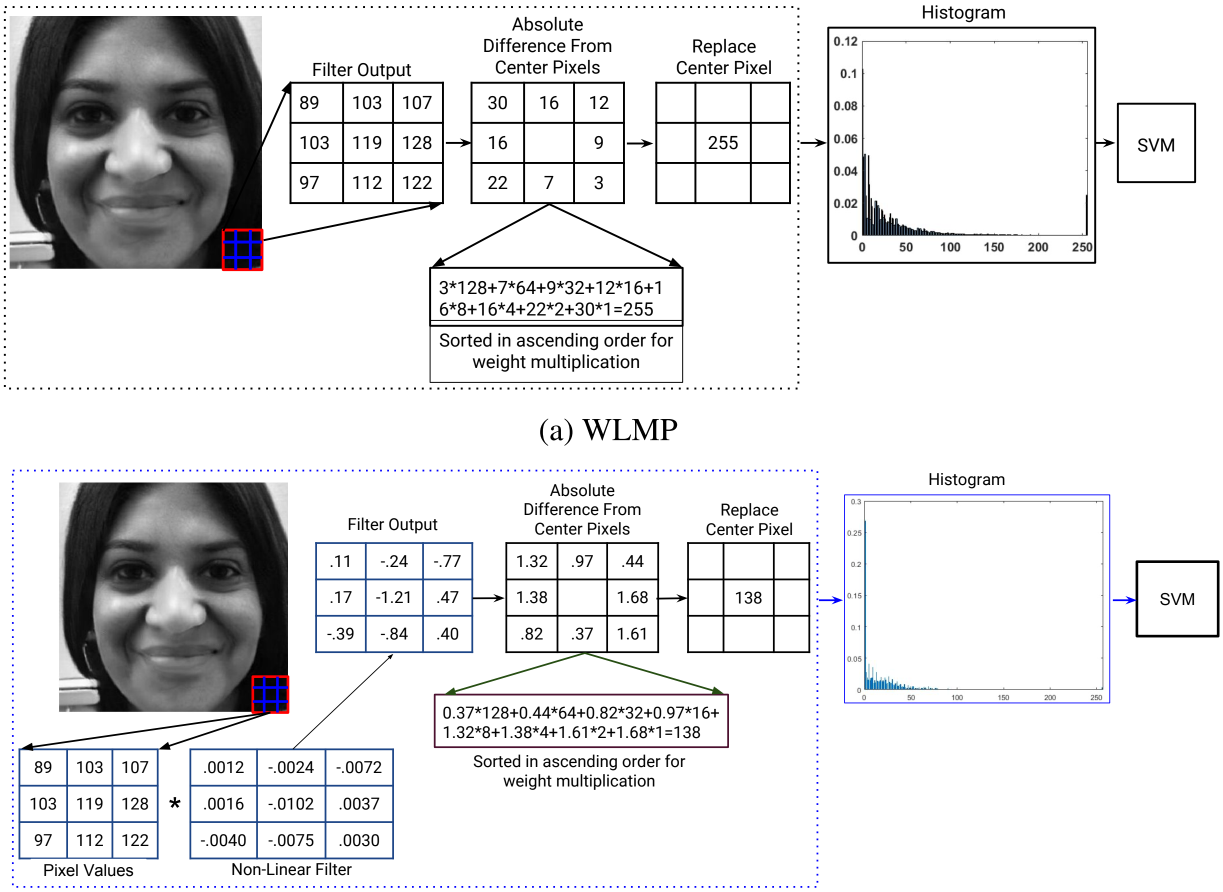
In 2014, Ferrera et al. in [97] showed the effectiveness of morphing attack to gain illegal access to the system. Morphed images were generated using genuine face images of two different individuals. GNU image manipulation program V4.0 software was used for morphing which first detects the facial landmarks from the faces to be morphed. These points were then mapped over each other and the output morphed image consists of the average features of both the images. In other words, the resulting morphed images contain equal facial properties of all the images used in its generation. Ferrera et al. [97] selected the best morphed images based on the match scores provided by the face recognition system. One major limitation of this research is the size of the database and the number of subjects used for evaluation. In 2016, Raghavendra et al. [268] prepared a relatively large database of morphed images using a process similar to [97]. The database contains 450 morphed images generated by morphing two and three different face images. Instead of choosing the best morphed image based on the score of the face recognition system, the mean output image is used as the best morphed image. However, the database is not released to the research community.

Gomez-Barrero [117] showed that not only face recognition systems are vulnerable to morphing, rather iris and fingerprint systems are also sensitive to these attacks. They showcased that when a morphed image is used as the enrollment image in the recognition system, even with higher thresholds on the match score, more than one individual can share an identity. While the e-passport and e-Pass renewal in countries such as New Zealand use a digital copy of the face in its application process, several other countries still use the print of the face image and a scanned copy as part of their application process. Inspired from this process, Scherhag et al. [294] prepared the morph attack database where first the morph images are printed using two different printers and then two different kinds of scanners are used to digitize the printed morphed images. The database consists of 693 morph images, out of which there are 231 digital attack images and 462 scan attack images. It was shown that 100% recognition is achieved when digital attack images are matched with its original images, while scan images yield more than 95% Impostor Attack Presentation Match Rate (IAPMR) using VeriLook SDK. They performed morph attack detection using the existing feature descriptors, BSIF, Local Phase Quantization (LPQ), Local Binary Pattern (LBP), and 2D Fast Fourier Transform (FFT), with SVM classification.

Robertson et al. [280] performed a detailed study to showcase that face morphing can be a serious threat on face recognition systems. They performed three experiments, in the first two experiments human examiners were asked to match two pairs of images and in the third experiment, smartphone face recognition is attacked by the morphed images. Experiment one showed that morph image which contains 50% features of both the genuine users are able to get more than 68% acceptance rate. This makes the problem serious because two different individuals can share the passport of one identity. In the third experiment, they performed face unlocking using three different images. With 100% morphing level, the identity of one person is completely change to second identity and 91.8% acceptance rate was reported. With 90% morph level, no significant drops in acceptance rate is reported. Neubert [235] applied multiple image degradation on the images with the intuition that the morphed images suffer heavy edge degradation to detect the attack. Similarly, for morph attack detection Makrushin et al. [216] have computed DCT coefficients from the JPEG compressed images and fit a logarithmic curve over the Benford features. Raghavendra et al. [269] developed two version (print and digital) of database by taking the average of two face images and morphing them together. The detection of morph attack is presented using LBP features in

YCbCr and HSV color space. Scherhag et al. [295] have proposed reference and no-reference image based morph detection using various texture descriptors such as LBP, BSIF, Speeded Up Robust Feature (SURF), and Histogram of Gradients (HoG). Scherhag et al. [293] proposed the detection of morphed attack on face recognition through landmarks difference between bonafide and morphed image.

Seibold et al. [298] have used three pre-trained deep CNNs (AlexNet, GoogLeNet, and VGG19) to detect the morph attack on face recognition systems. The training of all three networks from scratch yields 4.4% to 7.4% higher false reject rate (FRR) than the pre-trained networks. Raghavendra et al. [272] performed the fusion of the features of first fully connected layers of AlexNet and VGG19 to detect the morph attack. The proposed approach shows Bonafide Presentation Classification Error Rate (BPCER) value of 14.38%, 41.78%, and 28.76% at 5% Attack Presentation Classification Error Rate (APCER). Wandzik et al. [336] shows the vulnerability of the deep CNN based face recognition under morphing attacks. The ResNet v1 shows 99.97% acceptance rate on original images, the acceptance rate drops down to 34.66% on morphed images blended with 0.5% probability of images. Nguyen et al. [237] proposed the multi-task network for the detection of manipulated face images while cropping out the manipulated region. In place of using multi-task network, Dang et al. [80] have used the attention mechanism to highlight the important region for forgery detection. Matern et al. [222] and Yang et al. [359] have argued that in the computer generated images, several visual features are found missing at teeth, eyes, head pose, and facial contours. These visual features can be effectively used for morphed images detection. Nguyen et al. [238] have used the capsule network on the extracted features of VGG network for various fake faces detection. He et al. [138] have used multiple color spaces as input to CNN for effective features extraction. Recently, Jiang et al. [152] one of the largest face swapping database and presented a variational auto-encoder to enrich the quality of face swapped images. Similarly, Li et al. [187] have presented a high quality deepfake video database containing images of celebrities. De Lima et al. [86] and Mas Montserrat et al. [221] have used the combination of CNN and recurrent network to model the inconsistencies presented in the forged videos at the frame level. Liu et al. [195] enhanced the global texture features extraction capability of CNN through the insertion of Gram block. Datta and Murthy [82] have presented the approach of face generation based on the distribution of images belonging to same and different persons. Similarly, Galbally et al. [107]



(a) WLMP

(b) WLMP with convolution of image using non linearly learned filters (NL-WLMP)

Figure 4-3: Illustrating the components involved in the proposed Weighted Local Magnitude Pattern (WLMP) descriptor for digital attack detection.

have proposed the synthetic iris generation using genetic algorithm from the binary iris code. Li et al. [186] have proposed recurrent convolution network to detect the fake images generated using generative networks. They have analyzed the inconsistency in the eye region to detect whether the videos are real or fake. Neves et al. [236] have proposed removing GAN fingerprints from the synthetically generated images to fool the manipulation detector algorithm. Recently, Scherhag et al. [297] and Tolosana et al. [328] have presented a survey of the existing digital attack detection algorithms. Overall, existing works have showcased, beyond doubt, the threat of morphing on face recognition systems and showed the need of proper address for secure use of system specially the electronic documents.

4.3 Proposed Digital Presentation Attack Detection Algorithm

It is our assertion that digital alterations generally perform smoothing and blending to minimize the irregularities due to the differences in the source (and target) frames. This reduces the difference in neighboring pixel values in the resultant image. It is our hypothesis that it will be easy to detect digital alterations if we can highlight the irregularities between the pixels in the neighborhood and give weight according to the absolute values. Based on this hypothesis, we propose a novel feature encoding scheme for detecting digital alterations. As shown in Figure 4-3, the proposed algorithm is based on a novel feature encoding method termed as *Weighted Local Magnitude Patterns* (WLMP) for encoding digital alterations. The detailed description of the proposed WLMP and its variants is discussed below:

4.3.1 WLMP

The input image is first tessellated into multiple patches of size 3×3. For each patch, the absolute difference between the center pixel and its neighborhood pixels are calculated. Since there are eight neighborhood pixels, there are eight difference values. The difference values are sorted in ascending order. Instead of binarizing the absolute differences, the sorted values are multiplied with 2^p, where $p = 0, ..., 7$ for eight neighborhood values. The motivation for sorting and multiplying is to give higher weight to the pixel which has a value similar to the center pixel. The final output value is then mapped to a value in the range of 0 to 255 (i.e., any value greater than 255 is set of 255). Finally, a histogram feature vector is calculated based on the weighted local magnitude patterns of the image. The output images using the proposed feature descriptor are shown in Figure 4-4 along with the corresponding output obtained by LBP. It can be observed that the output images of the proposed feature retain the high-frequency information while reducing the low-frequency information. With images obtained from Snapchat's swapped/morphed feature, facial keypoint regions such as eye, mouth, and nose are most affected, while the central region is well blended. This is clearly highlighted in the output images of the proposed feature extractor. Therefore, we postulate that for morphing related attacks, the proposed feature is better at detecting alterations than existing feature descriptors.

As shown in 4-3(a), we can visualize that the computation of WLMP descriptor is with identity

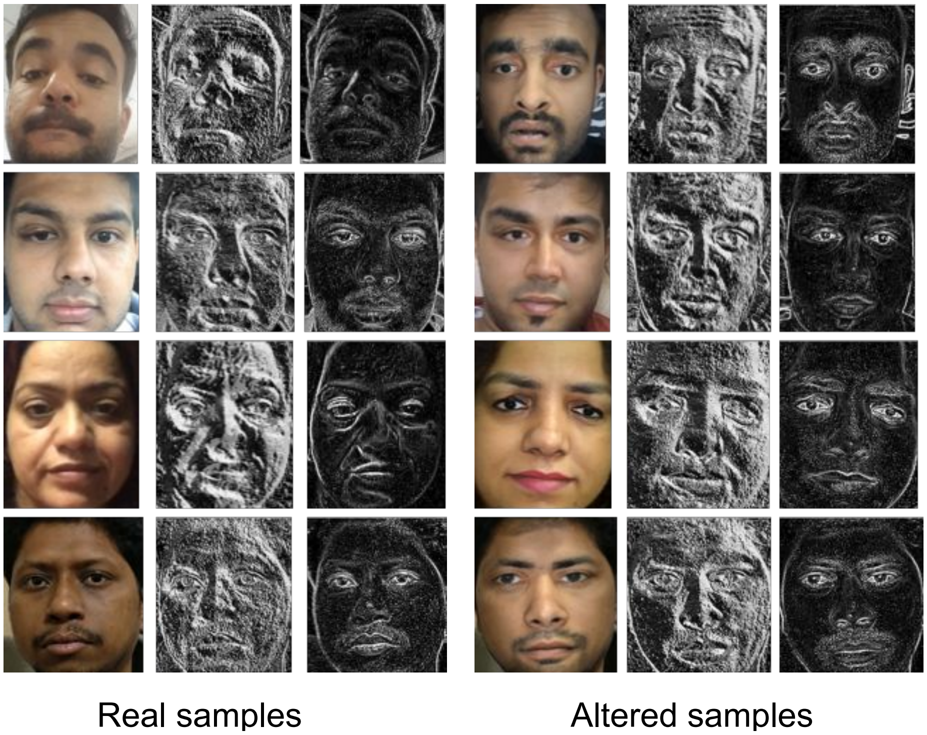


Figure 4-4: Illustrating the features obtained for real and altered samples from Snapchat database. In both real and altered samples, first column is the input images, second column is the LBP features, and last column is proposed feature images, respectively.

filter (i.e. convolving the patch with identity filter and then computing WLMP values). However, convolution with linear/non-linear filters can help in extracting the features from the locally connected regions which can better differentiate original with altered images. For instance, morphing changes the local features of the face so that certain landmarks of the face exhibit the features of the persons used for morphing. Further, while applying morphing, different facial structures undergo varying spatial changes to create an output image. A convolution operation with a filter before computing WLMP can help in detecting robust feature by providing the spatial invariance. In this research, convolution from filters obtained using non-linearly learned filters using deep learning model, GoogLeNet [318].

4.3.2 WLMP with Non-Linear Filter

Face morphing and digital alteration change the micro texture property of the face region and hence convolution of the input image with filters makes a strong case for encoding changes in the texture. For instance, while the software used for morphing blends two images together nicely, some minute/micro-level artifacts can be observed around key facial such as eye and mouth. Convolution with a learned filter can enhance these micro artifacts and help in computing representative WLMP descriptor.

The non-linear filters used in this research are obtained from GoogLeNet model [318]. The filters at layer two which are of the size 3×3 from the pre-trained model[5] are utilized. The initial layer filter can result in highlighting the lower level features such as edges [366] which might be one of the most reliable information in detecting digital alteration. At the same time convolution with non-linear filters can help in boosting the detection of high frequency information in the proposed WLMP descriptor. The pre-trained model provides 4 filters each with dimension $3 \times 3 \times 64$, we have taken the average around the third dimension to get the single filter of size 3×3. Therefore, four filters i.e. average filter response from each of the 4 outputs of the model, are used in this research for convolution and features extraction. Figure 4-3(b) shows the WLMP feature descriptor computation using a GoogLeNet filter and termed as Non-Linear Weighted Local Magnitude Pattern (NL-WLMP).

4.3.3 MagNet: Proposed Algorithm for Digital Presentation Attack Detection

WLMP, and its variant, provides feature descriptor which can be fed into a 2-class (i.e. original vs attacked) classifier such as Support Vector Machine (SVM) [76]. Figure 4-5 illustrates the steps involved in the proposed digital presentation attack detection algorithm using WLMP. In the proposed algorithm, termed as MagNet, WMLP and NL-WLMP are individually used to compute the classification scores which are then combined using score fusion. The score of each test frame is computed as the weighted score fusion of the scores computed from two different SVM classifiers, i.e. WLMP+SVM and NL-WMLP+SVM. Specifically,

- For each test image, WLMP feature descriptor is computed and a score value is computed using the WLMP trained SVM classifier;

[5] http://www.vlfeat.org/matconvnet/pretrained/

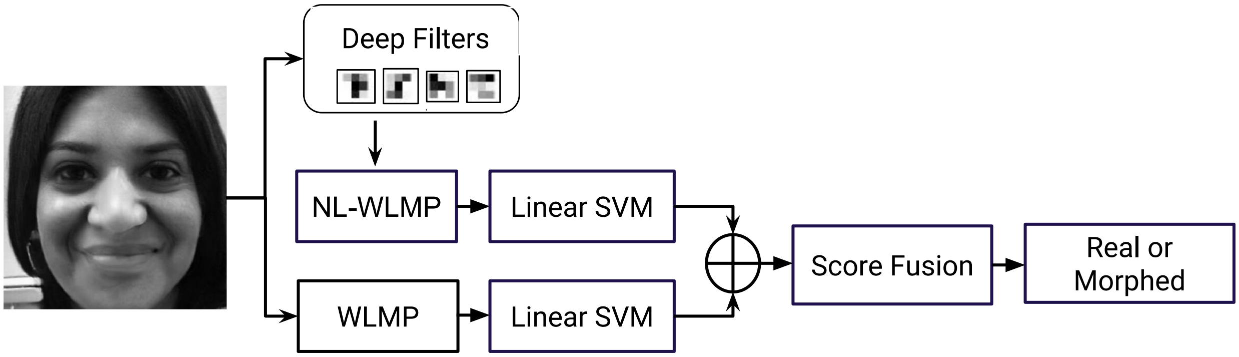


Figure 4-5: Proposed MagNet algorithm using fusion of WLMP and NL-WLMP.

- Similarly, NL-WLMP feature descriptor is computed from the test image followed by score is obtained through NL-WLMP trained SVM classifier;
- Final score of a test image is computed using weighted sum of the above two scores. The weights for fusion are computed over the training/development set of each database using grid search.

4.4 *IDAgender*!! Proposed Digital Attack Databases

The second contribution of this research is ***IDAge**nder*, the proposed digital presentation attack databases. ***IDAge**nder*[6] contains different subsets corresponding to different digital operations and it is unique in terms of digital mediums, number of subjects, and type of alterations. There are several open source algorithms and tools available for creating digitally altered images. However, Snapchat and *morphthing.com* are one of the most popular and easily accessible tools for morphing or swapping face images. FaceApp, a mobile application, has recently become popular within few months of its development for morphing gender and age characteristics. Since it is easy to navigate through these apps, non-technology savvy users can also efficiently use it to create various kinds of altered images. For example, a video where a woman swaps her face with Kardashians had been viewed more than $21,000$ times in a week[7]. The face swap feature is effective to change the

[6] This database will be made available to the research community via http://iab-rubric.org/resources.html.

[7] https://tinyurl.com/k6nfly9

Table 4.1: Summary of the proposed *IDAgender* databases.

Database	Real	Morphed	Unconstrained	Subjects
Snapchat	129 Videos	612 Videos	✓	130
Identity Morphing	545 Images	1,200 Images	✓	545
FaceApp	250 Images	375 Images	✓	125

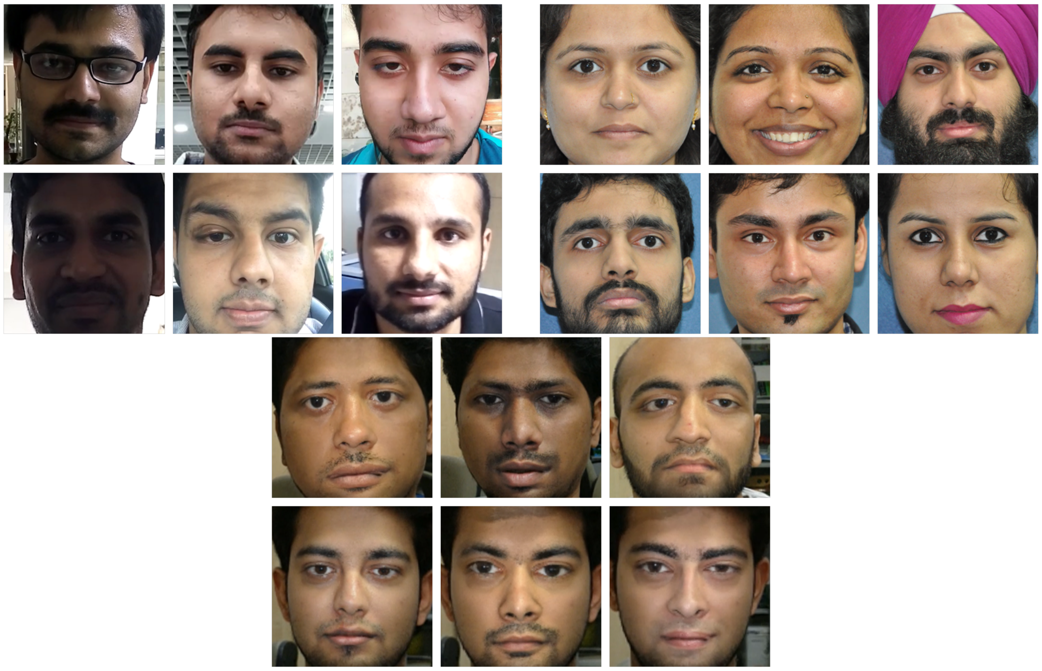


Figure 4-6: Sample images from the proposed Snapchat face swap database. (a) Sample bonafide images set (left), (b) sample images used for face swap (middle), and (c) morphed images from Snapchat (right).

properties of the face completely and by just looking at the altered face, it is difficult to determine whether it is real or not. The proposed *IDAgender* database consists of three different subsets: morphing, swapping, and FaceApp. The details of the three subsets are discussed in the following subsections and Table 4.1 lists all three components prepared in this research.

4.4.1 Proposed Snapchat Face Swap Database

This subset of *IDAgender* is a video database and consists of two parts: bonafide faces and morphed faces. Since morphing is applicable in both images and videos and the Snapchat feature is more prevalent on mobile phones, the bonafide/genuine videos are captured using mobile phones. For every user, at least one video of around six seconds is captured using the front camera. In

Table 4.2: Characteristics of the proposed Snapchat face swap database.

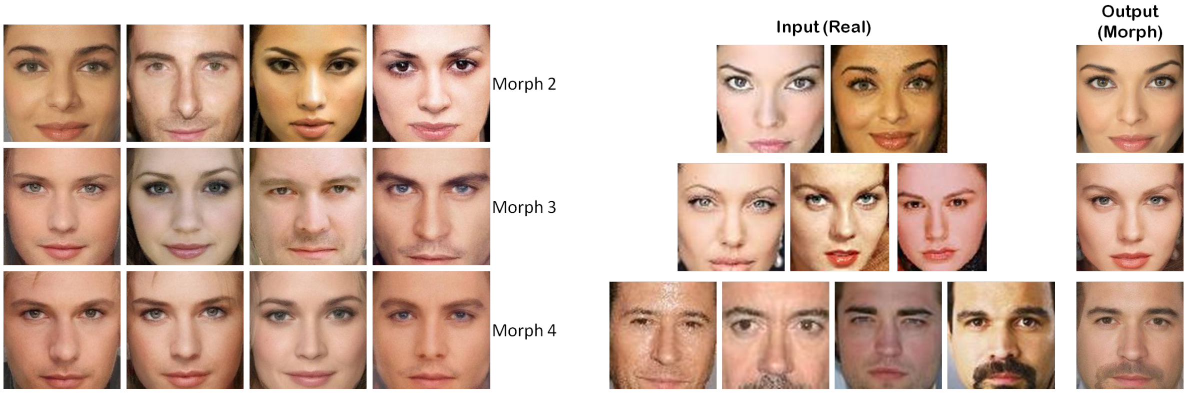
Data Type	**Subjects**	**Videos**	**Detected Faces**
Real	110	129	30,728
Attack	31	612	1,04,052

total, 129 bonafide videos are captured from 110 individuals over a period of two months. These videos are captured in unconstrained environment such as natural outdoor, hallway, and inside office premises. Faces present in the videos are detected using Viola-Jones face detector and normalized to 296×296 pixels. As summarized in Table 4.2, after face detection, the bonafide subset contains more than $30,000$ face frames. Figure 4-6(a) shows sample images from the bonafide set captured in different illumination and background conditions.

For generating the morphed faces, face swapping feature of Snapchat[8], a popular social messaging app, is used. The steps involved in the process are as follows: first the face is detected using Viola-Jones face detector [335]. To make the change more accurate and precise, key point location of the facial features such as eyes, mouth, and face boundary are detected using Active Shape Model (ASM) [75]. Once the facial keypoints are detected, a 3D mesh is generated which fits the face properly and can move in real time with changes in the face. The facial keypoints are detected from both the faces and the central region is morphed from one image to the other. The boundary is then seamlessly blended to create the new morphed face image.

To prepare the morphed videos using Snapchat, two good quality frontal face images of 84 subjects (different from the bonafide faces) are captured in a semi-controlled environment. Samples of these images are shown in Figure 4-6(b). These images are termed as the input gallery for face swapping/morphing. To create a morphed video, Snapchat application requires the users to select the host video/image, and an image with which they want to perform the face swap/morph. Using host videos from 31 subjects and input images from 84 subjects, 612 presentation (face swap) attack videos are prepared. Samples of morphed faces are shown in Figure 4-6(c). Similar to bonafide faces, the detected morphed faces are normalized to size 296×296. Table 4.2 summarizes the characteristics of this subset of the proposed database.

[8] https://www.snapchat.com/



(a) Sample morph image set

(b) Input and output image from the website

Figure 4-7: Sample images from the proposed IdentityMorphing face swap database.

4.4.2 Proposed IdentityMorphing Face Swap Database

The second subset is prepared using the morphthing.com website by morphing a number of face images together. The morphing tool requires the users to select the number of face images to be morphed together, and it has the facility of morphing a maximum of four face images together. The face images in this database are captured in an unconstrained environment and have varying image quality. The real images used in preparing the morph images are publicly available images of celebrities on this website.

Table 4.3 shows the comparison of the proposed IdentityMorphing database with the existing morph databases. The proposed database has $1,200$ morph images, out of which 500 images are generated by morphing two faces, 450 by morphing three faces, and 250 images are generated through morphing of four faces together. The proposed database is at least three times larger than existing databases in terms of the number of subjects and images, respectively. Sample images generated by morphing 2-4 faces together are shown in Figure 4-7. Figure 4-7 (a) shows the morphed output images while Figure 4-7 (b) shows the input real and output morph/swap images. The output morph faces have both visual and facial specific characteristics of all the faces used in its generation.

Table 4.3: Characteristics of the proposed and existing related morph databases.

Database	Subjects	No. of Faces Morphed	No. of Morphed Images
Raghvendra et al. [268]	110	2 and 3	450
Scherhag et al. [294]	–	2	231
Proposed IdentityMorphing	545	2, 3, 4	1,200

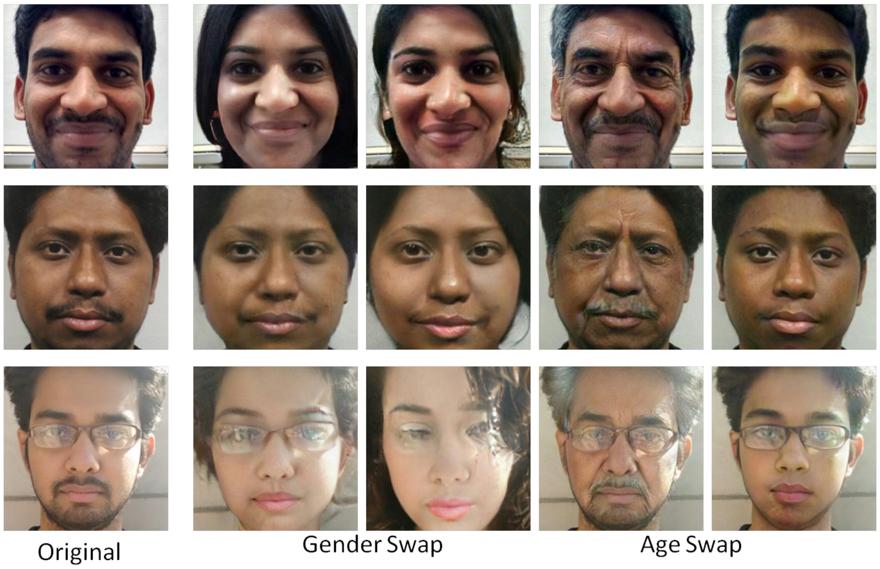


Figure 4-8: Sample images showing age and gender swap using FaceApp. Each row represent different subject.

4.4.3 Proposed FaceApp Database

The third subset of the proposed digitally morphed database is prepared using a mobile application, “FaceApp"[9]. FaceApp provides filters for gender morphing and age addition or subtraction. In this research, the database is prepared by morphing the gender of the person, adding the age to look older, and subtracting age to look younger.

To the best of our knowledge, this is the first work which presents a database having face images with altered age or gender. To create the digitally morphed images, first, good quality frontal face images of 125 subjects are captured in controlled illumination. To create the morph images, face image of each user is given to FaceApp and the image is morphed as per a given chosen filter. The filter represents the operation that will be performed on the input images; the operation can be gender morphing and age addition/subtraction. The proposed database contains

[9] https://play.google.com/store/apps/details?id=io.faceapp&hl=en

Table 4.4: Characteristics of the Proposed FaceApp database.

Type of Swap/Morph	**Digital**
Number of Alterations	2 (Age and Gender)
Number of Images	625
Types of Images	250 Real, 375 Altered
Number of Subjects	125

Table 4.5: Experimental protocol of each of the proposed databases.

Database	**Videos/Images**	**Real Folds**	**Attack Folds**	**Iterations**	**Metrics**
Snapchat	129 Real and 612 Attack	3	10	30 (i.e. 3×10)	ACER & EER
Identity Morphing	545 Real and 1,200 Attack	3	6	18 (i.e. 3×6)	
FaceApp	250 Real and 375 Attack	2	3	6 (i.e. 2×3)	

375 digitally morphed face images. Characteristics of the FaceApp database is given in Table 4.4.

4.5 Experimental Protocol and Performance Metrics

We also define a benchmark protocol that can be used to report and compare results on ***IDAgender***. In place of one particular train and test sets, multiple fold cross-validation is performed for evaluating the performance of the algorithms.

Protocol for Snapchat Database: As explained earlier, the bonafide (real) subset of the Snapchat database contains 129 videos from 110 subjects and the presentation attack subset contains a total of 612 videos from 31 subjects. Out of these videos, the real subset is divided into three random folds, where two folds contain 40 videos from 40 subjects each. The third fold contains 49 videos from 30 subjects. In the attack subset, the number of videos are large and hence it is divided into 10 folds. Each fold of the attack subset contains 60 videos corresponding to three subjects except the last fold which contains 72 videos from four subjects. Unlike three fold cross validation where two folds are used for training and one for testing, on this database, one fold is used for training, and the remaining folds are used for testing. The training set is reduced to evaluate the performance with limited training samples.

Protocol for IdentityMorphing Database: IdentityMorphing database contains 574 bonafide images and $1,200$ morphed images. Out of these images, the real (bonafide) images are divided into three folds and morphed images are divided into six random folds. Similar to the Snapchat database, one fold at a time is used for training while the remaining folds comprise the testing set.

Protocol for FaceApp Database: FaceApp database contains 250 bonafide images and 375 age and gender morph images. 375 morphed images are divided into three folds where each fold contains 125 images, and 250 bonafide images are divided into two folds with 125 images in each fold. Similar to the previous two databases, one fold is used for training and the results are reported with remaining as the test set.

The protocol of each of the proposed databases is listed in Table 4.5. Real and attack subsets are divided in a manner such that equal samples (approximately) from both the classes can be used for training. For example, IdentityMorphing database contains 545 real images which are divided into 3 folds, where each fold contains 180 images. Similarly, the attack set is divided into 6 folds with each fold containing 180 samples (approximately). FaceApp database is divided into two real folds and three attack folds, where both type of folds contains 125 samples of two classes.

Performance Metrics: The performance of presentation attack detection is reported in terms of the Equal Error Rate (EER) and Average Classification Error Rate (ACER). EER is defined as the point where the Bonafide Presentation Classification Error Rate (BPCER) is equal to the Attack Presentation Classification Error Rate (APCER). BPCER is the percentage of bonafide faces which are incorrectly classified as attack/altered faces while APCER is the percentage of attack faces which are incorrectly classified as bonafide faces. To calculate the BPCER and APCER on the test set, a threshold value is selected based on the EER of the development set. In this research, half of the training set is used as the development set. ACER $(\%)$ is then computed as the average of BPCER and APCER.

$$ACER(\%) = \frac{BPCER + APCER}{2} * 100 \tag{4.1}$$

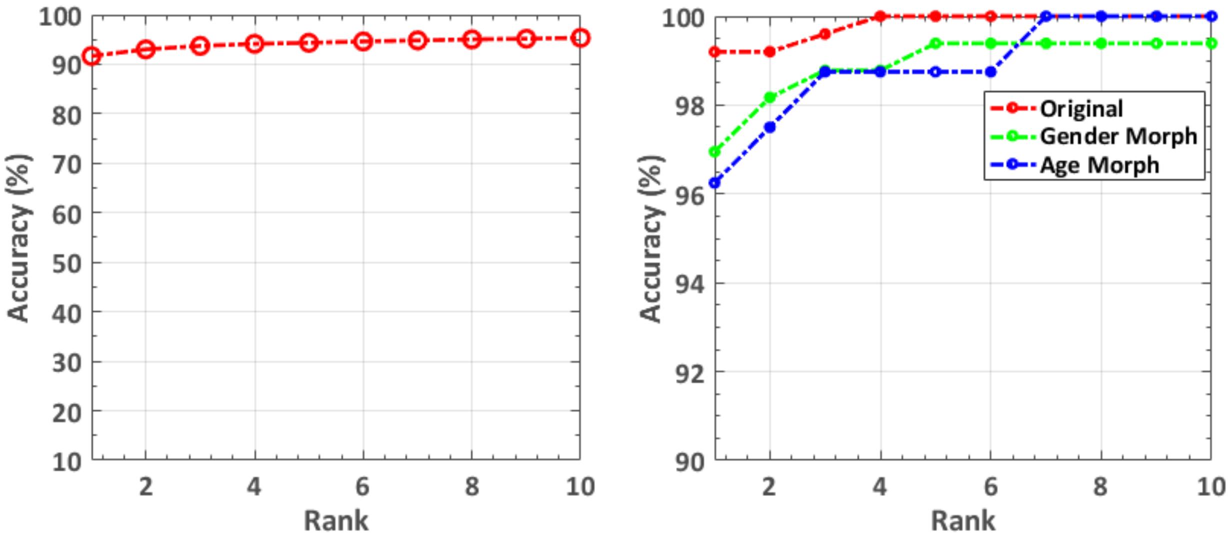
4.6 Effect of Swapping Attack on Face Recognition

To evaluate the effectiveness of the face swap feature as an attack on the face recognition system, we have performed two different experiments: 1) Various iOS devices are now equipped with the face unlock feature. Thus, the first experiment is face unlocking on iPhone/Android and 2) face identification using a Commercial-Off-The-Shelf System (COTS), FaceVACS[10]. In the first experiment to unlock the iPhone/Android, video of the morphed face prepared using an image of the genuine person who is enrolled in the mobile device is displayed in front of the mobile camera. It is interesting to observe that the face recognition algorithm in iPhone is unable to detect the attack and hence gets unlocked every time. This shows the vulnerability of face recognition in mobile devices towards digital attacks. In the second experiment, face identification is performed using a COTS system on all three databases and results are summarized in the next subsection.

