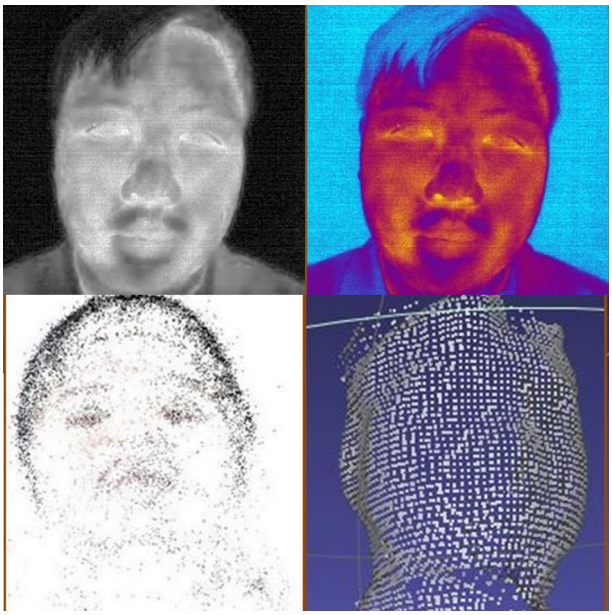
Face Recognition and Face Spoofing Detection Using 3D Model

Kim Trong Nguyen

Thèse
de doctorat
de l'UTT

Kim Trong NGUYEN

Face Recognition and Face Spoofing Detection Using 3D Model



Champ disciplinaire :
Sciences pour l'Ingénieur

2019TROY0012 **Année 2019**

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

THESE

pour l'obtention du grade de

DOCTEUR
de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

EN SCIENCES POUR L'INGENIEUR

Spécialité : OPTIMISATION ET SURETE DES SYSTEMES

présentée et soutenue par

Kim Trong NGUYEN

le 6 mai 2019

Face Recognition and Face Spoofing Detection Using 3D Model

JURY

M. B. EL HASSAN	PROFESSEUR	Président
M. C. CHARRIER	MAITRE DE CONFERENCES - HDR	Rapporteur
M. D. FOFI	PROFESSEUR DES UNIVERSITES	Rapporteur
M. F. RETRAINT	PROFESSEUR DES UNIVERSITES	Directeur de thèse
Mme C. ZITZMANN	ENSEIGNANTE CHERCHEURE	Directrice de thèse

Acknowledgements

This work has been carried out within the laboratory of Systems Modeling and Dependability (M2S) at University of Technology of Troyes (UTT). It is funded by Troyes Champagne Métropole through the Graduate School of Engineering of EPF.

This work has been accomplished under the supervision of Mr Florent RETRAINT and Miss Cathel ZITZMANN. I would like to express my deepest appreciation to them for their highly professional guidance and incessant support. Florent has accompanied me from my master's intership at UTT and encouraged me to continue the research in the field. They gave me a lot of ideas for my thesis and my work. I highly appreciate the friendly and professional environment they have created for me during my three-years doctoral.

I would like to express my special thanks to Mr Christophe CHARRIER and Mr David FOFI for accepting to review my PhD thesis.I would also like to thank Mr Bachar Ahmed ELHASSAN for agreeing to examine this thesis. Valuable remarks provided by the respectful experts in this field like them would improve the thesis quality.

Most of all, I would like to thank my parents, my sister and my girlfriend for always loving and supporting me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and problematic

Human faces are the most characteristic features which can be used to distinguish one person from the others. Recognizing the parents and the family is the very first lesson for each human being. Since the development of imagery technology, human brains are charged with another task: face recognition from a photo. This task is now the most critical identification solution in our society when our face appears in many papers such as passports, ID card, driver license, student card ... At this era of digital technology, the task of face recognition is more and more entrusted to automatic systems. The powerful computer is able to accomplish many complex tasks including people detection and authentication, movement tracking and predicting, illness detection and classification... Face recognition is now applied in a wide range of use-cases with different levels and constraints of security. There are passive systems using by the authority to inspect the activity in sensitive regions. There are active calibrated systems for access control in airports, companies ... and many other facilities. There are also many distributed uncalibrated systems like laptop unlock program, smartphone identification application which gives users more initiatives to decide the environmental conditions of their attempt.

However, visible face recognition is theoretically and practically vulnerable to face spoofing attack. A well-performing authentication system can be easily bypassed using a photo of a genius user's face presented in front of the system's camera. The threat is especially dangerous since many people let their pictures be public on the internet, in particular, on social networks. An intruder can find many high-quality photos without needing anything more than the user's name and exploits them to operate the attack. In order to strengthen the authentication process, the administration can add to the system a new security layer which can reduce this vulnerability as known as the liveliness detector or face spoofing detector. The existent solutions for such layer are wide large in terms of technology but almost all the precise methods require a complex system with two or more cameras and even other types of sensor. The more complex the system, the less applicable the solution, these methods are useless outside its use-cases. Therefore, in this study, we propose a new way for liveliness detection dedicated to simple mono-camera systems like

smartphones and tablets. This novel method exploits their movability to make use of not only one picture but a whole video of user's face in many poses in order to rebuild the face in 3D coordination which is used later to distinguish a genius face from the attack attempts.

Another crucial problem of visible face recognition is linked to the fact that all the light and color that can be observed from human faces is only a reflection of the light from other sources such as the sun, the lamps. Obviously, visible imagery is highly dependent on the illumination conditions. Some research has proposed methods that can function correctly across the change of light intensity, but there is always a decrease in terms of precision when the modification is so brutal. No visible face recognition method can be processed in the lack of light, but the necessity of authentication in darkness is not overrated. In this context, infrared and particularly thermal imagery has become a promising alternative and complement method for face recognition. However, until this day, thermal face recognition does not achieve the required mature level to be applied widely and distributively. In fact, the application of infrared images in face recognition process is challenged by the lack of distinguishable feature in these images. IR spectrum has its own problems which can affect the precision of the identification program. In order to deal with this problem, we aim to introduce a new method of thermal face recognition solution employing 3D models of the head which contains information of vascular network measured from a thermal video. The process is dedicated to functioning in different use-cases where the only required equipment is a single thermal camera.

1.2 Outline

This first chapter of the Thesis introduces, in general, the context and the problems that lead to this study. Il also describes the structure of this Thesis. The second chapter is dedicated to present an overview of the automatic face recognition domain which, in this study, is divided into two principal sections: Visible face recognition and thermal face recognition. In the first section of this chapter, the Thesis introduces the advantage and problematic of visible imagery. It also provides some representative methods from the beginning of computer vision to this day. Infrared technology is presented in the next section where we emphasize how it can bypass the problematic of visible imagery. But, as having been stated above, infrared technology is not void of its challenges which is also described in the second section.

The third chapter introduces a novel approach to model the head of users by its 3D features. The chapter starts by reviewing the advantage of representing users' face data using its 3D model. Then, the next section is dedicated to giving an overview of the 3D reconstruction method using different sensors. The last part of the chapter presents the details process to make this 3D model from a single video of users' face which is applied in two principal directions of this Thesis.

The fourth chapter constructs and examines a new method for the detection of one of the major problems in the visible face recognition domain: face-spoofing attack. It starts with a recall of face-spoofing attack and how it can affect the authentication process. The next section describes the detailed scheme to detect this attack. In the last part, we provide the performance of this method using the result of our experiment.

In the fifth chapter, we propose a novel face recognition solution using a 3D model of the head computed from a thermal video which contains information of vascular network. By its nature, the thermal imagery can obviously detect the face-spoofing attack and stay invariant to illuminant conditions. However, there is less distinguishable information in a thermal image than a visible image which reduces the precision of the face recognition process. The 3D reconstruction provides geometric data of the face which can be mixed with vascular network information from thermal imagery to improve its performance.

Chapter 2

Literature Review

Face recognition is a task so common to humans that the individual does not even notice the extensive number of times it is performed every day. Nowadays, face recognition has been studied as a specific case of object recognition. It has received special attention in recent years due to a great variety of applications such as robot-human interaction, control by gesture, surveillance, security, and people tracking. A lot of face modeling techniques and classification methods have appeared and have progressed in the last three decades. The very first section of this chapter is dedicated to introducing some features of visible face recognition with its advantages and problematics. Some representative methods of visible face recognition are also mentioned and compared with each other inside a brief overview. The second section analyzes the potential of applying infrared technology in face recognition to solve at the same time two major problems in visible imagery linked to illumination conditions and face spoofing attacks.

2.1 Visible Face Recognition

2.1.1 State of the art

Introduction

Authentication by biometric characteristics is the most important demand which makes evolute face recognition. This technique has become more and more popular and plays a key role in many security systems such as passport identification for some airports, access control for a lot of companies and even unblocking application for some smartphone.

The reason of this phenomenon can be explained by several advantages that face recognition holds over other biometric technique. Face recognition is above all, a very natural, very human-friendly process which scores the highest percentage of compatibility with machine-readable travel documents among the 6 biometric techniques (face, finger, hand, voice, eye, signature).

However, facial features are not considered as the most reliable biometric technique because of its limits in terms of performance in unconstrained environment [1]. Facial expression, various illumination conditions, face spoofing attack, twin

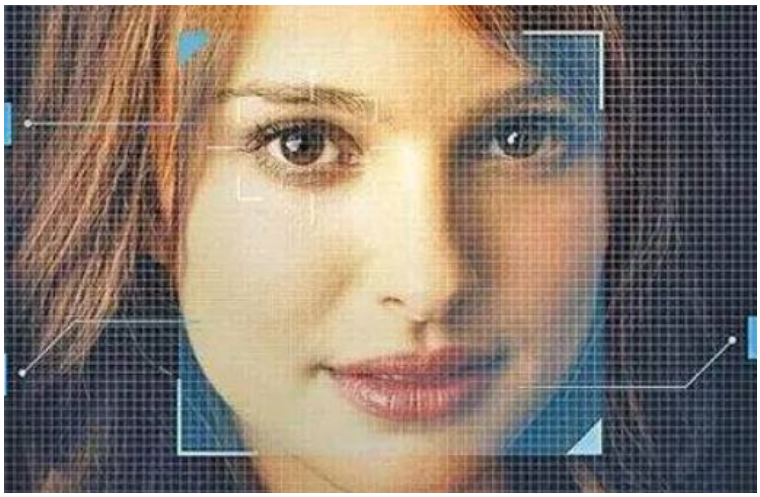


FIGURE 2.1: Face recognition example

faces problems, disguise or makeup, many data imperfection continues to challenge the computer vision community for further progress in face recognition method.

The following part is dedicated to highlighting some typical solutions for face-recognition challenge from the early state of this domain to now. However, visible imagery is always vulnerable to face spoofing attacks as we asserted this problem in the next subsection. A study of state of the art about face spoofing detection can be found in the third part of this section.

Methods

Eigenfaces Identifying the different face from an image is the primary purpose of face recognition. Little noise (unconstrained environmental conditions) exists in every photo, but the presence of these noises does not make the image totally random. There are some patterns which help in recognizing the different features of the image. A lot of patterns which can be seen in face recognition are in the neighborhood of the nose, eye, and mouth or the distance between the facial features. In facial recognition field, these characteristics are known as eigenfaces [2]. Most of these experiments are processed on the frontal view of the face. Principle component analysis (PCA) is a method for extracting these eigenfaces from an image. It is a vice-versa, that is if a system is including a set of eigenfaces then the original image of the face can be restored. This process is efficient and practical compared to others techniques in constrained conditions [3, 4, 2].

But this approach has a few weaknesses over unconstrained environment and behavior such as facial expressions in which there are some changes in facial feature and shape. Besides, the variety in pose heads to the distortion of the distance of the elements. Thirdly, the changes in the illumination conditions, for example, the bright light will make image saturate. But the illumination conditions can be overcome using the Fisher face method which is an enhancement of eigenfaces, but it uses the Fisher's linear discriminant analysis (LDA) [5, 6].

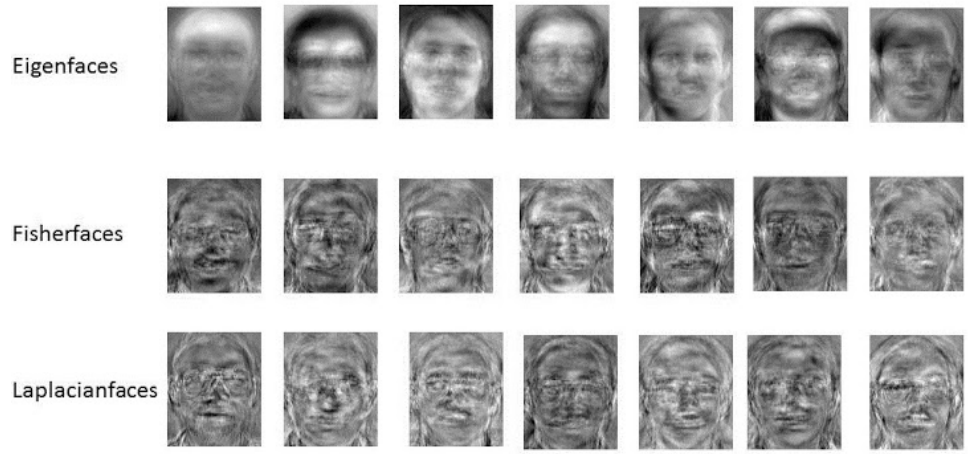


FIGURE 2.2: Face description using Eigenface and Fisherface

Neural Networks Neural network gives a nonlinear method to the face recognition demand. Its principal advantage over the linear techniques is that it reduces the rate of bad classification within the neighborhood classes. The first neural network solution used for face recognition is WISARD; it includes the dedicated network for each individual [7]. The development of a neural network is essential for face recognition. The training time of the technique takes about 4 hours and the classification time is up to 0.5 seconds.

There are multiple techniques of face recognition which employs the neural network approach for face authentication. PDBNN is a probabilistic decision-based neural network and it applies the idea of decision based neural network (DBNN) [8, 9]. The network approach is not entirely connected with this method. The network is split into K subnets; each subnet identifies one individual from the set. Its neurons use Gaussian activation function. The "face-subnet" is the summation of neuron outputs..

This method primarily consists of two stages. Firstly, the subnets are trained by their own face models, and after that, the subnet features are trained by some other appropriate samples from other face class, this stage is called as decision-based learning scheme. Only misclassified features are utilized by the decision-based learning scheme and not the whole of the training samples for the training. If the samples are classified to the wrong subnet, then the parameters of the legitimate subnet will be attuned, so that its decision region will be shifted closer to the misclassified sample. PDBNN classification has the benefits of both statistical methods and neural network techniques [8]. It is simple to implement its distributed computing process on parallel machines.

Geometrical feature matching The overall geometrical feature of a face is enough for identifying a person. The facial form is represented by the location and the dimension of the facial elements like nose, mouth, eyebrows, eye, and the face outline [10, 11, 12]. Mixture-distance-based method was proposed by I.J.Cox et.al [12],

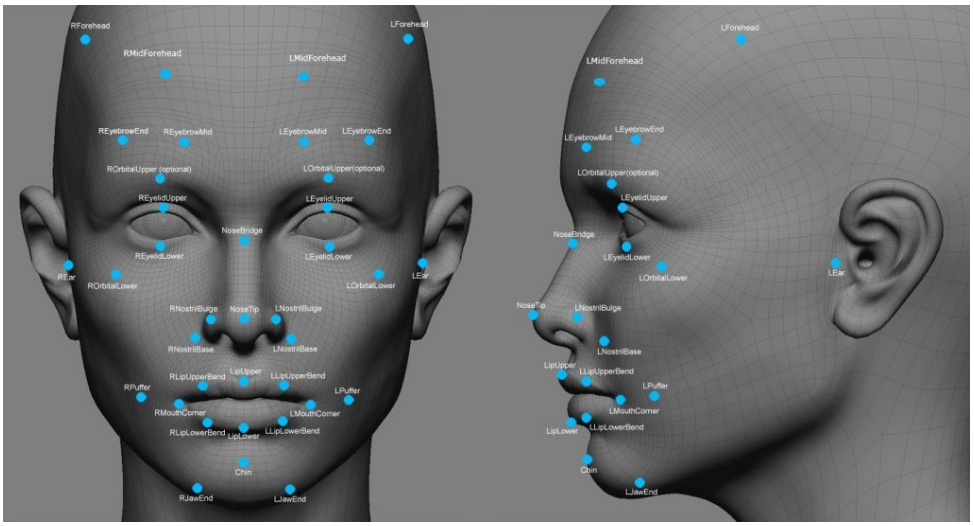


FIGURE 2.3: Geometrical features detected in the face

which scored the accuracy of 95%. They used 30 manually extracted distances to describe each face. B.S. Manjunath et. al. created 35-45 characteristic points for each face by applying Gabor Wavelet Decomposition [13]. After counterbalancing the various centroid position, two cost estimations, the similarity index, and topological index were calculated. The identification accuracy, which is the best match to the right person, was 86% and 94% of the right individual faces. The distance between characteristics points may be convenient for finding the matches in the data set. But they are more dependent upon the accuracy of the feature-dependent algorithm.