4.6.1 Face Recognition with SnapChat and FaceApp Databases

To perform the Face Recognition (FR) experiment, another set of frontal images is collected from the individual whose images are used for creating the morphed videos. These images comprise the gallery for face identification. From each of the attack videos in Snapchat database, 30 random frames are used as the probe set for face identification experiment. Figure 4-9(a) shows the CMC curve obtained for this experiment. It can be observed that 90% of the time, attack images are matched to enrolled gallery images at rank-1. Similarly, from the FaceApp database, one frontal image of each person, over which the gender and age morphing has performed, is used as the gallery image. All the morphed images in the FaceApp database are used as the probe images. Figure 4-9(b) shows the effect of gender and age morph on face identification. When real images are matched to the gallery images, COTS shows more than 99% identification accuracy while morph images suffer a drop of 2-3% accuracy at rank-1. For the gender morph and age morph experiment, the digitally altered images corresponding to that particular subset are used as probe. Figure 4-10 shows the identification score of the probe images with respect to the gallery image of that subject. On matching real images without morphing, the highest score value of 0.999 is recorded while with age and gender morphed images, the score reduces to 0.187 and 0.397, respectively. The low matching scores show the successful identity evasion using digital attack.

[10] http://www.cognitec.com



(a) Using SnapChat database.

(b) Using FaceApp database.

Figure 4-9: CMC plot for face identification.

4.6.2 Face Recognition on the IdentityMorphing Database

Similar to the SnapChat and FaceApp databases, IdentityMorphing database contains the morph images of two, three, and four different identities that can be utilized for identity fraud. In this case, the identity fraud can be described as the scenario where one probe image can match to different gallery images or multiple individuals can share an identity in the gallery database. To perform this experiment, real images of two individuals are used as gallery images and morphed image generated using images of those individuals are used as the probe image. Figure 4-11 shows the sample images used in the identity fraud experiments. All the probe images shown in the figure yield high match score value of 0.999 to both the gallery images.

4.7 Digital Presentation Attack Detection Results

The performance of the proposed algorithm is shown on the ***IDAgender*** digital attack face database. In the literature, face presentation attack detection [105] using texture features has shown state-of-the-art performance (discussed in Section 4.2). Therefore, we have compared the performance of the proposed algorithm with the following state-of-the-art texture feature based algorithms along with CNN model.

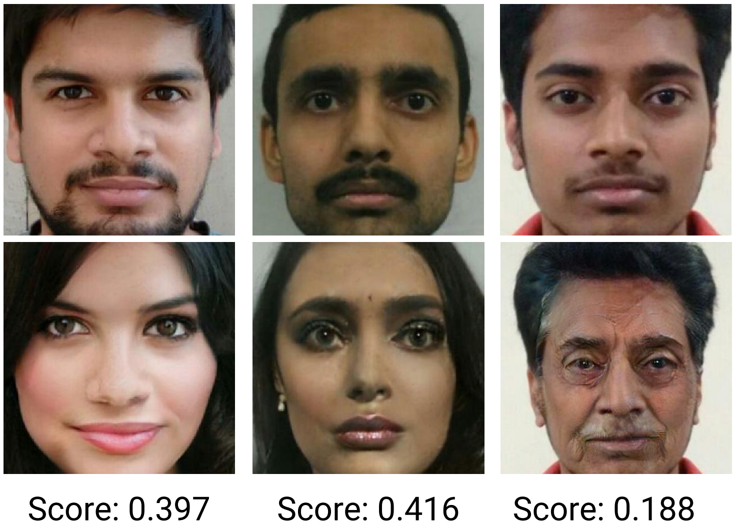


Figure 4-10: Gallery and probe images with corresponding match scores obtained using COTS on the images from FaceApp database. The first row contains sample gallery images and the second row contains the corresponding age/gender morph probe image of that subject.

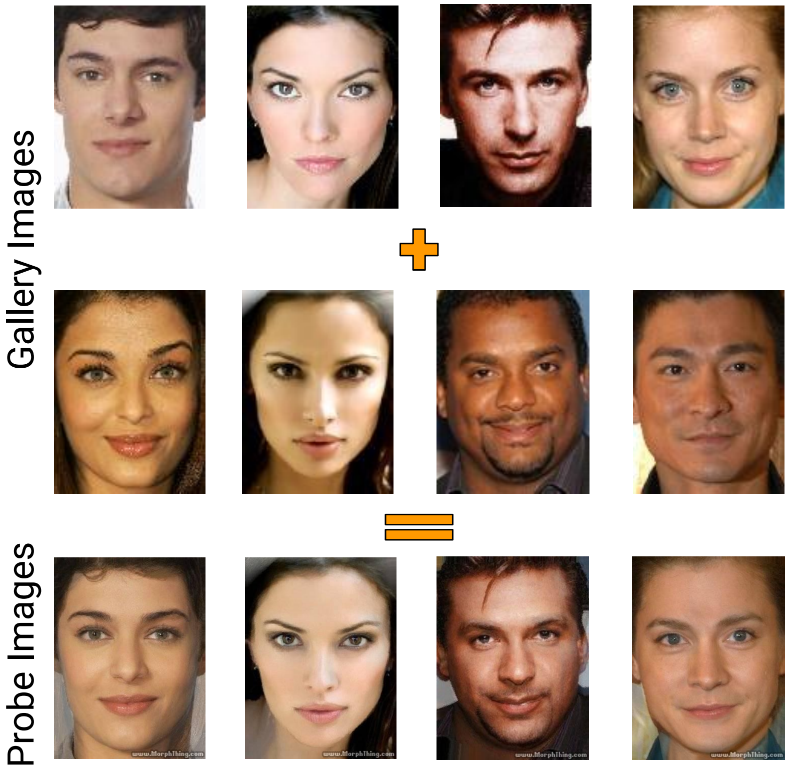


Figure 4-11: Identity duplication experiment using real gallery and morph probe images.

- Local Binary Pattern (LBP) [211]

- Rotation Invariant Uniform LBP (RIULBP) [239],
- Complete LBP (CLBP) [133],
- Uniform LBP (ULBP),
- Local Phase Quantization (LPQ) [240],
- Binarized Statistical Image Features (BSIF) [157], and
- Combination of Redundant Discrete Wavelet Transform (RDWT) [100] with Haralick features [136], [11],
- Agarwal et al. [14],
- Pre-trained VGG16 Convolutional Neural Network (CNN) [306],
- Fine-tuned GoogLeNet CNN [318],
- XceptionNet [285],
- ResNet-18 [166],
- Sharp multiple instance learning (S-MIL) [183]

4.7.1 Results and Analysis on Snapchat database

The protocol defined in Section 5 is used for experiments on the Snapchat swap database. Since the proposed database contains videos, the results can be measured both in terms of video classification and frame classification. In case of videos, the entire video is classified as bonafide or attack whereas for frame based, every frame is classified as bonafide or attack. The score of the video is calculated as the average of all the scores corresponding to frames of that video.

First, the performance of the proposed MagNet algorithm is evaluated on the SnapChat database and compared with state-of-the-art algorithms in literature [297]. Figure 4-5 shows the proposed algorithm with the combination of micro-texture encoding using GoogLeNet filters and local magnitude pattern. On the Snapchat database, the fusion of WLMP and NL-WLMP yields the average

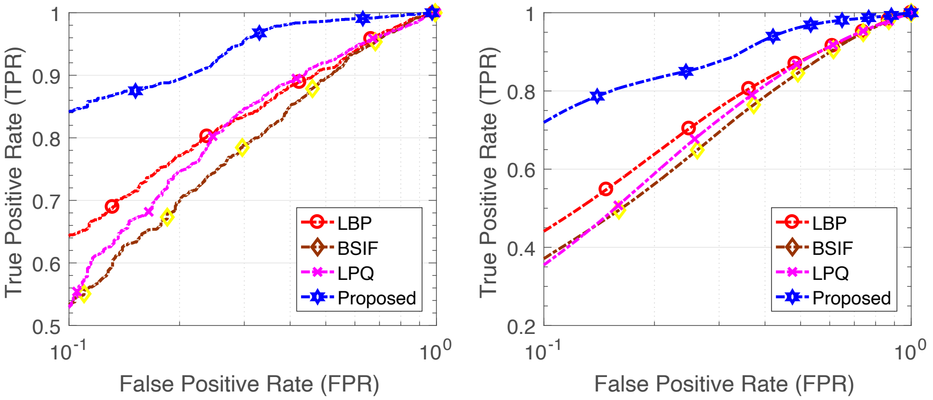


Figure 4-12: ROC curve for video (left) and frame (right) based presentation attack detection on the proposed Snapchat database.

EER of 13.2% and 18.0% for video and frame based detection, respectively. However, the combination of WLMP and L-WLMP yields the EER of 14.3% and 21.0% for video and frame based detection, respectively. Table 4.6 and Figure 4-12 show the results of the proposed and existing features for digital presentation attack detection by face swapping. The results are summarized in Table 4.6 and the analysis is summarized below:

- The proposed MagNet algorithm which is a combination of WLMP and NL-WLMP shows an improvement of 34% in terms of EER from the second best-performing feature, i.e., LBP (i.e., hand-crafted filter based algorithm) for video-based attack detection;

- The performance of L-WLMP in detecting digital attacks is similar without filtering (WLMP) and with linear filtering;

- The lower bit linear filters yield lower detection performance. One possible reason lies in the length of the feature vector. The feature vector of n bit linear filter is 2^n. Hence, the higher bit filters have a large feature dimension;

- The non-linear filters which are learned using the deep CNN model, i.e., GoogLeNet performs better than linear filters. The primary reason might be the richness of edge features preserved in the initial layers of the CNN model;

- From the experimental results, we observe that the filtered WLMP versions yield better results than original WLMP. Among the three variants, NL-WLMP outperforms WLMP and

Table 4.6: Classification performance (%) of the proposed and existing algorithm for video and frame based presentation attack detection on the proposed SnapChat database. The results are reported in terms of the average equal error rate and classification error rates along with standard deviation (±).

Input	**Features**	**EER**	**ACER**	**APCER**	**BPCER**
Video	LBP [212]	21.7 ± 6.1	21.3 ± 5.9	16.51 ± 5.7	26.28 ± 6.1
	ULBP [239]	24.5 ± 6.0	22.7 ± 5.8	12.17 ± 6.0	33.45 ± 5.9
	RIULBP [239]	24.7 ± 4.9	23.5 ± 4.7	16.12 ± 6.2	32.62 ± 3.2
	CLBP [133]	24.5 ± 6.1	24.8 ± 5.9	12.60 ± 5.2	37.16 ± 6.6
	Haralick+RDWT [12]	25.6 ± 7.2	24.5 ± 7.3	13.99 ± 8.5	35.24 ± 6.1
	BSIF [157]	25.2 ± 9.1	24.9 ± 9.3	29.96 ± 13.0	20.07 ± 5.6
	LPQ [240]	22.9 ± 5.2	23.9 ± 5.0	33.33 ± 5.4	**14.58 ± 4.6**
	[14]	18.2 ± 5.6	18.1 ± 5.5	6.71 ± 5.9	29.51 ± 5.1
	Proposed (MagNet)	**13.2 ± 3.4**	**12.9 ± 3.2**	**5.62 ± 2.8**	20.15 ± 3.6
Frame	LBP [212]	27.1 ± 4.3	27.3 ± 4.1	21.83 ± 5.7	32.80 ± 2.5
	ULBP [239]	29.0 ± 3.4	28.6 ± 3.3	17.68 ± 4.2	39.70 ± 2.4
	RIULBP [239]	28.7 ± 3.7	28.7 ± 3.9	20.94 ± 5.1	37.06 ± 2.7
	CLBP [133]	28.7 ± 3.8	28.8 ± 3.6	18.88 ± 4.7	38.80 ± 2.5
	Haralick+RDWT [12]	28.9 ± 4.8	28.4 ± 4.6	21.82 ± 4.3	35.08 ± 4.9
	BSIF [157]	30.2 ± 7.0	30.2 ± 6.9	31.45 ± 8.8	29.19 ± 5.0
	LPQ [240]	28.7 ± 4.0	30.4 ± 3.8	40.30 ± 3.6	**20.50 ± 4.0**
	[14]	24.5 ± 5.1	25.4 ± 4.9	10.60 ± 5.9	40.26 ± 3.9
	Proposed (MagNet)	**18.0 ± 0.4**	**17.6 ± 0.3**	**8.72 ± 0.4**	26.47 ± 0.2

L-WLMP by at least 2.2% (in terms of EER). After score fusion of WLMP and NL-WLMP, the EER further reduces by over 2%;

- It is interesting to observe that the combination of Haralick+RDWT [11], which yields low EER on physical spoofing databases, provides the highest EER value of 25.6% in video-based digital attack detection. In case of frame-based attack detection, BSIF feature (linear filtering based algorithm) yields the lowest performance;

- The proposed WLMP feature histogram incorporates sorting in ascending order. However, when the difference values are sorted in descending order (i.e., higher weight to the least discriminant neighbor, reverse of the WLMP), the EER increases from 18.2% to 25.5% and 24.5% to 29.3% for video and frame based detection, respectively.

It is our observation that the alterations performed via SnapChat are seamless in nature, and therefore, it is a challenging task to differentiate between the bonafide and attack data. Since

Table 4.7: Classification performance of the proposed L-WLMP algorithm for video and frame based presentation attack detection on the proposed Snapchat database. The results are reported in terms of the equal error rate and average classification accuracy (%).

Input	BSIF Filter Bit	EER	ACER
Video	5	19.9	22.3
	6	21.8	23.2
	7	18.7	21.9
	8	18.2	20.6
	7 and 8	**17.6**	**19.5**
	5, 7, and 8	17.9	19.6
	5, 6, 7, and 8	18.0	20.0
	5, 6, 7, and 8 + PCA	23.2	24.5
Frame	5	26.2	27.3
	6	27.3	28.1
	7	25.7	27.2
	8	24.9	26.3
	7 and 8	**24.6**	**25.9**
	5, 7, and 8	24.7	26.0
	5, 6, 7, and 8	24.7	26.0
	5, 6, 7, and 8 + PCA	28.7	29.6

every video has a large number of frames, the performance is better for videos, compared to frames/images.

Instead of using a handcrafted filter (or identity filter) and non-linear filters in the original WLMP (described in section 4.3.1) and NL-WLMP respectively, a set of linear filters learned on patches of natural images are used. In this research, BSIF filters [157] of size 3×3 are adopted for convolution. The reason for selecting the BSIF filters over other linear filters is the effectiveness in texture feature extraction. Linear filters used in this research are trained on 50,000 natural image patches [145]. The learning of BSIF filters have two major steps: (1) whitening, and dimensionality reduction using PCA [347], referred as canonical preprocessing and (2) selection of statistical independent component of the filters by Independent Component Analysis (ICA) [72]. The WLMP algorithm with linear filter is termed as L-WLMP and is shown in Figure 4-13.

The L-WLMP algorithm uses different number of filters with respect to bit sizes, therefore the first experiment is performed to analyze the effect of bit length on the classification performance. Multi-bit BSIF magnitude patterns are extracted by concatenating the individual patterns obtained from each filter. The analysis in terms of multi-bit magnitude pattern is reported in Table 4.7.

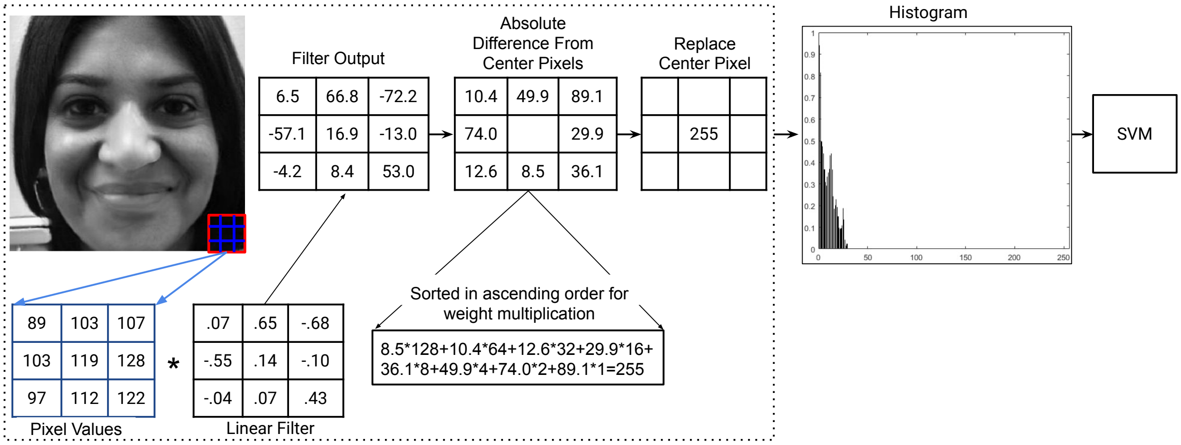


Figure 4-13: WLMP with convolution of image using linearly learned filters (L-WLMP).

For both video and frame based countermeasures, we observe that higher multi-bit filters yield lower EER and ACER. Feature fusion of two higher bit filters such as bits 7 and 8, yields the EER value of 17.6% and 24.6% for video and frame based attack detection, respectively. Based on the performance of higher bit filters and combination experiments, we observed that fusion of features from bit 7 and 8 yields the best results.

The detection performance of L-WLMP is 4-5% lower in comparison to the performance of NL-WLMP. In place of GoogLeNet filters, when we have used VGG-Face filters in NL-WLMP, the performance is 2-3% lower.

4.7.2 Results and Analysis on IdentityMorphing database

The experiments on the IdentityMorphing database are performed based on the evaluation protocol provided in Section 4.5. The performance of the proposed algorithm is compared with LBP, BSIF, and LPQ, which were the top three features on the Snapchat database and most popular in the literature of digital attack detection. Table 4.8 shows the performance of the proposed and top three existing algorithms. The trend in performance is similar to the SnapChat database and the observations are summarized below:

- The proposed algorithm yields an average EER value of 0.0% for frame based attack detection. While the bonafide presentation classification error rate is 0.4%, which is the least among all the algorithms, the attack presentation classification error rate is 0.0%;

Table 4.8: Results of the proposed MagNet and existing algorithms on the proposed *IDAgender* databases using frames/images as input. Results of two best performing algorithms are highlighted.

Database	**Features**	**EER (%)**	**ACER (%)**
Snapchat	LBP [212]	27.1 ± 4.3	27.3 ± 4.1
	LPQ [240]	28.7 ± 4.0	30.4 ± 3.8
	BSIF [157]	30.2 ± 7.0	30.2 ± 6.9
	VGG-16 [306]	17.7 ± 2.4	18.4 ± 2.3
	GoogLeNet [318]	28.1 ± 5.1	29.1 ± 4.9
	S-MIL [183]	**16.9 ± 3.6**	**18.2 ± 2.7**
	XceptionNet [285]	19.7 ± 4.7	23.6 ± 3.1
	ResNet-18 [166]	30.0 ± 5.9	31.6 ± 5.7
	Proposed (MagNet)	**18.0 ± 0.4**	**17.6 ± 0.3**
Identity Morphing	LBP [212]	**0.6 ± 0.2**	**0.9 ± 0.1**
	LPQ [240]	6.1 ± 0.3	6.2 ± 0.2
	BSIF [157]	6.2 ± 0.4	6.2 ± 0.2
	VGG-16 [306]	4.7 ± 1.1	9.7 ± 1.0
	GoogLeNet [318]	12.3 ± 2.1	11.5 ± 0.9
	S-MIL [183]	9.4 ± 1.2	11.7 ± 1.8
	XceptionNet [285]	7.9 ± 2.4	9.1 ± 1.1
	ResNet-18 [166]	8.5 ± 1.8	10.6 ± 1.2
	Proposed (MagNet)	**0.0 ± 0.0**	**0.2 ± 0.0**
FaceApp	LBP [212]	1.3 ± 0.8	2.7 ± 0.7
	LPQ [240]	**1.2 ± 0.4**	**1.3 ± 0.3**
	BSIF [157]	30.3 ± 4.4	30.5 ± 4.5
	VGG-16 [306]	18.3 ± 2.5	21.4 ± 2.3
	GoogLeNet [318]	23.7 ± 2.1	24.5 ± 2.7
	S-MIL [183]	8.6 ± 1.8	12.3 ± 1.2
	XceptionNet [285]	10.2 ± 3.6	14.7 ± 1.8
	ResNet-18 [137]	16.2 ± 3.3	14.8 ± 1.6
	Proposed (MagNet)	**0.4 ± 0.7**	**2.5 ± 0.4**

- The proposed features show an improvement of 78% in terms of ACER from the second best-performing feature i.e. LBP;

- The ACER of the proposed algorithm is 0.2%, whereas, the ACER of the CNN model is 9.7%;

- Micro texture features BSIF and LPQ, provides high EER value of 6.2% and 6.1% on the IdentityMorphing database. BSIF shows the highest EER and ACER;

- Even at lower False Positive Rate (FPR), the proposed algorithm yields high True Positive

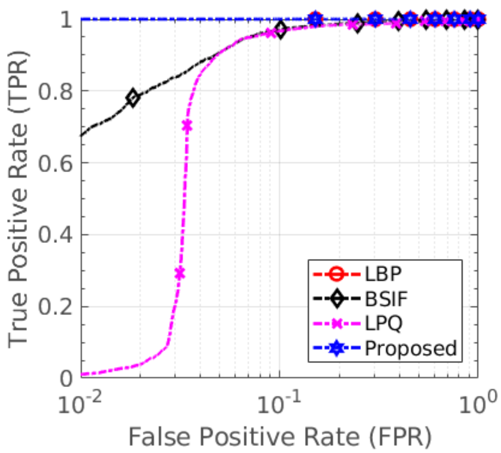


Figure 4-14: ROC curve for frame based presentation attack detection on the proposed Identity-Morphing database.

Rate (TPR). The TPR of proposed, BSIF, and LPQ feature is 100%, 68%, and 1.1% receptively at 1% FPR (Figure 4-14);

- Poor performance of existing texture features and CNN model shows the challenge in morph attack detection and importance of proposed efficient algorithm.

Similar to the Snapchat database results, the results on Identity Morphing also show that the features that have achieved high performance on physical attacks might not necessarily be the best set of digital attacks.

4.7.3 Results and Analysis on FaceApp database

The results on the FaceApp database using six fold cross validation as defined in Section 4.5 are given in Table 4.8.

- The proposed algorithm yields an average EER of 0.4% for frame/image based attack detection. On the other hand, BSIF histogram feature provides the highest EER and ACER values;

- The proposed algorithm shows the lowest APCER among all the features used for comparison, which is highly required in the high-security systems;

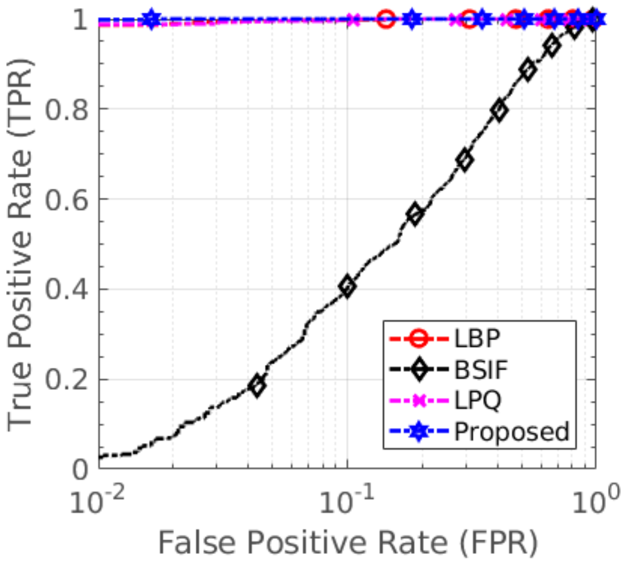


Figure 4-15: ROC curve for frame based presentation attack detection on the proposed FaceApp database.

- The EER and ACER of the CNN model is 17.9% and 18.9% (relatively) higher than MagNet, respectively;
- TPR of the proposed algorithm, LPQ, and BSIF features at 1% FPR is 99.8%, 98.8%, and 2.7%, respectively (from Figure 4-15). The results indicate that the proposed algorithm is robust to different kinds of face swap such as gender and age.

4.7.4 Findings from Each Subset of Proposed Database

The results show that the detection of morphed images generated using the Snapchat image is challenging compared to other social media platforms, including morphthing.com and FaceApp. The prime reason for comparably lower detection performance might be attributed to the post smoothing performed by Snapchat to make the morphed images look ready for upload on the social media accounts. In contrast, morphthing.com does not bother about any post-processing by itself. The images on the website cover a broad spectrum of faces that differ in race and ethnicity; the swapping of faces of different groups left the artifacts higher than the same group's morphing faces. On the other hand, FaceApp is a neural transfer-based app used for facial attribute conversions such as age and gender manipulation. That leads to the drastic change in the facial appearance and, hence, through the proposed algorithm to magnify these changes to capture it. Consequently leads to high detection performance. The other existing algorithms, including deep neural networks, do not aim to magnify such minute and facial attribute conversion artifacts and therefore perform significantly lower than proposed *'MagNet'*.

4.7.5 Statistical Significance of Results

We next evaluate the statistical significance of results obtained using WLMP, L-WLMP, and NL-WLMP. McNemar test [223] is performed on the test labels provided by the various descriptors mentioned above with null hypothesis being the results are statistically same. The output of different descriptors can be reported in terms of the confusion matrix which is then used to calculate the McNemar test statistics. The test statistics are given below:

$$H_0 : P_b = P_c, H_1 : P_b \neq P_c \tag{4.2}$$

where, H_0 is the null hypothesis, P_b and P_c are the probabilities of error yields by two different classifier. The McNemar test is performed on four sets of descriptor pairs:

- WLMP and L-WLMP,
- WLMP and NL-WLMP,
- WLMP and Fusion of WLMP + L-WLMP, and
- WLMP and Fusion of WLMP + NL-WLMP

Using the McNemar test, we observe that the accuracies provided by these different descriptors are statistically significant by rejecting the null hypothesis at 5% significance level.

4.7.6 Cross-Database Experiments

Apart from that, we have also performed experiments under cross-database settings utilizing proposed databases. When the proposed algorithm, namely '*MagNet*' and VGG-16 model is trained on the Snapchat database and tested on FaceApp, the EER of 12.0% and 46.1% is observed, respectively. The proposed algorithm surpassed the CNN model in terms of area under the curve (AUC) as well by a significant value and yielded 94.18%, whereas VGG-16 CNN yields only

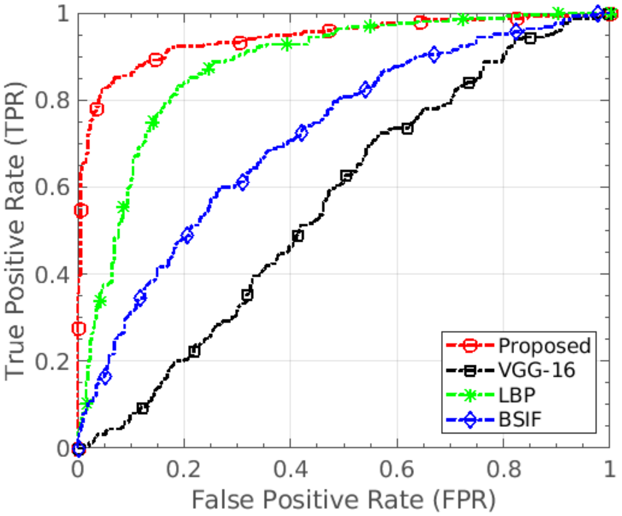


Figure 4-16: ROC of the cross-database experiment where the algorithms are trained on Snapchat and tested on FaceApp database. The proposed algorithm shows generalizability in handling unseen distortion type effectively.

56.02%. Compared to image features such as LBP, BSIF, and RDWT+Haralick [12], the AUC of the proposed algorithm is at-least 7.0% better. Fig. 4-16 shows the ROC curve of the experiment depicting the generalizability of the proposed algorithm in the handling of unseen manipulations. Similar higher performance has been observed when the proposed algorithm and CNN models such as VGG-16 and ResNet-18 are trained and tested on cross-database settings, including Snapchat vs. IdentityMorphing and IdentityMorphing vs. FaceApp.

Besides utilizing the proposed databases for cross-database evaluation, we have examined the robustness of the proposed detection algorithm on the existing database. The DeepFake [162] database contains the face morphed images generated using generative adversarial networks (GANs). In total, the database contains 640 tampered videos in low and high quality. When the proposed algorithm trained on the Snapchat database is tested on the morphed images of Deepfake, presentation attack detection error (APCER) of value 0.0% is achieved. In other words, on both quality subsets, the proposed algorithm yields perfect detection performance. The performance of the proposed algorithm is 47.80% and 28.2% better than the deep neural network namely VGG-Face [245] and recently proposed PFTD algorithm [215].

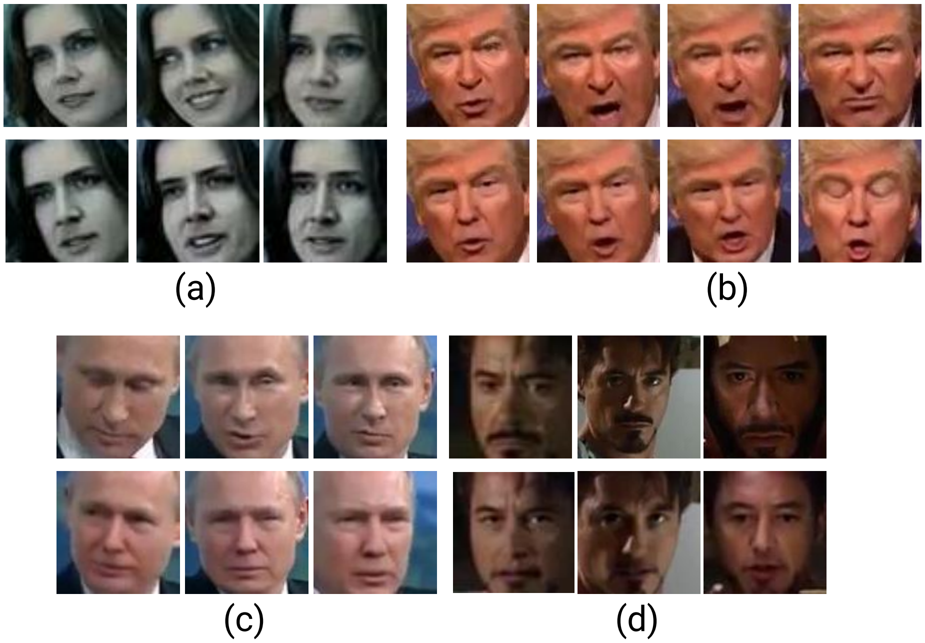


Figure 4-17: Real and morphed faces from the youtube collected videos. First row shows real samples and second row shows morphed samples.

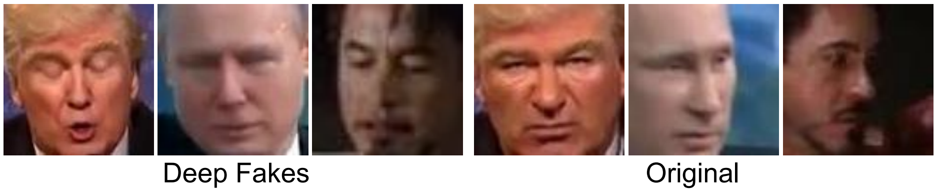


Figure 4-18: Samples of original and deepfake images mis-classified by the proposed algorithm.

4.7.7 Real World Evaluation

We have also evaluated the performance of the proposed detection algorithm on artificially generated fake videos. The success of deep learning algorithms in various computer vision related tasks has garnered significant interest. These technologies can be efficiently used to create artificial images and videos [121]. The popularity and easy availability of these tools, such as *Deep Fakes and Fake APP*[11], have led to increase in fake videos on Internet by multiple folds. Pursuing the misuse or the dark side of these effective algorithms, people have created fake porn video of a famous female celebrity[12].