Template Matching In the template matching approach, images are described by the 2D matrix of intensity, and these values are compared by metrics methods such as Euclidean distance using a single template which describes the whole face. There are multiple techniques based on the template matching, in which the entire face is represented by more than one model [14].

Bruneli and Poggio chose four feature regions (eyes, mouth, nose and the whole face) [15]. While comparing the performance between the template matching solution and the geometrical matching solution on the same face database, they observed that the geometrical matching solution is less accurate than the template matching algorithm. However, the requirement of high capacity in computation is the most crucial limit of template matching and, furthermore, the difficulty also situates in the template representation. There are a few discrepancies between the test image and the template. This tolerance might influence the achievement of the face recognition method.

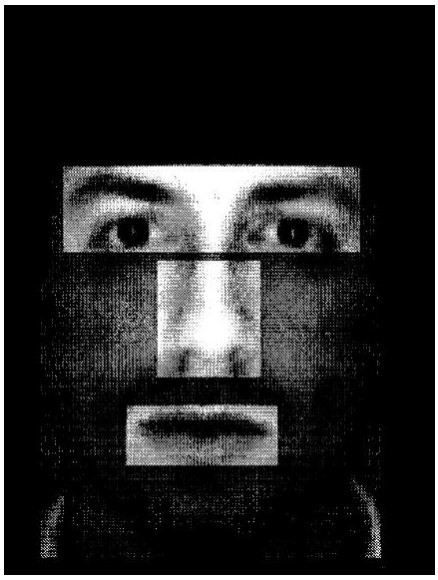


FIGURE 2.4: Regions for template matching

Thermal image based technique The thought of a thermal image was introduced to overcome the distortions and the illumination variation. It takes the subsurface characteristics of the face which may be admitted as a biometric feature. In this method the obtained thermal images are given to morphological processes and filtration, selecting the features of the face which can be employed for face recognition. Principally the outlines of the face are taken into account.

A study was conducted by Chen in which they compared the visible image methods and infrared methods [16]. In the experiments, it was shown that the visible image methods were outperformed by the infrared technique when they were done in an environment where the illumination condition was not constrained.

Selinger and Socolinsky affirmed that the combination of the two (infrared and visible imagery techniques) improved the performance when the experiments were taken outdoor whereas the thermal image based techniques have a few drawbacks such as the temperature of the skin [17, 18]. Although, it has been remarked that unlike thermal imagery, the hyperspectral information of the face is least influenced by the temperature than the thermal radiance. Before it, the spectroscopy has also been studied broadly in the remote sensing applications and biomedicine, assessing that various people show a high variation on the hyperspectral features of the facial texture, but these characteristics do not vary for the same person under different brightness condition and over the time.

Problematic

Illumination A face is a 3D object, thanks to which various light source on the face can cause different obscurations and various brightness [19]. The variation in face images of different individuals can be less notable than the variation of the image of

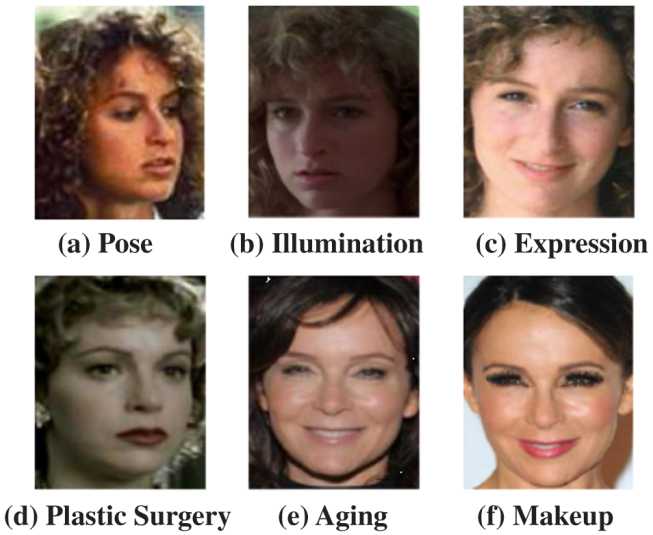


FIGURE 2.5: Face recognition's problematics.

the same individual in different illumination conditions. There have been multiple studies to develop facial features which are invariant against lighting variations [20].

Another problem associated to the absence or the lack of illumination which makes visible imagery useless. In this dark environment, even face detection cannot be accomplished. The presence of another camera type such as thermal sensor is necessary in this case.

Face spoofing attack Spoofing attack is trying to get a false acceptance of the authentication system using fake evidences. In the case of face spoofing, the attacker can use a photo, a video or a 3D mask of legitimate user as fake proof. In this digital era, a photo of normal person can be easily found in social network which makes the whole system become vulnerable.

Twin Faces Due to security purposes, the issue of twin faces was introduced. Even the human eyes get a lot of difficulty in recognizing the twins. There have been multiple studies performed on twin faces, but those are under calibrated environmental conditions i[21]. To recognize identical twins these methods either uses the entire face or different facial elements such as eyes, nose, and mouth.

Disguise Disguise is the most significant security menace of a system, or we can assume that disguise is still a significant problem for face-recognition methods to recognize a person when he or she seeks to cover its own identity to imitate someone else. There are few proposed solutions by researchers in which the difficulty of disguise can be resolved [22].

However, this problem is not necessarily considered in an authentication system which can demand the person to take off his or her disguise when it detects one. It is the make-up, a type of unintentional disguise, which is the real problem for the identification process.

Facial expressions The expression is an internal movement of the image which creates huge intra-class variations. Various facial expressions which hinder the performance of face identification are joy, astonishment, anger, anxiety, sorrow, excitement and many more. To handle these expression problems, there are 3D model-based approach and local features based approach [23, 24].

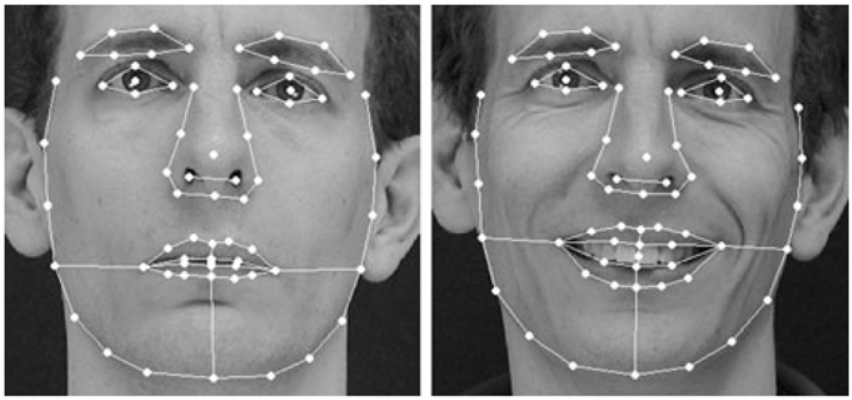


FIGURE 2.6: Facial expression changes the face model.

In our first study, we concentrate on the research of a new method that can detect the face spoofing attack using a single uncalibrated normal camera such as smartphone. The concept of this type of attack will be introduced in the next section.

2.1.2 Face Spoofing Attack

Introduction

Face spoofing is an active attack against the authentication system by face recognition [25, 26, 27, 28, 29, 30]. The notion active emphasizes the real intention of the attacker instead of a normal user's mistake like the case of makeup or natural problems such as illumination conditions of facial expressions. In this case, the attacker is supposed to have sufficient knowledge of the system mechanism.

Attackers have many ways to attack a facial recognition system [31, 32, 33]. They can utilize a photo of legitimate user printed on a piece of paper or displayed on an LCD screen and present it in front of the camera in operation. They can also replay a video which filmed the victim previously or evenly use a 3D mask to mislead the face detection process. Face spoofing can be classified into two types of attack by their false proof of authentication: 2D attack and 3D attack. The 2D attack can be further divided into photo attack of video attack (replay attack). Since the attackers are supposed to know the authentication method, they can choose and customize their attack for each system.

Photo attack

In the early age, face spoofing could be accomplished using a single printed photo of a legitimate user. The attacker shows this photo to the biometric sensor (in our

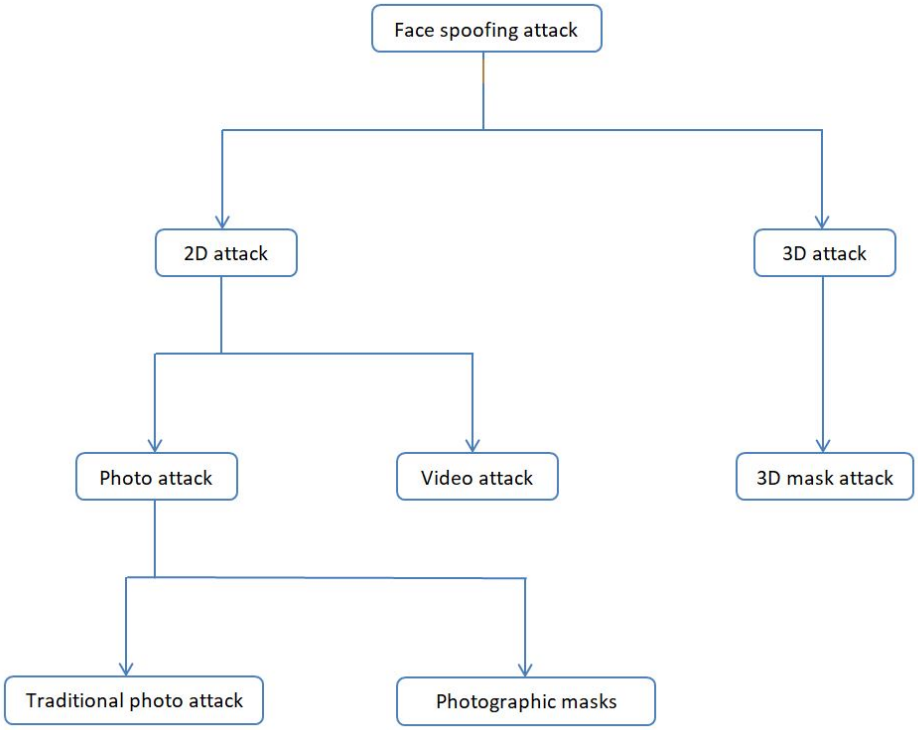


FIGURE 2.7: Different types of face spoofing attack.

case, the biometric modality is usually a camera) which considers this one as a proof of authentication and grants access to the owner. This type of simple process bears the name of its material: photo attack. It is the easiest and also the most widespread method especially as users photo can be simply retrieved from the social networks such as Facebook, Twitter, and Instagram [34] .

An advanced type of photo-attack in which high-resolution prints of eyes and mouth are morphed is developed under the name photographic masks. During the attack, the impostor placed himself behind, so that certain facial movements like eyes blinking or random face expression is reproduced [35]. Photographic masks become a crucial threat to a lot of face liveliness detection methods which uses these facial movements as the ultimate information for their classification.

Video attack

Video attack is another advanced version of photo attack in which the attacker replays a video of the genuine user in front of the camera [36]. The video can be taken by a smartphone, a tablet, a surveillance camera, with or without the cooperation of the user. The screen that displays that video can also be diverse in terms of size and resolution. The main purpose of replacing a photo with a video is also to reproduce some facial movement in order to deceive the authentication system [37, 38].

For the same idea to mimic certain face movement video attack is closer to the original version than photographic masks. However, the video replayed is a fixed sequence which cannot interact with the system. In some case of an active system

FIGURE 2.8: Photo attack illustration.

which demands the user to generate an unexpected expression, the video attack may be totally inefficient. Furthermore, a video which focuses the most on the users' face is more difficult to retrieve than a single photo.

3D mask attack

The 3D mask attack is the most advanced attack due to the depth elements in the facial features [39]. In 3D mask attack, attackers have to focus on their target and do firstly manage to construct a 3D mask or maybe a sculpture of the target. The 3D masks are usually made of different materials and sizes, i.e., paper, plastics and silicon. If the mask is constructed perfectly, there is less chance to detect it. However, the achievement of this type of attack is quite difficult and expensive.

2.1.3 Face Spoofing Detection

The issued research studies through the former 20 years have shown that important protection against the spoofing of biometric proof has taken place making biometric method safer and more robust thanks to the intense efforts of researchers [40, 41]. Many approaches have been proposed in the literature to deal with face spoofing attacks using different features like texture, liveliness, structure, etc. These types of method of detection can be introduced as follows:

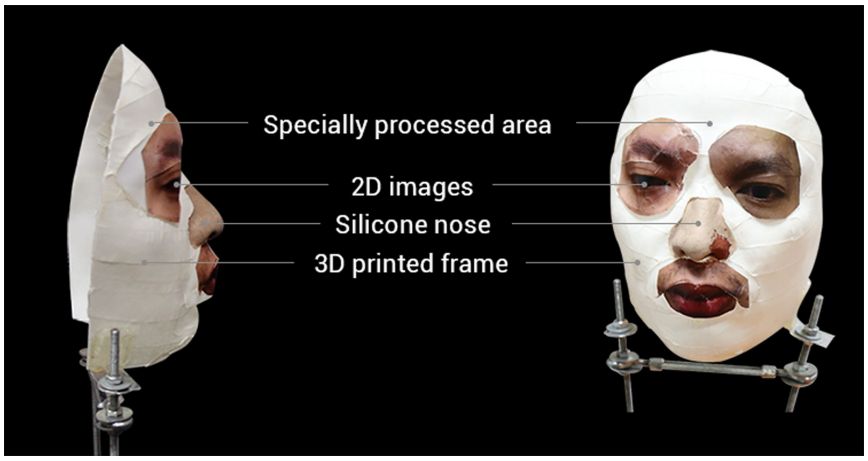


FIGURE 2.9: 3D-mask attack enhanced by painting.

Textural information

Textural information, which manages to be different between real-face images and fake ones, can be exploited for face spoofing detection. From a single image, Matta et al. [42] propose to analyze the texture of facial images using multi-scale Local Binary Pattern (LBP). In the same spirit, Kim et al. [43] , also utilized LBP, but in fusion with frequency analyses by using the power spectrum. Other researchers exploited the Local Graph Structure (LGS) [44] or its improved versions (ILGS, SLGS) as texture descriptors to conceptualize their face spoofing detection method. Another method proposes to exploit the statistic behavior of the distribution of noises local variances to detect face spoofing attacks [45].

Zhang et al. [46] proposed a texture based technique in which Lambertian model was employed to recognize the human skin and SVM (Support Vector Machine) was applied to classify the real person and imposter. This liveness detection method has some limit dueing with real-time spoofing attempt in unconstrained conditions. The low-resolution webcam (320x240 pixel frames at 25 fps) used in this experiment gave an accuracy of 7% error rate.

Zinelabidine Boulkenafet et al. [47] described a face spoofing detection technique using color texture analysis. In this algorithm, the researcher concentrated on luminance data of the face photo and the chrominance data was rejected so that this method could recognize the legitimate users. This novel approach is tested on CASIA-FASD database, Replay Attack database, MSU mobile face spoof database. The results showed by experimentation is n accuracy of 0.4% error rate.

Ivana Chingovska et al.[48] suggested a texture feature technique in which the Local Binary Pattern (LBP) algorithm was applied. Tests were made various sorts of databases: NUAA Photograph Imposter Database, Replay Attack Database, Public Database, and CASIA Face Antispoofing Database. This strategy presented 15% total error rate of different databases.

Jukka Komulainen et al. [49] presented a context-based face anti-spoofing method in the year of 2013. The experiment was tested with two publically accessible databases,

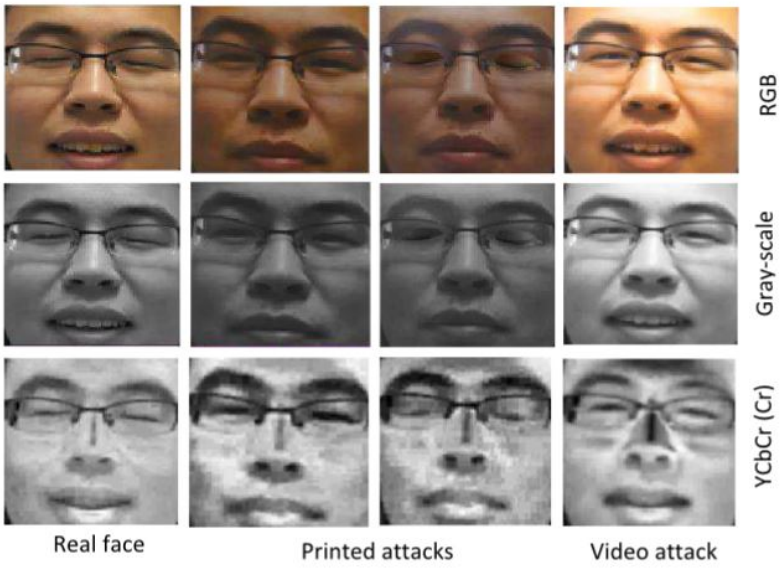


FIGURE 2.10: Face spoofing detection algorithm based on color texture analysis. [47]

and it made 3% error rate to recognize the genuine person and the imposter.

Nesli Erdogmus et al. [50] proposed a 3D mask spoofing attack detection method in 2014. The researcher applied Linear Discriminant Analysis (LDA), Modified Local Binary Pattern (MLBP), and Support Vector Machine (SVM) to distinguish the genuine user and the imposter in biometric techniques. Morpho database and 3D Mask databases were employed to make the experiments. According to the suggested process, the test showed a 3% error rate.

Santosh Tirunagari et al. [51] introduced a visual dynamic method. The researcher suggested a classification scheme including Principal Component Analysis (PCA), Local Binary Pattern (LBP), Dynamic Mode Decomposition (DMD), and Support Vector Machine (SVM). All these combinations of classification techniques provided more precise results to detect the spoofing attack. This scheme was tested employing three accessible databases: Replay Attack Database, Print Attack Database, and CASIA-FASD Database. The test shows an error rate of 9.50%.

Shervin Rahimzadeh et al. [52] introduced various descriptor mixing method. This kernel mixing method was constructed based on a fast kernel discriminant analysis (KDA). The test was executed on another publically accessible database. In this method, multi-scale, dynamics binarized statistical image patents were utilized. The researcher assessed the method on Replay Attack database, CASIA database, NUAA Photograph Imposter and Cross database which generated an error rate of 1.67%. The obtained results display an improvement in recognizing the different spoof attacks as compared to other methods.

Allan Pinto et al. [53] suggested a face spoofing technique based on visual codebooks of spectral-temporal cube methods. This texture-based technique was based

on support vector machine. Mid-level feature extractors were used rather than low-level feature extractors. This suggested method was realized on databases of all types such as photos, video, and 3D mask database and the achieved results were a 0.6% error rate on different databases.

Akshay Agarwal et al. [54] introduced a novel strategy using hard lick method on Face antispoofing. The technique yielded an error rate of 1.1% on video database. However, the technique demonstrated less effective performance on photo databases and 3D databases.

Liveliness Detection

Some approaches manage to distinguish real faces from spoofed faces by seeking proofs of liveness from a sequence of images or from a video capturing the face. Kollreider et al. [55] proposed an approach in which lip movements are exploited for face spoofing detection while the user is asked to speak some numerical digits. Huyng-Keun Jee et al. [56] , in their approach, proposed to study uncontrollable movements of eyes regions, such as the eye blinking or pupil movement.

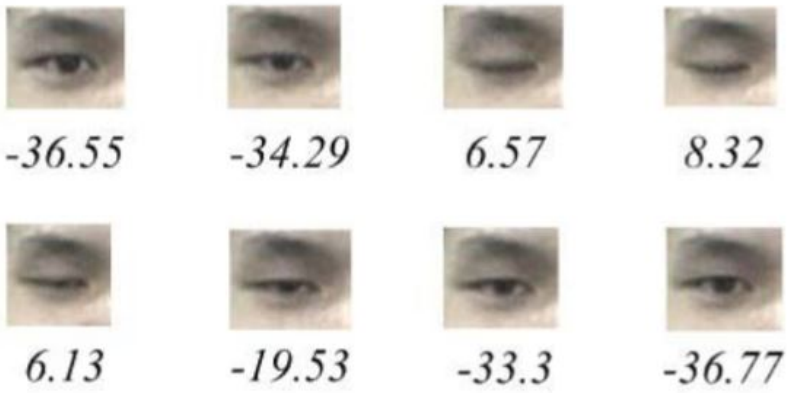


FIGURE 2.11: Score the eye closity to detect eye blinking.

Lin Sun et al. [39, 57] described a real-time liveness detection method against photo attack in 2006. This method detects involuntary eye blinking. This method demands no additional device besides a webcam. Adaboost classifier and HMM methods are applied for eye blinking detection which yields high accuracy results with 3% error rate. In these studies, the researchers exploit eyes movements by modeling and detecting the two principal states of the eyes: opened-state and closed-state.

Mihai Gavrilescu et al. [58] proposed soft biometric methods using the neural network and principal component analysis. In this video-based face recognition scheme, the researchers employed several facial expressions on people in various frames. The results showed an error rate of 5.5%.

Structure and motion study

Some other methods exploit the differences between 2D objects and a 3D face in their structure, their moving features or the depth information that they provide. For instance, Kim *et al.* [59] proposed to compare images captured in different focusing. For a 3D object, due to depth information, the difference between images of different focusing will be clearer than the one in the case of 2D objects. The approach permits to identify efficiently spoofing attacks using a 2D display support. Studying the difference in the behavior of optical flow generated by 2D spoofed face and real face have been also envisaged.

K.Kollreider et al. [60] described a holistic liveness detection technique. In this integral method, K.Kollreider proposed a lightweight novel optical flow method using score based technique. This method based on Gabor and SVM (Support Vector Machine) yields an error rate of less than 0.5%.

Anjos et al. [35] proposed a motion-based method which uses the monotone motion of a photo attack. In this method, the Print-Attack database, which includes 200 videos for 200 real-time attempts using 50 photogs were employed. The experiment demonstrates a moderate success of the technique with an error rate of 10%.

TABLE 2.1: Existing methods comparing.

Author	Methods	Attacks	Database	Accuracy
Pan et al. (2007)	Eyeblink Detection using conditional random field (CRP)	Photo	Public Database and NUAA Photograph Imposter Database	Error Rate= 2%
Kollreider et al. (2008)	Motion Detection	Photo and Video	Proprietary	Error Rate= 3.5%
Kollreider et al. (2009)	Face motion Detection using optical flow of lines	Photo	Proprietary	Error Rate= 0.5%
Tan et al. (2010)	Face texture using the Lambertian model	Photo	Public Database and NUAA Photograph Imposter Database	Error Rate= 15%
Anjos et al. (2011)	Context-Based using correlation between face motion v/s background motion	Photo	Public Database and Print Attack Database	Error Rate= 10%
Zhang et al. (2011)	Reflectance using multispectral lighting in 2D images	Photo, Video and 3D Masks	Public Database and Print Attack Database	Error Rate= 7%
Chingovska et al. (2012)	Face Texture using Local Binary Pattern (LBP)	Photo and Video	Public Database, Replay Attack Database, NUAA Photograph Imposter Database and CASIA Face Antispoofing Database	Error Rate= 15%
Komulainen et al. (2013)	Context based using upper body & spoof support Detection	Photo and Video	Public Database, NUAA Photograph Imposter Database and CASIA Face Antispoofing Database	Error Rate= 3%
Nesli Erdogmus et al. (2014)	Modified Local Binary Pattern (MLBP) + Linear Discriminant Analysis (LDA) + Support Vector Machine (SVM)	Photo and 3D Mask	Morpho Database and 3D Mask Attack Database	Error Rate= 3%
Santosh Tirunagari et al. (2015)	Principal Component Analysis (PCA) + Local Binary Pattern (LBP) + Support Vector Machine (SVM)	Photo and Video	Print Attack Database, Replay Attack Database and CASIA-FASD Database	Error Rate= 9.5%

Infrared technology is also known as a robust solution against face spoofing attack. In the next section, we focus on thermal face recognition methods in recent years and assert its capacity to compete with visible imagery in some particular cases.

TABLE 2.2: Existing methods comparing.

Author	Methods	Attacks	Database	Accuracy
Shervin Rahimzadeh Arashloo et al. (2015)	Spectral Regression kernel discriminant analysis (SR-KDA)	Photo and Video	Replay Attack Database, CASIA Face Antispoofing Database, NUAA Photograph Imposter Database and Cross-Database Evaluation	Error Rate= 1.67%
Mihai Gavrilescu et al. (2015)	Neural Network (NN) + Principal Component Analysis (PCA)	Photo and Video	Honda/UCSD Video Database, Youtube Faces Database, Replay Attack Database, NUAA Photograph Imposter Database and CASIA Face Antispoofing Database	Error Rate= 5.5%
Allan Pinto et al. (2015)	Partial Least Square (PLS) and Support Vector Machine (SVM)	Photo, Video and 3D Masks	Replay Attack Database, CASIA Face Antispoofing Database, UVAD Database, 3DMAD Database	Error Rate= 0.6%
Zinelabidine Boulkenafet et al. (2016)	YCbCR + HSV	Photo and Video	Replay Attack Database, CASIA-FASD Database and MSU mobile face spoof Database	Error Rate= 0.4%
Akshay Aggarwal et al. (2016)	Haralick Texture Features + Discrete Wavelet Transformed + Principal Component Analysis + Support Vector Machine	Photo and Video	3DMAD Database, CASIA-FASD Database and MSU-MFSD Database	Error Rate= 1.1%

2.2 Thermal Face Recognition

For the last few years, Infrared imagery has attracted particular attention, principally thanks to its robustness against the changes in illumination by visible light and its capacity of liveliness detection. This section is divided into three parts. The first subsection analyzes the thermal spectrum with its features, its advantages, and its limits. The second part gives a brief overview of thermal face recognition solutions. A very potential solution using the vascular network is described in the last subsection.

2.2.1 Introduction

Thermal Spectrum

Infrared radiation (IR) is a type of electromagnetic radiation (EMR) of which wavelengths are longer than wavelengths of visible light. Infrared radiation is ordinarily invisible to the human eye and hardly distinguishable by human perception. A detailed report of its physical characteristics, which is outside the border of this thesis, can be found in [61]. Infrared imagery can be used in many applications as follows:

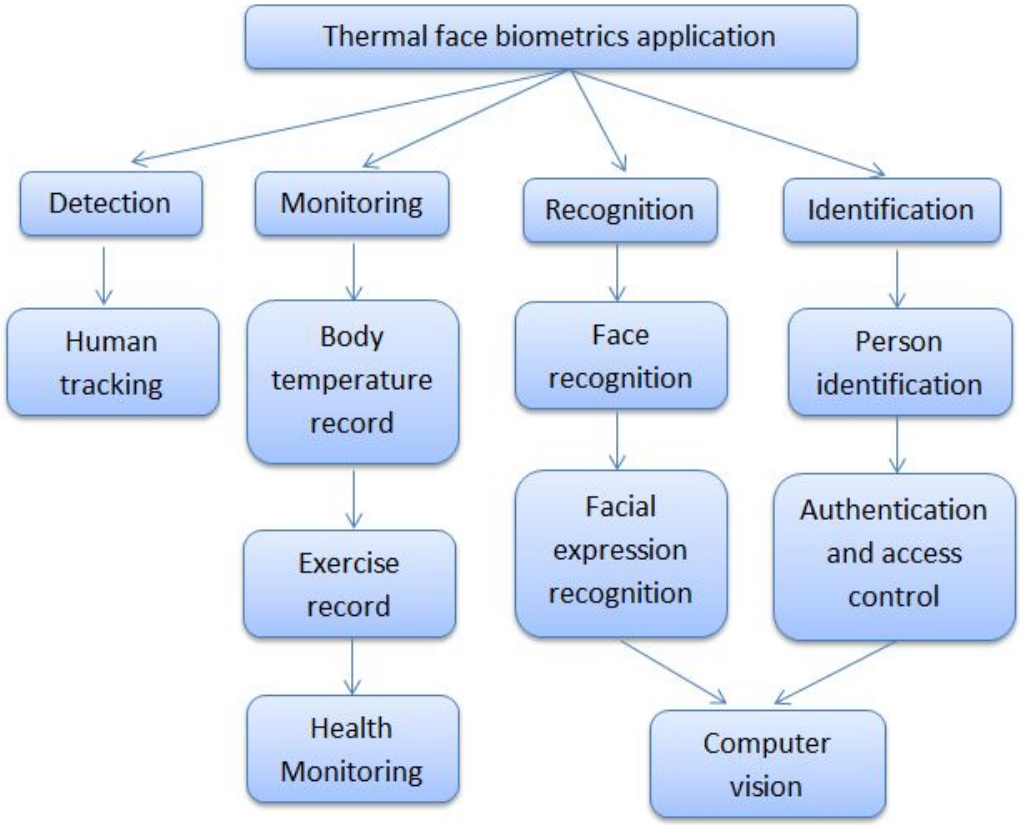


FIGURE 2.12: Infrared imagery application

In the scope of face recognition, information obtained by an infrared sensor has particular advantages over standard cameras which are used in the visible spectrum. For instance, thermal data of the faces can be acquired under any illumination condition, even in the absence of any lighting source or absolute dark environments, and there is some study which proves that infrared face image may exhibit a higher degree of robustness to facial expression change [62].

Infra-red radiation is also less affected by scattering and absorption by dust or smoke than reflected visible light [63, 64]. Another advantage of infrared imagery is based on its capacity to detect any disguise in the face of which the material may not radiate the same way as human skin. In contrast to visible spectrum imagery, infrared imagery can be used to extract not only exterior features but also useful subcutaneous anatomical information, such as the blood perfusion or the vessels

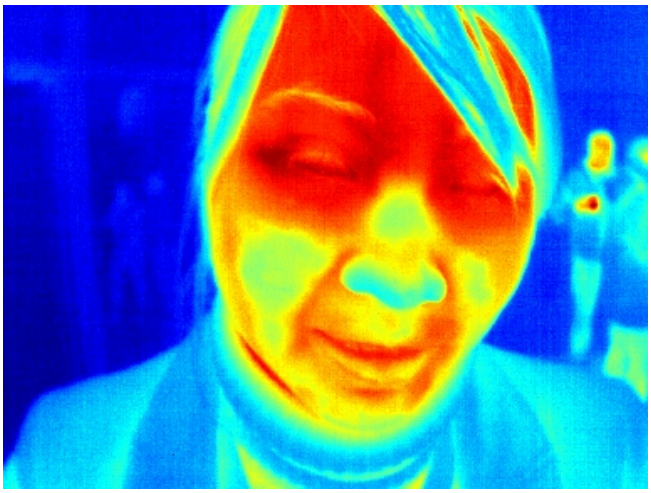


FIGURE 2.13: Infrared image for face recognition

feature of a face [65]. Finally, unlike visible spectrum imaging, thermal vision can naturally detect face spoofing attack.

Spectral Composition: In the existing literature, it has been customary to divide the infrared spectrum into four sub-bands :

- Near IR (NIR): wavelength 0.75 - 1.4 μm.
- Short wave IR (SWIR): wavelength 1.4 - 3 μm.
- Medium wave IR (MWIR): wavelength 3 - 8 μm.
- Long wave IR (LWIR) : wavelength 8 - 15 μm.

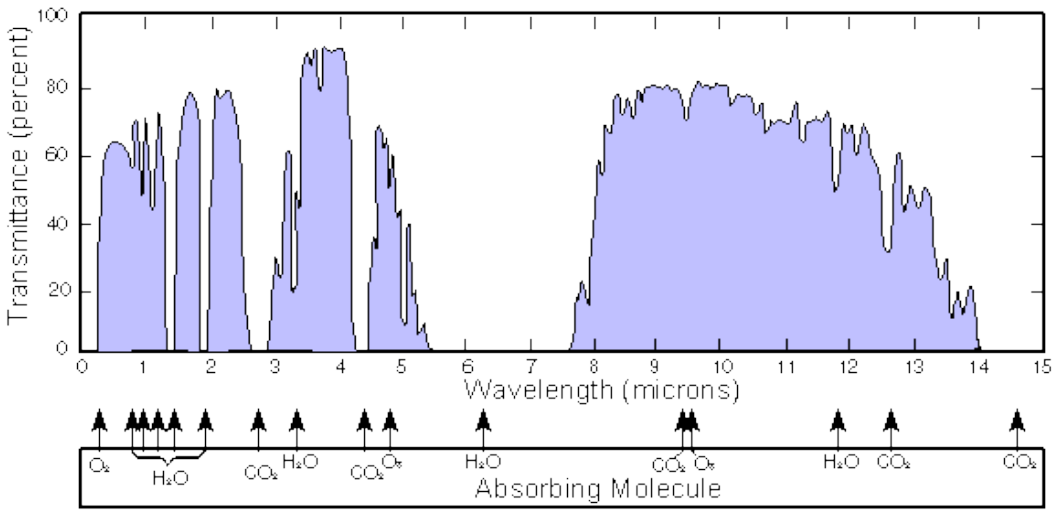


FIGURE 2.14: Transmitting rate in atmosphere of IR spectrum

This classification of the IR spectrum is also observed in the manufacturing of infrared cameras, which are usually made with sensors that correspond to energy radiation constrained to a specific sub-band.