To evaluate the performance of the proposed algorithm on this attack, we collected real and swapped faces from online YouTube videos[13]. In total, $2,338$ real and deep-fakes swap faces are

[11] https://www.fakeapp.org/

[12] https://tinyurl.com/y9ydm4yy/

[13] As part of the database release, we will also release the YouTube links from where this database is collected

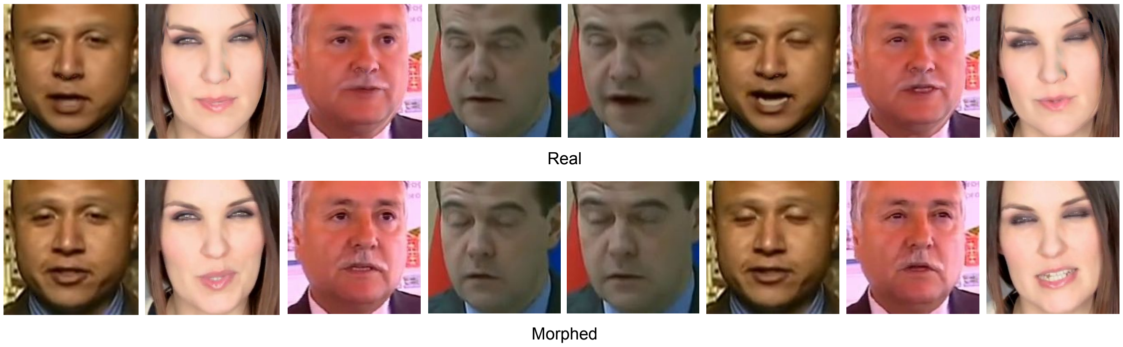


Figure 4-19: Real and morphed faces from the FaceForensics videos.

collected. Figure 4-17 shows the real and deep fake generated morphed samples collected from the YouTube. For example, in Figure 4-17 (c) Trump's face is morphed on Putin's face and similarly in (d) Elon's face is morphed over Downey Jr.'s face.

Protocol and Result: $2,388$ faces pertaining to each class are divided into five random folds, where each time four folds are used for training the classifier, and the images from the remaining one fold are used for testing. The average EER value with standard deviation is calculated to report the performance of the proposed algorithm. The proposed algorithm yields $2.05 \pm 0.49\%$ EER on this new online collected database. While majority of the samples are correctly classified, further analysis of mis-classified samples shows that poor image quality is a covariate in attack detection. As shown in Figure 4-18, if the image is of poor quality, it becomes challenging to determine whether the image is attacked or not. In comparison, the second best EER of $6.89 \pm 0.10\%$ is achieved using CNN based digital presentation attack detection. Finally, computationally, the proposed algorithm requires 0.2 seconds to process an input image whereas, CNN based approach requires 3.0 seconds (using the same the computing platform).

4.7.8 Experiments on Existing Database

In the previous sections, the experiments on either proposed databases collected in our lab are conducted or on the videos collected from YouTube are performed. However, the tremendous improvement of the machine learning algorithms such as generative adversarial networks (GANs) have made the generation of synthetic images, transferring from facial attributes among faces, or morphing two faces together an easy task. By utilizing these algorithms, several challenging digitally tampered face databases are prepared in literature. Therefore, to further show the strength

Table 4.9: Classification accuracy of the proposed and existing algorithm for video based morphed attack detection on the FaceForensics database.

Algorithm/Network	Accuracy %
Steganalysis Features + SVM [104]	99.40
Cozzolino et al. [78]	99.60
Bayar and Stamm [34]	99.53
Rahmouni et al. [271]	98.60
Raghavendra et al. [272]	97.70
Zhou et al. [376]	99.93
XceptionNet [69]	99.93
MesoNet [10]	96.80
VGG-16 [306]	99.50
ResNet-50 [137]	99.93
ResNet-152 [137]	99.89
Multi-patch ResNet-18 [167]	99.96
Proposed (MagNet)	**100.00**

of the proposed face morphed detection algorithm, one of the challenging large-scale morphed databases namely FaceForensics [284] is also used along with database prepared by Jain et al. [149]. We have used the FaceForensics database [284], which contains half a million edited face images. In total, the database contains 704 and 150 training videos of real and morphed classes each. For evaluation, subject independent real and morphed videos of 150 subjects. For the collection of morphed videos Face2Face [325] reenactment technique is used. Figure 4-19 shows the real and morphed images of the FaceForensics database. The morphed image detection results of MagNet using the pre-defined database protocol along with existing algorithms range from handcrafted features to deep CNN features are reported in Table 4.9.

In Steganalysis + SVM, the co-occurrence features from horizontal and vertical edge images are captured. This technique wins the first challenge of image forgery detection [77]. Cozzolino et al. [78] extracted the handcrafted features from the deep learning architecture for the detection of morphed images. Bayar and Stamm [34] propose 18 layers convolutional neural network for face morphing detection. The network consists of a constrained convolutional layer that is designed to suppress the high-level information. Rahmouni et al. [271] extracts the four statistics features from the CNN architecture. The features of the first fully connected layer of VGG-19 and AlexNet are concatenated by Raghavendra et al. [272]. The concatenated elements are then passed to the Probabilistic Collaborative Representation classifier for real and morphed images classification. Zhou et al. [376] have performed the fusion of the scores obtained from two CNNs: GoogLeNet Inception V3 and triplet CNN. The state-of-the-art Xception network is fine-tuned for face morphing detection in a transfer learning fashion. The results of the existing algorithms are taken from [284]. The algorithm [272], which fused the features from two CNNs, i.e., AlexNet and VGG-19, has 2.3% lower detection accuracy than the proposed MagNet. The proposed algorithm either outperforms the algorithms or perform comparable the algorithms utilize deep CNN features, statistics, and steganalysis features. These extensive experiments on multiple proposed databases, including a variety of morphing techniques, real-world databases, and existing challenging databases, show the capability of the proposed MagNet algorithm in face morphing detection.

On frame based evaluation: where every single frame is classified as real or morphed, the proposed MagNet algorithm yields 1.00%, 1.07%, and 0.93% ACER, BPCER, and APPCER respectively. BPCER represents the bonafide (real) presentation classification error rate, and APPCER represents the attack presentation classification error rate. ACER is the average of BPCER and APPCER. The MagNet algorithm shows the EER value of 1.67% and 0.00% for the frame and video-based evaluation on the FaceForensics database, respectively.

Jain et al. [149] have performed the attribute transfer using StarGAN [68] and generated $18,000$ face images. Nine different attributes, such as brown hair, gender, aging, hair, and gender together, are transferred. For the classification of GAN vs. real images, similar protocol mentioned by Jain et al. is followed. The database is divided into training, validation, and testing set. The testing set contains random $1,500$ real and $1,000$ GAN generated images, while the validation set contains 500 images of both classes (i.e., real and GAN). The detection performance of the proposed and existing algorithms is given in Table 4.10. The proposed algorithm outperforms the existing state-of-the-art algorithms by at-least 0.3%. Other than the detection accuracy, the EER, APCER, and BPCER of the proposed algorithm are 0.0%, 0.0%, and 0.4%, respectively.

The proposed algorithm is also evaluated on the images prepared by Jain et al. [148] using super-resolution GAN [175]. Similar to StarGAN, nine facial attributes are transferred on the CelebA database [196]. The detection performance is measured using a similar protocol used for starGAN images. The database is divided into training $(27,500$ images), validation $(1,000$ images), and testing images $(2,500$ images), respectively. The proposed algorithm yields 99.88% detection accuracy with 0.08% EER, 0.06% BPCER, and 0.20% APCER values.

Table 4.10: Classification accuracy of the proposed and existing algorithm for GAN generated images.

Algorithm	Accuracy %
Bharati et al. [38] (Unsupervised DBM)	81.90
Bharati et al. [38] (Supervised DBM)	87.10
Jain et al. [149] (Thresholding)	99.48
Jain et al. [149] (SVM)	99.65
Proposed (MagNet)	**99.96**

We have also tested the generalizability of the proposed **'MagNet'** algorithm using unseen GAN type images. In this setting, the images generated using one type of GAN is used for training while testing is done on another type of GAN images. When StarGAN images are used for training and SRGAN images are used for testing, the MagNet yields 95.32% detection accuracy with 3.52% EER, 2.27% BPCER, 6.90% APCER, and 4.58% ACER values.

4.8 Summary

Rich literature on physical presentation attack detection shows the maturity of algorithms in protecting the face recognition systems. However, these systems are still vulnerable to digital attacks such as morphing. With the advancements in deep learning and computer vision, several easy to use applications are available where with few taps, an image can be easily altered. While most of the times, these applications are for entertainment purposes, they can be also used for digitally attacking face (biometric) recognition systems. This chapter extends the research on digital attacks and presents a ***IDAge**nder* database of morphed faces using three sources of alterations, Snapchat, FaceApp, and MorphThing.com. These face attacks show the vulnerability of the face recognition system both in mobile phones and a commercial system. To address these attacks, a new computationally efficient digital presentation attack detection algorithm is proposed using a novel descriptor, termed as Weighted Local Magnitude Patterns. The proposed algorithm achieves lower error rates compared to several existing texture feature based approaches on the proposed databases. Apart from that, the strength of the proposed algorithm is also evaluated on the face swap images developed using generative adversarial networks. The GANs generate the high quality face swap images that by just looking any human can fail in identifying that whether an image in question is real or morphed. The superiority of the proposed algorithm over several complex deep learning based algorithms on such complex databases showcase its efficiency. In future, we plan to extend the proposed algorithm by designing a unified algorithm for different kinds of digital attacks.

Chapter 5

Image Transformation based Defense Against Adversarial Perturbation on Deep Learning Models

5.1 Introduction

Deep learning algorithms are making unprecedented improvements in a variety of tasks [9], [141], [146], [206], [230], [312]. Autonomous driving, natural language understanding, and playing complex games such as Atari are successful examples of modern-day artificial intelligence approaches. However, with good things, there are always some perils as well. Several researchers have shown that deep learning algorithms are prone to attacks [20, 243, 319]. As shown in Figure 5-1, highly popular VGG-16 architecture based models can be fooled by adversarial algorithms such as Elastic Net [64], L_2 [57], and universal adversarial perturbations [231]. The research in adversarial attacks ranges from image based attacks to learning sophisticated attacks for individual deep learning models. Image processing attacks involve adding artificial lines, occlusions, or noise in the image in a manner that confuses the classification algorithm [122, 124, 168]. Learning based attacks leverage the singularities of the deep neural network classifier and generate noise patterns that lead to misclassification of the input data.

Two main research directions that are being pursued to address this challenge are (i) attack

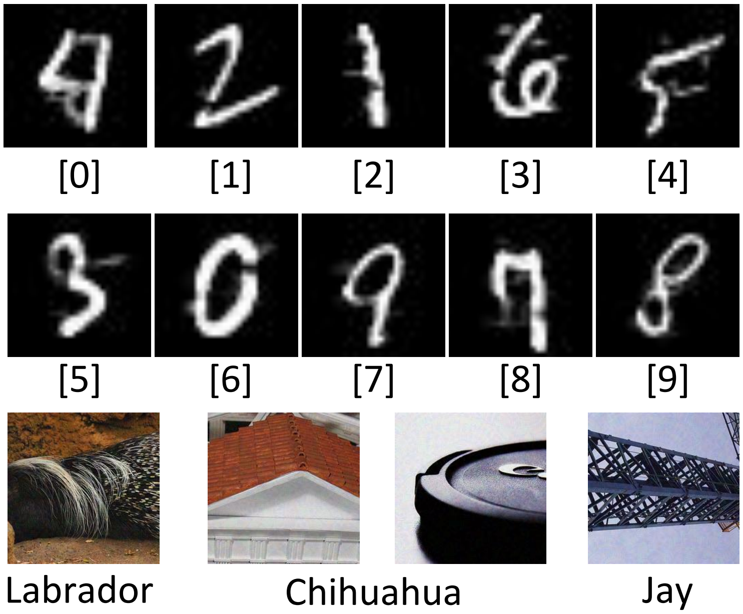


Figure 5-1: Adversarial images and corresponding classification results (using VGG-16 model) on MNIST and ImageNet databases. The adversarial images are generated using Elastic-Net, L_2, and Universal perturbation. The predicted (incorrect) class label of the images are written below it.

detection and (ii) mitigating the effect of attacks. Since the nature of these security attacks is different, the research on attack detection and mitigation is still focused on designing separate algorithms for different models, attacks, and databases [18]. However, it is not pragmatic to create a detection or mitigation algorithm for every attack. Often times, there are new or unseen attacks on the system, and it will be difficult for such focused algorithms to detect and mitigate unseen attacks. Therefore, in this research, we aim to design detection and mitigation algorithms that are generalizable across different kinds of attacks and deep architectures.

Inspired from the watermarking literature in which detection algorithms are targeted towards detecting imperceptible watermarks embedded in the image [258], we hypothesize that the adversarial perturbations can be considered analogous to imperceptible watermarks. In this research, we first propose a transformation-based adversarial example detection algorithm that transforms the input image using ‘Sine’ or ‘wavelet’ transforms and then encodes the spatial envelope of the transformation output. The effectiveness of the proposed algorithm is evaluated on multiple attacks, databases, and deep models. The results show that while most of the existing detection algorithms provide defense against simple adversarial attack generation algorithms such as FGSM, L-BFGS, and black-box based, they are ineffective against sophisticated attacks such as C&W’s L_2

[57], Elastic-Net [64], and Universal perturbation [231]. The proposed algorithm efficiently detects both simple and sophisticated attacks, including Iterative Fast Gradient Sign Method (IFGSM), L_2, Elastic-Net, Universal perturbation, and Fast Feature Fool (F3).

The second contribution of this research is designing a mitigation algorithm. We propose a wavelet denoising-based algorithm with soft thresholding to mitigate the effect of adversarial perturbations. The results show that the proposed algorithm is not only reduces the noise embedded in the image by increasing the structural similarity between the images but it also improves the face verification performance. The key highlights can be summarized as follows:

- a novel adversarial detection algorithm is proposed utilizing input transformation, image features, and support vector machine classifier;
- extensive experiments pertaining to **'intra'** and **'inter'** database, attack, and DNN models shows the strength of the proposed detection algorithm;
- a novel adversarial mitigation algorithm is developed to remove the noise and restore the accuracy of DNN models.

5.2 Literature Review

Researchers have proposed different kinds of attack algorithms in the literature. Akhtar and Mian [18], Yuan et al. [364], Bulusu et al. [52], and Ren et al. [279] have presented surveys of algorithms for adversarial examples generation as well as algorithms to defense against these attacks. Goodfellow et al. [119] and [120] have also reviewed some existing adversarial machine learning algorithms and the possible future direction towards it. Table 5.1 summarizes the adversarial perturbation algorithms and the deep models which they attack. In this section, we provide a detailed review of detection and mitigation algorithms available in literature.

5.2.1 Attack Generation

The field of adversarial example generation on deep learning system started with the introduction of minimal perturbation vector (ρ) which can fool the classifier [319]. The minimal perturbation

Table 5.1: Summarizing the literature of attack generation algorithms.

Authors	Attack Algorithm	Database	Deep Models
Szegedy et al. [319], 2014	L-BFGS	MNIST, ImageNet	AlexNet
Goodfellow et al. [122], 2015	FGSM	MNIST, ImageNet	GoogLeNet
Kurakin et al. [168], 2016	Basic Iterative	MNIST, CIFAR-10	AlexNet
Moosavi-Dezfooli et al. [232], 2016	DeepFool	MNIST, CIFAR-10, ILSVRC 2012	LeNet, NiN, GoogLeNet, CaffeNet
Su et al. [314], 2019	One Pixel	CIFAR-10, ImageNet	NiN and VGG
Carlini and Wagner [57], 2017	Logit Layer Regularized Loss	MNIST and CIFAR-10	7 layer CNN
Moosavi-Dezfooli et al. [231], 2017	Universal	ILSVRC 2012	VGG, GoogLeNet, ResNet-152, CaffeNet
Mopuri et al. [233], 2017	Fast Feature Fool	ILSVRC 2012	VGG, GoogLeNet, CaffeNet
Baluja and Fischer [30], 2017	Transformation Networks	MNIST, ImageNet	AutoEncoder based
Poursaeed et al. [259], 2017	Generative Models	ImageNet	VGG-16, VGG-19, Inception-v3
Kwon et al. [170], 2018	Multi-target adversarial	MNSIT	Custom
Han et al. [135], 2019	Multi-Target Attack (MAN)	CIFAR-10, ImageNet	VGG16, ResNet (32 and 152)
T. Co et al. [71], 2019	Gabor Noise	ImageNet	Inception-v3, ResNet-50, VGG-19
Brunner et al. [51], 2019	Boundary attack using image patch as starting point	Subset of ImageNet	Inception-v3
Zhao et al. [375], 2019	Gradient free zeroth order and bayesian optimization attack	MNIST, CIFAR-10, ImageNet	Custom and Inception v3
Sharif et al. [304], 2019	Adversarial Generative Nets	Custom	VGG, OpenFace

vector is optimized using the box-constraint algorithm L-BFGS.

$$||\rho||_2 \;\; s.t. \;\; C(I_c + \rho) \neq l; \;\; I_c + \rho \in [0,1]^m \tag{5.1}$$

where I_c denotes the clean image, l is the true label of the image, $||\cdot||_2$ denotes the L_2 norm, and C represents the deep learning classifier.

Goodfellow et al. [122] proposed adding the gradient computed over each image to fool the classifier. The perturbation can be computed using following equation:

$$\rho = \epsilon \; sign(\nabla \jmath(\theta, I_c, l)) \tag{5.2}$$

where θ denotes the parameters of the deep classifier and $\nabla \jmath$ computes the gradient w.r.t. θ. The magnitude of ρ is controlled using ϵ. The above method computed the perturbation for each image and added a single time into the input image. The algorithm of adding gradient is referred to as the 'Fast Gradient Sign Method (FGSM)'. Later, Kurakin et al. [168] proposed the iterative version of FGSM, which aims to increase the loss of the classifier. The variant of FGSM based on minimization of L_2 norm of the perturbation can be defined as:

$$\rho = \frac{\nabla \jmath(\theta, I_c, l)}{||\nabla \jmath(\theta, I_c, l)||_2} \tag{5.3}$$

While the above perturbation generation algorithm is targeted towards minimizing the L_∞ or L_2

norm, Papernot et al. [243] proposed the Jacobian-based Saliency Map Attack (JSMA) method based on minimization of L_0 norm. The process is based on the computation of a saliency map of the image pixels and modifies specific pixels participating in classification. Kurakin et al. [168] introduced the iterative version of FGSM by iteratively computing the gradient and clipping the ε ball. IFGSM method for adversarial image generation minimizes the L_∞ distortion between x and x_0.

$$x'_1 = x_0 - \varepsilon * sign(\triangledown J(x_0, t)) \tag{5.4}$$

Here, $\triangledown$ defines the gradient of the network concerning the current set of parameters on image x_0. The other variants based on minimization of L_1 and L_2 distances (denoted by FGSM-L_1 and FGSM-L_2) between the original and adversarial examples are:

$$x_1 = x_0 - \varepsilon \frac{\triangledown J(x_0, t)}{\left\| \triangledown J(x_0, t) \right\|_q} \tag{5.5}$$

where $q = 1$ and 2 for L_1 and L_2 distortions respectively. Similar to FGSM, variants of IFGSM attack are also used for adversarial example generation. The L-BFGS, FGSM, and JSMA either modify each pixel of the image or a few pixels of the image. Su et al. [314] proposed the extreme case in pixel perturbation which modifies a single pixel for perturbation. Later, Carlini and Wagner [57] proposed three variants of adversarial generation algorithm based on the minimization of L_∞, L_2, and L_0 of the logit-layer representation. The algorithm based on the L_2 norm is found to be the most sophisticated attack in literature, resistant to various defense techniques [244]. On a similar concept, Chen et al. [64] have developed the Elastic-Net adversary, which is a combination of L_1 and L_2 norms.

The above-mentioned adversaries have a limitation that the perturbation vector is computed over each image and hence requires solving the optimization function for every image. To solve this problem and to increase the complexity/efficiency of adversarial defense algorithms, Moosavi-Dezfooli et al. [231] and Mopuri et al. [233] have proposed data-dependent and independent universal perturbation algorithms, respectively. The ‘universal’ perturbation algorithm aims to compute a single perturbation vector, which can fool the classifier on ‘any’ image. The attacks mentioned above are computed for each image specifically. Moosavi-Dezfooli et al. [231] and

Mopuri et al. [233] presented universal perturbation, which learns a function to fool the classifier for ‘any’ image. The universal perturbation is defined as:

$$P_{I_c \sim \Im_c}(C(I_c) \neq C(I_c + \rho)) \geq \delta \; s.t. \; \|\rho\|_p \leq \xi \tag{5.6}$$

where $C(\cdot)$ is the deep classifier, I_c is the clean image, P$(\cdot)$ is the probability, ξ is the constant, and $\delta \in [0, 1]$ is the fooling ratio.

Other generation algorithms based on classifier decision boundary [232], spatial transformation [351], and training of neural network [30] have also been proposed in the literature. While the attacks mentioned above modify the pixel of the input image, there exists some work that proposes the existence of attacks in the physical world through the development of 3D objects. Sharif et al. [303] have proposed the development of 3D eyeglasses, which, after wearing, can fool the face recognition system. Similarly, Athalye et al. [25] and Brown et al. [50] have proposed the generation of transformation invariant adversarial examples which, when captured back, can fool the deep learning models. Han et al. [135] have proposed the multi-targeted adversarial attack using the image and label encoder-decoder network. Co et al. [71] showed that CNN models are even sensitive towards Gabor noise and can also act as a universal perturbation. Brunner et al. [51] have used the patch on another image as a starting point to craft the boundary-based attack. Zhang et al. [375] proposed the gradient regime based zeroth order and Bayesian optimization-based black-box attacks. Table 5.1 summarizes the attacks along with the databases and models they are evaluated on.

5.2.2 Attack Detection

Table 5.2 summarizes the research efforts that have undergone for detecting adversarial perturbations. As the attacks are based on both image transformations and learned from the architecture, the detection techniques have also explored both the directions: *image processing-based* and *learning-based*. Image processing-based techniques are related to data compression, augmentation, and distribution, while learning-based algorithms constitute classification-based detection algorithms.

Lu et al. [200] analyzed the outputs of the ReLU layers of the network, and found that they are different for original and adversarial images. To model this observation for differentiating between

Table 5.2: Summarizing the literature of adversarial detection algorithms.

Authors	Algorithm	Database	Attacks	Limitations
Lu et al. (2017) [200]	Radial basis function with SVM	CIFAR-10, ImageNet	FGSM, IFGSM, DeepFool	Fails for complex attacks
Metzen et al. (2017) [229]	Binary classification based sub-network	CIFAR-10	FGSM, IFGSM, DeepFool	Not robust
Li and Li (2017) [184]	Statistics of Convolution filters	ImageNet	LBFGS and EA	Fails on complex attacks such as L_2
Grosse et al. (2017) [128], Hosseini et al. (2017) [143]	Additional class augmentation	MNIST, DREBIN, MicroRNA, GTSRB	FGSM, JSMA, black box	No suitable for complex attacks
Lee et al. (2017) [176]	Generative Adversarial Networks training	CIFAR-10, CIFAR-100	FGSM	Needs adversarial training
Meng and Chen (2017) [227]	External detectors	MNIST, CIFAR-10	FGSM, IFGSM, DeepFool, C&W	Not robust [56]
Feinman et al. (2017) [95]	Density feature space of dropout networks	MNIST, CIFAR-10	FGSM, IFGSM, JSMA, C&W	Fails for complex attacks
Tramèr et al. (2018) [330]	Ensemble training with adversarial data while training	ImageNet	FGSM, IFGSM	Scalability on complex attacks
Akhtar et al. (2018) [17]	Perturbation Rectifying Network	ImageNet	Universal	Only evaluated on universal perturbation
Goswami et al. (2018) [124]	Filter response of hidden layers	MEDS and PaSC	Black-box, universal, DeepFool	Not robust for complex attacks
Zhang et al. (2018) [367]	Binary CNN Detector	MNIST, CIFAR-10, subset of ImageNet	FGSM, Iterative, and C&W	Not generalize when trained on weak adversaries
Ma et a. (2018) [208]	Local Intrinsic Dimensionality	MNIST, CIFAR-10, SVHN	FGSM, Iterative, Saliency, C&W	Not evaluated on ImageNet, and Universal attacks
Lee et al. (2018) [177]	Mahalanobis distance score	SVHN, CIFAR-10	FGSM, BIM, DeepFool, CW	Not extensively evaluated on ImageNet, Universal attacks, and under unseen attack scenarios
Samangouei et al. (2018) [287]	Generative Network	MNIST, F-MNIST	FGSM, RAND+FGSM, C&W	Not evaluated on complex databases
Goswami et al. [123] (2019)	CNN Filter Values	MEDS, PaSC, MBGC	Black-box, EAD, l_2	Not generalize against optimization based attacks
Zhao et al. [374] (2019)	Eigen Values Features	MNIST, CIFAR-10	FGSM, Iterative, and Proposed	Not tested against complex attacks
Taran et al. [322] (2019)	Randomization	MNIST, F-MNIST, CIFAR-10	C&W	Not evaluated on multiple attacks and used limited testing samples

the non-perturbed and adversarial images, they trained an SVM classifier with RBF kernel on the output produced by the ReLU layer. Metzen et al. [229] introduced a binary classifier at the end of the network for adversarial example detection. Li and Li [184] computed the statistical features by applying PCA on the features of the intermediate layers of CNN. Grosse et al. [128] and Hosseini et al. [143] proposed adversarial detection algorithm with the introduction of an extra class label in the targeted model. Lee et al. [176] used the Generative Adversarial Networks

(GANs) to generate the perturbation images and used these images to learn the defense against the adversarial perturbations. Meng and Chen [227] proposed the MagNet network by using one or more external classifiers for the detection of adversarial examples. Feinman et al. [95] used kernel density estimation as the features followed by a binary classifier for detecting adversarial and clean images. Goswami et al. [124] have shown that the intermediate representations of the hidden layers of the DNN model are different from clean and adversarial input. For easy adversary, Kolter & Wong [348], Sinha et al. [307], and Raghunathan et al. [270] proposed verified adversarial defense on MNIST database. Recently, Prakash et al. (2018) [260] proposed a pixel deflection technique for mitigating the effect of perturbation. However, if not integrated with an efficient detection algorithm, it also reduces the performance of original images. Lu et al. [201] and Tian et al. [326] have argued that the adversarial examples can be easily detected in transformation space. They have performed transformations such as rotation and shift for the detection of adversarial examples. Lu et al. [201] have only shown the results for physical adversarial attacks. Whereas, Tian et al. [326] have not extensively evaluated the detection algorithm against multiple easy and challenging attacks and highly dependent on the number of transformed images. Ma and Liu [207] have used the internal characteristics of the CNN models to craft an effective attack detection algorithm. Tow different probabilistic entities, namely provenance invariant and value invariant, are measured. Liu et al. [191] have drawn the comparison of adversarial attacks with steganalysis and proposed the algorithm by modeling the dependency of adjacent pixels using a hidden Markov model.

5.2.3 Attack Mitigation

Table 5.3 summarizes the research efforts towards attack mitigation. Dziugaite et al. [90], Das et al. [81], and Guo et al. [131] present the data compression based defense against adversarial attacks. They have used JPEG compression to mitigate the effect of adversarial perturbation based on the Fast Gradient Step Method (FGSM) attack. Inspired by JPEG compression based defense, Bhagoji et al. [37] presented the compression using PCA, which in turn increases the corruption in the input image. Similar to data compression, Xie et al. [353] proposed random resizing/ subsampling of the input sample to reduce the effect of adversarial noises. Training deep learning

Table 5.3: Summarizing the literature of adversarial mitigation algorithms.

Authors	Algorithm	Database	Attacks	Limitations
Gu and Rigazio (2014) [129]	Deep Contractive Networks	MNIST	LBFGS	Simple attack and database
Luo et al. (2016) [205]	Foveation	ImageNet	LBFGS, FGSM	Needs retraining and not evaluated on complex attacks
Papernot et al. (2016) [244]	Defensive distillation	MNIST, CIFAR-10	Saliency Map	Needs additional training
Dziugaite et al. (2016) [90], Das et al. (2017) [81]	JPEG based Data Compression	ImageNet	FGSM, DeepFool, C&W	Either not effective or need adversarial retraining
Bhagoji et al. (2017) [37]	Principal Component Analysis	MNIST	FGSM, IFGSM, C&W	Not robust
Zantedeschi et al. (2017) [365]	Data augmentation based	MNIST, CIFAR-10	FGSM, DeepFool, C&W	Not robust [56]
Liang et al. (2018) [188]	Scalar quantization and spatial filtering	ImageNet, MNIST	FGSM, DeepFool, C&W	Not robust
Guo et al. (2018) [131]	Input transformations such as cropping, rescaling, compression	ImageNet	FGSM, IFGSM, DeepFool, C&W	Needs additional training
Goswami et al. (2018) [124]	Selective dropout	MEDS and PaSC	Black-box, Universal, DeepFool	Not robust
Prakash et al. (2018) [260]	Pixel Deflection	ImageNet	FGSM, LBFGS, JSMA, Deepfool, C&W	Small subset of database
Goswami et al. (2019) [123]	Filter dropout	MEDS, PaSC, MBGC	Black-box, EAD, l_2, Universal, DeepFool	Not robust
Wang et al. (2019) [342]	Random Switching of CNN blocks	MNIST, CIFAR-10	FGSM and PGD	Highly depend on switching parameters
Ghosh et al. (2019) [110]	Variational Autoencoder	MNIST	FGSM	Not evaluated on complex attacks
Zhang et al. (2019) [371]	Randomization and Discretization of Images	MNIST, subset of ImageNet	PGD	Not robust against significant amount of noise

models using the samples generated using Gaussian distribution can also enhance the performance of the systems [365]. Luo et al. [205] presented the 'foveation' based defense mechanism, which involves applying neural networks in different regions of the images. Gu and Rigazio [129] learned the Deep Contractive Network (DCN) based on the advantage of denoising auto-encoders, where the smoothness penalty is introduced while training the model. This defense shows the effectiveness against box constraint-based attacks. Liang et al. [188] treated the adversary as the noise in the input images and used spatial quantization and filtering techniques as a defense against them. Goswami et al. [124] have proposed the mitigation algorithm by switching off the adversarially affected hidden nodes while performing recognition. Guo et al. [131] proposed various image transformations based adversarial mitigation algorithms. The random cropping and rescaling, compression, total variance minimization, and image quilting are applied on the ImageNet database to reduce the effect of adversarial noise. Due to the non-differentiable nature of these image transformations, the proposed method is said to be robust to defeat. Recently, Shao et al. [302] have shown the drawback of existing defenses against open-set classes and proposed a defense network for unseen image categories. Wang et al. [338] have performed the study in learning the relationship between high-frequency information processing by the CNNs and their robustness. It is argued that CNN filters over adversarial images are smoother than learned on clean images. The adversarial robustness is presented over simpler attacks and found less effective for lower strength perturbation. The introduction of one single perturbation, which can fool the target classifier on any image, increased the complexity of adversarial defense algorithms. The availability of these perturbation in the physical world using adversarial patch [50] and ineffectiveness [24, 56] of recent defense algorithms [227, 244, 281, 310, 352] shows the urgent need of effective defense against them. Reddy et al. [277] have shown that the robustness can be achieved by inspiration for biological neurons, which utilizes non-uniform sampling and multiple receptive fields.

5.3 Proposed Adversarial Detection Algorithm using Image Transformation

Attack generation algorithms have primarily focused on incorrectly predicting the label while maintaining visual appearance. This kind of constrained optimization can be defined as adding noise to the image.

$$x' = x \oplus r, \quad s.t. \; minimize \; ||r||_2 \tag{5.7}$$

where, r is the adversarial noise optimized with the constraint of imperceptibility and x is the input data. This noise can be embedded into the image using different kinds of techniques. Therefore, the algorithm must be generalizable to different attack generation algorithms. The adversarial

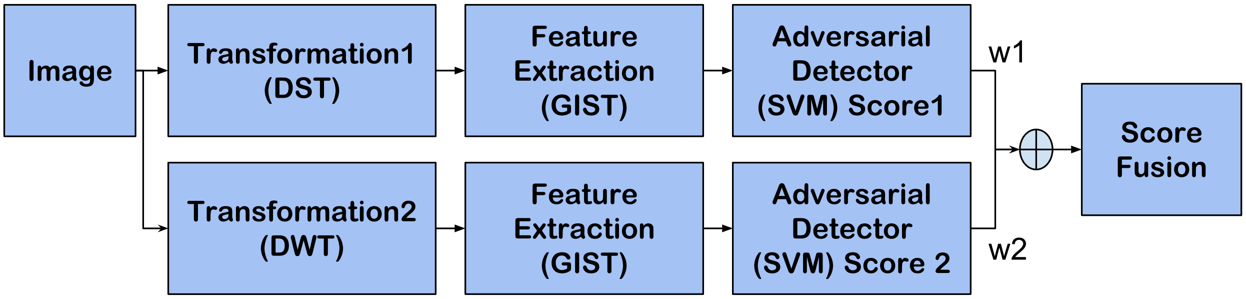


Image
Transformation1 (DST)
Feature Extraction (GIST)
Adversarial Detector (SVM) Score1
w1
Transformation2 (DWT)
Feature Extraction (GIST)
Adversarial Detector (SVM) Score 2
w2
Score Fusion

Figure 5-2: Steps involved in the proposed adversarial detection algorithm.

noise added into the original image can be perceived as embedding watermark into the source images. The adversarial noise optimized using the DNN model is added with the aim of visual imperceptibility. Motivated by this observation, the proposed algorithm first computes the image transformation followed by encoding the global signature of the image (Figure 5-2).

5.3.1 Image Transformations

Image transformations map the image from the spatial domain to another domain, which encodes certain specific properties. For instance, discrete Fourier transform (DFT) decomposes an image into constituent frequencies; discrete wavelet transform decomposes an image into four approximation, diagonal, horizontal, and vertical sub-bands. It is our assertion while the noise is imperceptible in the spatial domain, it may be easier to detect in the transformed domains. Therefore, we explore the effectiveness of different transformation techniques summarized below. Figure 5-3 shows the visualization of varying image transformations applied on both original and perturbed images.

Discrete Cosine Transform (DCT): DCT provides two advantages: (i) the representation of visually significant frequency information using only a few coefficients and (ii) discrimination based on the compression provided. DCT of an input ‘A’ can be written as following where, N is the dimension of ‘A’.

$$S(k) = \sum_{n=0}^{N-1} A(n) cos \frac{\pi k n}{N+1} \tag{5.8}$$

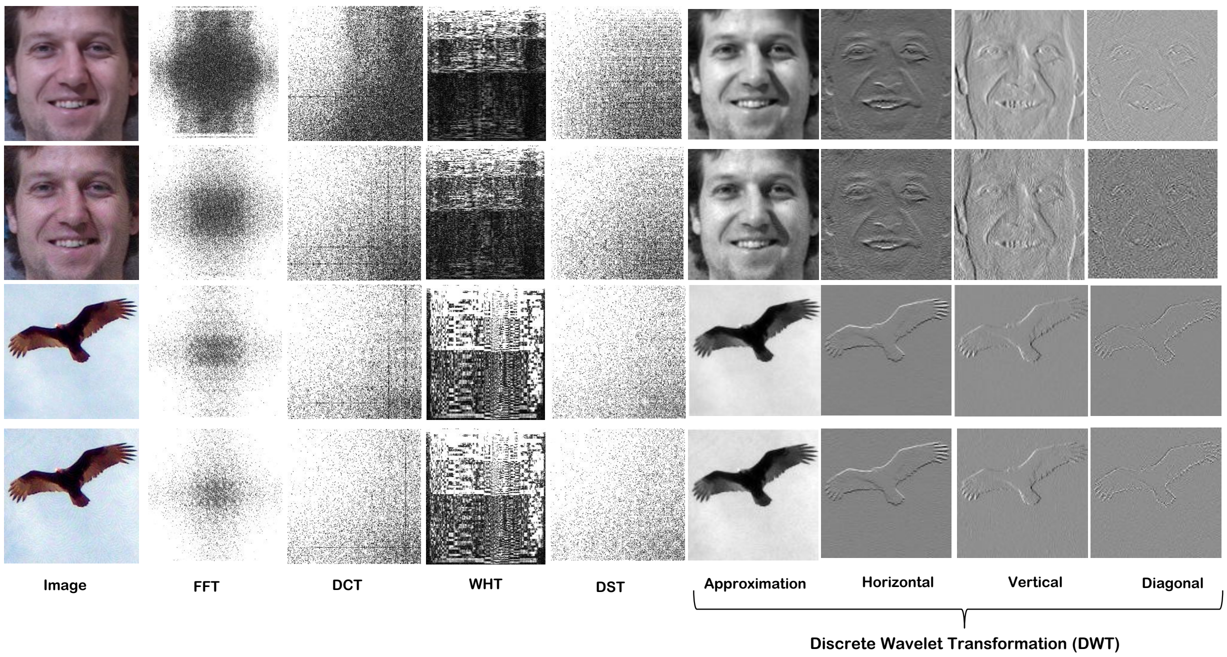


Figure 5-3: Visualization of different image transformations. The first and third rows comprise the original images and the second and fourth rows comprise universal adversarial perturbed image.

Fast Fourier Transform (FFT): FFT decomposes the signal in both time and frequency. Decomposing the values into different frequencies can help in identifying the noise on the spatial frequencies present in the non-perturbed images. FFT of an input 'A' can be written as:

$$S(k) = \sum_{n=0}^{N-1} A(n) e^{-2\pi i n k / N} \tag{5.9}$$

Discrete Sine Transform (DST): DST is a variant of Fourier transform which produces only real coefficients. The advantages of DST are high energy compaction and a sparse representation which can help in distinguishing the noise embedded in the clean image. DST of an input 'A' can be written as:

$$S(k) = \sum_{n=0}^{N-1} A(n) sin \frac{\pi k n}{N+1}. \tag{5.10}$$

Discrete Wavelet Transform (DWT): DWT provides the multi-resolution decomposition of the input image by passing it from a set of low and high pass filters. The decomposition offers low frequency and high-frequency components present in the input image in terms of four different subbands (as shown in Figure 5-3). In this research, we have used the 'Haar' filter for single-level wavelet decomposition.

Walsh Hadamard Transform (WHT): WHT transform is the generalized version of FFT and contains only +1 and -1 values in the kernel matrix. WHT has a wide spectrum of applications in power spectrum analysis, filtering, and speech processing.

To illustrate the importance of using transformations for differentiating between original and perturbed, the first row of Figure 5-3 shows that for a clean image, the high-frequency component of the FFT transform is thick at the center, followed by horizontal and vertical directions. Similarly, the DWT high-frequency components, especially the vertical and diagonal sub-bands, are noisy as compared to the clean image. Another illustration can be observed with DST coefficients that have higher absolute values in the down right corner for a clean image.

Some other recently proposed approaches have also claimed that input transformation shows strong defense against various adversaries [131, 352]. However, the input transformations explored for security are mainly based on random sub-sampling, scaling, cropping, and compression that might be compromised using adversaries applied at multiple scales and orientations. The proposed defense applies image transformations which also filter the image at various scales and orientations, and makes the proposed algorithm applicable to a broad spectrum of attacks.

5.3.2 GIST Feature Extraction and Classification

With the hypothesis that the spatial distribution of both original and perturbed images are different in the transformed domain, we propose the use of GIST features [241] on transformed images to encode the characteristics of both perturbed and original images. GIST features encode the edge information by computing the gradients in multiple scales and orientations. It estimates the spatial envelope present in the natural and human-made (un-natural) images. The spatial envelope of the image pixels is defined using the degree of naturalness, openness, and expansions.

We observe that the perturbation, which is even minimal in magnitude, changes the pixel compactness and pixel encoding structure in the local neighborhood region of the images. As shown in the last three columns of Figure 5-3, DWT horizontal, vertical, and diagonal subbands of the adversarial images accentuate noise information, which is missing in the non-perturbed images. The degree of openness in adversarial images is generally higher because of an increase in the high-frequency information due to noise. The original image is composed of the excellent distribution of the edge information across the image, whereas, the perturbed images contain extra edge information developed because of the random insertion of high-frequency content. It is important to note that the adversarial noise can be embedded at any location and independent of input transformations such as rotation and scale. Therefore, the adversarial detection algorithm should be robust to the filtering of the input image at multiple scales and orientations using filters.

GIST features provide these characteristics, and they are computed by convolving the image with Gabor filters at multiple scales and orientations [362]. The similarity of Gabor filters to the human visual system makes a strong case for texture discrimination. The Gabor decomposition, which highlights the high-frequency information embedded due to the noise in different directions, is highlighted and further encoded using GIST features. Once the GIST features are extracted, a Support Vector Machine (SVM) classifier [76] is trained for binary class (original and adversarial) classification.

5.3.3 Fusion

Experimentally[1], we observed that DST yields the best results on low-resolution images while DWT performs the best on high-resolution images. Therefore, to develop a generalized detector across low and high-resolution images, we have combined the classification scores of the two best performing input transformations. The final detection results are obtained by computing the weighted fusion ($w_1 = 0.5$ and $w_2 = 0.5$) of scores obtained by matching GIST obtained from DST and DWT transformations. The $scores \geq 0$ are classified as *real class* while others are classified as *adversarial class*. The weights are learned using a grid search over the training set; however, the best results are found with equal weight fusion.

[1] The experimental results of the proposed adversarial detection algorithm are described in Section 5.6.

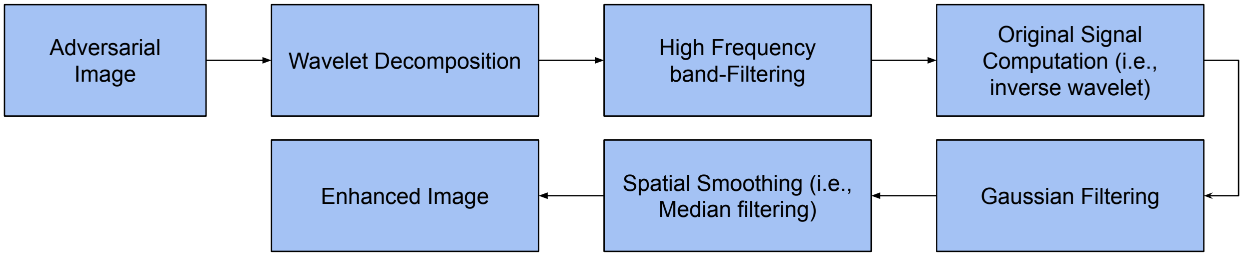

Adversarial Image
Wavelet Decomposition
High Frequency band-Filtering
Original Signal Computation (i.e., inverse wavelet)
Gaussian Filtering
Spatial Smoothing (i.e., Median filtering)
Enhanced Image

Figure 5-4: Steps in the proposed adversarial mitigation pipeline.

5.4 Proposed Adversarial Mitigation Algorithm

The aim of defense against the adversarial attack can be two folds: (i) the system can discard an adversarial image and ask for a new image or (ii) the image can be “enhanced" to remove the adversarial noise from the input image. In real-world applications, discarding too many images may affect the usability of the system. Hence, it is desirable to remove the adversarial noise embedded in the images. Based on the observations from different image transformations (shown in Figure 5-3), we propose to utilize image preprocessing based algorithm to mitigate the effect of adversarial noise. The steps involved in the proposed adversarial mitigation algorithm (shown in Figure 5-4) are listed below:

- The image is decomposed into an approximation and three high-frequency subbands using Wavelet decomposition.

- Adaptive thresholds are learned over each high-frequency sub-band and sub-band pixel values are thresholded. The images are divided into multiple local regions and based on the first-order statistics of the region, a threshold is computed to enhance the image. The size of the local region is computed using the following equation: 2×floor(size(I)/16)+1 [48], where I is the input image. The pixel values greater than the threshold are set to one, and other values are set to zero.

- Inverse wavelet transform is computed over the filtered wavelet sub-bands.

- Gaussian filtering with a sigma value of 0.5 is applied to further remove the noise. Inverse wavelet image is smoothed using a median filter of size 3×3 to obtain the enhanced image.

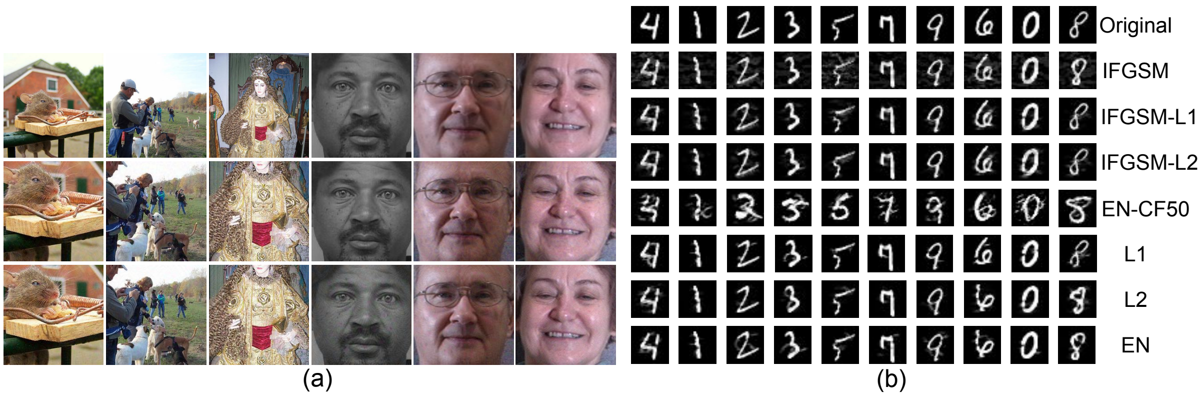


Figure 5-5: (a) Original, universal and fast feature fool adversarial images of ImageNet, MEDS, and Multi-PIE database. First row are original samples, second row are universal perturbation images, and last row are the fast feature fool adversarial images. (b) Original and Adversarial samples generated using various perturbation algorithms on the MNIST database.

In the classification pipeline, the images detected as perturbed are processed through the mitigation pipeline; otherwise, they are directly given as input to the recognition module.

5.5 Attacks, Databases and Protocol

This section summarizes the adversarial attack generation algorithms, the databases, and the experimental protocols used for evaluation.

5.5.1 Attack Generation Algorithms

To evaluate the robustness of the proposed detection algorithm, we have selected a wide variety of adversarial perturbations.

- IFGSM and variants [168]
- Regularized loss based attack [57, 64] including C&W's L_2, which is one of the most complex attacks to be defended.
- Universal [231] attack
- Fast Feature Fool [233]

Table 5.4: Summarizing the list of databases and attacks along with the number of original and perturbed images used for performance evaluation.

Database	**Attacks**	**Original**	**Perturbed**	
			Per Attack	**Total**
MNIST	L_1, L_2, EN, EN-CF50, PGD, IFGSM, IFGSM-L$_1$, IFGSM-L$_2$, and DeepFool	9,000	9,000	81,000
CIFAR-10	PGD ($\epsilon = 0.03, 0.05$), IFGSM ($\epsilon = 0.03, 0.05$), FGSM ($\epsilon = 0.03, 0.05$), and Universal	10,000	10,000	70,000
ImageNet*	Universal and Fast Feature Fool : VGG-16	10,000	5,000	10,000
MEDS	Universal: VGG-16, CaffeNet and GoogLeNet	836	836	2,508
	Fast Feature Fool: VGG-16, CaffeNet and GoogLeNet	836	836	2,508
Multi-PIE	Universal: VGG-16, CaffeNet and GoogLeNet	1,680	1,680	5,040
	Fast Feature Fool: VGG-16, CaffeNet and GoogLeNet	1,680	1,680	5,040
MBGC (Query)	Universal: VGG-16, CaffeNet and GoogLeNet	24,042	24,042	72,126
	Fast Feature Fool: VGG-16, CaffeNet and GoogLeNet	24,042	24,042	72,126

*Also evaluated on complete validation set of size $50,000$ original and $1,00,000$ adversarial images.

5.5.2 Databases and Experimental Protocol

To show the effectiveness with different kinds of images, the results are reported with following six databases (DBs) pertaining to digits, objects, and faces. MNIST [7] is a handwritten digit database. CIFAR-10 [164] and ImageNet [87] are widely used object databases. MEDS [99] is a multi-encounter face database released by NIST for face recognition research. MultiPIE [127] is a popular large face database and MBGC [251] is a large scale database used for face recognition challenges. The samples of faces, images from ImageNet, and MNIST digit images are shown in Figure 5-5.

Table 5.4 summarizes the attacks and models used for evaluation along with the number of original and perturbed images, and the adversarial generation algorithms. Across different attacks, the results are shown on more than $2,31,015$ adversarial images. To the best of our knowledge, this is one of the most significant numbers of images used to showcase the detection and mitigation performance.

Two sets of experiments are performed: intra-database and inter-database/inter-model. Intra-database (familiar) conditions are defined where the training and testing images belong to either the same database or are generated using the same adversarial generation algorithm. The inter-database experiment simulates the environment where training and testing images correspond to different databases or adversarial attack generation algorithms or DNN models. In intra-database conditions, images pertaining to both the classes (corresponding to each attack and database) are divided as $50\% - 50\%$, one part is used for training the detector and the second half is used for

Table 5.5: Range and resolution of distortion parameters to generate IFGSM adversarial samples.

Method	**Grid Search**	
	Range	Resolution
IFGSM-L_∞	[10^{-3}, 1]	10^{-3}
IFGSM-L_1	[1, 10^3]	1
IFGSM-L_2	[10^{-2}, 10]	10^{-2}

Table 5.6: Experimental setup of C&W, EN, and L_1 attacks.

Parameter	**Value**
Initial Learning Rate	0.01
Iterations	1,000
Initial Regularization	0.001
Steps	9
Optimizers	ADAM and projected FISTA

evaluation. In the inter-database/inter-model experiments, one kind of images (regarding database or adversarial attack) are used for learning the detector while the remaining type of images (pertaining to the database or attack) are used for testing.

5.5.3 Implementation Details

Adversarial Examples Generation: We have used the original implementation of the attacks with default settings. For both C&W L_2 and Elastic-Net attacks, $1,000$ iterations are run over nine binary search steps to search the regularization parameter (initially 0.001). The initial learning rate is set to 0.01. C&W and EN attacks find the best adversarial examples by solving the ADAM and projected FISTA with the square-root decaying learning rate optimization functions, respectively. For C&W L_2, EN, and L_1 attacks, the adversarial examples containing the least distortion among all the successful examples are selected. Similarly, for IFGSM attack[2], fine-grained search is used to search for the best distortion parameter. To implement IFGSM, 10 iterations of FGSM with ϵ-ball clipping is used. The distortion parameter is divided by 10 in each FGSM iteration as found most effective in [330]. In total, $10,000$ (i.e., $1,000$ FGSM operations $\times 10$ iterations) gradients values are computed to generate the iterative-FGSM adversarial samples from a single input image. The attack strength parameter is selected via grid search; the smallest strength, which leads to a

[2] CleverHans package is used https://github.com/tensorflow/cleverhans

Table 5.7: Adversarial detection accuracy (%) of the proposed and existing algorithms on the MNIST database.

Attack	**Adaptive Noise [188]**	**Bayesian Uncertainty [95]**	**ODIN [189]**	**CNN Filter Response [123]**	**ESRM [191]**	**Proposed (GIST Features + SVM)**					
						FFT	**DCT**	**WHT**	**DWT**	**DST**	**DST+DWT**
L_1	78.6	77.3	78.7	82.6	76.7	54.3	76.5	70.1	79.6	88.3	**89.6**
L_2	79.2	78.5	72.0	80.8	67.5	54.3	60.6	74.1	96.6	98.8	**99.2**
EN	79.6	78.1	75.6	78.5	65.6	52.5	67.4	68.0	85.7	92.5	**93.7**
EN-CF50	80.6	79.5	75.2	79.0	64.2	65.6	90.8	91.3	97.8	99.3	**99.5**
PGD	74.6	77.8	73.3	74.9	71.5	71.4	80.2	86.1	97.8	99.0	**99.2**
IFGSM	85.9	85.3	90.2	92.1	78.9	68.7	96.7	98.9	99.9	99.9	**99.9**
IFGSM-L_1	85.7	84.9	86.2	89.7	81.5	53.9	76.1	81.9	98.1	99.7	**99.9**
IFGSM-L_2	85.1	84.4	84.1	86.2	76.4	52.5	75.1	81.5	98.3	99.7	**99.9**
DeepFool	72.1	76.3	81.2	84.5	68.4	63.7	74.1	79.0	88.8	95.3	**96.4**

Table 5.8: Adversarial detection accuracy (%) of the proposed and existing algorithms on the CIFAR-10 database.

Attack	**ESRM [191]**	**NIC [207]**	**ODIN [189]**	**AN* [188]**	**DCT**	**DST**	**DWT**	**DST + DWT**
PGD ($\epsilon = 0.03$)	66.4	75.6	63.1	59.2	90.9	78.1	82.7	**84.2**
PGD ($\epsilon = 0.05$)	71.6	87.9	65.7	62.3	97.3	90.2	92.8	**93.6**
FGSM ($\epsilon = 0.03$)	61.4	79.3	81.8	83.2	90.8	75.2	83.3	**85.1**
FGSM ($\epsilon = 0.05$)	66.0	85.1	86.1	85.8	97.4	90.1	92.8	**93.4**
IFGSM ($\epsilon = 0.03$)	62.9	84.8	80.3	87.1	94.3	81.3	89.4	**92.5**
IFGSM ($\epsilon = 0.05$)	73.2	90.5	85.4	89.0	98.5	92.0	96.5	**97.6**
Universal ($\epsilon = 0.3$)	67.1	76.4	74.9	67.7	97.0	95.7	98.9	**99.2**

*AN= Adaptive Noise

successful attack is selected along with corresponding adversarial example. Tables 6.2 and 6.3 list the experimental parameters used to generate the adversarial images. Universal and F3 perturbed images on ImageNet and face databases are crafted using three different deep networks: VGG-16, GoogLeNet, and CaffeNet. VGG-16 and GooeLeNet are 16 and 22 deep layers networks and yield state-of-the-art object recognition performance on ImageNet the database in 2014 and 2015, respectively. Further details of attack parameters are available in the original papers.

Adversarial Examples Detection: We evaluated the performance of proposed adversarial example detection with multiple SVM kernels, including linear, polynomial, and radial basis function (RBF). The cost parameter of the kernels is optimized over the training set using a grid search with range [2^{-1}, 2^4]. For RBF kernel, the gamma value is optimized in the range [2^{-5}, 2^2], while polynomial kernels of degree 2 to 4 are evaluated. We observed that the linear kernel yields the best results. Therefore, for testing, the SVM classifier with a linear kernel is utilized over the transformed GIST features.

5.6 Adversarial Detection Results

This section presents the results of the proposed detection algorithm. The results are segregated according to the databases and the attacks that have been performed on the images. The performance of the detection algorithm is compared with existing detection algorithms, Adaptive Noise [188], Bayesian Uncertainty [95], Out of distribution (ODIN) [189], CNN Response [124], Selective Dropout [123], and Randomization [352]. Along with existing algorithms, the proposed algorithm is also compared with two CNN models (VGG-16 [306] and DenseNet [144]) used for adversarial examples detection. The accuracies are reported in terms of average class-wise detection accuracy, where the classes are original and perturbed.

5.6.1 Intra-Database Testing

The results for intra-database testing are segregated in three parts according to the three kinds of databases.

Results on MNIST Database: The results are evaluated for nine different attacks. Table 5.7 summarizes the performance of the proposed algorithm with the existing detection algorithms. The comparison shows that the existing algorithms yield at most 82.6% and 78.5% detection accuracy on the most complex L_1 and EN attack algorithm, while the proposed algorithm shows 89.6% and 93.7% detection accuracy, respectively. IFGSM attack and its variants based on L_1 and L_2 are almost perfectly detected. A similar performance trend is observed across all other attacks and the proposed algorithm yields the best results. The challenges in performance trends can also be observed in Figure 5-5 where the images corresponding to the EN attack seem to have the least amount of perceptible noise, whereas the perturbations in samples corresponding to the other attacks are comparatively more visible. On another sophisticated attack in literature [364], i.e., DeepFool, the proposed algorithm yields 96.4% detection accuracy, which is at-least 15.2% better than existing algorithms.

Next, to understand the effectiveness of individual transforms, the performance is computed by applying GIST with individual transformations, and the results are summarized in Table 5.7. The results show that FFT, DCT, and WHT do not yield good results. DST yields the best results followed by DWT.

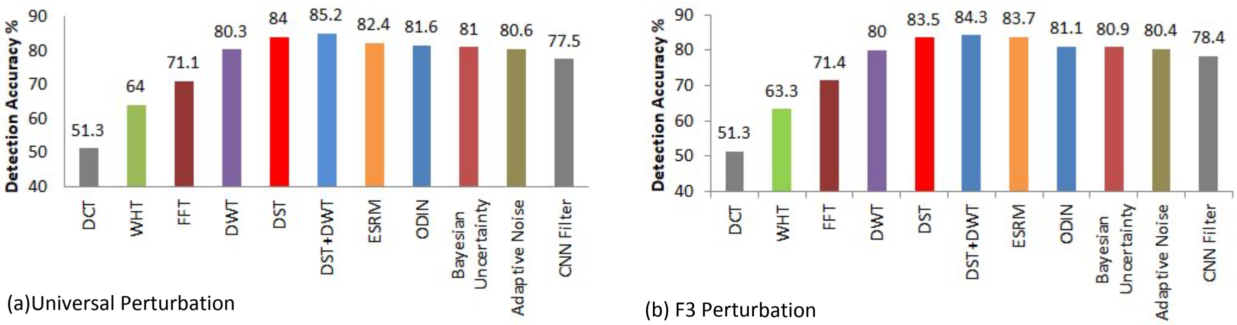


Figure 5-6: Adversarial detection performance of the proposed and existing algorithms on the **'ImageNet'** database.

Results on CIFAR-10 Database: The detailed results are summarized in Table 5.8. The proposed algorithm yields 84.2% detection accuracy on PGD attack with $\epsilon = 0.03$ and the accuracy improves with the increase of noise strength parameters. Since PGD is one of the strongest first-order iterative attacks, high detection accuracy on PGD shows the efficacy of the proposed algorithm in handling complex attacks. Similarly on other iterative attacks the detection performance is at-least 85.1% which is 23% better than ESRM [191] and 5.8% better than NIC [207]. Other than that, on DeepFool, the proposed algorithm is 6% better than ESRM [191]. It is interesting to note that on the CIFAR-10 database, DCT yields the best results followed by the fusion of DST and DWT, except for universal perturbation detection.

Results on ImageNet Database: Based on the predefined protocol, we have used $5,000$ original images to generate $10,000$ adversarial images using *universal* and *F3* perturbation algorithms. Figure 5-6 summarizes the results of the proposed adversarial detection algorithm on the ImageNet database using the proposed transformed features and the existing algorithms. On the universal perturbation, the proposed algorithm with DST transform yields 84% detection, and it is further improved to 85.2% when DST and DWT are combined. On the other hand, existing algorithms yield the detection accuracy in the range of 77.5% to 82.4%. As shown in Figure 5-5, the classes in the ImageNet database have a lot of variability, and the images are captured in varying backgrounds with different sensors. Therefore, compared to other database, it is relatively challenging to determine whether an image is attacked or not, and therefore, the best results are around 84.0% for the two attacks. We have also performed the experiments using weighted fusion of all transformations, i.e. $w_1 \times DST + w_2 \times DWT + w_3 \times DCT + w_4 \times FFT + w_5 \times WHT$. We have

Table 5.9: Adversarial detection performance (%) of the proposed algorithm on face databases.

Database	**DNN Model**	**Attack**	**Adaptive Noise [188]**	**Bayesian Uncertainty [95]**	**ODIN [189]**	**CNN Filter Response [123]**	**ESRM [191]**	**Proposed (GIST Features + SVM)**					
								DCT	**WHT**	**DST**	**FFT**	**DWT**	**DST+DWT**
MEDS	VGG-16	Universal	80.2	80.3	82.0	80.4	82.5	57.4	78.5	94.3	96.3	98.3	**98.7**
		F3	79.6	79.9	81.3	80.6	81.9	61.6	88.0	96.5	96.9	98.3	**98.9**
	GoogLeNet	Universal	79.2	79.9	80.8	81.2	79.7	60.5	85.3	97.0	97.1	99.4	**99.6**
		F3	77.0	77.3	81.1	78.1	80.4	60.3	83.4	93.1	97.8	97.2	**97.5**
	CaffeNet	Universal	78.9	78.4	77.5	80.8	84.3	59.0	82.3	95.1	92.9	98.2	**98.4**
		F3	78.8	78.5	78.3	80.5	83.1	67.5	88.6	99.2	97.6	99.8	**99.9**
Multi-PIE	VGG-16	Universal	75.5	74.7	84.4	86.6	87.8	57.7	93.0	99.6	100.0	100.0	**100.0**
		F3	76.0	75.0	84.9	85.9	87.2	61.8	98.9	99.9	100.0	99.9	**100.0**
	GoogLeNet	Universal	69.4	69.8	83.9	87.4	85.0	61.8	98.9	100.0	100.0	99.9	**100.0**
		F3	70.2	70.5	84.2	87.0	83.6	59.8	99.0	99.9	100.0	99.9	**100.0**
	CaffeNet	Universal	71.1	70.3	79.6	85.2	89.1	58.2	97.4	99.9	100.0	99.9	**100.0**
		F3	70.2	69.6	80.5	84.8	88.7	67.1	99.0	100.0	100.0	100.0	**100.0**

observed that when equal weights ($w_{1 \ldots 5}$) are provided to each transformation, the final detection performance is about 5% lower than the fusion of two best transformations, i.e., DST and DWT. With weight optimized fusion (based on the training error, weights are assigned), the performance is slightly lower compared to the fusion of DST and DWT.

We also evaluated the performance when state-of-the-art deep networks are used for adversarial example detection. Interestingly, the fine-tuned VGG-16 model yields only 75.8% accuracy for detecting 'Universal' attack on the ImageNet database. Similar lower performance is observed with DenseNet model for adversarial sample detection. We have also evaluated the performance of the proposed adversarial detection algorithm on the **'complete validation set'** $(50,000)$ of the ImageNet database. We have observed slightly improved performance on the complete set as compared to a random subset, thus validating the scalability of the proposed algorithm. For example, the performances of DWT and DST transformation for F3 perturbation detection are 80.7 and 84.1, respectively. For Universal attack detection, it is 82.9 and 87.4 as compared to 80.3 and 84.0 on DWT and DST transformations, respectively.

Results on Face databases: As summarized in Table 5.4, images from multiple face databases are used, and two different perturbation algorithms (universal and F3) are selected to generate the adversarial images. Since both the attacks are learning-based, the perturbations for both universal and F3 are generated using three different DNN models: VGG-16, GoogLeNet, and CaffeNet. Adversarial detection results obtained from different input transformations and existing algorithms are reported in Table 5.9.

With the face images generated using both *universal* and *F3* perturbation algorithms, the proposed detection algorithm with DWT+DST fusion yields the highest detection accuracies. With *universal* attack, the accuracies on MEDS are 98.7%, 99.6%, and 98.4%, for VGG-16, GoogLeNet [318], and CaffeNet [165] models, respectively. For adversarial images generated using F3 [233] attack on the MEDS database, the proposed algorithm is at-least 97.5% successful in adversarial detection across DNN models. On the CMU Multi-PIE database, perfect detection accuracy is achieved for each DNN model and attack. Although the detection accuracies on both MEDS and Multi-PIE databases are more than 97.5%, it is observed that the performance on the MEDS database is slightly lower than the Multi-PIE database. It is our assertion that the lower performance on MEDS could be attributed to the high-resolution images of MEDS. Since the same amount of noise is added to both the databases, the percentage of perturbation is comparatively lower for MEDS, thus making it challenging to detect. In terms of comparison, with universal perturbation, the detection performance of adaptive noise reduction [188], Bayesian uncertainty [95], ODIN [189], CNN response [123], and ESRM [191] is in the range of 69.4% to 89.1% which is at-least 10.9% less than the proposed algorithm on MEDS and Multi-PIE databases.

5.6.2 Inter-Database, Attack, and Network Testing

Adversarial attack generation research has shown that the attack generation algorithms are generalizable and transferable across networks (CNN models) and database images. The importance of unseen testing can also be observed from the real-world where a newly developed attack can be performed on an already built security mechanism. To address such cases, the defense mechanism or detection algorithms must be generalizable across attacks, images, and networks. Therefore, in this section, the transferability or generalization of the proposed detection algorithm is evaluated in four scenarios: (i) Unseen Database, (ii) Unseen Attack Algorithm, (iii) Unseen DNN Model, and (iv) Unseen Database and DNN Model. Tables 5.10 to 5.12 and Figure 5-8 showcase the results of these four experiments. It is to be noted that these results are reported only with the best performing algorithms.

Unseen Database

It refers to the scenario where one database is used for training, and another database is used for testing. Since we have used multiple independent face databases for evaluation, the results are

Table 5.10: Detection performance (%) of the proposed algorithm (DST+DWT) for universal and deepfool adversarial attack on face databases in cross-database scenarios.

Train DB	Test DB	Universal	DeepFool
MEDS	Multi-PIE	99.2	98.2
	MBGC	84.5	80.1
Multi-PIE	MEDS	93.5	94.1
	MBGC	86.2	87.5

reported on the three face databases with universal and DeepFool attacks. As shown in Table 5.10, the proposed detector yields more than 80.1% detection accuracy across different databases and attacks, and the best performance is 99.2%. The high efficiency shows the transferability and generalizability of the algorithm for adversarial detection.

Comparing the performance with intra-database settings reveals that there is a significant difference in the performance for inter-database training testing, as the lowest detection rate for intra-database is 97.5%. Interestingly, it is observed that the best performing detection algorithm on intra-database yields the best results for inter-database experiments as well. We assert that the performance drops more on Multi-PIE - MEDS (train-test) combination because the image resolution and the variability present in Multi-PIE are lower than MEDS. The detection model trained for Multi-PIE may not have learned to detect perturbed images with lower perturbation percentage, while the reverse combination helps the detector learn on more challenging cases from the MEDS database, thus providing higher performance on Multi-PIE test database.

Comparison with existing algorithms [95] and [188] shows that the detection performance on the MEDS - Multi-PIE combination is in between $77.4 - 79.3\%$ whereas, the results of reverse combination are in the range of $70.1 - 70.9\%$. For both the combinations, Adaptive Noise yields higher results than Bayesian Uncertainty, and both of them are at least 19.9% lower than the proposed detection algorithm. Similarly, CNN Response [124], ODIN [189], and ESRM [191] yield at-least 6%, 14%, and 21.3% lower detection performance on the MEDS database. On the MBGC database, existing algorithms such as Bayesian Uncertainty, Adaptive Noise Reduction, and CNN filter response show 59.1%, 59.6%, and 76.1% detection accuracy, respectively. Other existing detection algorithms, such as based on intermediate CNN layer features, CNN based classification network, and density estimation based methods, are ineffective against complex optimization-based adversarial attacks such as L_2 [55, 57]. The detection performance of the proposed algorithm

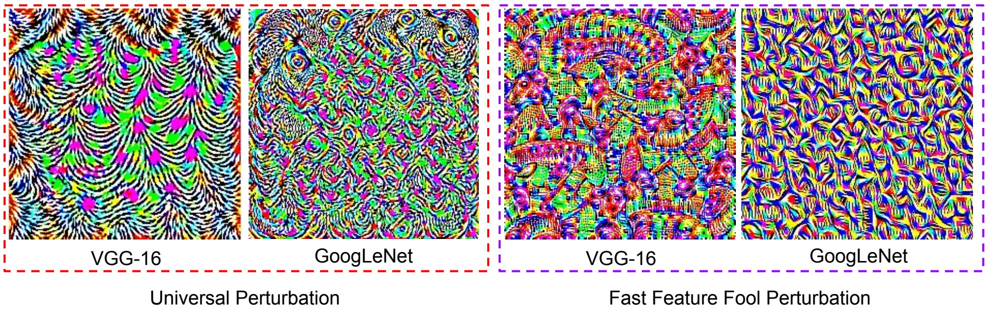


Figure 5-7: Samples of noise added across different perturbation algorithms with VGG-16 and GoogLeNet architectures.

Table 5.11: Detection performance (%) of the proposed algorithm on face databases in cross-model scenarios with universal attack.