It should be highlighted that this classification of the IR spectrum is not arbitrary. Instead, different sub-bands correspond to continuous frequency chunks of the solar spectrum which are divided by absorption lines of different atmospheric gasses [61]. In the scope of face recognition, one of the most significant differences between the 4 IR sub-bands emerges as a result of the human skin's heat emission spectrum which is, in ideal condition, shown in Figure 2.15.

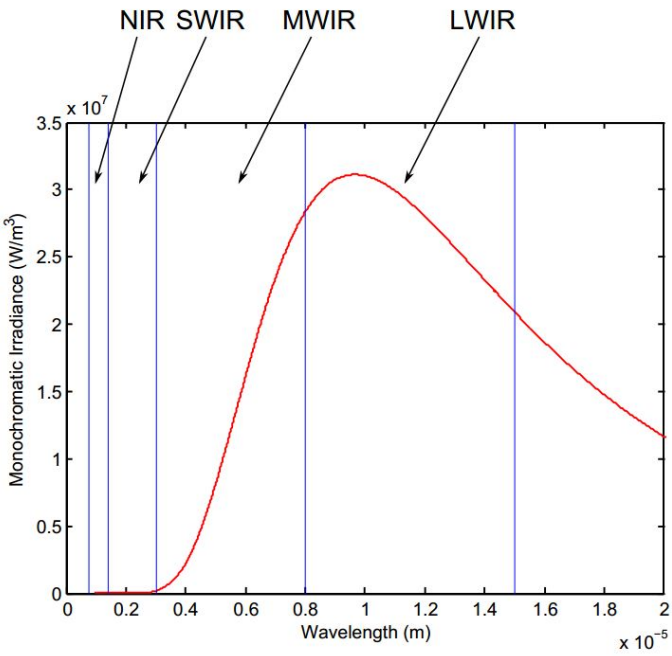


FIGURE 2.15: Infrared spectrum

Usually, the NIR and SWIR bands can be referred to as "the reflected infrared radiation." The human's body does not emit a significant amount of electromagnetic energy in these sub-bands. In fact, the NIR and SWIR bands are dominated by the reflected radiation. Just like the visible imagery, these two sub-bands require an additional energy source (which can be in the form of light or heat.

The MWIR and LWIR bands are often labeled as "thermal infrared radiation". Unlike NIR and SWIR bands, they do not require any additional source of infrared energy since the human's body emits strong electromagnetic radiation in these spectra. Since IR sensors in these bands depend mostly on the amount of emitted energy of a recorded object, they are, in the contrast of visible light camera, invariant to the change of lighting conditions, robust against a lot of problems like weather conditions and can operate in complete darkness.

Especially, since most of the heat radiation is emitted in LWIR sub-band, MWIR sub-band plays a lower role in the thermal spectrum. However, heat energy emitted in MWIR sub-band is usually strong enough to overcome any other type of radiation. Both of these sub-bands can be used in a passive thermal system. That is one of the reasons why thermal sub-band like LWIR or MWIR have received the most attention in the context of face recognition. Unlike them, body heat emission in the SWIR and

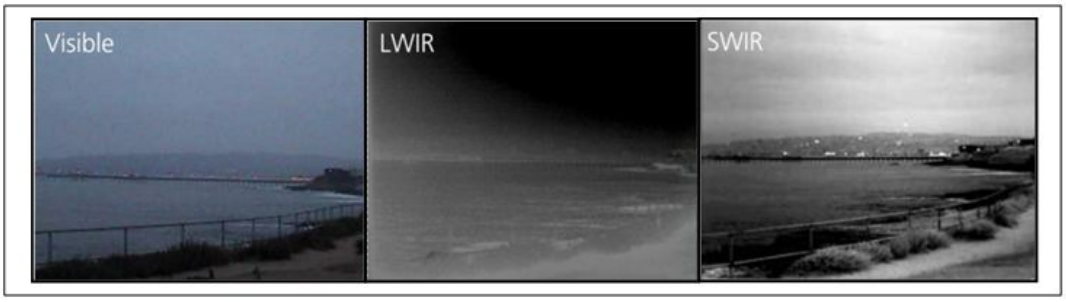


FIGURE 2.16: SWIR image is almost as details as visible image

NIR sub-bands is minimal, and face recognition systems operating on data acquired in these sub-bands require appropriate illuminators. The sensors used in NIR sub-band is very close to the ones used in visible imagery which make it start to receive a lot of attention from the face recognition community, while the utility of the SWIR sub-band has yet to be studied in depth.

FIGURE 2.17: Face image in the visible spectrum (left), SWIR, MWIR, LWIR(right)

Infrared images captured under different sub-bands tend to have different properties. Infrared images provide fewer details than visible cameras since they are monochrome. The color obtained in the visible spectrum can be naturally interpreted and contains more information. The one taken under SWIR or NIR is very detailed, close to visible imagery. The details level of the infrared image decreases when the wavelength increases. LWIR camera provides the least details among these sub-bands, and it is also the only sub-band which achieves full invariance to illumination conditions.

In general, infrared sensor does not distinguish various wavelengths inside a sub-band. Color sensors require a highly complex structure to classify wavelength which is very difficult to be applied into infrared modality. The infrared spectrum is much larger than the visible spectrum and cannot map uniformly into the human color system. Outside the visible spectrum, color does not have a natural interpretation.

Infrared images are typically displayed in the form of grayscale images which describe the intensity of infrared energy captured by each pixel. Lighter area represents

TABLE 2.3: Infrared sub-bands comparing.

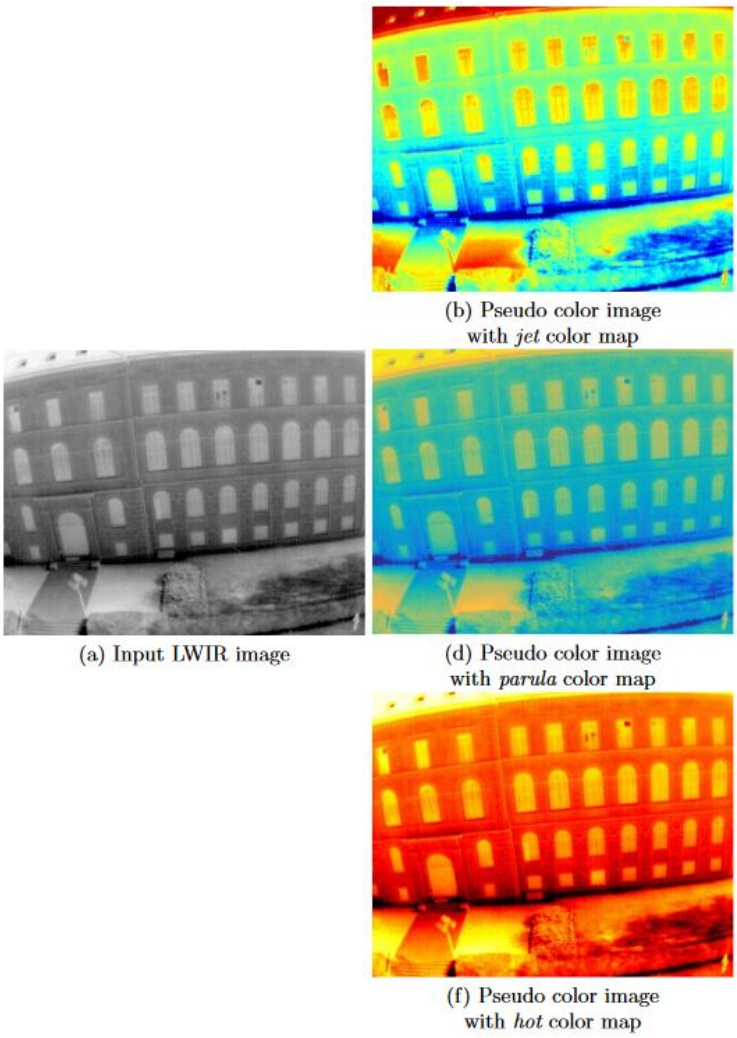
	NIR	SWIR	MWIR	LWIR
Wavelength	0.75 - 1.4 μm	1.4 - 3 μm	3 - 8 μm	8 - 15 μm
Illuminator	+	+	-	-
Details	++	+	-	–
Lighting Invariance	–	-	+	++
Type	Reflected	Reflected	Thermal	Thermal

higher temperature (eyes, lips) However, in some particular case, these monochromatic infrared images can be shown in pseudo-color where different colors represent the change in intensity. This technique favors the human vision of which color conception can be easier to distinguish different intensities. The pseudo-color is not absolute but relative, that means in different images, one temperature may be represented by different colors depending on the temperature range of the object and the background.

The use of infrared imagery for automatic face recognition has its own problems and challenges. For example, thermal images are sensitive to the environmental heating condition, as well as the emotional, physical and health condition of the person. LWIR images are even affected by alcohol intake. Another problem source is the eyeglasses which are totally opaque to many of the IR spectrum (LWIR, MWIR and SWIR) [66, 67]. This means that a large area of the face wearing eyeglasses may be occluded, causing the loss of important discriminative features around the eyes. Unsurprisingly, each of the problem has begun a new research direction. Some researchers have proposed fusing the information from IR and visible modalities as a hybrid solution to the problem of eyeglasses opaqueness. Others have suggested methods which use infrared images to extract a map of invariant features such as facial blood perfusion data[[65] or vascular network [68] in order to overtake the temperature dependency of thermal "appearance".

Thermal Sensors

Cooled infrared detectors: The conventional cooled infrared sensors detect and convert electromagnetic energy in the same way as standard visible-light camera (indeed, they are made of different materials). Without any cooling system, the infrared radiation of the object is mixed with the energy emitted by the sensors themselves. The cooling system can keep the sensor's temperature at a prefixed level so that the energy captured by sensors does not vary by detector status.



(b) Pseudo color image with *jet* color map

(a) Input LWIR image

(d) Pseudo color image with *parula* color map

(f) Pseudo color image with *hot* color map

FIGURE 2.18: Infrared image in pseudo-color

Cooled infrared sensors are mainly contained in a vacuum-sealed case or Dewar and cryogenically cooled. The cooling system is requisite for the operation of the semiconductor materials used in these sensors. Depending on the sensor technology, operating temperatures can vary from 4^o K to a little lower than room temperature. However, a great part of cooled detectors functions in the 60 ^{o} K to 100 ^{o} K range. The most commonly used cooling systems are rotary Stirling engine cryocoolers.

The disadvantage of cooled infrared cameras is their expensive cost to produce and to operate. Cooling is not only energy consuming but also take a lot of time. These cameras may need several minutes to cool down before they can begin to run. Although the cooling apparatus is comparatively bulky and expensive, cooled infrared sensors provide images in higher quality compared to uncooled detector thanks to their superior capture rate.

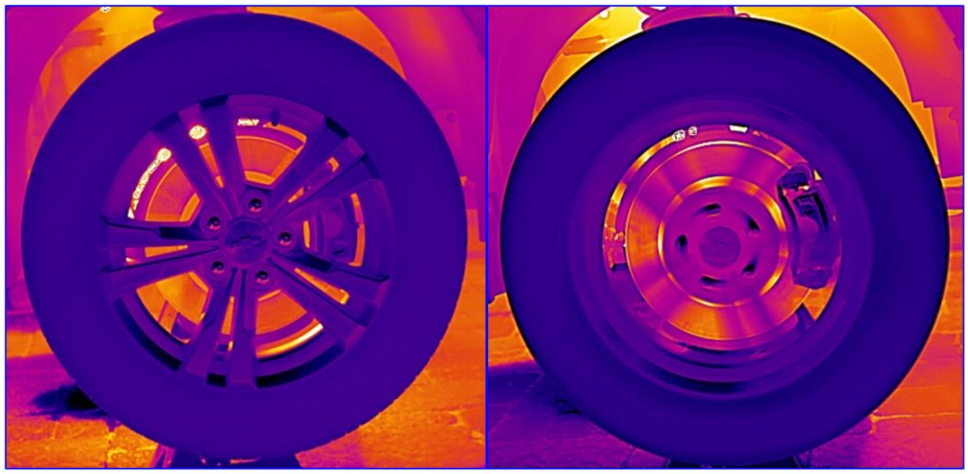


FIGURE 2.19: Image taken by cooled Infrared camera(left) in high capture rate compared to one taken by uncooled camera(right)

Uncooled infrared detectors: Unlike the cooled camera, uncooled thermal cameras use a sensor running at ambient temperature or a sensor which could be stabilized at a temperature level just below ambient temperature using small temperature control elements. Modern uncooled detectors all use microbolometer sensors that work by the change of electric aspects like resistance, voltage or current when heated by infrared radiation. Unlike photons detectors, energy detectors like microbolometer do not directly count the amount of input photons but measure only a total captured energy level. The variation of these electric elements is measured and compared to the original values of the sensor.

Uncooled infrared sensors do not require any bulky, expensive, energy consuming cryogenic coolers. They also need to be stabilized to a fixed operating temperature to reduce image noise, but this temperature is not as low as the one required by a cooled detector. These advantages make infrared cameras smaller and cheaper than cooling technology so that an uncooled camera can be added into any individual machine such as smartphone or drone.

However, in uncooled detectors, the temperature variance at the sensor pixels are miniature; a 1^{o}K variation at the object produces merely a 0.03 ^{o}K difference at the detector. The pixel response rate is also much lower than a cooling system, at the level of tens of milliseconds. This makes their resolution and image quality more moderate than a cooled camera. This is because of differences in their fabrication material which is limited by currently available technology.

Uncooled sensors are principally based on pyroelectric and ferroelectric materials or microbolometer technology. These materials are used to form pixels with a high level of temperature-dependency in some electrical properties.

Advantages

A significant part of the early work on the capacity of infrared images as identity proof was studied by Prokoski et al. [69]. They were the first to propose the idea that

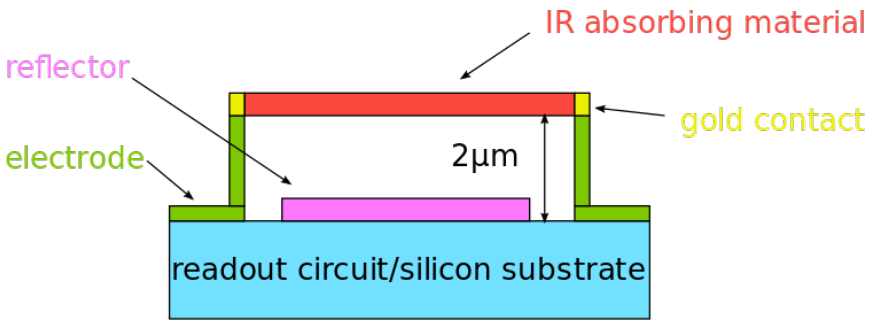


FIGURE 2.20: Microbolometer used in Uncooled infrared detectors

thermal "appearance" of a face could be used to retrieve distinct biometric features which contain a high level of repeatability and uniqueness.

The invariance to complex Ilumination condition The invariance to complex Ilumination condition is the most essential advantage of using thermal face recognition compared to using standard visible imagery. The lightning has very little affectation upon the infrared signature. The longer the wavelength, the weaker the variance. Under LWIR sub-band, thermal radiation does not depend on any illumination factor. This is also the main purpose of using thermal imagery in face recognition [70].