Train Model	Test Model	Database	DST	DWT	DST+DWT
VGG-16	GoogLeNet	MEDS	87.5	92.4	**93.1**
		Multi-PIE	90.5	95.6	**95.8**
GoogLeNet	VGG-16	MEDS	88.1	94.5	**95.8**
		Multi-PIE	93.4	97.4	**98.5**

is at least 16.4% better than VGG-16 based classification model when trained using the Multi-PIE database. These results support the limitations of existing defense algorithms reported in Table 5.2.

Unseen DNN Model

The adversarial images used for training the classifier are generated using one model (i.e., VGG-16) while the detector is tested with the adversarial images created using another model (i.e., GoogLeNet). Figure 5-7 shows the noise patterns generated by the two adversarial algorithms with VGG-16 and GoogLeNet models. It can be observed that the patterns with the two models have visible differences in them. The results of this experiment are summarized in Table 5.11. The results showcase a reduction of around $1.5 - 6.5\%$ in the detection performance in comparison to the same DNN model generated adversarial examples detection (Table 5.9). Comparing the performance of the proposed results with the existing algorithms shows that there is a difference of at least 30% between them across the four cases.

Table 5.12: Detection performance (%) of the proposed algorithm on the MNIST database in cross-attack scenarios.

Train Attack	Test Attack	DWT	DST	DST+DWT
IFGSM-L_∞	IFGSM-L_1	89.9	95.5	**96.0**
	IFGSM-L_2	89.0	95.0	**95.4**
IFGSM-L_1	IFGSM-L_∞	93.2	99.7	**99.8**
	IFGSM-L_2	98.1	99.7	**99.8**
IFGSM-L_2	IFGSM-L_∞	92.6	99.8	**99.9**
	IFGSM-L_1	98.3	99.7	**99.9**
L_1	IFGSM-L_∞	71.3	92.4	**92.6**
	L_2	84.4	92.4	**93.2**
	EN	91.4	92.3	**94.3**
EN	IFGSM-L_∞	67.3	93.7	**93.7**
	L_1	76.9	85.7	**86.0**
	L_2	87.2	93.6	**93.8**
L_2	IFGSM-L_∞	93.2	99.0	**99.4**
	EN	84.9	88.3	**89.4**

Unseen Attack Algorithm

In this scenario, the adversarial images used for training the classifier are generated using one attack algorithm (for instance, L_1 and L_2), while the detection algorithm is tested with samples generated from another attack algorithm L_∞. The results of unseen attack algorithm scenarios are summarized in Table 5.12. The proposed DST+DWT detection algorithm trained with adversarial images generated using L_∞ norm IFGSM attack and tested with images generated from L_1 and L_2 IFGSM attacks yields at least 95.4% detection rate. On the other hand, the detector trained using L_1 or L_2 variants of IFGSM and tested on the remaining two yields near-perfect detection performance (99.8% at-least). Similarly, when the proposed algorithm is trained using logit-layer loss based attacks (such as L_1, L_2, and EN) and tested on the remaining, the lowest detection accuracy is 86.0%.

To compare the proposed algorithm with the existing algorithms, we computed the average detection accuracy across all the combinations of attacks. The results show that the proposed algorithm yields an average detection accuracy of 94.8% while the accuracy of Adaptive Noise is 64.8% [188], Bayesian Uncertainty [95] is 64.3%, ODIN [189] is 76.2%, and ESRM [191] is 67.6%.

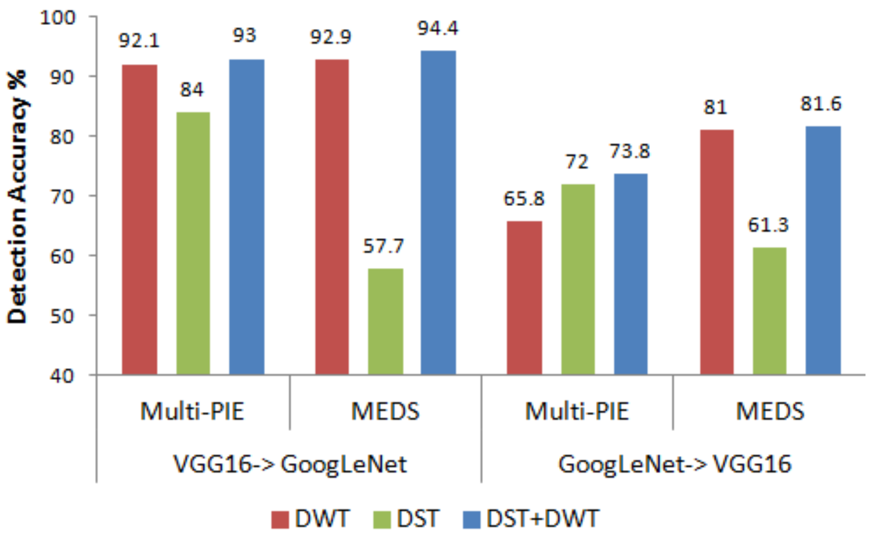
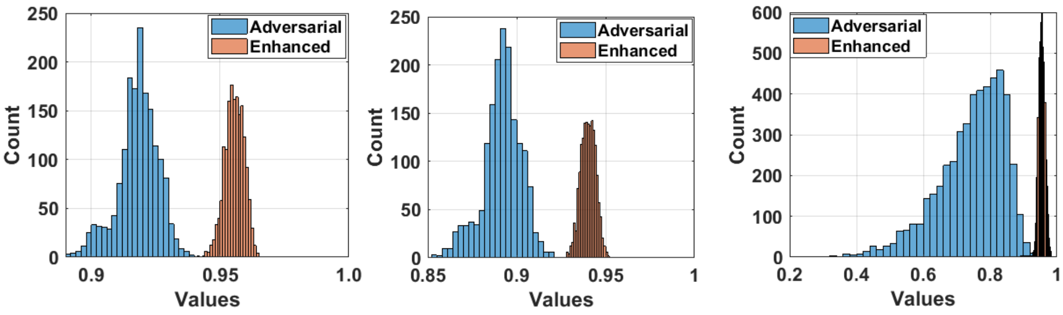


Figure 5-8: Adversarial detection performance (%) on face databases in 'doubly' unseen scenarios i.e., cross-database and cross DNN model, using the proposed algorithm for universal attack. (VGG16 -> GoogLeNet) means VGG16 generated adversarial images are used for training and GoogLeNet adversarial images for testing. The results are reported on the test face database while the other database is used for training.

Unseen Database and DNN Model

To further strengthen the defense of the proposed algorithm, 'doubly' unseen scenarios are tested: where the unseen database and DNN model adversarial images are used for testing.The adversarial images used for training the classifier are generated using one model (i.e., VGG-16) of one database while the detector is tested with the adversarial images created using another model (i.e., GoogLeNet) of another database, and vice-versa.

The results of 'doubly' unseen database and DNN model scenarios are summarized in Figure 5-8. On using the Multi-PIE database and adversarial images generated via GoogLeNet for testing and images of the MEDS database with adversarial images of the VGG16 model for training, the proposed algorithm yields 93% detection accuracy. When the detector is trained on adversarial examples generated from GoogLeNet on Multi-PIE and tested on adversarial examples generated from VGG-16 on MBGC, the accuracy of DST+DWT is 80.6%.



(a) MEDS (GoogLeNet) (b) Multi-PIE (VGG-16) (c) ImageNet (VGG-16)

Figure 5-9: Illustrating the effect of adversarial mitigation using SSIM on GoogLeNet and VGG-16 universal adversarial images.

General Observations and Resiliency

Across the experiments, we have observed that existing algorithms [95, 123, 124, 188, 191] and CNN models fail in unseen training and testing conditions. However, the proposed algorithm yields better results, thus showcasing better generalization capabilities. The results show that these images transformations and generalized image descriptors are effective for adversarial detection and do not need complex deep neural network architectures.

Additional experiments to understand the effectiveness of the classifier suggested that in comparison to the SVM classifier, neural network yields lower performance across adversarial generation algorithms and databases. Carlini and Wagner [55] also made a similar observation where they claimed that deep CNN and neural network classifier based defense fails for sophisticated attacks.

An additional experiment is performed to evaluate the **resiliency** of the proposed detection algorithm towards attacks using the VGG-16 model and PGD attack with different attack strength (distortion) on the CIFAR-10 database. Since the proposed algorithm is not based on a neural network, generating the samples to attack the detection algorithm is not trivial. To understand this phenomenon, we performed PGD attack with varying attack strength values. Noise is visible in the higher attack strength while the embedded adversarial noise is “invisible” for lower strength values. We have computed both attack success rate as well as attack detection rate and results are documented in Table 5.13. Interestingly, we have observed that the adversarial examples which can bypass the detection algorithm have minimal noise added. Therefore, they do not affect the classification performance of the target VGG-16 model. If the perturbation is significant, there is

Table 5.13: Detection accuracy of PGD attack with different distortion values on CIFAR-10 generated using the VGG-16 model.

Attack strength (ϵ)	Attack Success Rate %	Attack Detection Accuracy %
0.03	94.87	84.2
0.02	93.64	83.7
0.01	69.80	81.5
0.009	65.33	80.9
0.007	54.63	79.3
0.005	40.79	77.9

a very high probability that the proposed detection algorithm detects the attack.

To further evaluate the effectiveness of the proposed algorithm, we have also performed experiments with secondary adversarial attack. Carlini and Wagner [55] claim that the detection network can be fooled using the attack they termed as "secondary adversarial attacks". Inspired by this, Liu et al. [191] performed the attack by removing the untargeted 10% perturbations to evaluate their algorithm's resiliency (their detection algorithm also does not involve neural networks). To make the work in-line with such previous practices, we have also performed experiments using the C&W l_2 [57] attack on the ImageNet database [87] using VGG-16 network. It involves removing the untargeted 10% perturbations to fool the detection algorithm. We have observed that the performance of DST based detection algorithm reduces by 5-6% while the performance of DWT degrades by 3-4%. The proposed fusion algorithm is more robust, and the performance reduces by only 2%. Moreover, the success rate of the attack drops by more than 54%. Similar experiments are performed on the CIFAR-10 database using the VGG-16 model. The removal of untargeted adversarial shows a drop of 45% in terms of attack success rate. On the other hand, the adversarial examples detection accuracy of the proposed algorithm suffers a loss of 1%.

5.7 Results of Proposed Adversarial Mitigation

The potential bottleneck of existing adversarial defense algorithms is that either they are ineffective against complex attacks or they require re-training the entire deep learning network for adversarial training [169, 330]. However, as mentioned earlier, the proposed mitigation algorithm improves the quality of image so that it can be directly used for classification. To present the effectiveness,

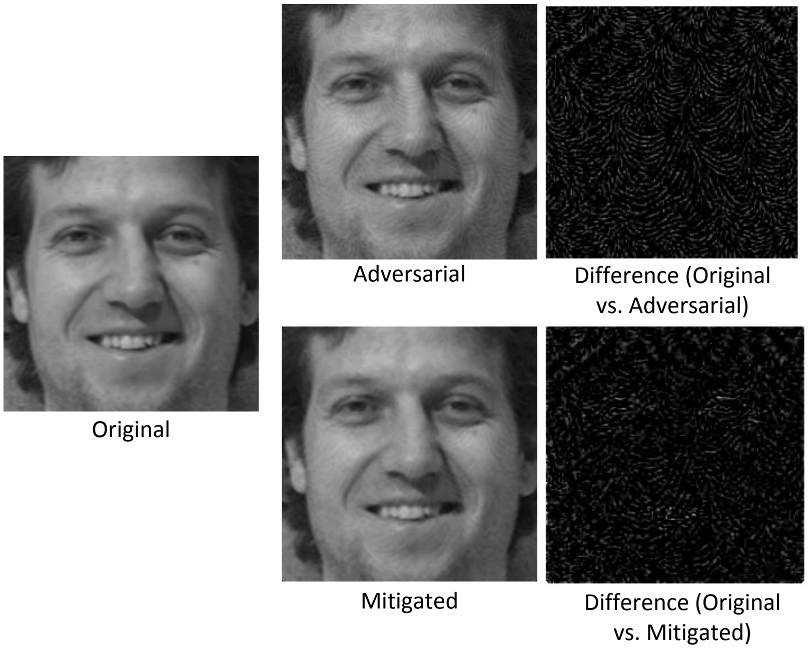

Original
Adversarial
Difference (Original vs. Adversarial)
Mitigated
Difference (Original vs. Mitigated)

Figure 5-10: Shows the original, adversarial, mitigated, and noise (difference) image, which helps in illustrating the effect of the proposed adversarial mitigation algorithm. Difference images showcase the impact of mitigation (Best viewed with 4X zoom).

the following two metrics are used:

- *Structural Similarity (SSIM)*: It is a measure of computing the quality difference between the original/adversarial images and original/enhanced images. $SSIM \in [0, 1]$ and should be close to 1 for similar images. To establish good mitigation performance, the value of SSIM between original and enhanced images should be closer to one.

- *Classification accuracy*: Compared to the perturbed image, the recognition/classification accuracy should be increased, and it should be closer to the performance of original images.

Figure 5-9 illustrates the SSIM scores for different adversarial DNN models. It can be observed that the SSIM scores between original and enhanced images for VGG-16 and GoogLeNet models shift towards one for all three databases. implying increased similarly between the image obtained after mitigation and the original image, compared to the perturbed image. This observation can be visually validated from Figure 5-10 which shows the perturbed and the mitigated images along with the difference from the original image.

To further showcase the impact of the proposed mitigation algorithm, we have performed face

Table 5.14: Face verification results to evaluate the mitigation performance on the MEDS database. The results are reported in terms of true positive rate (TPR) (%) at 1% false positive rate (FPR).

Images		**Light CNN**	**VGG**
Original		89.3	78.4
Adversarially Perturbed		41.6	30.5
Mitigated	Randomization (2018) [352]	46.8	35.6
	Selective Dropout (2019) [123, 124]	61.3	40.6
	Kernel Smoothing (2020) [338]	48.0	38.9
	Proposed	**71.4**	**60.8**

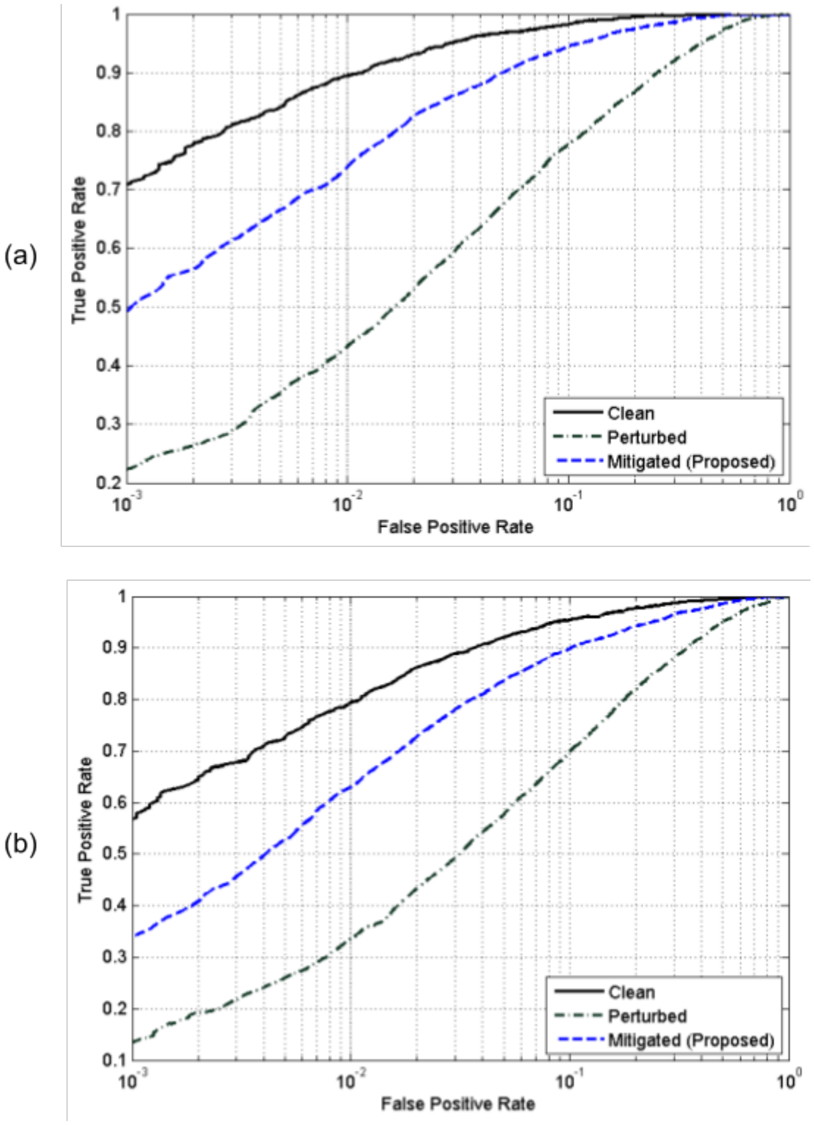


Figure 5-11: Face verification using (a) Light CNN and (b) VGG-Face.

verification using two DNN models: Light CNN [349] and VGG-Face [245], on the protocols defined by Goswami et al. [123, 124]. VGG-Face is trained on 2.6 million face images pertaining to $2,622$ subjects, whereas LightCNN is trained on images of $99,891$ individuals. The database contains original images while the test set contains either the original, perturbed, or mitigated images. The features from original, adversarial, and mitigated images are computed from these model and matched with the features of every image in the database. Out of L_1 and L_2 norms, we have found that the L_2 norm yields better performance, therefore the results are reported using L_2 norm.

The verification produced a score matrix of size 836×836, where 836 is the number of images in the MEDS database. The results are compared with selective dropout, randomization, and kernel smoothing based mitigation algorithms by Goswami et al. [123, 124], Xie et al. [352], and Wang et al. [338], respectively. Table 5.14 summarizes the true positive rate (TPR) at 1% false positive rate (FPR) and Figure 5-11 shows the ROC plots. It can be observed from the table that Selective dropout [123, 124] yields an improvement of 19.7% and 10.1% for LightCNN and VGGFace, respectively, the kernel smoothing defense [338] showcases improvements of only 6.4% and 8.4%, while the proposed algorithm shows an improvement of 29.8% and 30.3%, respectively. Randomization algorithm [352] performs poorly on face verification experiments and shows a small improvement in the accuracy. A similar observation regarding the ineffectiveness of Randomization is reported in [115]. Analyzing the results at 0.1% FPR from Figure 5-11 shows that the proposed mitigation algorithm yields 20% to 26% improvement in the verification accuracy.

5.8 Summary

Deep learning algorithms provide state of the art results on a multitude of applications. However, it is also well established that they are highly vulnerable to adversarial perturbations. This chapter presents novel detection and mitigation algorithms inspired by Occam's razor principles. It is often believed that the solution to this vulnerability of deep learning systems must come from deep networks only. Contrary to this common understanding, in this chapter, we propose a non-deep learning approach that outperforms many deep network approaches by searching over a set of well-known image transforms such as Discrete Wavelet Transform and Discrete Sine Transform, and classifying the features with a support vector machine based classifier. Existing deep networks based defense have been proven ineffective against sophisticated adversaries, whereas image transformation based solution makes a strong defense because of the non-differential nature, multiscale, and orientation filtering. The proposed approach, which combines the outputs of two transforms, efficiently generalizes across databases as well as different unseen attacks and combinations of both (i.e., cross-database and unseen noise generation CNN model). The proposed algorithm is evaluated on large scale databases, including object database (validation set of ImageNet) and face recognition (MBGC) database. The proposed detection algorithm yields at-least 84.2% and 80.1% detection accuracy under seen and unseen database test settings, respectively. Across all the algorithms, the proposed detection algorithm yield the best results and support the detection of unseen attacks, attack models generated by unseen databases, and deep learning models. Besides, we also show how the impact of the adversarial perturbation can be neutralized using a wavelet decomposition based filtering method of denoising. The mitigation results with different perturbation methods on several image databases illustrate the effectiveness of the proposed method.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

DAMAD: Database, Attack, and Model Agnostic Adversarial Perturbation Detector

6.1 Introduction

High accuracies of deep learning networks have motivated the development of automated solutions for a variety of tasks ranging from information retrieval to disease prediction, and surveillance. However, recent research efforts [57, 319] have showcased that the singularities of deep networks can be exploited to design attacks for corresponding networks. The widespread popularity of deep learning algorithms and their vulnerability to adversarial examples has motivated research in detecting such attacks. Detection can act as the first crucial stage of defense against adversarial attack and the detected examples can be discarded or processed further to remove the adversarial effects for correct classification.

Perturbation detection algorithms generally assume that the model, attack, and data characteristics are known and focus on intra-database, intra-attack, and intra-model training-testing. However, in real world settings, an attacker may be using *unseen* model trained on *unknown* database to generate perturbed samples. As shown in Figure 6-1(a), this introduces three main challenges in adversarial perturbation detection that may affect the performance of perturbation detection algorithms:

Cross-database variations refer to the scenario when the perturbation detection model is trained on one database and tested on a different database. For instance, when training is performed using

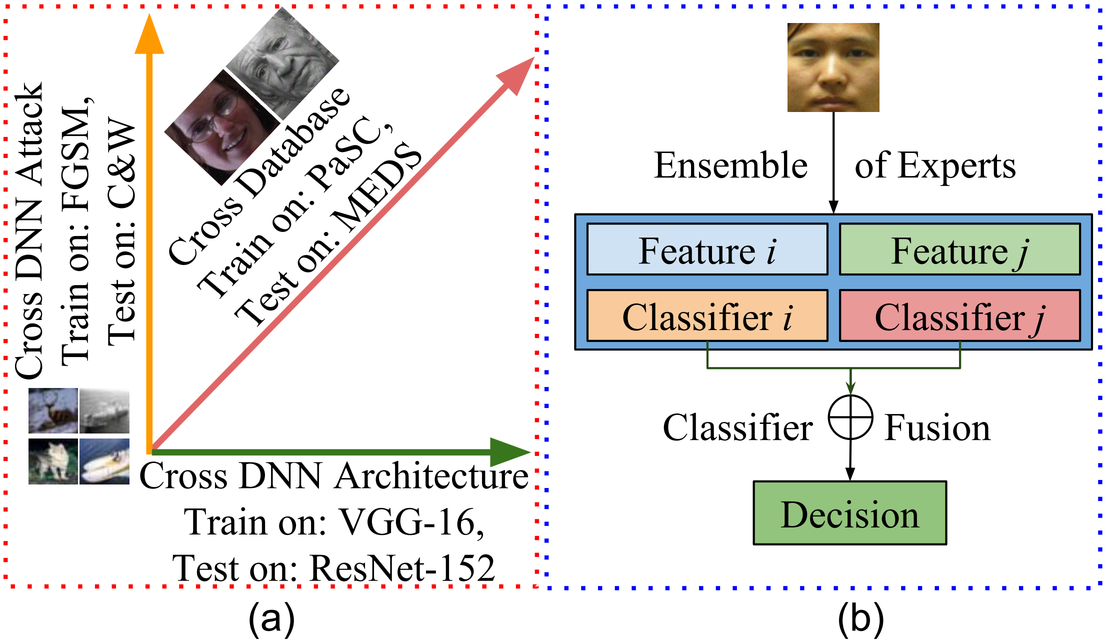


Figure 6-1: (a) Three challenges in adversarial perturbation detection: (i) cross-database, (ii) cross-architecture, and (iii) cross-attack (i.e. cross DNN Loss). (b) represents the motivation towards generalized adversarial detection approach.

CMU MultiPIE [127] database and testing is performed on PaSC [36] face database.

Cross-model variations refer to the scenario when the perturbation detection model is trained on adversarial images generated from one DNN architecture while the test cases are generated from another attack DNN model. For instance, perturbed images for training are generated using VGG-16 [306] model but the test images are generated using ResNet-152 model [137].

Cross-attack variations refer to the scenario where the perturbation detection algorithm is trained on one attack and tested on another. For instance, perturbation detection is trained with l_1 loss and tested with l_2 loss based attack.

Section 5.2 presents the comprehensive details of the existing adversarial attack generation and defense algorithms. While many detection methodologies have been proposed in the literature, a common and significant limitation of all the existing detection algorithms is the ineffectiveness [55] in detecting challenging adversarial examples such as generated using C&W's (l_2) attack [57]. For instance, the detection algorithm proposed by Li et al. [184] is able to achieve perturbation detection accuracy of only 8% when used to detect adversarial samples generated using Carlini and Wagner (C&W's) attack (l_2) on the MNIST database. Similarly, on the CIFAR-10 database only 1% adversarial examples are successfully detected [55, 57]. Other defense algorithms based on gradients hiding [244, 281], input transformations [131, 352], generative networks [310], CNN based classifiers [118, 128, 229], and single classifier [227] are also proven ineffective across stronger attacks [24, 55, 56]. Recent algorithms [270, 307] are able to provide certified defense for small perturbations on the MNIST attack database but they are not effective [330] against multiple attacks and databases. Wang et al. [339] show without any extra computational cost that trade-off between the clean and adversarial robustness of the CNN models. While the above defenses are evaluated on the images, some research works are proposed to tackle the adversarial videos. Lo et al. [198] over and under complete representation for the purification of adversarial features to mitigate its effect on recognition. Lo and Patel [197] have performed the adversarial training through multiple distorted videos in training for possible adversarial robustness. Not directly related to adversarial attacks, Tao and Cao [321] presents the resilient learning against erroneous database labels through noisy labels.

In this research, we have developed a generalized defense algorithm termed as ***DAMAD:*** Database, Attack, and Model Agnostic Detector. In order to achieve generalizability, the proposed approach follows "ensemble of experts" or fusion approach and combines different features from multiple "experts" (algorithms) (as shown in Figure 6-1(b)). The key highlights of this research are:

- a novel adversarial perturbation detection algorithm is proposed which is an amalgamation of a non-linear embedding obtained from an auto-encoder and statistical texture attributes obtained from DenseNet feature-maps;
- the proposed detection algorithm is evaluated on object, face, and digit recognition problems. Extensive experiments with multiple publicly available databases and deep networks showcase the efficacy of the algorithm in detecting different kinds of attacks;
- experiments pertaining to cross-database, cross-model, and cross-attack (DNN loss) scenarios showcase the effectiveness of DAMAD; and the strength of DAMAD is also evaluated against a white-box attack[1] and the proposed fusion approach shows resiliency against

[1] In adversarial attack research, white-box attack refers to the condition in which an attacker has access to the defense/detection mechanism and classifier. On the other hand, black-box attacks are defined where an attacker does not have access to both classifier and defense mechanism.

these attacks. The proposed algorithm also outperforms recent detection algorithms such as Adaptive Noise Reduction (ANR) [188], Bayesian Uncertainty (BU) [95], CNN response approach [123], LID [208], Base-OOD [139], ODIN [189], ESRM [191], and Mahalanobis [177] based algorithms.

To the best of our knowledge this is the first work where adversarial detection algorithm is proposed which is agnostic to multiple attack algorithms, CNN models, and databases. Based on the generalizability analysis across various unseen conditions, it is our assertion that the DAMAD algorithm can effectively be used against any adversarial attacks.

6.2 Proposed Algorithm

The adversarial image generation approaches embed an imperceptible “adversarial noise” in the original image. Across different attacks, we observed that the attack algorithms differ in the kind of noise (say gradient-based or magnitude), magnitude of noise (say single step or iterative), and the region where it is embedded (say, every pixel or salient regions only). Based on the literature and limitations discussed in Section 5.2, we hypothesize that a single algorithm may not be able to detect different kinds of adversarial noise. Inspired from the multimodal biometrics research [282], we postulate that an “ensemble of experts” or multi-classifier fusion approach, which combines linear and non-linear features obtained from distinct sources, can alleviate the limitations of single feature classification approaches. In other words, the multi-classifier fusion approach can better model the original and adversarial noise classes to provide better generalizability.

Figure 6-2 shows the block diagram of the proposed *DAMAD* algorithm to detect the presence of adversarial attack in an image. In the proposed algorithm,

- statistical Haralick features from the intermediate layers of DenseNet are extracted and probability confidence scores from Support Vector Machine (SVM) is computed;
- features from intermediate layers of an autoencoder are extracted and SVM probability confidence scores are computed; and
- the two probability scores are combined to obtain a final decision.

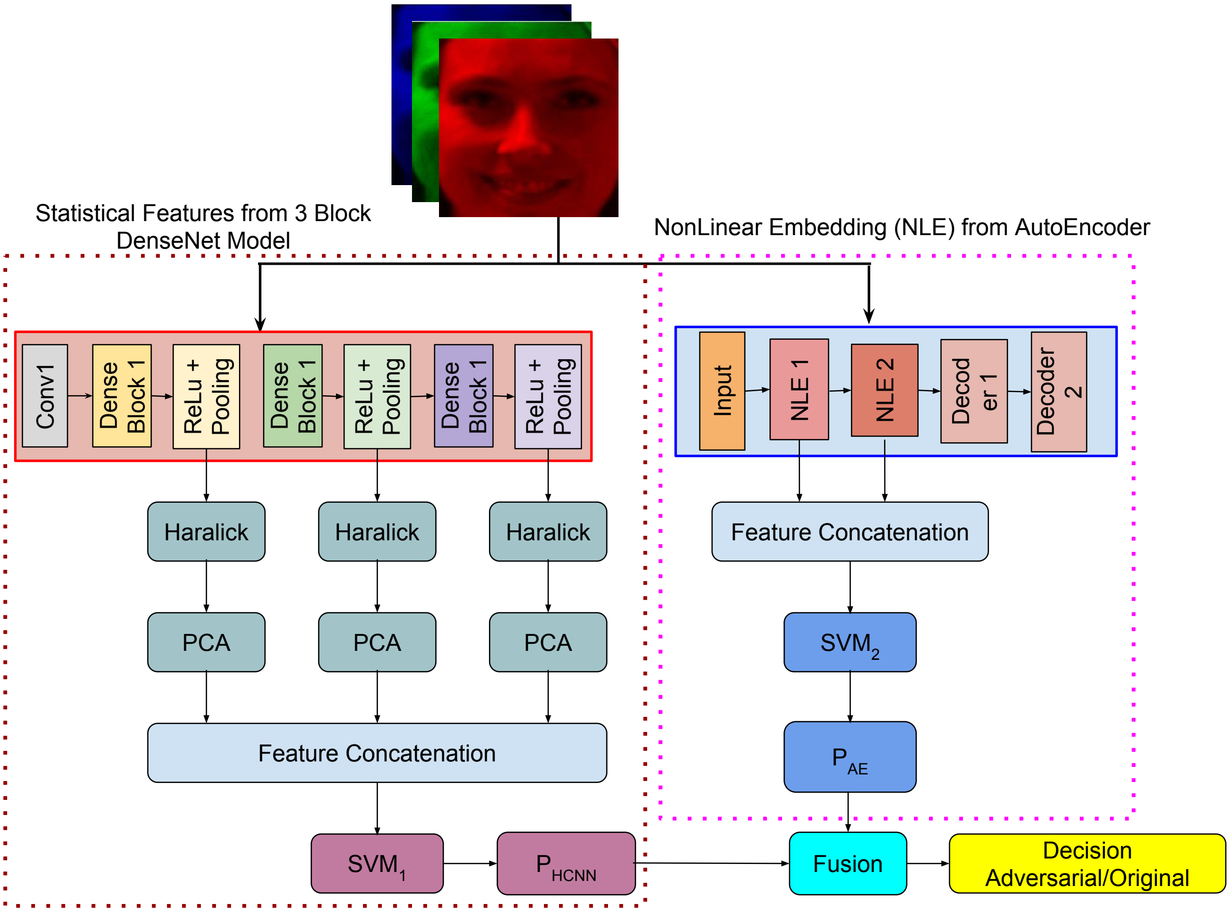


Figure 6-2: Proposed *DAMAD* adversarial perturbation detection algorithm combines statistical texture attributes obtained from DenseNet-121 feature maps and autoencoder embedding.

Statistical Features from DenseNet: Goswami et al. [123, 124] have showcased that filter responses of original and adversarial examples have significant differences, i.e. CNN filters are sensitive towards adversarial noise. From this observation, we propose to use intermediate layers of a deep network to learn the differences between original and attacked samples. In place of standard CNN, this research uses DenseNet which has stronger feature propagation and substantially fewer parameters. Further, we extract statistical features from the intermediate feature maps using Haralick features [136]. The Haralick features used are: Angular Second Moment, Contrast, Correlation, Sum of Squares: Variance, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Entropy, Information Measure of Correlation 2, Difference Entropy, Information Measure of Correlation 1, and Difference Variance. It encodes statistical properties of an input signal and extracts global attributes such as context, correlation, and entropy. Agarwal et al. [11] have proposed a combination of wavelet and Haralick for the detection of face presentation attack. However, the generalizability against unseen attack is a concern of the algorithm.

The proposed *DAMAD* utilizes three blocks of DenseNet-121 CNN model [144] to learn the filters that can accentuate the differences between the set of original and perturbed images. The dense blocks are initialized from the weights of the DenseNet-121 model trained on ILSVRC [87]. The dense block 1 consists of 6 dense layers, whereas, blocks 2 and 3 consist of 12 and 24 dense layers, respectively. Each dense layer has 2 convolution layers with filter size 1×1 and 3×3. The convolution block contains batch normalization, ReLU non-linearity followed by a convolution operation. The feature maps at any layer are the concatenation of all feature maps computed before that layer. After each dense block pooling, the transition layer is used to reduce the size of the feature maps which also helps in equating the size of each feature map. These new connectivities between the layers help in better encoding the patterns present in the input data and therefore motivated this research to compute the statistical features over the maps computed from DenseNet.

As shown in Figure 6-2, in the first convolution layer, a three-channel RGB image is convolved with 12 filters followed by the first block of DenseNet model. Before passing the output to the next DenseNet block, ReLU activation and 2×2 spatial pooling are applied to the feature maps. Similar process is repeated for the next two blocks of DenseNet network and filters are learned by using a fully connected (FC) layer (with two class classification). Once the network is trained, FC layer is removed and the filtered outputs at-the-end of each ReLU + Pooling is used to compute the Haralick features, i.e., 13 Haralick feature vector is computed for each filtered output in ReLU + Pooling layer. If the number of filtered output in any ReLU + Pooling layer is n, then the size of the Haralick feature vector is $13 \times n$. To reduce the dimensionality, Principal Component Analysis (PCA) is applied [347] and 99% Eigen energy is preserved. The reduced dimensional feature is then combined and classified into 2-classes (*original* and *perturbed*) using an linear SVM classifier (SVM_1 in Figure 6-2). From the literature, it has been observed that SVM and PCA are less susceptible to adversarial attacks [37, 172, 184, 274, 377]. Therefore, in context to the proposed DAMAD, PCA applied on the statistical features obtained from CNN followed by SVM classification helps in better discrimination between real and attack classes.

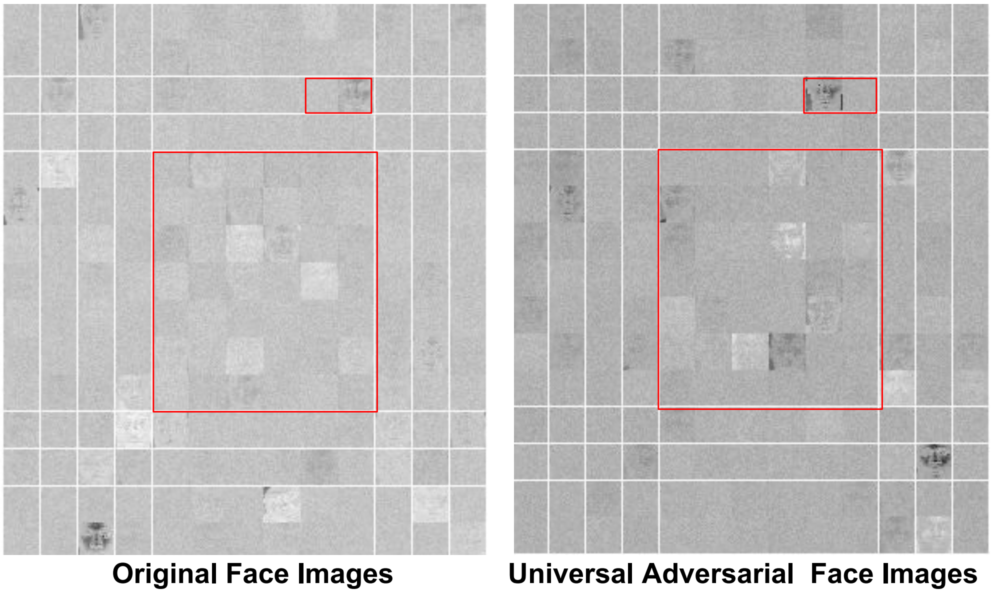


Figure 6-3: Hidden layer visualization of Auto-Encoder. The embeddings learned on adversarial examples are more noisy as compared to clean images which might help in detecting the attack.

AutoEncoder Model for Perturbation Detection: As the second classifier, a denoising AutoEncoder (AE) model is trained which can help discriminate between perturbed and original images. An AE captures the intrinsic properties of data and learns to abstract image properties by learning the latent space representation. Visualization of hidden layer encoding of an AE (original and adversarial images), as shown in Figure 6-3, shows different spatial distribution of both classes in a non-linear space. This property is explored for detecting adversarial noise in the input images. An unsupervised autoencoder has a reconstruction loss function as

$$argmin_{W,W'} \, ||X - W'\phi(WX)||_2^2 + \lambda R \tag{6.1}$$

where, W and W' are the encoding and decoding weights, ϕ is the non-linear activation (e.g. ReLU), λ is the regularization constant, and R is the regularizer (e.g. $||.||$ norm and dropout). The stacked autoencoder extends equation 6.1 to

$$argmin_{W_1,\cdots,W_n,W_1',\cdots,W_n'} \, ||X - g \circ f(X)||_2^2 + \lambda R \tag{6.2}$$

Table 6.1: Number of original images and adversarial (perturbed) images generated for each database.

Database		Original	Adversarial Model	Perturbed	Classes
Face	MEDS	836	1. DeepFool: VGG-16 2. Universal: VGG-16, GoogLeNet, and ResNet-152	2,508	518
	PaSC	7,443		22,329	293
	Multi-PIE	1,680		5,040	336
Object	ILSVRC 2012	5,000	Universal: VGG-16, GoogLeNet, and ResNet-152	15,000	1000+
	CIFAR-10	9,000	FGSM-l_1, IFGSM-l_1, FGSM-l_2, IFGSM-l_2, FGSM, IFGSM, DNN Loss (l_1, l_2, EN), PGD	100,000	10
Digit	MNIST	9,000		100,000	10

$$g = W_1'\phi(W_2' \cdots \phi(W_n' f(X)), f = \phi(W_n \cdots \phi(W_1(X))$$

With two encoding layers, the feature can be represented as $H_x = \phi(W_1\phi(W_2X))$. Given an input original image X and a perturbed image Y, the features H_x and H_y are fed into a 2-class support vector machine (SVM) [76] with radial basis function kernel to distinguish between *original* and *perturbed* classes (SVM_2 in Figure 6-2).

Multiclassifier Fusion: The two feature networks, Haralick features from DenseNet and AE, are combined using a late fusion approach. The classification probability scores obtained from two SVM classifiers (P_{AE} and P_{HCNN}) are combined using *sum* rule, i.e. $P_{fusion} = \frac{P_{AE}+P_{HCNN}}{2}$ and the fused score is used to classify an input image as original or perturbed.

Implementation details: The *DAMAD* algorithm is implemented in Theano environment with K40 GPU and Adam optimizer. For the autoencoder model, given an input image of size N, two hidden layers are of size $[\frac{N}{2}, \frac{N}{2}]$. For CNN model, 12 filters of size 3×3 are used in first convolutional layer followed by 3 DenseNet blocks along with 2×2 max pooling layer. Further, for both the models, learning rate is 0.0001, dropout rate is set to 0.5, and number of epochs is 500. Geometric transformation ($1 - 3^\circ$ rotation) and reflection of the image is used for data augmentation.

6.3 Databases and Evaluation Protocol

This section summarizes the databases and attacks considered for evaluation along with the experimental protocol and existing algorithms for comparison.

Attacks: To evaluate the performance and generalizability of the proposed DAMAD algorithm, we

Table 6.2: Range and resolution of distortion parameters to generate IFGSM adversarial samples.

Method	**Grid Search**	
	Range	Resolution
IFGSM-L_∞	[10^{-3}, 1]	10^{-3}
IFGSM-L_1	[1, 10^3]	1
IFGSM-L_2	[10^{-2}, 10]	10^{-2}

have performed the experiments with different attacks: optimization based (EN [64], C&W [57]), Universal Perturbations [231], PGD [213], gradient based algorithms [122, 168], and DeepFool [232].

Universal attack or image agnostic attacks [231]: We have used three different deep neural network models to generate the universal perturbed images. The DNN models used are: VGG-16 [306], ResNet-152 [137], and GoogLeNet [318]. Since these are among the best performing networks for tasks such as object recognition and face recognition, we have selected these networks to generate the universal adversary on face and ImageNet databases.

DNN loss based adversarial perturbations: Nine different types of DNN loss based attacks are selected to generate the adversarial images from the MNIST and CIFAR-10 databases. The selected attack generation algorithms are among the most challenging attacks [55]. Gradient based adversarial example generation algorithms are the most common in the literature and hence they are also utilized to generate the adversarial images. In this research, we have used a basic version of a gradient based algorithm known as Fast Gradient Sign Method (FGSM) and iterative version of FGSM (IFGSM). PGD is another stronger variant of FGSM attack which iteratively computes the adversarial noise. It is also considered the universal adversary among first order adversaries. Other than the basic version, l_1, l_2, and Elastic Net (EN) norm minimization based variants are also used to generate the adversarial examples.

DeepFool: The minimal norm perturbation is computed iteratively. The algorithms start with a clean image which resides in the decision boundary defined by the classifier. At each iteration subtle noise vector is added in input image which the aim to take this image outside the decision boundary.

Attack Parameters: For C&W and EN attacks, regularization parameter (initially c=0.001) is searched over nine binary steps where each step runs for 1000 iterations. The initial learning rate

Table 6.3: Experimental setup of C&W, EN, and L_1 attacks.

Parameter	**Value**
Initial Learning Rate	0.01
Iterations	1,000
Initial Regularization	0.001
Steps	9
Optimizers	ADAM and projected FISTA

is set to 0.01. ADAM optimizer, and projected FISTA with square-root decaying rate are used for C&W and EN, respectively. Similarly, for IFGSM and its variants, CleaverHans[2] package is used. The best distortion parameter is selected using fine-grained search. 10 FGM iterations are implemented with distortion parameter $\epsilon/10$ in each iteration. All other settings are kept as default for all the attacks. The experimental parameters used for adversarial examples generation are reported in Table 6.2 and 6.3. The original codes provided by the authors of Universal, DeepFool, and PGD attacks are used with quasi-imperceptible adversarial noise. The adversarial examples selected contains the lowest distortion.

Databases: The results are reported with six popularly used face, object, and digit recognition databases. The face databases are: Point and Shoot Challenge (PaSC) database [36], CMU Multi-PIE [127], and the Multiple Encounters Dataset (MEDS) [99]. From these three databases, more than 9500 frontal or nearly-frontal images are randomly selected. The object recognition databases used are CIFAR-10 [164] and ImageNet (i.e. ILSVRC-2012) [87]. We have selected 5000 images from the ImageNet database and 9000 images from the CIFAR-10 database. MNIST database [174] contains images of handwritten digits from 0-9. Similar to the protocol/code defined in [64], we have selected 9000 images from the MNIST database. In total, we have more than 32,500 original images pertaining to more 2150 classes, across these six databases.

We next created adversarially perturbed images corresponding to the adversarial attacks discussed above. In total, there are more than 29,000 perturbed face images and 177,000 perturbed images from the other three databases. Table 6.1 provides the number of images generated for each of the databases and the adversarial model used for image generation. It is to be noted that the codes and models for adversarial generation are taken from original papers in order to avoid any bias.

[2] https://github.com/tensorflow/cleverhans

Protocol: The evaluation protocol includes both positive and negative attack detection (i.e. original and perturbed images). The experiments are segregated according to intra-variations and cross-variations (architecture/attack/database). For all the scenarios related to **intra-database** (such as training and testing on MEDS) and **intra-attack** (such as l_1-l_1) experiments, 50% of the data from both classes is randomly selected for training and the remaining 50% for testing. In **cross-database** (such as MEDS-PaSC) and **cross-attack** (such as l_1-l_2) scenarios, original/adversarial images of one database/attack are used for training while original/adversarial images of another database/attack are used for the evaluation. Similarly, in the case of **'cross DNN architecture',** adversarial images generated using one DNN model (such as VGG-16) are used to train the classifier, while at the time of evaluation, adversarial images generated using another model (such as GoogLeNet) is used. We have even performed two fold unseen training-testing experiments namely **'cross DNN architecture and cross-database',** where not only DNN architecture from which universal adversarial images are generated is different but testing database is also different. It is to be noted that this is the first work to report results on 'cross' training-testing condition in three areas: cross-database, different DNN architectures, and different loss functions (l_1/l_2/l_1+l_2/GSM).

Evaluation Metric: The results are reported using the average detection accuracy of real and adversarial examples. The detection accuracy is the average of true positive rate (TPR) and true negative rate (TNR). TPR is defined as the rate of real examples being classified as real and TNR is defined as the rate of adversarial examples being classified as adversarial. In order to maintain the class balance, in each experiment, we have used an equal number of real and adversarial examples.

Algorithms for Comparison: Performance of *DAMAD* is compared with three recently proposed detection algorithms: Adaptive Noise Reduction (ANR) [188], Bayesian Uncertainty (BU) [95], Base-OOD [139], ODIN [189], ESRM [191], and CNN response [123]. The Base-OOD and ODIN use the softmax probabilities of DNN model to identify out-of-distribution (OOD) samples. The ODIN, an enhanced version of Base-OOD, uses the temperature scaling to softmax probabilities [142] and input perturbation to enlarge the softamx score gap between in-and-out distribution samples. ESRM uses the concept of staganalysis for the detection of adversarial attacks. In model the dependency between the adjacent pixel in certain neighborhood and model that dependency using hidden markov model. Other than the existing adversarial detection algorithms, *DAMAD* is

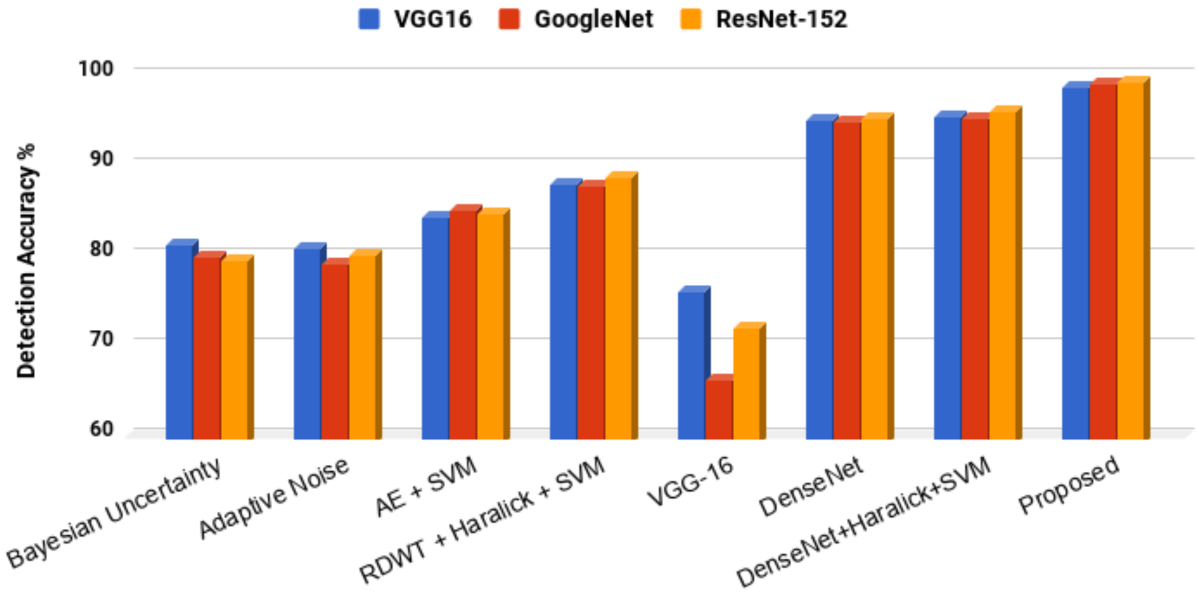


Figure 6-4: Detection performance of DAMAD and existing adversary detection algorithms on the **ImageNet** database with Universal adversarial perturbation (Best viewed in zoom and color.).

compared with two deep learning models: VGG-16 [306] and DenseNet [144]. The VGG-16 and DenseNet model (pre-trained on Imagenet) are fine-tuned using the adversarial and original images for perturbation detection. Along with these, detailed analysis is performed with individual components of *DAMAD*, RDWT (redundant discrete wavelet transform) + Haralick, and local binary pattern (LBP) handcrafted features. The SVM classifier is trained on the training set corresponding to each protocol and detection results are reported using features computed on the testing set. The comparison with recently proposed LID [208] and Mahalanobis [177] algorithms on complex l_2 attack on CIFAR10 database is also reported.

6.4 Results and Analysis

The results are divided into three parts. First the results are analyzed with respect to the intra-variations in database, model, and attack, followed by inter-variations. Finally, the general observations made across the intra-variations and inter-variations experiments are discussed.

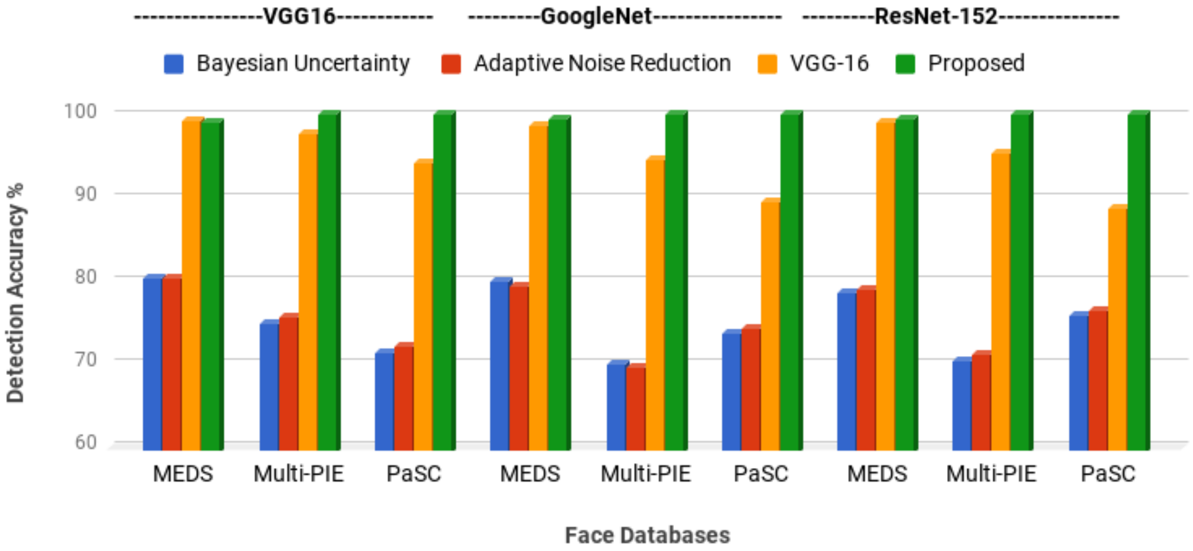


Figure 6-5: Detection performance of DAMAD and existing adversary detection algorithms on **face databases** with Universal adversarial perturbation (Best viewed in zoom and color.).

6.4.1 Results with Intra-Variations

Figure 6-4 and 6-5 summarizes the results on the ImageNet and the face databases in the intra-variations setting. On the ImageNet and face databases with Universal attack, the proposed DAMAD correctly classifies more than 98% samples, irrespective of the models used (VGG-16, GoogLeNet, and ResNet-152). The comparative results documented in Figure 6-4 show that the detection results of existing algorithms yield significantly lower performance. The performance of the detection algorithm proposed by Goswami et al. [123], which utilizes the intermediate filter response of VGG, is 13.2% and 16.8% lower than DAMAD on PaSC and MEDS databases, respectively. For C&W l_2 and PGD with ($\epsilon = 0.03$) attacks, the proposed DAMAD yields at-least 91% and 93% detection accuracy on face databases (MEDS, PaSC, and Multi-PIE). On the ImageNet database, in comparison to existing algorithms, there is a difference of at least 17% for all three models. On ImageNet, when VGG-16 fine-tuned model is used for universal adversarial sample detection, the accuracy is in the range of 65-75 which is significantly lower than DAMAD. Similarly, DenseNet only based detection model shows at-least 20 lower accuracy compared to the proposed DAMAD.

Table 6.4 summarizes the results on the MNIST and CIFAR-10 databases with different kinds of adversarial attacks. *DAMAD* yields more than 97.1% detection accuracy on optimization and

Table 6.4: Adversarial detection performance of the proposed DAMAD and existing algorithms on the CIFAR-10 and MNIST databases.

Databases	Algorithms	Attacks									
		l_1	l_2	EN	PGD	FGSM	FGSM-l_1	FGSM-l_2	IFGSM	IFGSM-l_1	IFGSM-l_2
MNIST	Bayesian Uncertainty [95]	77.3	78.5	78.1	74.6	83.7	82.9	81.2	85.3	84.9	84.4
	Adaptive Noise Reduction [188]	78.6	79.2	79.6	77.8	82.9	82.7	82.1	85.9	85.7	85.1
	Base-OOD [139]	72.5	68.9	65.6	66.0	87.8	82.3	78.5	88.3	82.9	80.0
	ODIN [189]	78.7	72.0	75.6	73.3	88.8	86.9	84.5	90.2	86.2	84.1
	RDWT + Haralick + SVM [11]	73.4	71.6	68.0	71.2	90.3	98.9	96.8	97.9	99.9	98.8
	Proposed DAMAD	**99.1**	**99.6**	**99.5**	**99.3**	**99.8**	**100**	**100**	**100**	**100**	**100**
CIFAR-10	Bayesian Uncertainty [95]	56.1	57.3	58.6	56.5	84.0	84.4	83.5	86.8	87.7	88.1
	Adaptive Noise Reduction [188]	55.9	57.8	57.2	59.2	83.2	83.5	83.8	87.1	88.3	88.5
	Base-OOD [139]	62.2	64.6	61.0	63.1	81.8	80.3	76.3	80.3	77.7	75.0
	ODIN [189]	64.3	62.9	63.6	65.7	82.1	81.6	82.0	86.7	86.0	82.5
	RDWT+Haralick+SVM [11]	61.7	62.3	61.3	60.4	62.9	57.8	56.1	72.2	64.0	62.8
	Proposed DAMAD	**98.3**	**98.1**	**99.0**	**97.8**	**97.4**	**97.5**	**97.5**	**97.1**	**97.3**	**97.5**

gradient based attacks on CIFAR-10 database. The detection performance of two existing algorithms (ANR: Adaptive Noise Reduction [188] and BU: Bayesian Uncertainty [95]) on gradient based attacks are in the range of $83.2\% - 88.5\%$ on CIFAR-10 database, which is at-least 8.6% lower than *DAMAD*. On the CIFAR-10 database, *DAMAD* is at-least 39.5% higher compared to existing algorithms on challenging l_1, l_2, and EN attacks. Similarly, on the MNIST database, the adversarial detection accuracy of *DAMAD* on gradient based attacks is in the range of 99%–100%, whereas the performance of two existing algorithms are in between 81.2% - 85.9%. The *DAMAD* improves the C&W l_2 attack detection performance of LID from 76.5% to 98.1% when ResNet model is used for LID [208]. Similarly, the detection performance of Mahalanobis [177] and ESRM [191] improves atleast by 6.3% and 30.6%, respectively when the proposed DAMAD is used for complex optimization based adversarial examples detection. The results are shown in Figure 6-6.

The performance is also evaluated on DeepFool adversary [232]. More than $9,500$ DeepFool adversarial images are generated from three face databases (MEDS, Multi-PIE, and PaSC) using the VGG-16 DNN architecture. Results of the *DAMAD*, both in ‘intra’ and ‘cross’ database scenarios, are reported in Table 6.5. The detection performance of the DAMAD algorithm is at least 4.6% and 15.6% better than DenseNet based classification model when trained using PaSC and Multi-PIE databases, respectively. The proposed algorithm outperforms the recently proposed algorithm based on CNN filter responses [123]. The performance of DAMAD is 29.8% and 39.3% better than CNN filter response algorithm on PaSC and MEDS databases, respectively. The results are shown in Figure 6-7.

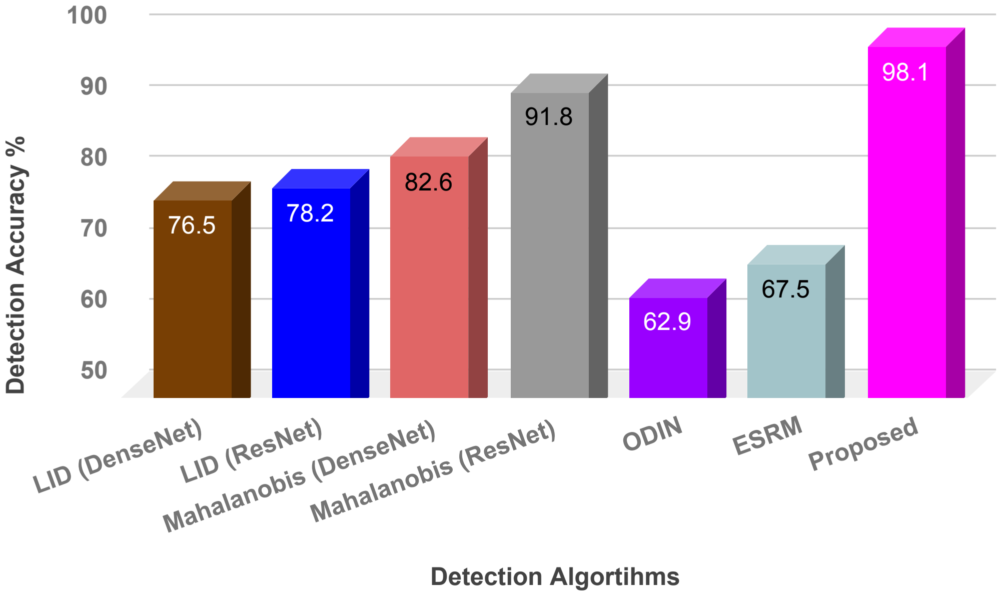
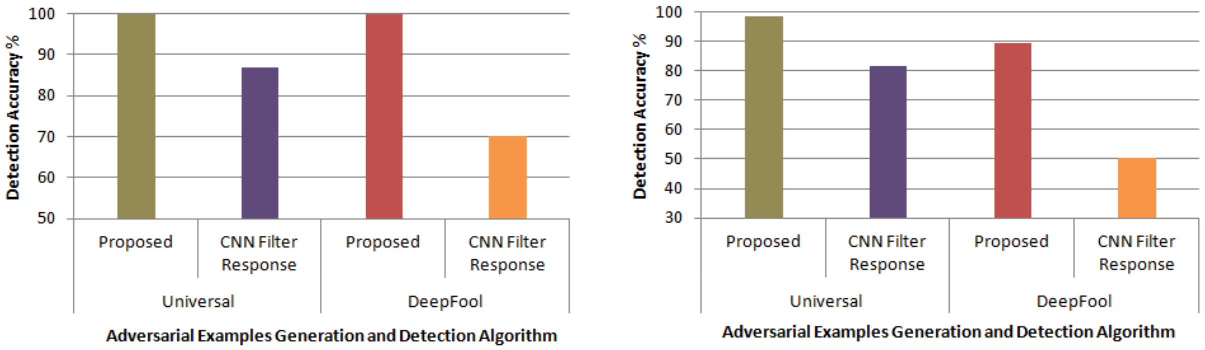


Figure 6-6: Comparison of the proposed algorithm with state-of-the-art detection algorithms (LID [208], Mahalanobis [177], ODIN [189], and ESRM [191]) on complex l_2 [57] attack on CIFAR-10 [164] database.



(a) Results on the PaSC database [36]

(b) Results on MEDS database [99]

Figure 6-7: Comparison of the proposed algorithm with state-of-the-art detection algorithm (**CNN Filter Response** [123]) on Universal [231] and DeepFool [232] attack on face databases. (Best viewed in color.)

Ablation Study: We next perform an ablation study and evaluate the effectiveness of individual steps of the algorithm on the ImageNet database. As shown in Figure 6-4, it is observed that individual components such as (AE+SVM) and (DenseNet+Haralick+SVM) are individually not effective. The combination of these components in the DAMAD algorithm yields the best results. Further, in place of DenseNet, RDWT+Haralick with SVM yields lower performance. This shows

Table 6.5: Detection performance of the **DAMAD** for DeepFool adversary on face databases with intra and cross database testing.

Train	Test		
	MEDS	Multi-PIE	PaSC
MEDS	90.2	87.6	86.2
Multi-PIE	89.5	95.0	100.0
PaSC	90.7	96.3	100.0

that all the components of the DAMAD algorithm are important for providing consistently accurate detection results across different models. Similar observations are noted for the three face databases in cross-database experiments (Table 6.6 in Section V.B.). The importance of DenseNet over ResNet as dicussed earlier is also proven through experiments. On ImageNet database, the accuracy of the DenseNet model is at-least 94.7% across all three universal perturbation generation CNN architecture. On the other hand, the performance of the ResNet-152 model is at-least 12% lower than DenseNet-121.

Table 6.6: Detection performance on face databases with cross database training-testing for the Universal attack. - represents the intra database conditions and corresponding results are reported in Figure 6-5.

DNN Model	Detection Algorithm	VGG-16			GoogLeNet			ResNet-152		
Train DB		MEDS	Multi-PIE	PaSC	MEDS	Multi-PIE	PaSC	MEDS	Multi-PIE	PaSC
MEDS	Bayesian Uncertainty [95]	–	79.3	78.5	–	73.1	74.1	–	72.9	75.0
	Adaptive Noise Reduction [188]	–	77.4	79.1	–	70.5	71.8	–	71.3	72.9
	Base-OOD [139]	–	82.3	81.5	–	76.4	78.1	–	74.0	77.9
	ODIN [189]	–	84.7	82.6	–	80.0	81.2	–	77.5	78.2
	VGG-16	–	50.0	50.7	–	50.0	50.5	–	50.5	52.0
	AE + SVM	–	78.6	79.1	–	71.9	72.2	–	73.4	73.8
	DenseNet	–	97.1	96.3	–	98.6	98.5	–	98.7	99.5
	DenseNet + Haralick + SVM	–	99.4	99.2	–	99.0	99.1	–	99.2	99.1
	RDWT + Haralick + SVM [11]	–	99.7	99.6	–	97.5	96.5	–	98.0	99.3
	Proposed DAMAD	–	**100**	**100**		**100**	**99.8**		**100**	**99.9**
Multi-PIE	Bayesian Uncertainty [95]	70.1	–	72.6	70.9	–	71.0	72.8	–	70.2
	Adaptive Noise Reduction [188]	70.9	–	71.7	70.3	–	69.1	70.5	–	69.0
	Base-OOD [139]	71.9	–	73.1	69.4	–	72.3	71.0	–	74.9
	ODIN [189]	73.3	–	75.1	68.0	–	69.9	70.5	–	76.6
	VGG-16	75.3	–	79.0	82.6	–	75.6	84.3	–	75.8
	AE + SVM	73.6	–	75.3	71.9	–	71.3	72.7	–	71.7
	DenseNet	96.9	–	92.8	93.1	–	98.0	91.2	–	98.9
	DenseNet + Haralick + SVM	97.0	–	94.4	93.8	–	99.5	92.1	–	99.7
	RDWT + Haralick + SVM [11]	71.3	–	99.9	70.9	–	99.9	71.4	–	99.9
	Proposed DAMAD	**98.4**	–	**100**	**99.0**	–	**100**	**99.5**		**100**
PaSC	Bayesian Uncertainty [95]	63.9	65.3	–	62.8	63.1	–	65.2	66.9	–
	Adaptive Noise Reduction [188]	65.3	67.8	–	63.2	63.0	–	66.8	68.7	–
	Base-OOD [139]	62.1	69.8	–	64.3	66.4	–	70.1	72.3	–
	ODIN [189]	64.3	77.7	–	68.2	72.3	–	78.9	74.9	–
	VGG-16	76.9	81.3	–	68.2	75.5	–	74.4	64.0	–
	AE + SVM	70.6	71.3	–	69.7	68.4	–	70.3	71.7	–
	DenseNet	90.3	90.7	–	85.9	87.4	–	88.6	89.1	–
	DenseNet + Haralick + SVM	91.5	92.9	–	88.1	90.8	–	90.1	90.6	–
	RDWT + Haralick + SVM [11]	68.7	100	–	67.9	100	–	68.4	100	–
	Proposed DAMAD	**98.2**	**100**	–	**97.4**	**100**	–	**98.8**	**100**	

6.4.2 Results with Inter-Variations

The next set of experiments are performed to test the generalizability of the proposed algorithm with variations in testing model, attack, and database, compared to the ones for training. The combination of attacks and databases are selected according to the research in literature. For instance, DNN loss based attacks have been performed on MNIST and CIFAR, while Universal attack with different models is implemented for ImageNet and all three face databases.

Cross-database Evaluation: Table 6.6 summarizes the results of cross-database testing of existing algorithms, *DAMAD*, and components of *DAMAD*. The detection performance of Bayesian Uncertainty (BU) and Adaptive Noise Reduction (ANR) is in the range of $62.8 - 79.3\%$ across different combinations of training and testing, whereas *DAMAD* lies in between $97.4 - 100\%$, thus showcasing the generalization capability of the algorithm. On universal adversarial images generated using the VGG-16 model, the CNN response [123] algorithm yields 53.4% and 63.2% detection accuracies on the PaSC and MEDS databases, respectively, which are 36.8% and 45.0% lower than the *DAMAD*, respectively. We have also used the DenseNet [144] as an adversarial model and generated the adversarial examples using face databases. We have generated the universal noise vector with different fooling rates (40%, 60%, and 80%) with input variation set to 0.03. When the DenseNet adversarial model is used, The proposed defense performed similar to other models such as VGG-16 and GoogLeNet. The accuracy ranges from 96.8% to 100% under cross-database scenarios which is similar to VGG-16, GoogLeNet, and ResNet-152 model reported in Table 6.6. Analyzing the performance of individual components (ablation study) of the algorithm shows that each component of the algorithm is important for high detection performance, and removing any component significantly reduces the performance on some cross train-test pairs.