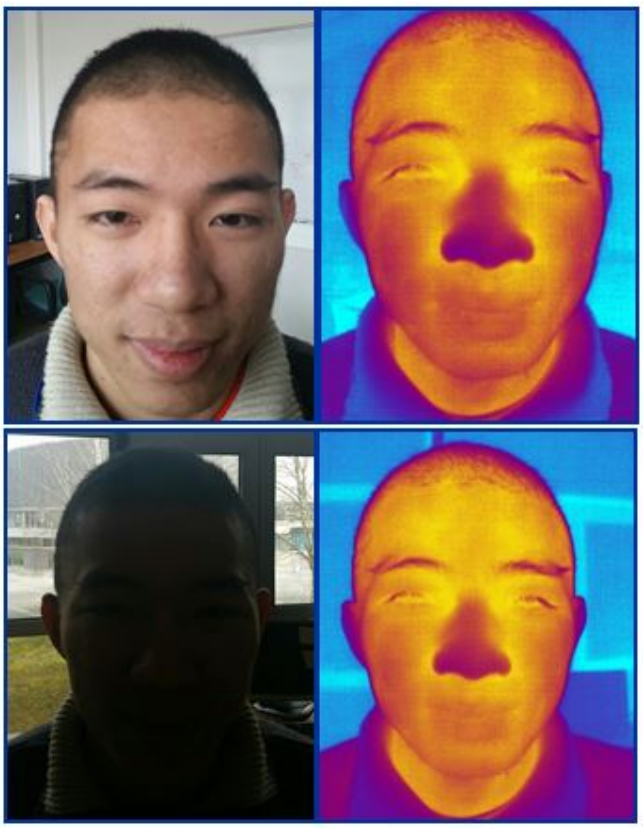


FIGURE 2.21: Thermal image is more robust against various illumination conditions than visible one

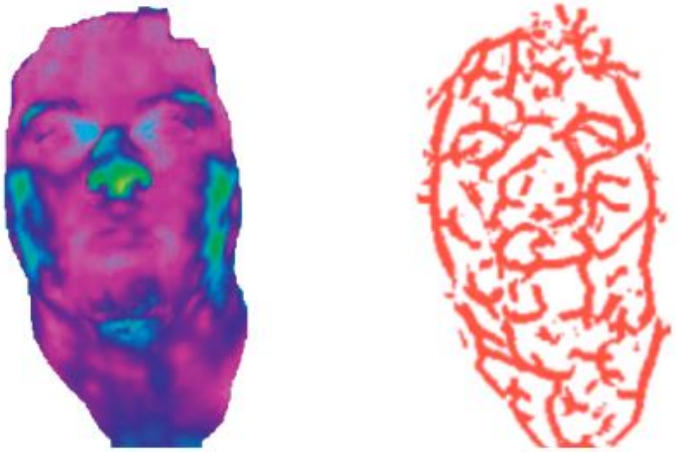


FIGURE 2.22: Early vascular network model

Facial expression and pose positions are two challenges that a face-recognition system must overcome in order to be useful in most real case application. Using the image space differences between infrared and visible images, Friedrich et al. [62] shown that infrared images are more robust against changes in facial expression or head pose than their standard visible imagery.

Face spoofing and disguise are also crucial risks that attack the authentication system by face recognition. The property of thermal imagery also opens the possibility of non-invasive extraction of superficial anatomical features for recognition such as blood perfusion and vessel patterns. Naturally, the blood vessel which transport circulating blood continuously , are warmer than the surrounding area. This property can be captured by the thermal camera and be extracted by processing technique in order to isolate the blood vessel from the face image. An essential characteristic of these patterns which makes them particularly interested in face identification is that the blood vessels are defined by young day and form a realative network which remains very little affected by ageing factors such as ageing. Furthermore, it seems that the human vessel feature can also respond for another key challenge: the scalability in large populations. Prokoski et al. assume that about 175 blood vessel based minor features can be retrieved from a complete facial image [71] which, they considered, can represent a far greater amount of possible setting than the number of the maximum human population. However, the authors did not propose a particular method to obtain the minutiae in question.

In the very same work, Prokoski et al. also indicated that spoofing attempts and disguises can both be detected naturally by infrared imagery. The critical proof is that the temperature signature of artificial hair or other facial mask differs from the heat distribution of natural skin and hair, allowing them to be distinguable one from another. This fact also provides thermal imagery immunity against the face spoofing attack. Making a mask which can match the vessel pattern of someone else is almost impossible for now.

Monozygotic twins: An interesting issue first proposed by Prokoski et al. [71] involves the thermal signature of monozygotic twins. The image of monozygotic twins is almost indistinguishable in the visible spectrum. Using a little number of infrared image of monozygotic twins which were evaluated for similarity, Prokoski et al. observed that the variation in appearance was significantly higher in the infrared imagery than in the standard visible representation, and provides sufficient proof to automatically distingue these twins.

Limits

In the scope of automatic face recognition and identification, the main problematic specific to the LWIR sub-band images, the only sub-band of the infrared spectrum can provide absolute invariance to illumination, arises from the fact that the heat model emitted by the object is affected by a lot of mixing variables, such as environmental temperature, atmosphere flow conditions, postprandial metabolism, exercise, sickness, alcohol and drugs [72]. Some of these variables create local, other global infrared appearance changes. Wearing make-up, enduring stress, blushing, having an infected tooth or even a headache are examples of issues which can affect the thermal appearance.

The great sensibility of the facial infrared image in no small number of external factors challenges the thermal face recognition community in finding persistent and discriminative features. It also provides proof for the ideas first proposed by Prokoski et al. [71] who discussed against the use of appearance-based methods for thermal face recognition in favor of superficial anatomical feature based approaches which is robust against many aforementioned factors.

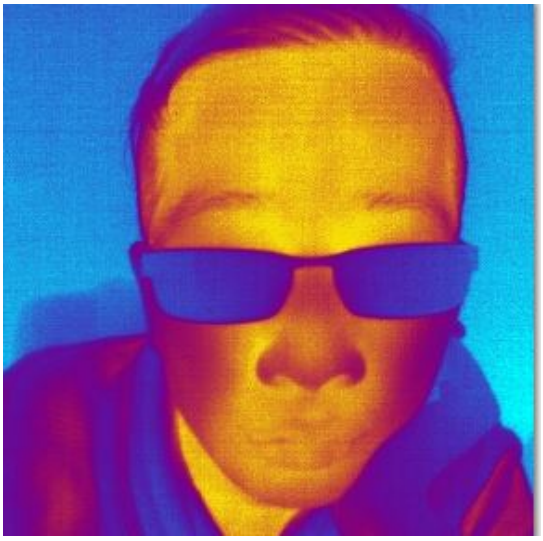


FIGURE 2.23: Thermal image perturbed by eyeglasses(They are not sun eyeglasses)

Another issue of using the thermal spectrum for face recognition is that glass, and indeed, eyeglasses are entirely opaque to wavelengths beyond the NIR sub-band, thus all the SWIR, MWIR, LWIR sub-bands. Consequently, an essential part of the face which is very rich in discriminative features may be occluded in the thermal images. In particular, the void of crucial information around the eyes can significantly decrease recognition accuracy. Many multimodal fusion based methods have been proposed to deal with this problem.

2.2.2 Methods

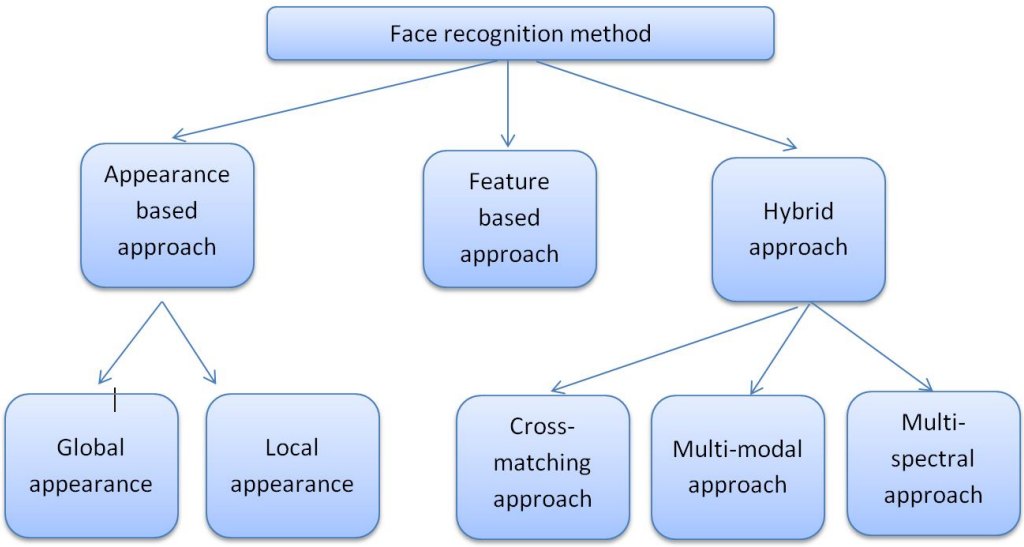


FIGURE 2.24: Categorization of thermal face recognition

The most critical challenge in face recognition is that facial thermograms provide different details compared to image in the visible spectrum. They contain less data in the structure and almost no information concern color and skin texture. It is crucial to define features that highlight thermal face characteristics in order to use in classification.

In the last decade, many studies in thermal face recognition have been realized using the same methodology as visible imagery. However, the result of these methods is far lower than the original one. A few other methods which devote only to thermal imaging are also developed by exploiting some unique feature of infrared images. A lot of methods use more than one step of feature extraction before the classification that makes categorizing thermal face recognition methods more complex.

In this study, we use the categorization of Ghissa el al in [73, 74], a comprehensive survey on infrared face recognition methods. These methods are regrouped according to their core feature extraction descriptor in three categories: Appearance-based method, Feature-based method, and Hybrid method.

Appearance-based approach

Inside the Appearance-based category, there are two types of approach: global appearance approach and local appearance method.

Global appearance-based approaches are the first developed methods for thermal face recognition. These approaches reuse essentially early technologies of visible face recognition like PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis) to project the thermal image into a high-dimensional vector space. Most of the early methods follow the work of Prokoski et al. and Socolinsky et al. [75] The approach was enhanced by Hermosilla et al.[76] and by Desa et al. [77] who applied KPCA (Kernel PCA) and KLDA (Kernel LDA) methods to thermal image respectively. One drawback of these approaches is that they require a huge number of samplings to maintain the accuracy of the covariance matrix.

Cutler described in his study an application of eigenfaces in infrared face recognition using a database of 24 persons at three viewpoints (frontal, left and right profiles) and two facial expressions (thus 288 images in total) [78]. This database is taken under SWIR and MWIR sub-bands. Moulay et al. [79] also proposed a face recognition framework using probabilistic Bayesian and SVM on Equinox and Laval University multispectral face databases. However, in the same study, the author reported that the best result is obtained by LDA.

Wu et al. proposed in [80] an architecture using CNN (convolutional neural network) for face recognition. The author wants to replace traditional methods using hand-crafted features that need a great work for feature selection and extraction by an auto process. CNN method is also applied experimentally on RGB-D-T face database . Simon et al. state that CNN architecture can achieve higher accuracy in face recognition than other traditional features like LBP or HOG (Histograms of Oriented Gradients) [81]. Using CNN method for NIR image, NIRFaceNet [82] proposed by Peng et al. reached 98,48% in terms of recognition rate.

Another work studied by Orji et al. [83] combined CNN and FWT (Fast Wavelet Transform) in order to form a deep neural network of 6 layers (one input, two convolution layers, two subsampling layers and an output). The author also compared the result between with and without the preprocessing using PCA and LDA. In their work, Kwasniewska et al. [84] used deep neural network Inception v3 for face detection and human tracking with a low-resolution thermal image acquired by a portable camera.

Also based on deep learning, Sarfraz et al. present in their work [85] a thermal/visible cross face recognition methods using DPM (Deep Perceptual Mapping). This study has the most extensive use case among all the thermal face recognition approach. The authentication system can match a thermal signature an object to its visible image database which is usually already available.

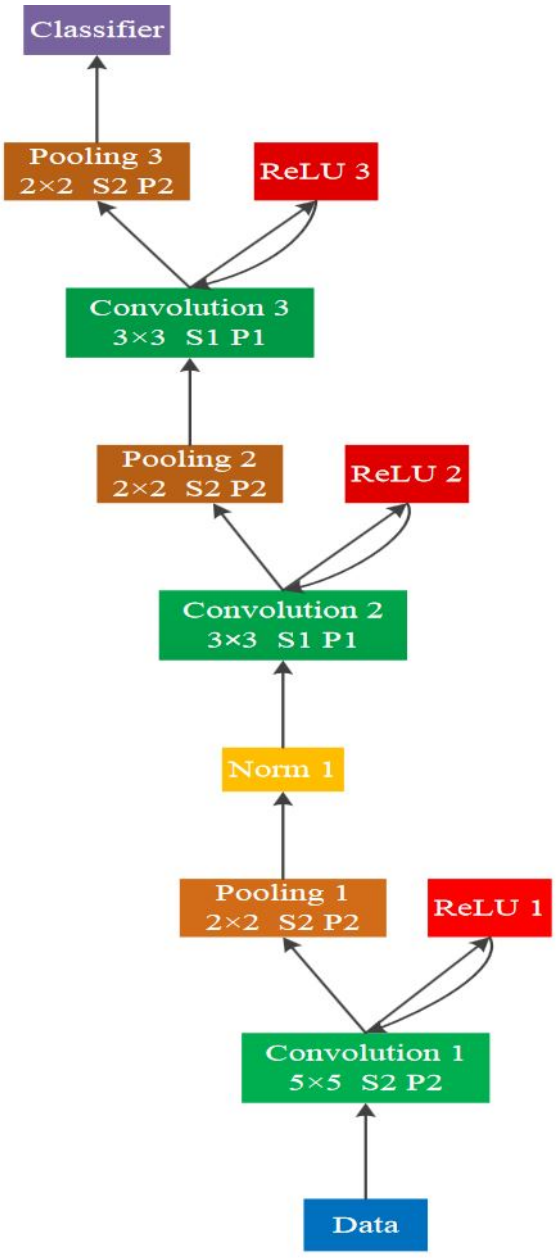


FIGURE 2.25: CNN structure for thermal face recognition

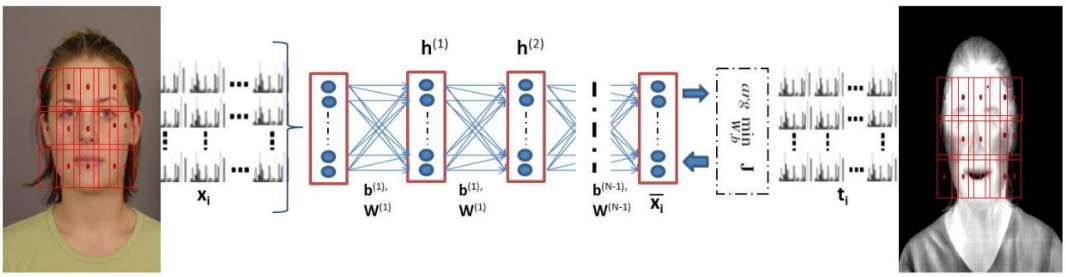


FIGURE 2.26: Thermal/visible cross matching using DPM

Local appearance approach considers the thermal image as a pseudo-type of visible images. It applies to some well-known local transformation like LBP (Local Binary Pattern), SIFT(Scale-Invariance Feature Transform), SURF (Speeded Up Robust

Features), WLD(Weber Linear Descriptor), GJD (Gabor Jet Descriptor) to describe the information of thermal images [76].

Li et al. [86] proposed an infrared face recognition method based on LBP under NIR sub-bands. In this study, the author trait NIR image as a pseudo-visible image and reuse standard face recognition technology to operate the classification. In order to deal with the illumination issue, the author proposed a schema of NIR image detector device which reduces the influence of lighting condition on face images.

In another study, Mendez et al. [87] also used LBP representation for LWIR images. They also indicated that LBP is robust against fixed-pattern noise so that not only no noise suppressing process is necessary, but also non-uniformity correction is not needed. Xie et al. [88] enhanced the method of applied joint encoding of multiscale LBP. This approach considers the correlation in divers microstructures using a co-occurrence matrix of multiscale LBP. This method can achieve 91.2% in accuracy under standard heating conditions which outperform classic LBP-based methods.

Other studies introduced by Wang [89] and by Majumder [90] used Gabor transformation as a feature generator. However, these methods are limited by the fact that thermal imaged have fewer details than visible ones and therefore it's hard to recognize the face.