Cross-attack Evaluation: In the cross-attack experiment, the training and testing adversarial images are generated using different attack types. Figure 6-8 shows the findings related to the cross-attack situation. The average (± standard deviation) adversary detection performances on cross attack scenario are $99.2{\pm}0.6\%$, $64.2{\pm}4.1\%$, $62.3{\pm}4.3\%$, $70.3{\pm}3.3\%$, $75.4{\pm}2.8\%$, and $68.7{\pm}2.7$ on the MNIST database using *DAMAD*, Adaptive Noise Reduction (ANR), Bayesian Uncertainty (BU), Base-OOD [139], ODIN [189], and ESRM [191] algorithms, respectively. Similarly, on the CIFAR-10 database *DAMAD*, Adaptive Noise Reduction (ANR), Bayesian Uncertainty (BU),

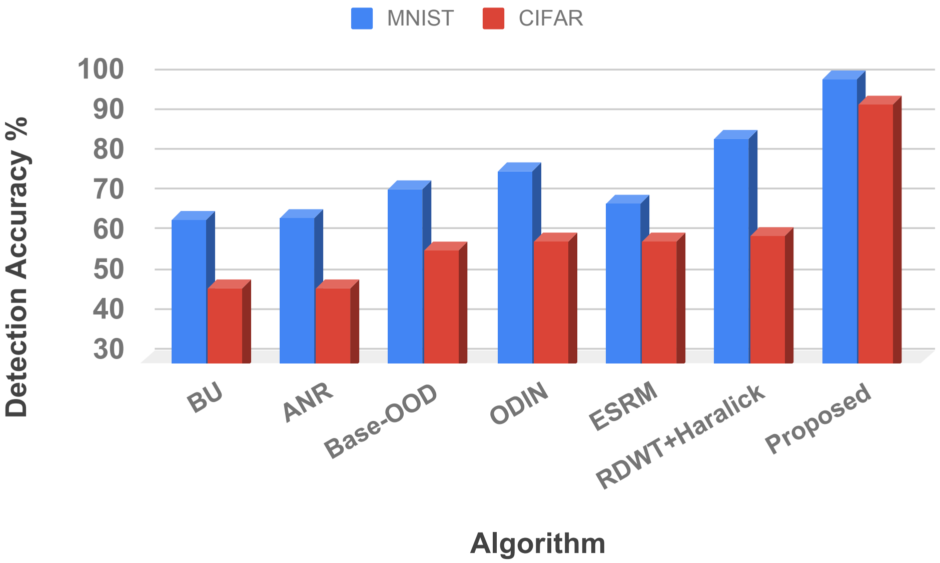
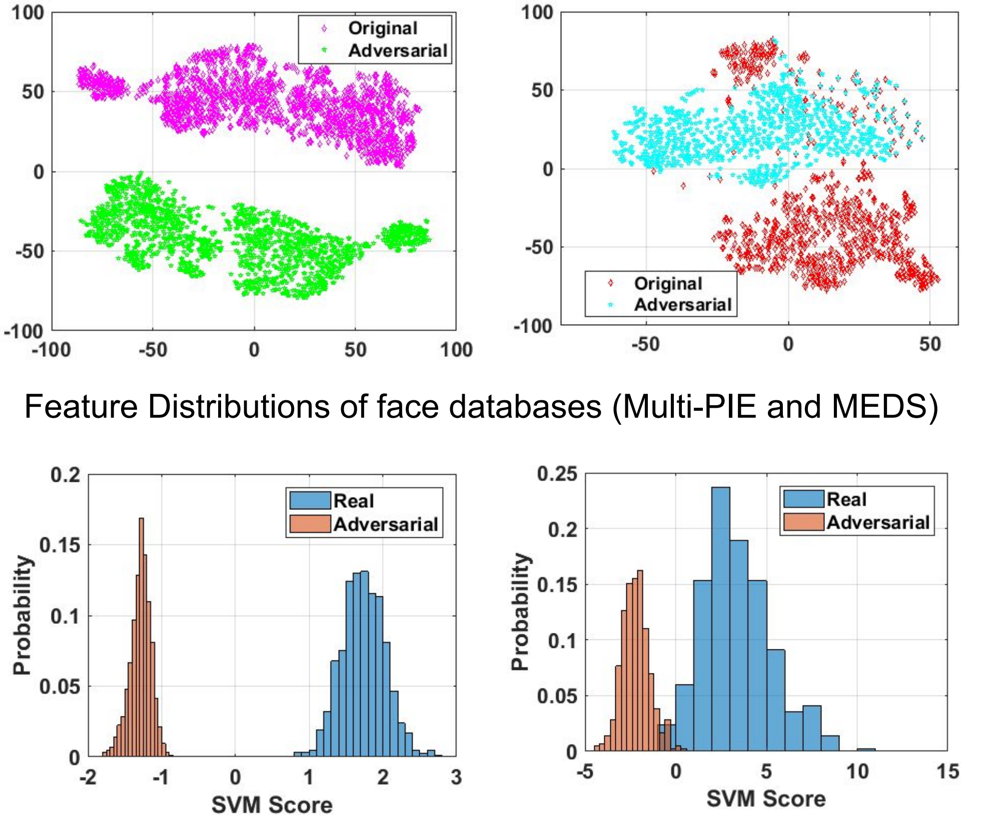


Figure 6-8: Attack detection results when the model is trained on one attack and tested with other attacks. Ten fold experiments are performed, each using only one attack for training and the remaining nine attacks for testing. The average detection accuracy is reported. The comparison with BU [95], ANR [188], Base-OOD [139], ODIN [189], ESRM [191], and RDWT+Haralick [11] algorithms is also reported.

Base-OOD [139], ODIN [189], and ESRM [191] algorithms yield an average detection accuracy of $93.7 \pm 1.2\%$, $46.7 \pm 3.1\%$, $47.5 \pm 3.2\%$, $56.8 \pm 2.2\%$, $58.9 \pm 1.5\%$, and 59.1 ± 1.9, respectively. This improved detection performance, which is more than 34% higher than existing algorithms, shows the generalizability and transferability properties of the proposed algorithm.

Cross-Databases and Cross-DNN-Architectures: To evaluate the generalizability in presence of more "unknown" factors, we performed another experiment as 'cross-database' and 'cross-architecture'. This experiment is performed using MEDS, Multi-PIE, and PaSC databases with VGG-16, GoogLeNet, and ResNet-152 architectures, where one database and one architecture is used for training while the other databases and architectures are used for testing. The proposed *DAMAD* achieves at least 99.97% accuracy which is significantly higher than the existing algorithms (less than 50% accurate). This experiment showcases that *DAMAD* is generalizable even in the case of both cross-databases and cross-architecture scenarios.



Feature Distributions of face databases (Multi-PIE and MEDS)

SVM Score Distributions of face databases (Multi-PIE and MEDS)

Figure 6-9: Haralick feature and classification score distribution of real and adversarial class images. Both feature and classification score distribution shows the high discriminability of original and adversarial images in statistical feature space.

6.4.3 Discussion

We have made the following observations across different experiments:

Without PCA: The proposed algorithm computes the Haralick features over each feature map which leads to high dimensional feature vector which is reduced using PCA. Without PCA, there is no significant difference in the classification performance; however, the computational load increases by multiple folds.

Universal perturbation can be detected easily in comparison to DNN loss based attacks. This observation is also made by [15] where PCA + SVM classification yields at least 93% detection accuracy on universal adversarial samples from multiple face databases. Testing universal adversarial perturbation with different parameter values (e.g. $\delta = 0.4$ and 0.2, $\epsilon = 0.5$, 1.0, and 10)

on the CIFAR and MNIST databases yields over 98% detection accuracy. Similarly, when C&W attack is tested on high resolution face images, over 95% detection accuracy is observed. Experiments with other CNN architectures (VGG and ResNet) are also performed and the results show that, on the ImageNet database, the detection accuracy of VGG-16 is 5-25% less than DenseNet in both intra and cross-variations testings.

Haralick Features on DenseNet: While the aim of an adversarial example is visual imperceptibility, they still modify the local pixel structure which can be detected using Haralick based statistical features. It is our hypothesis that if we detect these changes via statistical features, we should be able to detect the presence of adversarial noise. Haralick features measure the statistical characteristics such as homogeneity, entropy, contrast, correlation, and energy of the pixel distribution. From the experiments, it is evident that the statistical features computed over DenseNet maps outperform the DenseNet-only detection method. Further, the DNN loss based attacks are generated using non-linear CNN models, which may explain why the adversary detection learned over the DenseNet maps show higher performance than an AE based model.

Combination of Classifiers: We observe that DenseNet tries to learn feature maps which focus on low-level discriminative information. The statistical characteristics of Haralick features obtained from DenseNet maps and non-linear feature encoding using AE improve the strength of the proposed detector. The performance of the AE module suffers under the cross-database (Table 6.6) scenario in comparison to seen database performance. The accuracy of the AE module ranges from 68.4% to 79.1% under the cross-database scenario. Combination of statistical features and non-linear embedding shows the generalizability and transferability across databases, DNN loss functions, and DNN architectures. The high detection accuracy of the proposed adversarial detector can help make DNNs more robust in practical use by rejecting the adversarial examples.

Other Classifiers and Features: Figure 6-9 illustrates the t-SNE [210] scatter plots and SVM score distribution of real and adversarial face images. It is also found that the adversarial perturbation detection using SVM classification shows consistently higher performance as compared to other classifiers such as neural network (NNet). The accuracy of NNet on the face and ImageNet databases for 'intra-database' scenarios are in the range of 65-70%, which drops significantly for 'cross-database' (50–55%). We have also evaluated other traditional texture features such as Local Binary Patterns (LBP) [239] in place of Haralick features and the performance on MNIST and

CIFAR-10 databases is at least 3% *lower* than the Haralick texture features.

DAMAD algorithm is challenging to break because the algorithm is primarily utilizing the "ensemble of detectors" by combining DenseNet+Haralick+PCA+SVM and AE+SVM. It is our understanding that the proposed algorithm will be fooled in cases when the perturbation leads to minimal difference in features; however, we assert that in such cases, the object/face classification results will already be correct and will not require an attack detection algorithm.

6.5 Resilience of Detection Algorithm via White-Box Attack

In the real world settings, it might be possible that if the attacker has access to the detection algorithm, they might attack the detection algorithm itself. To evaluate the resiliency of DAMAD algorithm towards adversarial attacks, experiments with white-box attack scenario are performed. Since the attacker has access to the loss function of the network concerning its input and target labels, it can attempt to compute the perturbation to fool the network. Similar to Defense-GAN [287], FGSM (with $\epsilon = 0.3$) and C&W-l_2 (with $c = 2$) attacks are performed using projected gradient descent [213] for 100 iterations. DAMAD achieves more than 98% and 96% detection accuracy on the MNIST and CIFAR-10 databases, respectively. It is our assertion that one primary reason that DAMAD showcases resiliency against white-box attacks is that the decision is taken from multiple independent embeddings, AE and DenseNet features. Another important reason for it's resiliency is that shuffling of image parts or changes in pixel structure due to adversarial noise change the texture encoding (i.e., spatial relation) and it effectively gets captured by combination of DenseNet and Haralick features.

6.6 Summary

Adversarial perturbations have established the vulnerabilities of deep learning algorithms to adversarial attacks. Existing adversary detection algorithms attempt to detect the singularities; however, they are in general, loss-function, database, or model dependent. To mitigate this limitation, we propose *DAMAD* a generalized perturbation detection algorithm which is agnostic to model architecture, training dataset, and loss function used during training. The proposed adversarial perturbation detection algorithm is based on the fusion of autoencoder embedding and statistical texture features extracted from convolutional neural networks. The performance of DAMAD is evaluated on the challenging scenarios of cross-database, cross-attack, and cross-architecture training and testing along with traditional evaluation of testing on the same database with known attack and model. Comparison with state-of-the-art perturbation detection algorithms showcase the effectiveness of the proposed algorithm on six databases: ImageNet, CIFAR-10, Multi-PIE, MEDS, PaSC, and MNIST. Performance evaluation with *nearly a quarter of a million* adversarial and original images and comparison with recent algorithms show the effectiveness of the proposed algorithm. To the best of our knowledge, this is the first work addressing mismatched conditions in training and test database, loss functions, as well as the DNN architecture.

Chapter 7

Conclusion and Future Directions

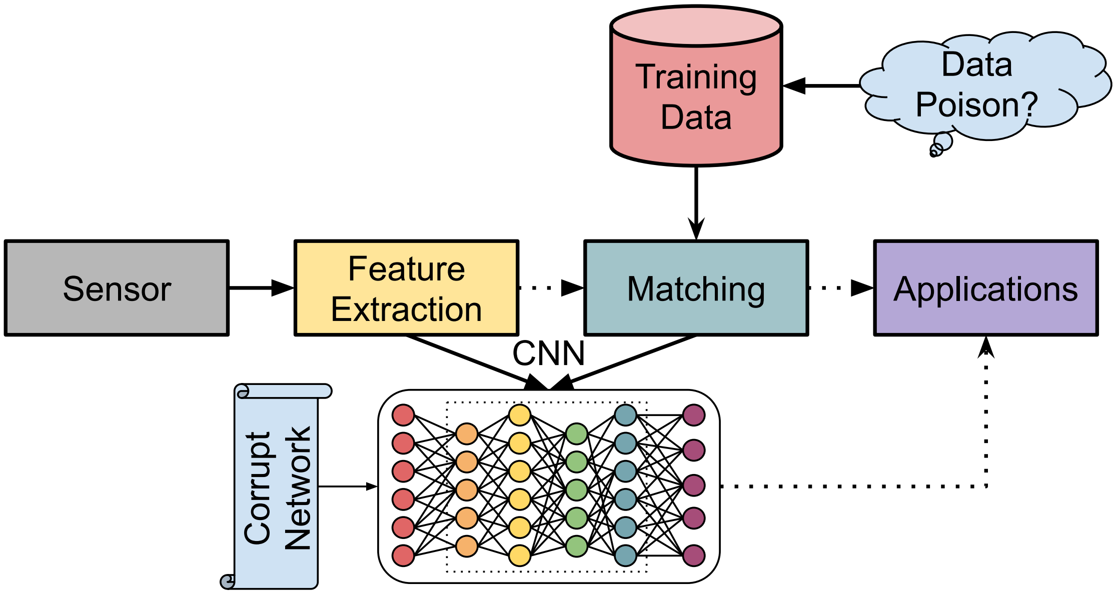


Figure 7-1: Shows the replacement of features extraction and classification step in the traditional computer vision system with the deep learning architecture. The future defense can be attributed to data poisoning in training or corruption deep learning network characteristics such as filters or decision functions.

Figure 7-1 shows the stages of a typical deep learning based computer vision system. As described in the *Introduction* section, each stage is vulnerable to various attacks. In this dissertation, we dealt with the attacks performed around one of the most important component of the system, i.e., data. The data attacks against which the security is provided in the dissertation can be broadly grouped into (i) presentation attacks and (ii) digital perturbation. The presentation attacks can be

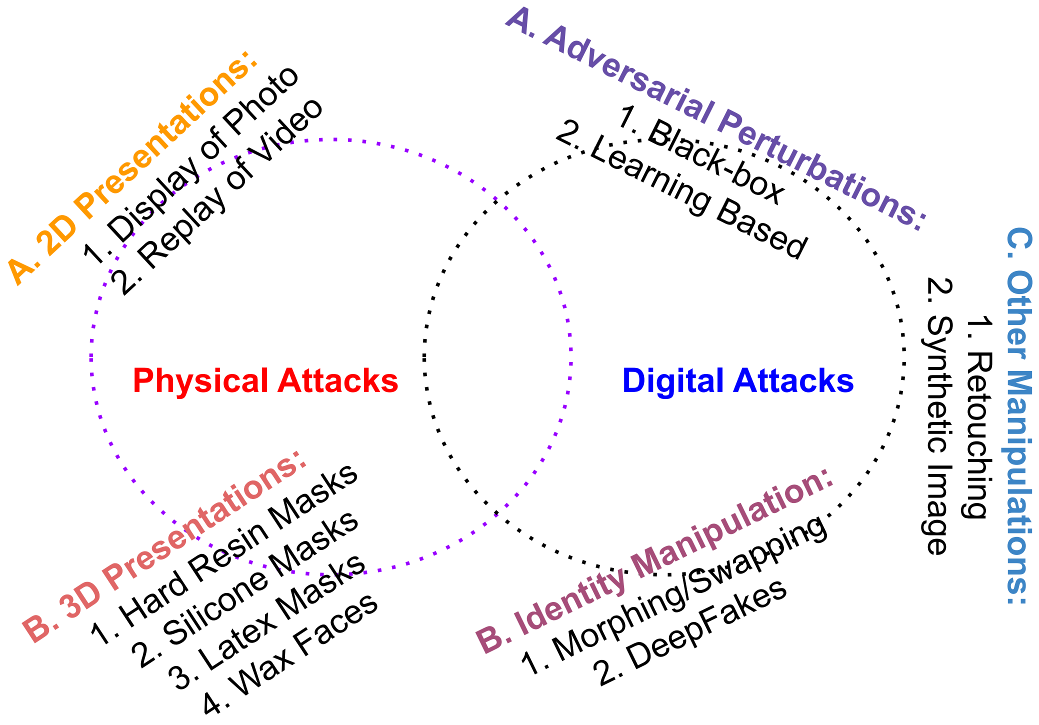


Figure 7-2: Categorization of the attacks on computer vision systems dealt with in this dissertation. The possible attacks are physical-based, performed at the sensor level, i.e., before the acquisition and digital attacks that perturb the data itself. Few anomalies can be utilized both for physical and digital attacks, such as adversarial perturbations and deepfake.

viewed as displaying fake data to the acquisition sensor and can be the printed photo or 3D masks. The goal of the presentation attacks can be two folds: (i) impersonation, i.e., a targeted attack to gain someone else's identity, and (ii) obfuscation, i.e., an untargeted attack where the aim is to hide one's own identity. The second attack contains two parts: one is related to biometric systems, specifically face morphing or swapping. Another is extensively applied against deep computer vision algorithms, including convolutional neural network (CNN) and is termed as adversarial attacks. It is shown that the morphed faces of different identities can share the identities; therefore, these attacks can be performed both in causative and exploratory form or can hamper the integrity and availability of the system. In the causative form, fake data can be directly provided to the machine learning systems. Counterfeit data can be placed in the training data for possible identity sharing or an increase of false positives in the exploratory attack. The adversarial attacks recently gained much attention on computer vision systems, especially deep learning algorithms, with intelligent, crafted, subtle noise. The adversarial attacks aim to fool the computer vision algorithm by adding intelligently crafted subtle noise in such a way that the noise is kept minimal so that human examiner can also be fooled in identifying whether input data, i.e., image or text, contains the noise. Similar to presentation and face morphing attacks, adversarial attacks can cover each category of attack. It can be causative or exploratory, targeted or untargeted, and can increase false negatives or false positives.

Figure 7-2 shows the division of the attack on computer vision algorithms ranging to be used for applications, including face verification to general-purpose object recognition. The attacks came into the picture as early as 2010; the first photo-based presentation attack databases were released. Later, several advancements have been made to improve the strength of the attacks by providing motion cues to the attacks by displaying video and developing sophisticated 3D masks. Similarly, another widespread attack shows the vulnerability of little adversarial noise came into the picture in 2014. Other than that, advancements in machine learning algorithms have made the generation of original images a reality. These synthetic images are now so powerful that by merely looking, any human can fail in identifying whether they are natural or not. Based on the continuous growing nature of attacks, updates in the defense algorithms are always required.

Apart from the attacks, the primary source of any computer vision algorithms is the camera. In large-scale projects, multiple source cameras are generally used, and each camera consists of its characteristics and generates entirely different images. The camera's probable aspects that affect the images are wavelength, processing algorithms, illumination sources, and resolution. Source camera identification is also one of the first steps in media forensics.

In this dissertation, we proposed an array of solutions for various computer vision problems, including identifying the source camera used in the acquisition of biometric images and segregating the clean images from the attack ones. First, an amalgamation of features is created for identifying image sources using multiple features such as statistical, textural, and image quality. Further, to reduce the curse of dimensionality of features, a novel feature selection algorithm is proposed. The proposed solution is not only tested against multiple biometric modalities but also against multiple imaging spectrum as well. Later, to protect the integrity of face recognition systems against physical and digital attacks, a novel feature engineering technique and the way it should be extracted for high efficiencies such as depth of extraction, i.e., global or local regions or both and input for feature extraction, i.e., raw images or transformed images or both. The intelligent way of feature extraction is useful against various attacks evaluated under *seen* and *unseen* database and attack settings. In the end, the *panoptic* defense algorithms are proposed to protect one of the most successful computer vision algorithms, i.e., deep convolutional neural networks (CNNs). The defense solutions correspond to *computational efficiency* and *agnostic* nature against various factors such as database, CNN architectures, and attack type.

7.1 Future Directions to Advance Secure Computer Vision

While the contributions of this dissertation protect the systems from multiple attacks, the continuously evolving nature of attacks requires ongoing upgrading in defense strategies. The possible future directions for advances in secure computer vision are shown in Figure 7-1. For example, in the era of deep learning, the components of deep learning systems can also be perturbed, such as the filters or entire layer or decision function [286]. Apart from that, the generation of synthetic images from generative networks also improves, and crafting high-resolution humans like face images or general object images is extremely easy. These synthetic generated images or adversarial images can be inserted in the training data through backdoor data poisoning [65, 185, 300, 313]. Therefore, the identification of all these new threats also needs to be addressed. Based on this, future directions can be described as:

- near infra-red and thermal spectrum videos/images can be used to improve the face presentation attack detection performance. A large scale database is required with challenging attacks including sophisticated silicone masks and latex masks in multiple imaging spectrum;

- because of the heavy use of pre-trained deep networks which might be trained on backdoor images, finding the patterns to determine whether it is trained on poisoned data requires attention;

- the digital forensic is a game of cat and mouse, i.e., if there is a security, there is one attacker also which is working on breaking the security algorithm. The need is to develop an adaptive algorithm which can counter the attack designed with the knowledge of security;

- the blocks of machine learning, such as feature extraction, matching, and network manipulation, also needs to be secured from attack. The combination of blockchain with a machine learning framework can further enhance the security mechanism. We have performed some preliminary work towards it and published in BTAS 2019 [113], CVPRW 2019 [112], and CVPRW 2020 [114].

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] CASIA Cross Sensor Databse. http://biometrics.idealtest.org/index.jsp. 33, 34

[2] CASIA version 2 database. Available: http://biometrics.idealtest.org/dbDetailForUser.do?id=2. Accessed: 2019-03-15. 30, 31

[3] CASIA version 3 database. Available: http://biometrics.idealtest.org/dbDetailForUser.do?id=3. Accessed: 2019-03-15. 31

[4] Miles Research Database. Available: http://www.milesresearch.com/. Accessed: 2019-03-15. 31, 32

[5] ND-Cross Sensor Database, University of Notre Dame. Available: http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html. Accessed: 2019-03-15. 31

[6] UPOL Iris Database. Available: http://www.inf.upol.cz/iris/. Accessed: 2019-03-15. 31, 32

[7] MNIST handwritten digit database. Available: http://yann.lecun.com/exdb/mnist/, 2010. 129

[8] Aditya Abhyankar and Stephanie Schuckers. Integrating a wavelet based perspiration liveness check with fingerprint recognition. *Pattern Recognition*, 42(3):452 – 464, 2009. 52

[9] Evan Ackerman. How drive. ai is mastering autonomous driving with deep learning. *IEEE Spectrum Magazine*, 1, 2017. 113

[10] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 109

[11] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using haralick features. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2016. XVIII, 97, 99, 152, 160, 162, 164

[12] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using haralick features. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–6, 2016. 99, 106

[13] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Fingerprint sensor classification via mélange of handcrafted features. In *International Conference on Pattern Recognition*, pages 3001–3006, 2016. 14, 16

[14] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore. SWAPPED! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE International Joint Conference on Biometrics*, pages 659–665, 2017. 97, 99

[15] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2018. 165

[16] Akshay Agarwal, Daksha Yadav, Naman Kohli, Richa Singh, Mayank Vatsa, and Afzel Noore. Face presentation attack with latex masks in multispectral videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 81–89, 2017. 50, 56

[17] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3389–3398, 2018. 119

[18] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 78, 114, 115

[19] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 2016. 69

[20] D. Anand, D. Tank, H. Tibrewal, and A. Sethi. Self-supervision vs. transfer learning: Robust biomedical image analysis against adversarial attacks. In *IEEE International Symposium on Biomedical Imaging*, pages 1159–1163, 2020. 113

[21] André Anjos and Sebastien Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2011. 72

[22] Sunpreet S Arora, Mayank Vatsa, Richa Singh, and Anil Jain. On iris camera interoperability. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 346–352, 2012. 6, 11, 14

[23] Sunpreet Singh Arora, Mayank Vatsa, and Richa Singh. Emerging challenges in iris recognition. Technical Report IIITD-TR-2012-002, 2012. 11

[24] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning*, 2018. 122, 149

[25] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *International Conference on Machine Learning*, 2018. 118

[26] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *IEEE International Joint Conference on Biometrics*, pages 319–328, 2017. 71

[27] Ismail Avcibas, Nasir Memon, and Bülent Sankur. Steganalysis using image quality metrics. *IEEE Transactions on Image Processing*, 12(2):221–229, 2003. 23

[28] Ismail Avcibas, Bulent Sankur, and Khalid Sayood. Statistical evaluation of image quality measures. *Journal of Electronic Imaging*, 11:206–223, 2002. 23

[29] Abdelkrim Ouafi Djamel Samai Mourad Oussalah Azeddine Benlamoudi, Kamal Eddine Aiadi. Face antispoofing based on frame difference and multilevel representation. *Journal of Electronic Imaging*, 26:26 – 26 – 14, 2017. 71, 74

[30] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint arXiv:1703.09387*, 2017. 116, 118

[31] Sudipta Banerjee and Arun Ross. From image to sensor: Comparative evaluation of multiple prnu estimation schemes for identifying sensors from nir iris images. In *IEEE International Workshop on Biometrics and Forensics*, pages 1–6, 2017. 14, 17

[32] Nick Bartlow, Nathan Kalka, Bojan Cukic, and Arun Ross. Identifying sensors from fingerprint images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 78–84, 2009. 14, 15, 43, 44

[33] Yashasvi Baweja, Poojan Oza, Pramuditha Perera, and Vishal M Patel. Anomaly detection-based unknown face presentation attack detection. *International Joint Conference on Biometrics*, 2020. 49

[34] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. 109

[35] Sevinc Bayram, Husrev Sencar, Nasir Memon, and Ismail Avcibas. Source camera identification based on cfa interpolation. In *IEEE International Conference on Image Processing*, volume 3, pages III–69, 2005. 13, 14

[36] J Ross Beveridge, P Jonathon Phillips, David S Bolme, Bruce A Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, P.J. Flynn, and Su Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, 2013. 148, 156, 161

[37] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Dimensionality reduction

as a defense against evasion attacks on machine learning classifiers. *arXiv preprint arXiv:1704.02654*, 2017. 120, 121, 152

[38] Aparna Bharati, Richa Singh, Mayank Vatsa, and Kevin W Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, 2016. 111

[39] Ishan Bhardwaj, Narendra D. Londhe, and Sunil K. Kopparapu. A spoof resistant multibiometric system based on the physiological and behavioral characteristics of fingerprint. *Pattern Recognition*, 62:214 – 224, 2017. 52

[40] Himanshu S Bhatt, Richa Singh, Mayank Vatsa, and Nalini K Ratha. Improving cross-resolution face matching using ensemble-based co-transfer learning. *IEEE Transactions on Image Processing*, 23(12):5654–5669, 2014. 77

[41] Sushil Bhattacharjee, Amir Mohammadi, André Anjos, and Sébastien Marcel. Recent advances in face presentation attack detection. In *Handbook of Biometric Anti-Spoofing*, pages 207–228. Springer, 2019. 51

[42] Sushil Bhattacharjee, Amir Mohammadi, Marcel, and Sébastien. Spoofing deep face recognition with custom silicone masks. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2018. 51

[43] Richard J Bolton and David J Hand. Statistical fraud detection: A review. *Statistical Science*, pages 235–249, 2002. 2

[44] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *IEEE International Conference on Image Processing*, pages 2636–2640, 2015. 73, 74

[45] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016. 56, 71, 72, 73, 74

[46] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2017. 71, 75

[47] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 612–618, 2017. 51

[48] Derek Bradley and Gerhard Roth. Adaptive thresholding using the integral image. *Journal of Graphics Tools*, 12(2):13–21, 2007. 127

[49] Rodrigo Bresan, Allan Pinto, Anderson Rocha, Carlos Beluzo, and Tiago Carvalho. Face-spoof buster: a presentation attack detector based on intrinsic image properties and deep learning. *arXiv preprint arXiv:1902.02845*, 2019. XX, 71, 74

[50] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *Advances in Neural Information Processing Systems*, 2017. 118, 122

[51] Thomas Brunner, Frederik Diehl, and Alois Knoll. Copy and paste: A simple but effective initialization method for black-box adversarial attacks. *CVPR Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems*, 2019. 116, 118

[52] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K. Varshney, and Dawn Song. Anomalous instance detection in deep learning: A survey. *arXiv:2003.06979v1*, 2020. 115

[53] Rizhao Cai and Changsheng Chen. Learning deep forest with multi-scale local binary pattern features for face anti-spoofing. *arXiv:1910.03850*, 2018. 72

[54] Raffaele Cappelli, Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Fingerprint verification competition 2006. *Biometric Technology Today*, 15(7âĂŞ8):7 – 9, 2007. 33

[55] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 136, 140, 141, 148, 149, 155

[56] Nicholas Carlini and David Wagner. Magnet and efficient defenses against adversarial attacks are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017. 119, 121, 122, 149

[57] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. XVIII, 113, 115, 116, 117, 128, 136, 141, 147, 148, 149, 155, 161

[58] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? In *arXiv preprint arXiv:1912.07398*, 2019. 49

[59] Oya Çeliktutan, Bülent Sankur, and Ismail Avcibas. Blind identification of source cellphone model. *IEEE Transactions on Information Forensics and Security*, 3(3):553–566, Sept 2008. 14, 15, 36, 40, 47

[60] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 5

[61] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M Robertson, and Stan Z Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593, 2019. 56, 71, 73, 74

[62] Mo Chen, J. Fridrich, M. Goljan, and J. Lukas. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3(1):74–90, 2008. 14

[63] Mo Chen, Jessica Fridrich, and Miroslav Goljan. Digital imaging sensor identification (further study). In *Electronic Imaging*, pages 65050P–65050P. International Society for Optics and Photonics, 2007. 14

[64] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In *AAAI conference on Artificial Intelligence*, pages 10–17, 2018. 113, 115, 117, 128, 155, 156

[65] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 172

[66] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *IEEE International Conference of Biometrics Special Interest Group*, pages 1–7, 2012. 8, 68, 72

[67] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. Face recognition systems under spoofing attacks. In *Face Recognition Across the Imaging Spectrum*, pages 165–194. Springer, 2016. 50, 61

[68] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 110

[69] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 109

[70] Debbrota Paul Chowdhury, Sambit Bakshi, Pankaj Kumar Sa, and Banshidhar Majhi. Wavelet energy feature based source camera identification for ear biometric images. *Pattern Recognition Letters*, 130:139–147, 2020. 14, 17

[71] Kenneth T. Co and Luis MuÃśoz-GonzÃąlez amd Emil C. Lupu. Sensitivity of deep convolutional networks to Gabor noise. *ICML Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019. 116, 118

[72] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. 100

[73] Ryan Connaughton, Amanda Sgroi, Kevin Bowyer, and Patrick J Flynn. A multialgorithm analysis of three iris biometric sensors. *IEEE Transactions on Information Forensics and Security*, 7(3):919–931, 2012. 6

[74] Ryan Connaughton, Amanda Sgroi, Kevin Bowyer, and Patrick J Flynn. A multialgorithm analysis of three iris biometric sensors. *IEEE Transactions on Information Forensics and Security*, 7(3):919–931, June 2012. 11

[75] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 89

[76] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 86, 126, 154

[77] Davide Cozzolino, Diego Gragnaniello, and Luisa Verdoliva. Image forgery detection through residual-based local descriptors and block-matching. In *IEEE International Conference on Image Processing*, pages 5297–5301, 2014. 109

[78] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164, 2017. 109

[79] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. 14, 16

[80] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020. 82

[81] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017. 120, 121

[82] Madhura Datta and C.A. Murthy. Two dimensional synthetic face generation and verification using set estimation technique. *Computer Vision and Image Understanding*, 116(9):1022 – 1031, 2012. 82

[83] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp-top based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pages 121–132. Springer, 2012. 62

[84] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *International Conference on Biometrics*, pages 1–8, June 2013. 74, 75

[85] Tiago de Freitas Pereira, Jukka Komulainen, André Anjos, José Mario De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1):1–15, 2014. 72

[86] Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*, 2020. 82

[87] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 8, 66, 129, 141, 152, 156

[88] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 77

[89] Mihal Dobeš, Jan Martinek, Dalibor Skoupil, Zdena Dobešová, and Jaroslav Pospíšil. Human eye localization using the modified hough transform. *Optik - International Journal for Light and Electron Optics*, 117(10):468 – 473, 2006. 31, 32

[90] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 120, 121

[91] Taiamiti Edmunds and Alice Caplier. Motion-based countermeasure against photo and video spoofing attacks in face recognition. *Journal of Visual Communication and Image Representation*, 50:314 – 332, 2018. 71, 74

[92] Susan El-Naggar and Arun Ross. Which dataset is this iris image from? In *IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2015. 14, 16

[93] Nesli Erdogmus and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2013. 8, 68, 69

[94] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 23

[95] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017. XVIII, 119, 120, 131, 132, 134, 135, 136, 138, 140, 150, 157, 160, 162, 164

[96] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication and Image Representation*, 38:451–460, 2016. 75

[97] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014. 78, 80

[98] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*, pages 195–222. Springer International Publishing, 2016. 78

[99] Andrew P Founds, Nick Orlans, Whiddon Genevieve, and Craig I Watson. Nist special databse 32-multiple encounter dataset ii (MEDS-ii). *NIST Interagency/Internal Report (NISTIR)-7807*, 2011. 129, 156, 161

[100] James E. Fowler. The Redundant Discrete Wavelet Transform and Additive Noise. *IEEE Signal Processing Letters*, 12(9):629–632, 2005. 18, 57, 59, 97

[101] Matthew Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015. 2

[102] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *{USENIX} Security Symposium ({USENIX} Security)*, pages 17–32, 2014. 2

[103] David Freire-Obregón, Fabio Narducci, Silvio Barra, and Modesto Castrillón-Santana. Deep learning for source camera identification on mobile devices. *Pattern Recognition Letters*, 126:86–91, 2019. 14, 17

[104] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 109

[105] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014. 95

[106] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, 2014. 72

[107] Javier Galbally, Arun Ross, Marta Gomez-Barrero, Julian Fierrez, and Javier Ortega-Garcia. Iris image reconstruction from binary templates: An efficient probabilistic approach based on genetic algorithms. *Computer Vision and Image Understanding*, 117(10):1512 – 1525, 2013. 82

[108] Javier Galbally and Riccardo Satta. Biometric sensor interoperability: A case study in 3d face recognition. In *International Conference on Pattern Recognition Applications and Methods*, pages 199–204, 2016. 11

[109] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 2019. 51

[110] Partha Ghosh, Arpan Losalka, and Michael J Black. Resisting adversarial attacks using gaussian mixture variational autoencoders. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 541–548, 2019. 121

[111] Soumyadeep Ghosh, Richa Singh, and Mayank Vatsa. Subclass heterogeneity aware loss for cross-spectral cross-resolution face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(3):245–256, 2020. 49, 77

[112] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. DeepRing: Protecting deep neural network with blockchain. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2019. 6, 173

[113] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Securing CNN model and biometric template using blockchain. *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2019. 6, 173

[114] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. DNDNet: Reconfiguring CNN for adversarial robustness. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2020. 173

[115] Akhil Goel, Anirudh Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2018. 144

[116] Miroslav Goljan and Jessica Fridrich. Camera identification from cropped and scaled images. In *Electronic Imaging*, pages 68190E–68190E. International Society for Optics and Photonics, 2008. 14, 15

[117] Marta Gomez-Barrero, Christian Rathgeb, Ulrich Scherhag, and Christoph Busch. Is your biometric system robust to morphing attacks? In *IEEEInternational Workshop on Biometrics and Forensics*, pages 1–6, 2017. 81

[118] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017. 149

[119] Ian Goodfellow. Defense against the dark arts: An overview of adversarial example security research and future research directions. *arXiv preprint arXiv:1806.04169*, 2018. 115

[120] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of ACM*, 61(7):56–66, 2018. 115

[121] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 107

[122] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015. 78, 113, 116, 155

[123] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 127(6-7):719–742, 2019. doi: 10.1007/s11263-019-01160-w. XVIII, 119, 121, 131, 132, 134, 135, 140, 143, 144, 150, 151, 157, 159, 160, 161, 163

[124] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. *AAAI Conference on Artificial Intelligence*, pages 6829–6836, 2018. 78, 113, 119, 120, 121, 122, 132, 136, 140, 143, 144, 151

[125] Diego Gragnaniello, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Local contrast phase descriptor for fingerprint liveness detection. *Pattern Recognition*, 48(4):1050 – 1058, 2015. 52

[126] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011. 8, 35

[127] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. 8, 35, 129, 148, 156

[128] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 119, 149

[129] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014. 121

[130] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 3

[131] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 120, 121, 122, 125, 149

[132] Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805, 2019. 49

[133] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010. 97, 99

[134] Aparecido Nilceu Marana Gustavo Botelho de Souza, JoÃčo Paulo Papa. On the learning of deep local features for robust face spoofing detection. *arXiv preprint arXiv:1806.07492v1*, 2018. 71

[135] Jiangfan Han, Xiaoyi Dong, Ruimao Zhang, Dongdong Chen, Weiming Zhang, Nenghai Yu, Ping Luo, and Xiaogang Wang. Once a man: Towards multi-target attack via learning multi-target adversarial network once. *IEEE International Conference on Computer Vision*, 2019. 116, 118

[136] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973. 20, 58, 97, 151

[137] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 66, 102, 109, 148, 155

[138] Peisong He, Haoliang Li, and Hongxia Wang. Detection of fake images via the ensemble of deep representations from multi color spaces. In *IEEE International Conference on Image Processing*, pages 2299–2303, 2019. 82

[139] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, pages 1–12, 2017. XVIII, 150, 157, 160, 162, 163, 164

[140] G. Heusch, A. George, D. GeissbÃijhler, Z. Mostaani, and S. Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 1–1, 2020. 51

[141] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. 113

[142] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 157

[143] Hossein Hosseini, Yize Chen, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Blocking transferability of adversarial examples in black-box learning systems. *arXiv preprint arXiv:1703.04318*, 2017. 119

[144] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 132, 152, 158, 163

[145] Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, volume 39. Springer Science & Business Media, 2009. 100

[146] Asim Iqbal, Romesa Khan, and Theofanis Karayannis. Developing a brain atlas through deep learning. *Nature Machine Intelligence*, pages 277–287, 2019. 113

[147] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007. 11

[148] Anubhav Jain, Puspita Majumdar, Richa Singh, and Mayank Vatsa. Detecting GANs and retouching based digital alterations via DAD-HCNN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 672–673, 2020. 110

[149] Anubhav Jain, Richa Singh, and Mayank Vatsa. On detecting GANs and retouching based synthetic alterations. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2018. 109, 110, 111

[150] Shan Jia, Guodong Guo, Zhengquan Xu, and Qiangchang Wang. Face presentation attack detection in mobile scenarios: A comprehensive evaluation. *Image and Vision Computing*, 93:103826, 2020. 51, 56

[151] Shan Jia, Chuanbo Hu, Guodong Guo, and Zhengquan Xu. A database for face presentation attack using wax figure faces. In *International Conference on Image Analysis and Processing*, pages 39–47. Springer, 2019. XX, 68, 69, 72, 76

[152] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020. 82

[153] Qingyue Jin, Yizhen Huang, and Na Fan. Learning images using compositional pattern-producing neural networks for source camera identification and digital demographic diagnosis. *Pattern Recognition Letters*, 33(4):381–396, 2012. 14, 15

[154] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *European Conference on Computer Vision*, pages 290–306, 2018. 74

[155] Oswaldo Ludwig Junior, David Delgado, Valter Gonçalves, and Urbano Nunes. Trainable classifier-fusion schemes: An application to pedestrian detection. In *IEEE International Conference on Intelligent Transportation Systems*, pages 1–6, Oct 2009. 40, 43, 44, 47

[156] Nathan Kalka, Nick Bartlow, Bojan Cukic, and Arun Ross. A preliminary study on identifying sensors from iris images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 50–56, 2015. 14, 16, 36, 40

[157] Juho Kannala and Esa Rahtu. BSIF: Binarized Statistical Image Features. In *International Conference on Pattern Recognition*, pages 1363–1366, 2012. 97, 99, 100, 102

[158] Nitin Khanna, Aravind K Mikkilineni, George TC Chiu, Jan P Allebach, and Edward J Delp. Forensic classification of imaging sensor types. In *Electronic Imaging*, pages 65050U–65050U. International Society for Optics and Photonics, 2007. 13, 14

[159] Nitin Khanna, Aravind K. Mikkilineni, and Edward J. Delp. Forensic camera classification: Verification of sensor pattern noise approach. *Forensic Science Communications*, 11(1):1–10, 2009. 14, 15

[160] Mehdi Kharrazi, Husrev T Sencar, and Nasir Memon. Blind source camera identification. In *IEEE International Conference on Image Processing*, volume 1, pages 709–712, 2004. 13, 14

[161] Jukka Komulainen, Abdenour Hadid, Matti Pietikäinen, André Anjos, and Sébastien Marcel. Complementary countermeasures for detecting scenic face spoofing attacks. In *International Conference on Biometrics*, pages 1–7, 2013. 72

[162] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 106

[163] Neslihan Kose and Jean-Luc Dugelay. Countermeasure for the protection of face recognition systems against mask attacks. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–6, 2013. 62

[164] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009. XVIII, 129, 156, 161

[165] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 135

[166] Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2589–2597, 2020. 97, 102

[167] Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2589–2597, 2020. 109

[168] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 113, 116, 117, 128, 155

[169] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 141

[170] Hyun Kwon, Yongchul Kim, Ki-Woong Park, Hyunsoo Yoon, and Daeseon Choi. Multi-targeted adversarial example in evasion attack on deep neural network. *IEEE Access*, 6:46084–46096, 2018. 116

[171] Nagashri N Lakshminarayana, Neeti Narayan, Nils Napp, Srirangaraj Setlur, and Venu Govindaraju. A discriminative spatio-temporal mapping of face for liveness detection. In *IEEE International Conference on Identity, Security and Behavior Analysis*, pages 1–7, 2017. 71

[172] Peter Langenberg, Emilio Balda, Arash Behboodi, and Rudolf Mathar. On the robustness of support vector machines against adversarial examples. In *IEEE International Conference on Signal Processing and Communication Systems*, pages 1–6, 2019. 152

[173] Ashref Lawgaly, Fouad Khelifi, and Ahmed Bouridane. Weighted averaging-based sensor pattern noise estimation for source camera identification. In *IEEE International Conference on Image Processing*, pages 5357–5361, 2014. 14, 15

[174] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 156

[175] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 110

[176] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, 2017. 119

[177] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Conference on Neural Information Processing Systems*, pages 7167–7177, 2018. XVIII, 119, 150, 158, 160, 161

[178] Ajian Li, Zichang Tan, Xuan Li, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing. *arXiv preprint arXiv:2003.05136*, 2020. 51

[179] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018. 71, 76

[180] Ruizhe Li, Yu Guan, and Chang-Tsun Li. PCA-based denoising of sensor pattern noise for source camera identification. In *IEEE China Summit & International Conference on Signal and Information Processing*, pages 436–440, 2014. 14, 16, 36, 40, 43, 44

[181] Stan Li, Dong Yi, Zhen Lei, and Shengcai Liao. The CASIA NIR-VIS 2.0 face database. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013. XX, 54, 66

[182] Stan Z Li, RuFeng Chu, ShengCai Liao, and Lun Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007. 54

[183] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *ACM International Conference on Multimedia*, pages 1864–1872, 2020. 97, 102

[184] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *IEEE International Conference on Computer Vision*, pages 5764–5772, 2017. 119, 148, 152

[185] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *arXiv:2007.08745*, 2020. 172

[186] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 83

[187] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 82

[188] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. Detecting adversarial image examples in deep networks with adaptive noise reduction. *arXiv preprint arXiv:1705.08378*, 2017. XVIII, 121, 131, 132, 134, 135, 136, 138, 140, 150, 157, 160, 162, 164

[189] Shiyu Liang, Yixuan Li, and R Srikant. Principled detection of out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2018. XVIII, 131, 132, 134, 135, 136, 138, 150, 157, 160, 161, 162, 163, 164

[190] Chen Lin, Zhouyingcheng Liao, Peng Zhou, Jianguo Hu, and Bingbing Ni. Live face verification with multiple instantialized local homographic parameterization. In *International Joint Conference on Artificial Intelligence*, pages 814–820, 2018. 71

[191] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. Detection based defense against adversarial examples from the steganalysis point of view. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4825–4834, 2019. XVIII, 120, 131, 133, 134, 135, 136, 138, 140, 141, 150, 157, 160, 161, 163, 164

[192] Jun Liu and Ajay Kumar. Detecting presentation attacks from 3d face masks under multispectral imaging. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–52, 2018. 50, 51

[193] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018. 51

[194] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 49, 51

[195] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020. 82

[196] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. Available: `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`. 110

[197] Shao-Yuan Lo and Vishal M Patel. Defending against multiple and unforeseen adversarial videos. *arXiv preprint arXiv:2009.05244*, 2020. 149

[198] Shao-Yuan Lo, Jeya Maria Jose Valanarasu, and Vishal M Patel. Overcomplete representations against adversarial videos. *arXiv preprint arXiv:2012.04262*, 2020. 149

[199] Giulio Lovisotto, Simon Eberz, and Ivan Martinovic. Biometric backdoors: A poisoning attack against unsupervised template updating. *arXiv preprint arXiv:1905.09162*, 2019. 3

[200] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *International Conference on Computer Vision*, pages 446–454, 2017. 118, 119

[201] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017. 120

[202] Oswaldo Ludwig and Urbano Nunes. Novel maximum-margin training algorithms for supervised neural networks. *IEEE Transactions on Neural Networks*, 21(6):972–984, June 2010. 26, 42

[203] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. Determining digital image origin using sensor imperfections. In *Electronic Imaging*, pages 249–260. International Society for Optics and Photonics, 2005. 13, 14, 15

[204] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006. 13, 14

[205] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015. 121

[206] Brendon Lutnick, Brandon Ginley, Darshana Govind, Sean D. McGarry, Peter S. LaViolette, Rabi Yacoub, Sanjay Jain, John E. Tomaszewski, Kuang-Yu Jen, and Pinaki Sarder. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nature Machine Intelligence*, pages 112–119, 2019. 113

[207] Shiqing Ma and Yingqi Liu. NIC: Detecting adversarial samples with neural network invariant checking. In *Network and Distributed System Security Symposium*, 2019. 120, 131, 133

[208] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018. XVIII, 119, 150, 158, 160, 161

[209] Yukun Ma, Lifang Wu, Zeyu Li, et al. A novel face presentation attack detection scheme based on multi-regional convolutional neural networks. *Pattern Recognition Letters*, 131:261–267, 2020. 56, 71, 76

[210] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008. 166

[211] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2011. 56, 62, 96

[212] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2011. 99, 102

[213] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representation*, 2018. 155, 167

[214] Dario Maio, Davide Maltoni, Raffaele Cappelli, James L Wayman, and Anil K Jain. Fvc2002: Second fingerprint verification competition. In *International Conference on Pattern Recognition*, volume 3, pages 811–814 vol.3, 2002. 33

[215] Puspita Majumdar, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Evading face recognition via partial tampering of faces. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–20, 2019. 106

[216] Andrey Makrushin, Christian Kraetzer, Tom Neubert, and Jana Dittmann. Generalized benford's law for blind detection of morphed face images. In *ACM Workshop on Information Hiding and Multimedia Security*, pages 49–54, 2018. 81

[217] Ishan Manjani, Snigdha Tariyal, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE Transactions on Information Forensics and Security*, 12(7):1713–1723, 2017. 68, 69, 70, 71, 72, 73, 76

[218] Sébastien Marcel, Mark S Nixon, Julian Fierrez, and Nicholas Evans. *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019. 3, 5, 49

[219] Sébastien Marcel, Mark S Nixon, Julian Fierrez, and Nicholas Evans. Handbook of biometric anti-spoofing presentation attack detection. 2019. 51

[220] Francesco Marra, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. A deep learning approach for iris sensor model identification. *Pattern Recognition Letters*, 113:46–53, 2018. 6

[221] Daniel Mas Montserrat, Hanxiang Hao, Sri K Yarlagadda, Sriram Baireddy, Ruiting Shao, Janos Horvath, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, et al. Deepfakes detection with automatic face weighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 668–669, 2020. 82

[222] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops*, pages 83–92, 2019. 82

[223] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. 105

[224] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. Face deidentification with generative deep neural networks. *IET Signal Processing*, 11(9):1046–1054, 2017. 77

[225] Suril Mehta, Anannya Uberoi, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Crafting a panoptic face presentation attack detector. In *IEEE International Conference on Biometrics*, pages 1–6, 2019. 49

[226] Alfred J Menezes, Jonathan Katz, Paul C Van Oorschot, and Scott A Vanstone. *Handbook of applied cryptography*. CRC press, 1996. 6

[227] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *ACM Conference on Computer and Communications Security*, pages 135–147, 2017. 119, 120, 122, 149

[228] David Menotti, Giovani Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falcao, and Anderson Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4):864–879, 2015. 71, 75

[229] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 119, 149

[230] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015. 113

[231] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. XVIII, 113, 115, 116, 117, 128, 155, 161

[232] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. XVIII, 116, 118, 155, 160, 161

[233] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. *arXiv preprint arXiv:1707.05572*, 2017. 116, 117, 118, 128, 135

[234] Priscilla BL Mukudi and Peter J Hills. The combined influence of the own-age,-gender, and-ethnicity biases on face recognition. *Acta psychologica*, 194:1–6, 2019. 49

[235] Tom Neubert. Face morphing detection: An approach based on image degradation analysis. In *Springer International Workshop on Digital Watermarking*, pages 93–106, 2017. 81

[236] Joao C Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020. 83

[237] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, 2019. 82

[238] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2307–2311, 2019. 82

[239] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):971–987, 2002. 97, 99, 166

[240] Ville Ojansivu and Janne Heikkilä. Blur insensitive texture classification using local phase quantization. In *International Conference on Image and Signal Processing*, pages 236–243. Springer, 2008. 97, 99, 102

[241] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 125

[242] Shi Pan and Farzin Deravi. Facial biometrie presentation attack detection using temporal texture co-occurrence. In *IEEE International Conference on Identity, Security, and Behavior Analysis*, pages 1–7, 2018. 71

[243] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy*, pages 372–387, 2016. 113, 117

[244] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597, 2016. 117, 121, 122, 149

[245] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 106, 144

[246] Kevin M Passino. Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Systems*, 22(3):52–67, Jun 2002. 26

[247] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016. 56, 68, 69, 71, 72, 75

[248] Ioannis Pavlidis and Peter Symosek. The imaging issue in an automatic face/disguise detection system. In *IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, pages 15–24, 2000. 50

[249] Fei Peng, Le Qin, and Min Long. Face presentation attack detection using guided scale texture. *Multimedia Tools and Applications*, pages 1–27, 2017. 71, 72, 73, 74

[250] Fei Peng, Le Qin, and Min Long. Face presentation attack detection based on chromatic co-occurrence of local binary pattern and ensemble learning. *Journal of Visual Communication and Image Representation*, 66:102746, 2020. 71, 73, 74

[251] P Jonathon Phillips, Patrick J Flynn, J Ross Beveridge, W Todd Scruggs, Alice J OâĂŹtoole, David Bolme, Kevin W Bowyer, Bruce A Draper, Geof H Givens, Yui Man Lui, et al. Overview of the multiple biometrics grand challenge. In *International Conference on Biometrics*, pages 705–714. Springer, 2009. 129

[252] Jaishanker K Pillai, Maria Puertas, and Rama Chellappa. Cross-sensor iris recognition through kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):73–85, 2013. 6

[253] Allan Pinto, Siome Goldenstein, Alexandre Ferreira, Tiago Carvalho, Helio Pedrini, and Anderson Rocha. Leveraging shape, reflectance and albedo from shading for face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 15:3347–3358, 2020. 71, 76

[254] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing*, 24(12):4726–4740, 2015. 71

[255] Allan Pinto, William Robson Schwartz, Helio Pedrini, and Anderson de Rezende Rocha.

Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Transactions on Information Forensics and Security*, 10(5):1025–1038, 2015. 74

[256] Norman Poh, Josef Kittler, Sebastien Marcel, Driss Matrouf, and Jean-Francois Bonastre. Model and score adaptation for biometric systems: Coping with device interoperability and changing acquisition conditions. In *IEEE International Conference on Pattern Recognition*, pages 1229–1232, 2010. 6

[257] Jouni Pohjalainen, Okko Rasanen, and Serdar Kadioglu. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language*, 29(1):145 – 171, 2015. 26, 41

[258] Vidyasagar M Potdar, Song Han, and Elizabeth Chang. A survey of digital image watermarking techniques. In *IEEE International Conference on Industrial Informatics*, pages 709–716, 2005. 114

[259] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 116

[260] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018. 120, 121

[261] Hugo Proença and Luís A Alexandre. UBIRIS: A noisy iris image database. In *International Conference on Image Analysis and Processing*, pages 970–977. 2005. 31, 32

[262] Hugo Proenca, Silvio Filipe, Ricardo Santos, Joao Oliveira, and Luis A Alexandre. The UBIRIS.v2: A Database of Visible Wavelength Iris Images Captured On-the-Move and At-a-Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535, Aug 2010. 31, 32

[263] Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994. 42

[264] Le Qin, Le-Bing Zhang, Fei Peng, and Min Long. Content-independent face presentation attack detection with directional local binary pattern. In *Biometric Recognition*, pages 118–126. Springer International Publishing, 2017. 71, 74

[265] Ramachandra Raghavendra and Christoph Busch. Robust scheme for iris presentation attack detection using multiscale binarized statistical image features. *IEEE Transactions on Information Forensics and Security*, 10(4):703–715, 2015. 62

[266] Ramachandra Raghavendra, Kiran B Raja, Sébastien Marcel, and Christoph Busch. Face presentation attack detection across spectrum using time-frequency descriptors of maximal response in laplacian scale-space. In *International Conference on Image Processing Theory, Tools and Applications*, pages 1–6, 2016. XX, 51, 61, 62

[267] Ramachandra Raghavendra, Kiran B Raja, Sushma Venkatesh, Faouzi Alaya Cheikh, and Christoph Busch. On the vulnerability of extended multispectral face recognition systems towards presentation attacks. In *IEEE International Conference on Identity, Security and Behavior Analysis*, pages 1–8, 2017. 50, 51

[268] Ramachandra Raghavendra, KiranB Raja, and Christoph Busch. Detecting morphed face images. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2016. 80, 91

[269] Ramachandra Raghavendra, KiranB Raja, Sushma Venkatesh, and Christoph Busch. Face morphing versus face averaging: Vulnerability and detection. In *IEEE International Joint Conference on Biometrics*, pages 555–563, 2017. 81

[270] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *International Conference on Learning Representations*, 2018. 120, 149

[271] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *IEEE Workshop on Information Forensics and Security*, pages 1–6, 2017. 109

[272] Kiran Raja, Sushma Venkatesh, RB Christoph Busch, et al. Transferable deep-cnn features

for detecting digital and print-scanned morphed face images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1822–1830, 2017. 82, 109, 110

[273] Raghavendra Ramachandra and Christoph Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys*, 50(1):8, 2017. 56

[274] Francesco Ranzato and Marco Zanella. Robustness verification of support vector machines. In *International Static Analysis Symposium*, pages 271–295. Springer, 2019. 152

[275] Okko Rasanen and Jouni Pohjalainen. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In *INTERSPEECH*, pages 210–214, 2013. 26, 41

[276] Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM systems Journal*, 40(3):614–634, 2001. 1

[277] Manish Vuyyuru Reddy, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *Advances in Neural Information Processing Systems*, 33, 2020. 122

[278] P Venkata Reddy, Ajay Kumar, SMK Rahman, and Tanvir Singh Mundra. A new antispoofing approach for biometric devices. *IEEE Transactions on Biomedical Circuits and Systems*, 2(4):328–337, 2008. 49

[279] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020. 115

[280] David J Robertson, Robin SS Kramer, and A Mike Burton. Fraudulent id using face morphs: Experiments on human and automatic recognition. *PloS One*, 12(3):e0173319, 2017. 81

[281] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI conference on Artificial Intelligence*, pages 1660–1669, 2018. 122, 149

[282] Arun Ross and Anil Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115 – 2125, 2003. 150

[283] Arun Ross and Anil Jain. Biometric sensor interoperability: A case study in fingerprints. In *Biometric authentication*, pages 134–145. Springer, 2004. 11

[284] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 109, 110

[285] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE International Conference on Computer Vision*, pages 1–11, 2019. 97, 102

[286] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. *International Conference on Learning Representations*, 2016. 172

[287] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. *International Conference on Learning Representations*, 2018. 119, 167

[288] Stamatis Samaras, Vasilis Mygdalis, and Ioannis Pitas. Robustness in blind camera identification. In *International Conference on Pattern Recognition*, pages 3874–3879, Dec 2016. 14

[289] Tiomotheos Samatzidis, Dirk Siegmund, Michael Goedde, Naser Damer, Andreas Braun, and Arjan Kuijper. The dark side of the face: exploring the ultraviolet spectrum for face biometrics. In *International Conference on Biometrics*, pages 20–23, 2018. 52

[290] Anush Sankaran, Aayush Jain, Tarun Vashisth, Mayank Vatsa, and Richa Singh. Adaptive latent fingerprint segmentation using feature selection and random decision forest classification. *Information Fusion*, 34:1 – 15, 2017. 25

[291] Anush Sankaran, Mayank Vatsa, and Richa Singh. Multisensor optical and latent fingerprint database. *IEEE Access*, 3:653–665, 2015. 33

[292] Riccardo Satta. Sensor pattern noise matching based on reliability map for source camera identification. In *International Conference on Computer Vision Theory and Applications*, pages 222–226, 2015. 14, 16

[293] Dhanesh Scherhag, Ulrich Budhrani, Marta Gomez-Barrero, and Christoph Busch. Detecting morphed face images using facial landmarks. In *International Conference on Image and Signal Processing*, pages 444–452, 2018. 82

[294] Ulrich Scherhag, R Raghavendra, Kiran B Raja, Marta Gomez-Barrero, Christian Rathgeb, and Christoph Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *International Workshop on Biometrics and Forensics*, pages 1–6, 2017. 81, 91

[295] Ulrich Scherhag, Christian Rathgeb, and Christoph Busch. Towards detection of morphed face images in electronic travel documents. In *IAPR International Workshop on Document Analysis Systems*, pages 187–192, 2018. 82

[296] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019. 5

[297] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019. 83, 97

[298] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. Detection of face morphing attacks by deep learning. In *Digital Forensics and Watermarking*, pages 107–120. Springer International Publishing, 2017. 82

[299] Alireza Sepas-Moghaddam, Fernando M Pereira, and Paulo Lobato Correia. Face recognition: A novel multi-level taxonomy based survey. *IET Biometrics*, 9(2):58–67, 2019. 49

[300] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on

neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018. 172

[301] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 14(4):923–938, 2019. 70, 71

[302] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense. *European Conference on Computer Vision*, 2020. 122

[303] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM Conference on Computer and Communications Security*, pages 1528–1540, 2016. 2, 118

[304] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security*, 22(3):16:1–16:30, 2019. 116

[305] C Sidney Burrus, Ramesh A Gopinath, and Haitao Guo. Introduction to wavelets and wavelet transforms. *A Primer; Prentice Hall: Upper Saddle River, NJ, USA*, 1998. 18

[306] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 97, 102, 109, 132, 148, 155, 158

[307] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018. 120, 149

[308] Xiao Song, Xu Zhao, Liangji Fang, and Tianwei Lin. Discriminative representation combinations for accurate face spoofing detection. *Pattern Recognition*, 85:220–231, 2019. 75

[309] Xiaoning Song, Qiqun Wu, Dongjun Yu, Guosheng Hu, and Xiaojun Wu. Face anti-spoofing detection using least square weight fusion of channel-based feature classifiers. Technical report, EasyChair, 2020. 71

[310] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *International Conference on Learning Representations*, 2017. 122, 149

[311] Leonidas Spinoulas, Mohamed Hussein, David Geissbühler, Joe Mathai, Oswin G Almeida, Guillaume Clivaz, Sébastien Marcel, and Wael AbdAlmageed. Multispectral biometrics system framework: Application to presentation attack detection. *arXiv preprint arXiv:2006.07489*, 2020. 52

[312] Kenneth O. Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, pages 24–35, 2019. 113

[313] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*, pages 3517–3529, 2017. 172

[314] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 116, 117

[315] Wenyun Sun, Yu Song, Changsheng Chen, Jiwu Huang, and Alex C. Kot. Face spoofing detection based on local ternary label supervision in fully convolutional networks. *IEEE Transactions on Information Forensics and Security*, 15:3181–3196, 2020. 74

[316] Wenyun Sun, Yu Song, Haitao Zhao, and Zhong Jin. A face spoofing detection method based on domain adaptation and lossless size adaptation. *IEEE Access*, 8:66553–66563, 2020. 73, 74

[317] Xudong Sun, Lei Huang, and Changping Liu. Multispectral face spoofing detection using vis–nir imaging correlation. *International Journal of Wavelets, Multiresolution and Information Processing*, 16(02):1840003, 2018. 51

[318] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 85, 86, 97, 102, 135, 155

[319] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014. 78, 113, 115, 116, 147

[320] Bozhao Tan and Stephanie Schuckers. Spoofing protection for fingerprint scanner by fusing ridge signal and valley noise. *Pattern Recognition*, 43(8):2845 – 2857, 2010. 51

[321] Feng Tao and Yongcan Cao. Resilient learning of computational models with noisy labels. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–10, 2019. 149

[322] Olga Taran, Shideh Rezaeifar, Taras Holotyak, and Slava Voloshynovskiy. Defending against adversarial attacks by randomized diversification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11226–11233, 2019. 119

[323] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaz Bortolato, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 2020. 2

[324] Thanh Hai Thai, R. Cogranne, and F. Retraint. Camera model identification based on the heteroscedastic noise model. *IEEE Transactions on Image Processing*, 23(1):250–263, Jan 2014. 14, 15

[325] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 109

[326] Shixin Tian, Guolei Yang, and Ying Cai. Detecting adversarial examples through image transformation. In *AAAI Conference on Artificial Intelligence*, pages 4139–4146, 2018. 120

[327] Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony TS Ho. Detection of face spoofing using visual dynamics. *IEEE Transactions on Information Forensics and Security*, 10(4):762–777, 2015. 71

[328] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020. 83

[329] Yoichi Tomioka, Yuya Ito, and Hitoshi Kitazawa. Robust digital camera identification based on pairwise magnitude relations of clustered sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 8(12):1986–1995, 2013. 14, 15

[330] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations*, 2018. 119, 130, 141, 149

[331] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *{USENIX} Security Symposium ({USENIX} Security)*, pages 601–618, 2016. 2

[332] Xiaoguang Tu, Jian Zhao, Mei Xie, Guodong Du, Hengsheng Zhang, Jianshu Li, Zheng Ma, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *arXiv preprint arXiv:1901.05602*, 2019. 56, 71

[333] Xiaokang Tu and Yuchun Fang. Ultra-deep neural network for face anti-spoofing. In *International Conference on Neural Information Processing*, pages 686–695. Springer, 2017. 71, 72

[334] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995. 25, 58

[335] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 54, 89

[336] Lukasz Wandzik, Raul Vicente Garcia, Gerald Kaeding, and Xi Chen. Cnns under attack: On the vulnerability of deep neural networks based face recognition to image morphing. In *Digital Forensics and Watermarking*, pages 121–135. Springer International Publishing, 2017. 82

[337] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, pages 707–723, 2019. 3

[338] Haohan Wang, Xindi Wu, Pengcheng Yin, and Eric P Xing. High frequency component helps explain the generalization of convolutional neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 122, 143, 144

[339] Haotao N Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, and Zhangyang Wang. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. *Advances in Neural Information Processing Systems*, 33, 2020. 149

[340] Ruoying Wang, Kexin Nie, Tie Wang, Yang Yang, and Bo Long. Deep learning for anomaly detection. In *International Conference on Web Search and Data Mining*, pages 894–896, 2020. 5

[341] Shun-Yi Wang, Shih-Hung Yang, Yon-Ping Chen, and Jyun-We Huang. Face liveness detection based on skin blood flow analysis. *Symmetry*, 9(12), 2017. 71

[342] Xiao Wang, Siyue Wang, Pin-Yu Chen, Yanzhi Wang, Brian Kulis, Xue Lin, and Peter Chin. Protecting neural networks with hierarchical random switching: Towards better robustness-accuracy trade-off for stochastic defenses. *International Joint Conference on Artificial Intelligence*, 2019. 121

[343] Hong Wei, Lulu Chen, and James Ferryman. Biometrics in ABC: counter-spoofing research. *FRONTEX Global Conference on Future Developments of Automated Border Control*, pages 10–11, 2013. 62

[344] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 56, 68, 71, 72, 75

[345] A Wayne Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, 100(9):1100–1103, 1971. 26, 41

[346] Peter Wild, Petru Radu, Lulu Chen, and James Ferryman. Robust multimodal face and fingerprint fusion in the presence of spoofing attacks. *Pattern Recognition*, 50:17 – 25, 2016. 51

[347] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 100, 152

[348] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295, 2018. 120

[349] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 144

[350] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 9005–9012, 2019. 49

[351] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 118

[352] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *International Conference on Learning Representation*, 2018. 122, 125, 132, 143, 144, 149

[353] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision*, pages 1369–1378, 2017. 120

[354] Guanshuo Xu and Yun Qing Shi. Camera model identification using local binary patterns. In *IEEE ICME*, pages 392–397, 2012. 14, 15

[355] Daksha Yadav, Mayank Vatsa, Richa Singh, and Massimo Tistarelli. Bacteria foraging fusion for face recognition across age progression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013. 26

[356] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 72

[357] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *International Conference on Biometrics*, pages 1–6, 2013. 72, 75

[358] Pengpeng Yang, Rongrong Ni, Yao Zhao, and Wei Zhao. Source camera identification based on content-adaptive fusion residual networks. *Pattern Recognition Letters*, 119:195–204, 2019. 14, 16

[359] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265, 2019. 82

[360] Chun-Hsiao Yeh and Herng-Hua Chang. Face liveness detection based on perceptual image quality assessment features with multi-scale analysis. In *IEEE Winter Conference on Applications of Computer Vision*, pages 49–56, 2018. 71, 74

[361] Dong Yi, Zhen Lei, Zhiwei Zhang, and Stan Z Li. Face anti-spoofing: Multi-spectral approach. In *Handbook of Biometric Anti-Spoofing*, pages 83–102. Springer, 2014. 50

[362] Jihao Yin, Hui Li, and Xiuping Jia. Crater detection based on gist features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(1):23–29, 2015. 126

[363] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019. 5

[364] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019. 115, 132

[365] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *ACM Workshop on Artificial Intelligence and Security*, pages 39–49, 2017. 121

[366] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. 86

[367] Chiliang Zhang, Zhimou Yang, and Zuochang Ye. Detecting adversarial perturbations with saliency. *arXiv preprint arXiv:1803.08773*, 2018. 119

[368] Le-Bing Zhang, Fei Peng, Le Qin, and Min Long. Face spoofing detection based on color texture markov feature and support vector machine recursive feature elimination. *Journal of Visual Communication and Image Representation*, 51:56 – 69, 2018. 71, 74

[369] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 50, 51, 61, 65

[370] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019. XX, 64, 65

[371] Yuchen Zhang and Percy Liang. Defending against whitebox adversarial attacks via randomized discretization. In *International Conference on Artificial Intelligence and Statistics*, pages 684–693, 2019. 121

[372] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *IAPR International Conference on Biometrics*, pages 26–31, 2012. 68

[373] Zhixin Zhang, Xuhua Pan, Shuhao Jiang, and Peijun Zhao. High-quality face image generation based on generative adversarial networks. *Journal of Visual Communication and Image Representation*, 71:102719, 2020. 8

[374] Chenxiao Zhao, P Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. The adversarial attack and detection under the fisher information metric. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 5869–5876, 2019. 119

[375] Pu Zhao, Sijia Liu, Pin-Yu Chen, Nghia Hoang, Kaidi Xu, Bhavya Kailkhura, and Xue Lin. On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method. *IEEE International Conference on Computer Vision*, pages 121–130, 2019. 116, 118

[376] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1831–1839, 2017. 109, 110

[377] Fei Zuo and Qiang Zeng. Erase and restore: Simple, accurate and resilient detection of l_2 adversarial examples. *arXiv preprint arXiv:2001.00116*, 2020. 152