Feature-Based Method

Feature-Based Method studies some unique literal features of thermal images which do not appear in visible images. The blood perfusion model developed is a very example of the Feature-based approach category. Seal et al. stated that thermal face recognition could exploit local temperature changes in the face image [91]. The heat imbalance region represents anatomical information because of the heating effect caused by the blood flow under the skin. The author indicated that this imbalance could be observed as texture features and could be extracted by Haar wavelet transform.

In another recent study [92], Xie et al. indicated that veins structure induces a unique thermal signature of the face which is similar to a fingerprint. A feed-forward back propagation neural network with five layers was used in classification phases to obtain 95.24% in terms of accuracy. The author also highlights that segmentation preprocessing such as DAD (directional anisotropic diffusion) or region growing is required to extract blood perfusion features.

The blood perfusion model is appreciated for its simplicity in implementation and its robustness against various changes such as aging or illness. This thermal signature is not only independent of face geometric but also impossible to be spoofed. The main drawback of this model is that the resolution required of the image has a minimum, under this limit, the process cannot operate normally. The infrared camera has to be closed to user's face that makes this technique more appropriate for authentication than passive identification.

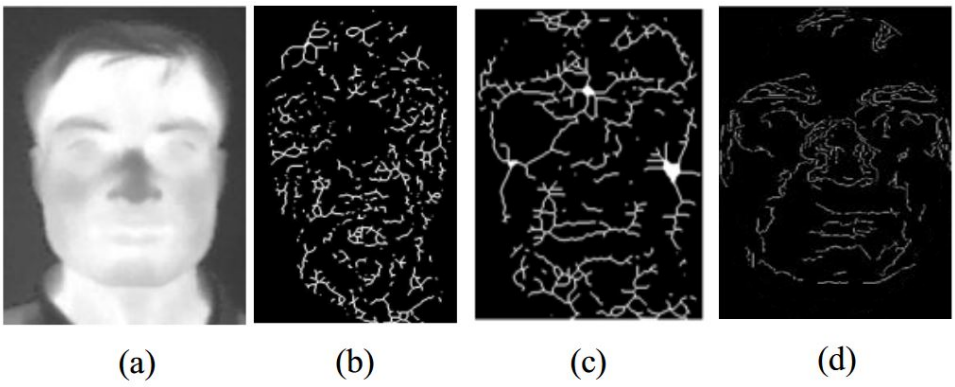


FIGURE 2.27: Blood perfusion model

The vascular network studied by Buddharaju [93] and later by Ghiass et al. [68] extracts blood vessels from an image using morphological filters. These approaches are proven to be effective and robust as a recognition method but they suffer from a high sensibility to normalization process.

Hybrid Method

Despite a lot of advantages, the accuracy rate of infra-red face recognition is far lower than the performance of standard visible imagery. In order to enhance the reliability of biometrics systems, research community tends to combine thermal imaging with other technologies and form hybrid solutions. In this category, many thermal/visible fusion methods, multi-spectral methods, multimodal methods are mentioned as follows.

In their work [94], Bourlai et al. investigate the advantages and limitations of cross-matching from SWIR, MWIR or NIR to visible imaging. The author state that long distant cast a severe consequence upon thermal face recognition rate. Their experiments indicated that cross-spectral matching is a tough challenge which demands further investigation.

Zhang et al. considered in their work [95] that direct application of visible face recognition model into thermal spectrum do not reach a satisfactory performance. Therefore, they proposed the TV-GAN (Thermal-to-Visible Generative Adversarial Network), a transformation technique allows obtaining a pseudo-visible image corresponding to the original thermal image. The transformation is said to be able to conserve enough of identification features to operate a visible face authentication.

Saxena et al. evaluated the possibility to used features from a CNN pre-trained on standard visible images in heterogeneous face recognition [96]. After having explored various learning strategies with different modalities, the author state that Near Infrared (NIR) image can recognize using the CNN pre-trained features of visible spectrum images.

FIGURE 2.28: TV-GAN generates Visible image from IR image

In a controlled environment, Kim et al. introduced a 3D multi-spectrum sensor system which contains three types of sensors: standard visible, thermal-IR and time-of-flight (ToF) [97]. They evaluated different sensor combinations to determinize the optimal one. With the proposed system, four different kinds of data could be obtained: 3D depth data from the ToF camera; near-infrared data from the ToF camera; visible (RGB) data from the color camera and thermal-IR data from the thermal-IR camera. The system could generate 3D multi-spectrum data including the 3D model of the head which can be used in the classification phase. The method reached 98.8% in experimentation with lighting variation and 98.4% with pose variation.

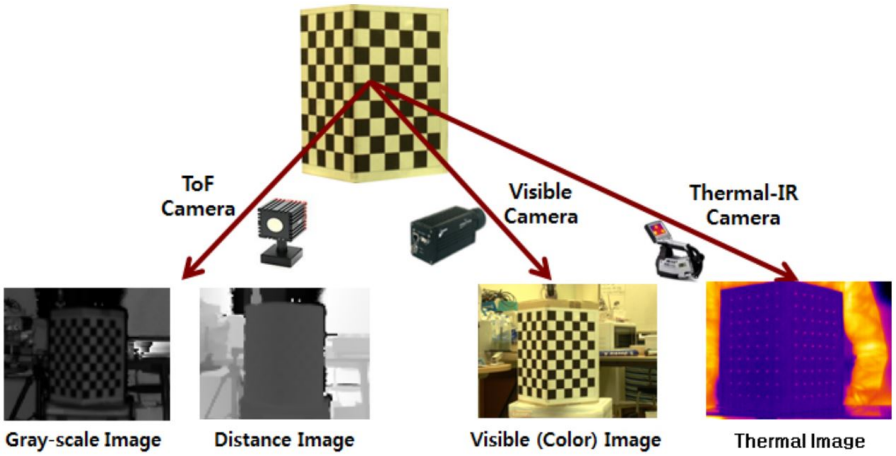


FIGURE 2.29: 3D multi-spectrum sensor system

2.2.3 Vascular Network

Introduction

The vascular network is the product of anatomical observations in thermal imagery. The key idea of this feature is the higher temperature of the blood vessel in comparing with neighbor region. The method is proposed by [93] and is enhanced by Reza Shoja Ghiass [68] . The so-called vascular network is a map of tubular structures extracted from a thermal image. This type of feature is proven to be a effective transformation in thermal face representation.

Method Details

For each frame F_i $(i = 1, ..., n)$ consider the two eigenvalues λ_1 and λ_2 of the Hessian matrix computed at a certain image locus and at a particular scale s. Without loss of generality let us also assume that $|\lambda_1| \leqslant |\lambda_2|$ The two key values used to quantify how tubular the local structure at this scale is are $\mathbf{R_A}$ and $\mathbf{S}$:

$$\begin{aligned} \mathbf{R_A} &= \frac{|\lambda_1|}{|\lambda_2|}, \\ \mathbf{S} &= \sqrt{\lambda_1^2 + \lambda_2^2} \end{aligned} \tag{2.1}$$

The former of these measures the degree of local "blobiness".If the local appearance is blob-like, the Hessian is approximately isotropic and $|\lambda_1| \approx |\lambda_2|$ making $\mathbf{R_A}$ close to 1. For a tubular structure $\mathbf{R_A}$ should be small. On the other hand, $\mathbf{S}$ ensures that there is sufficient local information content at all: in nearly uniform regions, both eigenvalues of the corresponding Hessian will have small values. For a particular scale of image analysis s, the two measures, $\mathbf{R_A}$ and $\mathbf{S}$, are then unified into a single vesselness measure:

$$V(s) = \begin{cases} 0 & if \quad \lambda_2 > 0, \\ (1 - e^{-\frac{R_A}{2\beta^2}}) \times (1 - e^{-\frac{S}{2c^2}}) & otherwise, \end{cases} \tag{2.2}$$

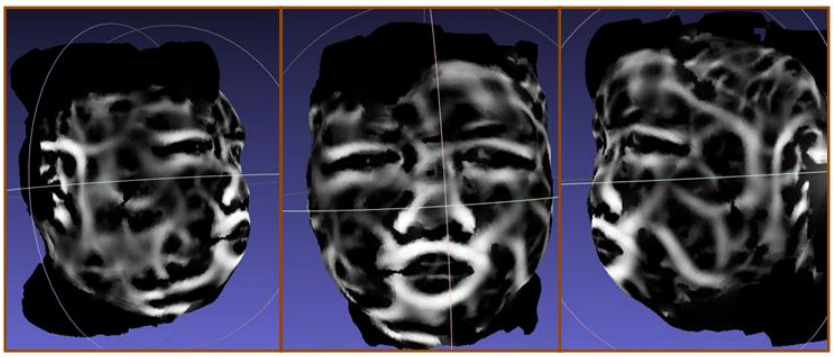


FIGURE 2.30: 3D vascular network model

where β and c are the parameters that control the sensitivity of the filter to R_A and S.

In fact, the "vessel value" of a pixel is represented by the measure $V(s)$ of the $(6s+1) \times (6s+1)$ block centered at this pixel. Finally, if an image is analyzed across scales from s_{min} to s_{max}, the vesselness of particular image locus can be computed as the maximal vesselness across the range:

$$V_0 = \max_{s_{min} \leq s \leq s_{max}} V(s) \tag{2.3}$$

In the end, each vertex is associated with a value V_0 which presents the vessel probability at this point. Another column V_0 can be added to matrix M:

$$\mathbf{M} = \begin{bmatrix} x_1 & x_2 & ... & x_N \\ y_1 & y_2 & ... & y_N \\ z_1 & z_2 & ... & z_N \\ V_0(1) & V_0(2) & ... & V_0(N) \end{bmatrix} \tag{2.4}$$

For each intensity image, the poses, positions and contributions to the 3D model is computed under a texture map. By using this texture map, these vascular networks can be projected to the 3D model in order to form a 3D vessel model which represents the 3D coordinates of vessel features (Fig.5.5).

Advantage

In our studie, Vascular Network is highly appreciated as a feature extraction approach thanks to its advantages compared to other methods.

The vessel features are robust against the change of image scale. The size of the user's face inside an input image or video cannot be predicted. Almost all the global and local appearance-based methods are highly dependent on image's resolution. A lot of features disappear when there are not enough pixels describing them.

The vessel features are also less affected by ageing or illness. There is not any significant change in a user's vascular network except the case of plastic surgery. But even in such extreme case, it is practically impossible to mimic the vessel map of someone else.

The Vascular Network is lately mentioned in the fifth chapter of this thesis as the primary feature extraction of the thermal image. The next chapter looks at the 3D model of user's face and its reconstruction by various techniques. A methodology using minimal equipment to obtain 3D data will be described at the end of that chapter. This little restraint allows a broad range of applications for this methodology.

Chapter 3

3D Reconstruction

3.1 Introduction

3D reconstruction is the domain appearing to respond for the need of capturing and recognizing the 3D geometric form of a subject. 3D reconstruction has numerous applications in various areas such as:

- Computer vision : object description for augmented reality, motion capture for body tracking, robotics mapping ...
- Medicine : organ scanning and modeling.
- Entertainment: filming and gaming.
- Archaeology: visualizing constructions and objects.
- Security: face recognition, fingerprint recognition, human tracking, video surveillance.

In this chapter, we will examine the capacity of representing users' face data by its 3D model. The following section synthesizes the state of the art of 3D reconstruction method using various sensor and technology. The last section introduces our scheme to obtain this model from a single video of user's face.

3.2 Existing Methods

3D reconstruction is the process of capturing the shape and appearance of real objects. This process can be accomplished either by active or passive methods. In passive methods, the number of cameras and images used in the process divides this type into: Monocular Cues Methods, Binocular Stereo Vision and Structure From Motion.

3.2.1 Active Method

Active methods, i.e. range data methods, using the depth map, rebuild the 3D surface by digital approximation method and reconstruct the object in scenario based on the model . These techniques actively interfere with the rebuild object, either

mechanically or radiometrically using rangefinders, to obtain the depth map, e.g. structured radiation, laser range finder, and other active sensing methods. A simple example of a mechanical technique would employ a depth gauge to estimate a gap to a rotating object placed on a turntable. More relevant active scanners release some radiation, beam or light and catch its reflection or diffraction passing through the object to probe an object or environment. Examples cover from colored visible light, time-of-flight lasers, moving light source to ultrasound and microwaves [98, 99, 100].

Time-of-flight : The time-of-flight laser scanner is an active scanner that employs laser beams to examine the subject. At the core of this kind of sensor, a time-of-flight laser range finder resides. The laser range finder spots the distance of a facade by measuring the round-trip period of pulsation of light. A laser is utilized to release a pulse of light and the amount of time before a detector recognizes the reflected light is measured. Because the speed of light c is constant, the round-trip interval is enough to calculate the travel distance of the light, which is twice the gap between the surface, the scanner. If t is the round-trip interval, then the distance can be calculated by $c \times t/2$. The exactitude of a time-of-flight 3D laser scanner relies on how accurate is the measured time t : 3.3 picoseconds (approx.) is the time needed for light to move 1 millimeter.

The laser range finder only estimates the distance of one object in its direction of view. Thus, the scanner examines point by point its entire field of view by turning the range finder's direction of view to examine various points. The view orientation of the laser range finder can be modified either by pivoting the range finder itself or by employing a set of turning mirrors. The latter method is usually applied since mirrors are much lighter and can thus be rotated much quicker and with higher precision. Standard time-of-flight 3D laser scanners are able to estimate the distance of 10,000$\tilde{1}$00,000 points per second.

Triangulation based 3D laser scanners are also active scanners that employ laser radiation to examine the environment. Concerning time-of-flight 3D laser scanner, the triangulation laser irradiates a laser on the object and utilizes a camera to localize the laser dot. Depending on the distance between the object's surface and the camera, the laser mark rises at various places in the camera's range of view. This method is named triangulation because of the triangle created by the laser emitter, the laser dot, and the camera. The length of one side of the triangle, the one between the laser emitter and the camera can be determined. The direction of the laser emitter corner is also defined. The angle of the camera corner can be defined by examining the position of the laser dot in the camera's range of view. These three sets of information entirely limit the contour and dimension of the triangle and provide the place of the laser dot angle of the triangle. In most circumstances, a laser line, rather than a single laser dot, is swept over the object to accelerate the scan process.

Time-of-flight and triangulation range finders each possess advantages and disadvantages which make them favorable for different circumstances. The strength of

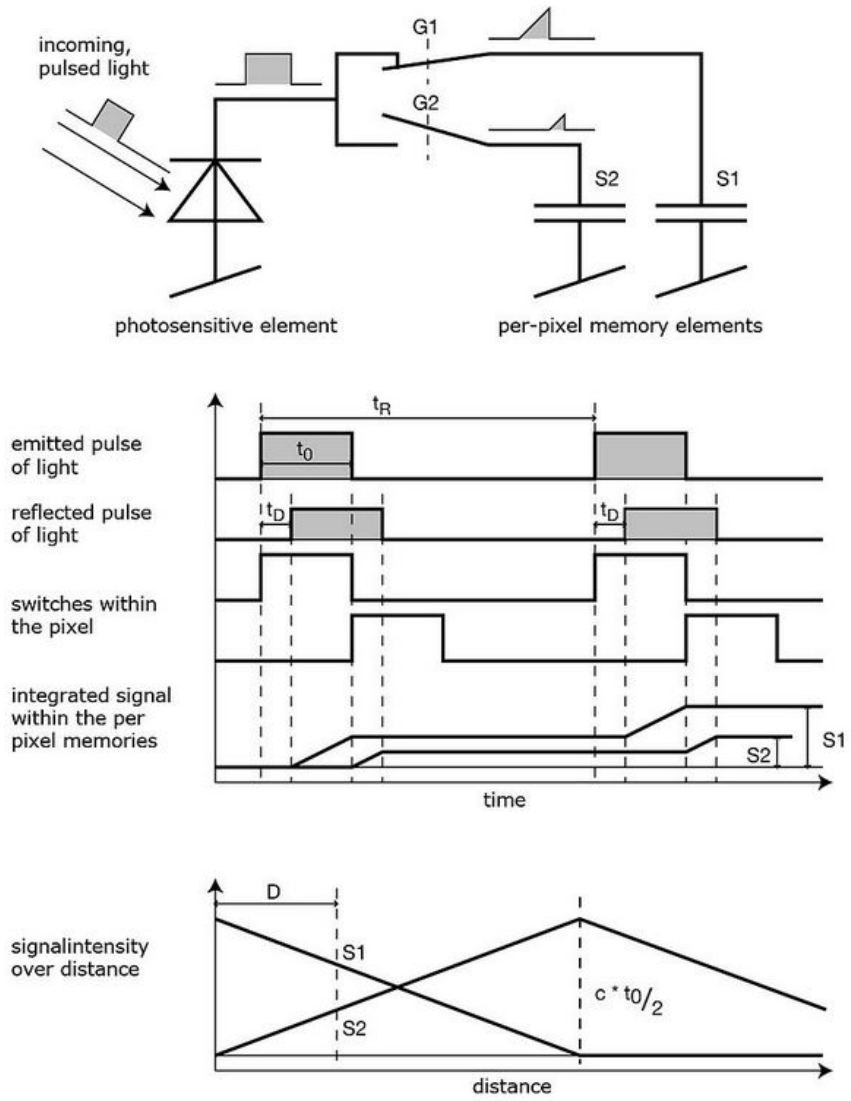


FIGURE 3.1: Function of Time-of-flight 3D laser scanner

TOF range finders is that they can operate across really long distances, on the scale of kilometers. These machines are thus fitting for examining huge structures such as buildings, houses. The problem of TOF range finders is their precision. Because of the high speed of the laser beam, measuring the round-trip time is challenging, and the exactitude of the length measurement is very limited, on the order of millimeters. Triangulation range finders are precisely the opposite. They hold a limited range, on the level of meters, but their precision is relatively high. The accuracy of the triangulation range finders is on the scale of some micrometers.

Time-of-flight scanner's precision can reduce when the beam knocks the edge of an object since the data which is transmitted back to the machine is from two distinct places for one laser pulsation. The coordinate corresponding to the machine's location for a point that has stricken the edge of an object will be determined using a mean and hence will set the point in an incorrect position. When utilizing a high-resolution scanner on a subject, the possibilities of the laser striking an edge are

grown, and the resulting information will expose noise just behind the sides of the object. Machines including a smaller laser width can solve this problem, but they are restricted by the range since the laser width increments over distance. The algorithm can also help by determining that the first object to be stricken by the beam should eliminate the second.

At a speed of 10,000 points per second, low-resolution scans can use less than one second, but high-resolution scans, demanding millions of samples, can use several minutes for some TOF Machine. The difficulty that produces is distortion from the movement. Since each point is examined at a different time, any movement in the object or the machine will distort the obtained information. Thus, it is regularly required to fix both the object and the device on stable stands and reduce vibration. Utilizing these machines to scan objects in movement is pretty tough.

Lately, there have been studies on counterbalancing for distortion from tiny amounts of vibration and distortions due to movement and rotation . When examining in one area for any length of time slight change can happen in the machine position due to variations in temperature. If the machine is arranged on a tripod and there is strong daylight on one view of the device, then that side of the tripod will expand and slowly distort the obtained information from one view to another. Some laser machines have scale compensators included inside them to offset any motion of the device during the scan time.

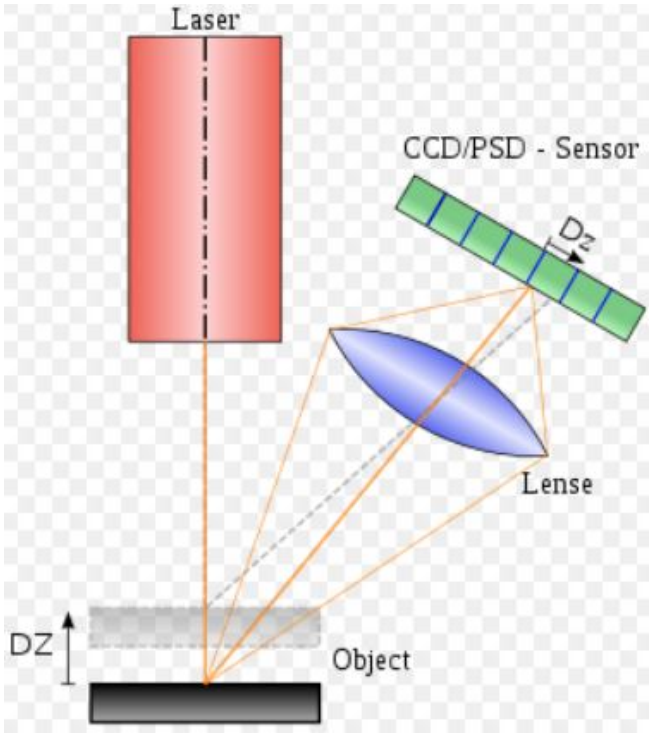


FIGURE 3.2: Principle of a laser triangulation sensor.

Hand-held laser scanners : Hand-held laser scanners generate a 3D model using the triangulation technique introduced before: a laser dot or line is projected onto

an object from a portable machine, and a sensor (typically a charge-coupled sensor or position sensitive sensor) estimates the gap to the surface [101, 102, 103]. Information is retrieved using an inner coordinate system and therefore to obtain information when the device is in motion the position of the device must be defined. The location can be determined by the device employing indicating points on the surface or by utilizing an external tracking technique. External tracking usually takes the appearance of a laser tracker (to give the device location) with integrated camera (to determine the direction of the device) or a photogrammetric method employing 3 or more cameras implementing the entire six degrees of freedom of the device. Both methods conduce to utilize infrared light-emitting diodes associated with the machine which are observed by the camera(s) through filters giving resilience to ambient lighting.

Information is retrieved by a machine and registered as points cloud in three-dimensional coordination, with processing this can be reformed into a triangulated mesh and then a computer design model, usually as non-uniform rational B-spline surfaces. Hand-held laser devices can join this information with passive, visible-light cameras which capture an object's textures and colors to reconstruct a complete 3D model.

Structured-light 3D scanner : emit a pattern of light beams on the object and study the deformation of this pattern. The pattern is projected onto the object employing either an LCD projector or another constant light source. A camera, offset slightly from the projector device, studies the appearance of the pattern and measures the distance of every dots in the range of view.

Structured-light scanning is still a very active area of research with many research papers published each year. Perfect maps have also been proven useful as structured light patterns that solve the correspondence problem and allow for error detection and error correction.

The advantage of structured-light 3D scanners is speed and precision. Instead of scanning one point at a time, structured light scanners scan multiple points or the entire field of view at once. Scanning an entire field of view in a fraction of a second reduces or eliminates the problem of distortion from the motion. Some existing systems are capable of scanning moving objects in real-time. VisionMaster creates a 3D scanning system with a 5-megapixel camera so 5 million data points are acquired in every frame. A real-time scanner utilizing digital fringe projection and phase-shifting method (certain types of structured light techniques) was developed, to catch, build, and render high-density features of a dynamically deformable subject (like facial expressions) at the speed of 40 frames every second. Lately, another device has been developed. Various patterns can be used to this method, and the frame rate for capturing and data processing reaches 120 frames every second. It is also able to examine isolated surfaces, for example, two swaying hands. By using the binary defocusing method, speed breakthroughs could achieve millions of

frames every tens second.

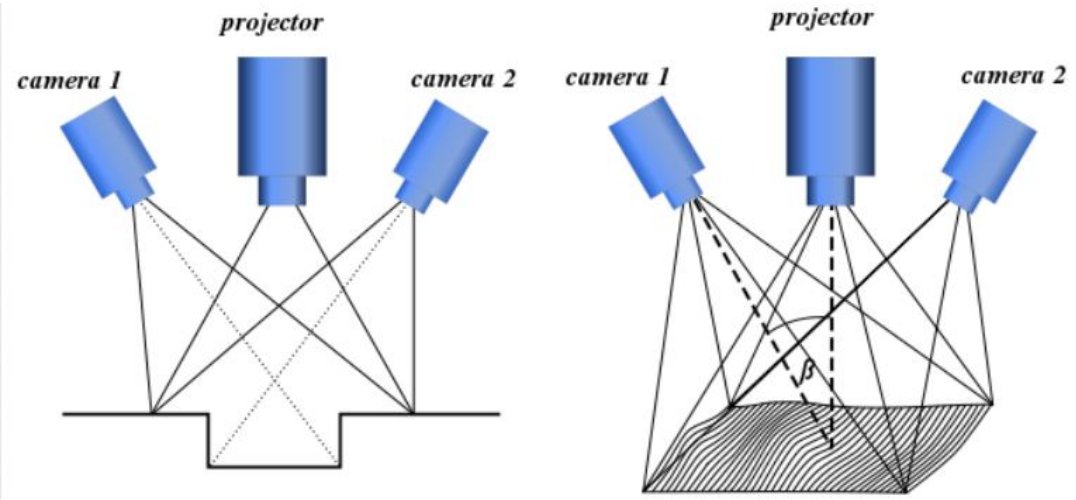


FIGURE 3.3: Fringe pattern recording system with 2 cameras (avoiding obstructions)

Two principal methods of stripe pattern generation have been developed: Laser interference and projection. The laser interference technique operates with two broad planar laser beam fronts. Their interference issues in regular, equidistant line patterns. Multiple pattern sizes can be generated by adjusting the angle between these lasers. The system allows for the precise and easy generation of very fine patterns with unlimited depth of field. Drawbacks are the high cost of implementation, complexities giving the ideal beam geometry, and common laser effects such as speckle noise and the potential self-interference with beam parts reflected from subjects. Usually, there is no means of modulating individual stripes, such as with gray codes.

The projection technique uses incoherent light and primarily operates as a video projector. Patterns are typically produced by emitting light through a digital spatial light modulator, usually using one of the three currently most popular digital projection method:digital light processing (DLP; moving micromirror) modulators, transmissive liquid crystal or reflective liquid crystal on silicon (LCOS) , which have many comparative strength and weaknesses for this purpose. Other techniques of projection could be and have been applied, however.

Patterns created by digital display projectors have little discontinuities due to the pixel boundaries in the displays. Adequately small boundaries, however, can reasonably be neglected as they are evened out by the smallest defocus. A standard measuring assembly including one projector and at least one camera. For several applications, two cameras on opposing sides of the projector have been established as useful.

Invisible (or imperceptible) structured light utilizes structured light without interfering with other computer vision tasks for which the projected pattern will be complicated. Example techniques involve the use of infrared light or extremely high frame rates shifting between two exact opposing patterns.

Geometric distortions by optics and perspective must be offset by a calibration of the measuring devices, using particular calibration patterns. An analytical model is applied for representing the imaging features of the projector and cameras. Primarily based on the simple geometric properties of a pinhole camera, the model also has to take into account the geometric distortions and optical irregularity of projector and camera lenses. The parameters of the camera, as well as its direction in space, can be defined by a series of calibration analyses, utilizing photogrammetric bundle adjustment.

There are various depth cues included in the detected stripe patterns. The displacement of any single stripe can immediately be transformed into 3D coordinates. For this purpose, the individual line has to be recognized, which can, for example, be achieved by tracking or counting stripes (pattern recognition technique). Another standard method projects alternating line patterns, resulting in binary Gray code chains classifying the number of each stripe striking the object. An important depth cue is also produced from the changing stripe widths along the subject surface. Stripe width is a function of the steepness of a surface part, i.e., the first derivative of the elevation. Finally, the wavelet transforms recently are considered for the same goal.

In many practical implementations, sets of measures combining pattern recognition, Fourier transforms, and Gray codes are obtained for a full and unambiguous reconstruction of shapes. Another approach also relating to the area of the fringe projection has been described, employing the depth of the field of the camera. It is also possible to apply projected patterns primarily as a method of structure insertion into scenes, for an essentially photogrammetric acquisition. The optical resolution of fringe projection techniques depends on the width of the stripes related and their visual nature. It is also restricted by the wavelength of light.

Popular optical stripe pattern profilometry hence permits for feature resolutions down to the wavelength of light, under one micrometer, and with broader stripe patterns, to approximate 1/10 of the stripe dimension. Concerning level precision, interpolating over many pixels of the received camera image can produce a stable high resolution and also accuracy, down to 1/50 pixel. Arbitrarily big objects can be measured with correspondingly broad stripe patterns and structures. Practical implementation is documented concerning objects of which size can be several meters.

As with all optical methods, reflective or transparent surfaces raise difficulties. Reflections cause light to be reflected either away from the camera or right into its optics. In both cases, the dynamic range of the camera can be exceeded. Transparent or semi-transparent surfaces also cause major difficulties. In these cases, coating the surfaces with a thin opaque lacquer just for measuring purposes is a common practice. A recent method handles highly reflective and specular objects by inserting a 1-dimensional diffuser between the light source (e.g., projector) and the object to be scanned. Alternative optical techniques have been proposed for handling perfectly transparent and specular objects.

Double reflections and inter-reflections can cause the stripe pattern to be overlaid with unwanted light, entirely eliminating the chance for proper detection. Reflective cavities and concave objects are therefore difficult to handle. It is also hard to handle translucent materials, such as skin, marble, wax, plants and human tissue because of the phenomenon of the subsurface scattering. Recently, there has been an effort in the computer vision community to handle such optically complex scenes by redesigning the illumination patterns. These methods have shown promising 3D scanning results for traditionally difficult objects, such as highly specular metal concavities and translucent wax candles.

3.2.2 Monocular Cues Methods

The monocular cues systems indicate to use images (one, two, three or more) from one viewpoint (camera) to proceed 3D reconstruction. It makes use of 2D characteristics(e.g. Silhouettes, shading, and texture) to measure 3D form, and that is the reason, for which it is also entitled Shape-From-X, where X can be silhouettes, motion, contour, shading, texture, etc. 3D reconstruction by monocular cues is quick and straightforward, and only one suitable numerical image is required thus only one camera is sufficient. Technically, it eludes stereo correspondence, which is moderately complicated.

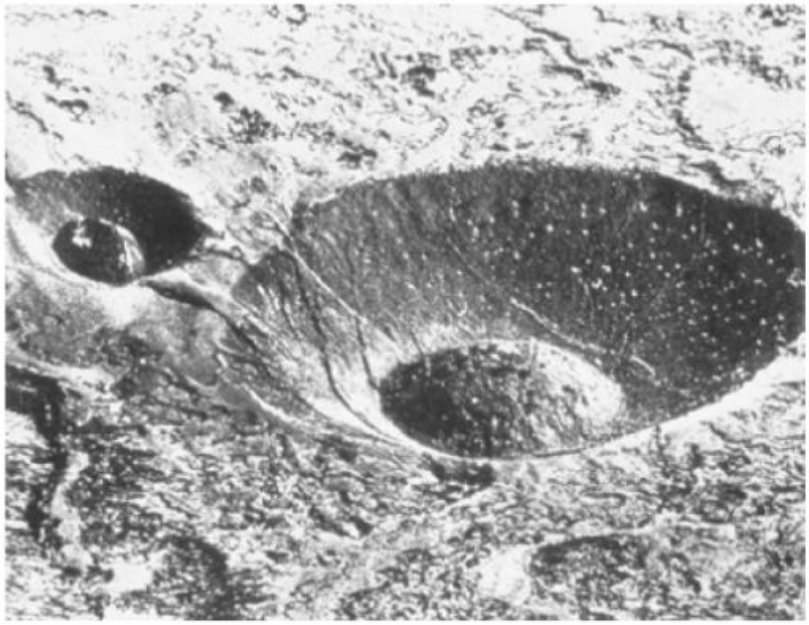


FIGURE 3.4: Example of Shape From Shading ambiguities

Shape From Shading is the method of estimating the three-dimensional shape of an object from one image of its surface. Contrary to most of the other three-dimensional reconstruction problems (such as stereo and photometric stereo), in the Shape of Shading problem, information is minimal.

Since 70's, the first study [104] by Horn introduced the Shape of Shading problem directly and rigorously as determining the solution of a nonlinear first-order Partial Differential Equation (PDE) named the illumination equation. In the later decade, the researchers concentrate on the computational section of the problem, attempting to calculate directly analytical solutions. Topics about the existence and uniqueness of such solutions were completely not even appeared at that moment with the critical exception of the studies of Bruss [105] and Brooks [106]. Because of the lousy quality of the issues, these problems, as well as those related to the convergence of digital schemes for computing solutions, became principal in the last decade of the 20th century.

Now, the Shape From Shading approach is considered as an ill-posed problem. For example, many articles prove that the existing solution is not unique. Such concave/convex ambiguities have usually represented the encountered problems like the one presented in Figure 3.4. In this figure, the ambiguity is due to a variation in the calculation of the parameters of the illumination. In fact, this sort of ambiguity can be usually generalized. Belhumeur et al. [107] show that when the lighting orientation and the Lambertian reflectance of the object are undefined, then the same image can be captured by a connected family of surfaces (relying linearly on three parameters). In other words, they prove that neither shading nor shadowing of a subject, observed from a single image shows its correct 3D structure.

Photometric stereo is a method in computer vision for evaluating the surface normals of a subject by observing that subject under varying illumination conditions. It is built on the basis that the quantity of light reflected from a surface is reliant on the direction of the surface concerning the light source and the observer system. By estimating the quantity of light bounced into a camera, the set of possible surface directions is restrained. Given sufficient light sources from various angles, the surface direction may be limited to a single direction.

Woodham originally proposed the method in 1980. Photometric stereo has since been generalized to multiple other circumstances, including extended radiation sources and non-Lambertian subject. Latter research tries to get the technique work in the appearance of projected shadows, highlights, complex and non-uniform lighting.

The first stage in the evaluation of the normal map is to calibrate the light source by estimating the light orientation. One method to do this is to utilize a chrome ball on which the brightest spot is employed to recognize the orientation of the light.

Shape-from-texture Regarding a pattern with some sort of regularity, or texture, converging on a receding surface, humans can readily recognize the 3D depth of the scene. This fact has long since intrigued many researchers, and studies have been made to reproduce, by a computer, this apparently highly intelligent human capacity. This topic is now generally remembered as 3D recovery of shape from texture.

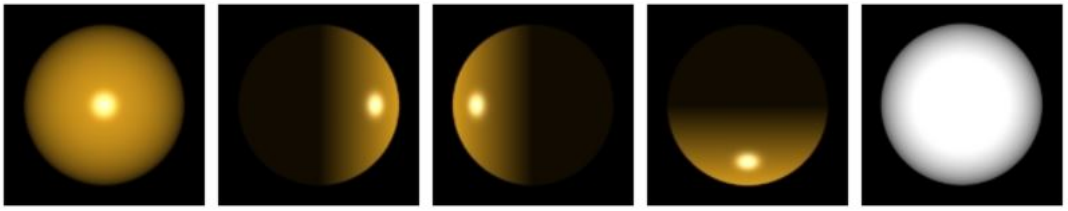


FIGURE 3.5: Synthetic data generated using OpenGL to verify light calibration

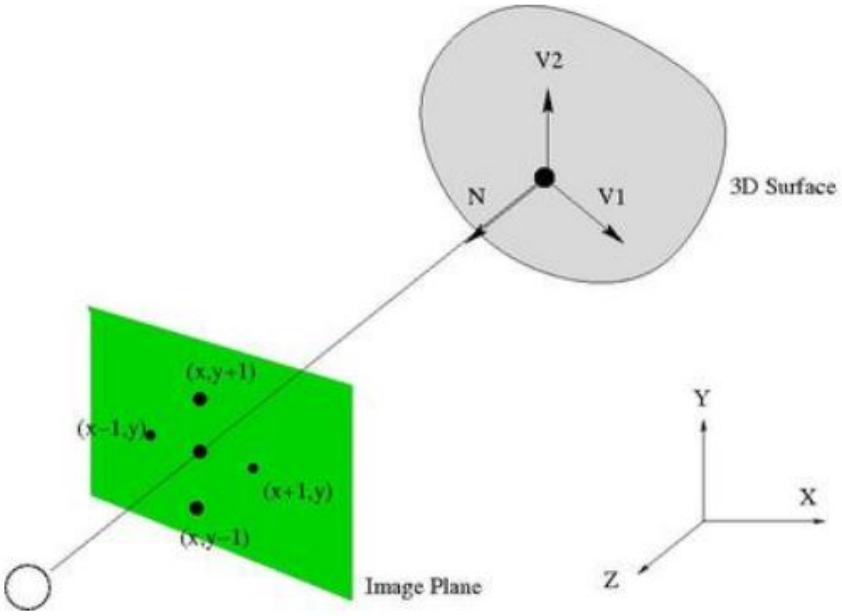


FIGURE 3.6: Depthmap by Photometric stereo

Typically, 3D reconstruction from texture is potential if we have some prior information about the right texture; if the observed surface has characteristics separate from those of the true texture, the 3D shape is calculated in such a way that the inconsistency is accounted for. For example, if the true texture is given to be an arrangement of components with a given shape, say circular, the patent gradient can be deduced from the perceived distorted shape, say elliptical, of the components. If the true texture components are given to be periodically distributed at periods of the same interval, the patent gradient can be evaluated from the rate of the converging period lengths. If the true texture components are aligned on parallel lines, or if individual texture components have parallel line sections, the surface gradient is induced from the fading points determined by pairs of such lines. Similar logic is reasonable if the true texture elements or their alignments are given to hold orthogonality or symmetry of some sort. One significant problem about these methods is that we must first identify the structure of the real texture regularity, periodicity, collinearity, parallelism, orthogonality or symmetry. This is in overall very challenging to automate by a computer as the perceived texture does not show the expected regularity,

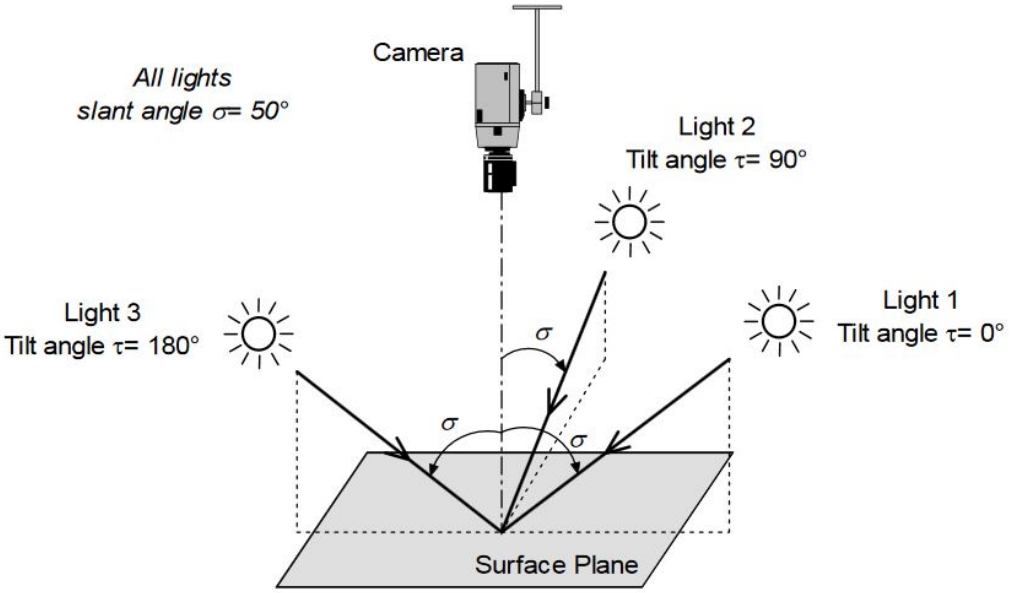


FIGURE 3.7: Photometric stereo's schema

periodicity, etc.

Despite this setback, these structure-based methods have been tried by many researchers. Perhaps this is because humans apparently seem to apply this type of reasoning: identification of such texture object is very natural for humans. Then, a new method which does not demand the identification of texture structure arrived. It is based on statistical hypotheses about the real texture population. For example, if the real texture is distributed isotropically, i.e., the line sections creating the texture possess no favored directions, the 3D surface shape can be calculated from detected favored directions. This method was first introduced by Witkin, and the process was developed by Davis et al. Kanatani provided an accurate numerical representation of the problem and explicit mathematical formulae by requesting tensor calculation and stereology.

Another potential statistical hypothesis is uniformity. When perceived, the texture seems dense on the surface part far away from the device and sparse on the region near the device. This phenomenon has also been acknowledged to play an essential role in human recognition of the external world (Gibson, Sedwick) and many tries have been done to mimic this effect by a computer. However, most of the reasons were based on natural inspiration or heuristics (e.g., Bajcsy and Lieberman, Rosenfeld, Zucker et al.). It was not until Aloimonos and Swain and Dunn that the question was handled in analytical expressions based on the imagery geometry of perspective projection. However, their creations include several ad-hoc approximations and hypotheses.

3.2.3 Binocular Stereo Vision

Binocular Stereo Vision receives the 3-dimensional geometric data of a subject from multiple images based on the study of the human visual system. The results are displayed in the form of depth images. Images of a subject obtained by two cameras concurrently in separate viewing angles, or by one single camera at separate times in various viewing angles, are used to reconstruct its 3D geometric data and restore its 3D form and position. This is more direct than Monocular techniques like shape-from-shading.

Binocular stereo vision system needs two similar cameras with parallel optical axis to perceive one same subject, obtaining two photos from different positions of view. In words of trigonometry connections, depth data can be determined from the variation. Binocular stereo vision technique is well matured and contributes to beneficial 3D reconstruction, heading to a more significant performance when compared to other 3D construction techniques. Regrettably, it is computationally expensive, besides it works rather inadequately when baseline distance is considerable.

2D digital image acquisition is the data source of 3D reconstruction. Usually used 3D reconstruction is based on two or more images, although it may use only one image in some circumstances. There are different types of techniques for image acquisition that depends on the circumstances and objectives of the specific system. Not only must the specifications of the system be met, but also the visual variation, brilliance, production of camera and the feature of the scene should be considered.

Camera calibration in Binocular Stereo Vision relates to the measurement of the mapping relationship between the image points P_1 and P_2, and space coordinate P in the 3D acquisition. Camera calibration is a fundamental and necessary part in 3D restoration using Binocular Stereo Vision.

Feature extraction is the method which intends to obtain the properties of the photos, through which the stereo correspondence performs. As a consequence, the properties of the photos approximately connect to the selection of matching techniques. There is no such globally appropriate theory of features descriptor, heading to a large diversity of stereo correspondence in Binocular Stereo Vision study.

Stereo correspondence this is to build the correspondence between fundamental elements in photos, such as to match P_1 and P_2 from two photos. Some interference elements in the background should be remarked, like light, noise, surface physical property, etc.

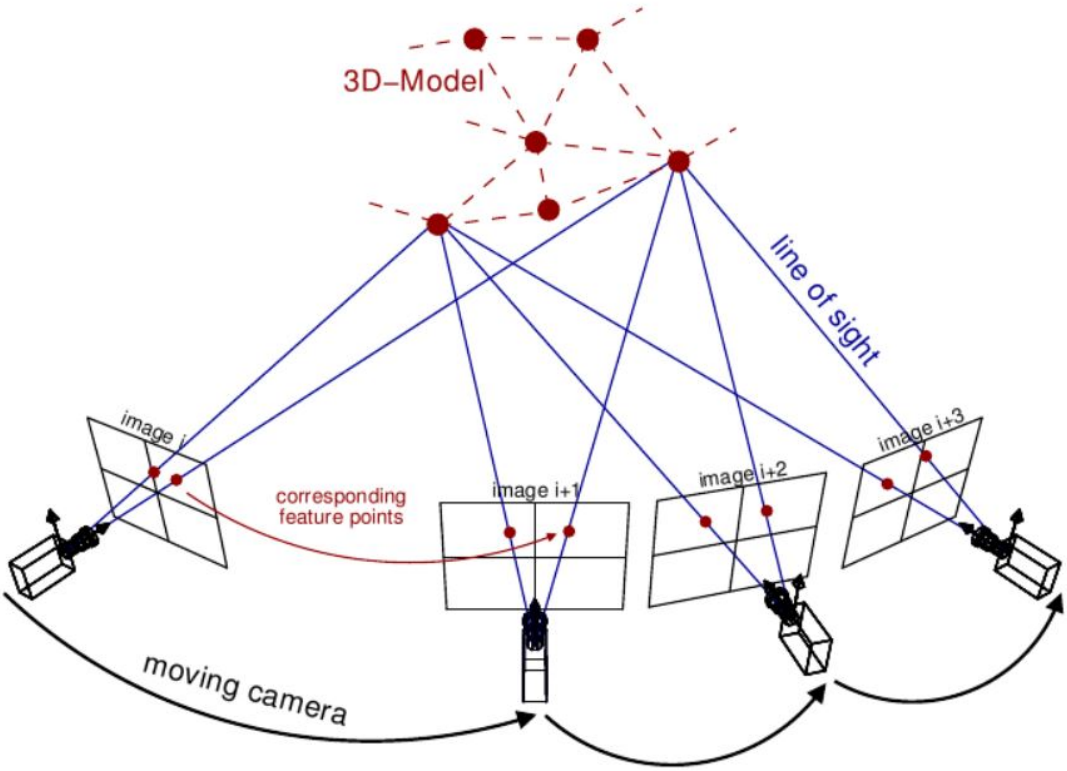


FIGURE 3.8: Structure from motion's principles

3.2.4 Structure From Motion

Structure from motion (SfM) is an imaging method for determining three-dimensional object from two-dimensional photo arrays that may be joined with local movement signals. Humans recognize a lot of data about the three-dimensional object in their surroundings by moving over it. When the person moves and the objects around the person move, data is received from images seen over time. Obtaining structure from motion represents a similar question to obtaining structure from stereo vision. In both cases, the correspondence between photos and the reconstruction of the 3D structure must be discovered.

The structure of an image is a projection from a 3D scene onto a 2D plane, during which the depth data is dropped. The 3D point corresponding to a particular image pixel is constrained to be on the line of view. From a single photo, it is improbable to decide which point on the line corresponds to the image pixel. If two images are accessible, then the position of a 3D point can be located at the intersection of the two projection line. This method is related to as triangulation. The core of this method is the connections between multiple views which conduct the data that corresponding sets of points must include some structure and that this structure is linked to the position and the orientation of the camera.

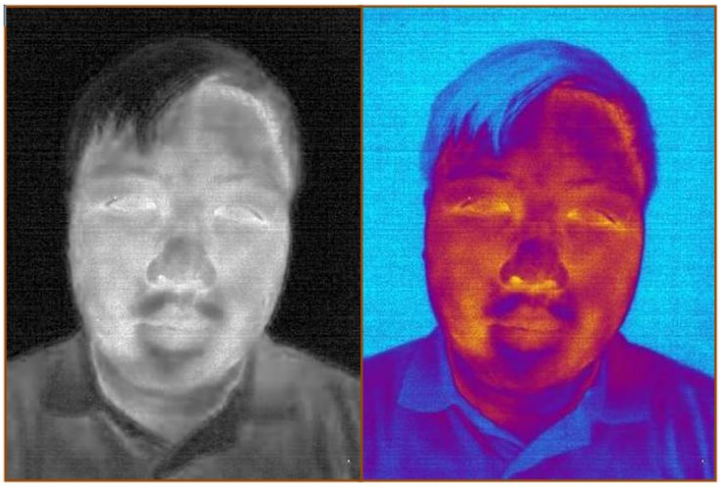


FIGURE 3.9: Gray-scale and iron-palette version

3.3 Scheme

In our study, the input thermal video is supposed to contain many frames from various poses of one head. An algorithm of 3D reconstruction is used to compute a 3D point cloud which describes the head filmed in the video. In this scenario, we use VisualSFM - Structure from motion [108], developed by Changchang Wu , a robust and stable reconstruction algorithm which uses not only the shape but also the color of each pixel to compute the coordination of the face. Therefore, instead of using the gray-scale version of intensity, the experiment observes another color representation of thermal image: the Iron-palette which can be computed from the gray-scale one (Fig.3.9).

These features(common edges and points) are tracked from one frame to next so the position and orientation of each frame can be estimated by geometric calculation. Different views of one point which can be obtained from many consecutive frames are extracted to estimate its deep coordination and then its 3D position.

From the original video, a set of frame F_i ($i = 1, ..., n$ where n is the number of frames) can be extracted. Each frame is an image ($p \times q$ pixels) which contains a view of the face:

$$\mathbf{F_i} = \begin{bmatrix} x^i_{1,1} & x^i_{1,2} & ... & x^i_{1,q} \\ x^i_{2,1} & x^i_{2,2} & ... & x^i_{2,q} \\ ... & ... & ... & ... \\ x^i_{p,1} & x^i_{p,2} & ... & x^i_{p,q} \end{bmatrix} \tag{3.1}$$

In fact, each frame is compared to all other frames by method SIFT (scale-invariant feature transform). Two frames (F_{c_1}, F_{c_2}) which maximize the similarity index are chosen to form the base of 3D object. These common points of F_{c_1} and F_{c_2} will form

a first model which is called ST_2 (ST_1 do not exist).

$$(\mathbf{c_1}, \mathbf{c_2}) = \underset{i \neq j}{\text{argmax}} SIFT(F_i, F_j) \tag{3.2}$$

Then, the process sorts all images in the decreased order of $SIFT(F_{c_1}, F_i)$ which make a complete sequence $c_3, c_4, ..., c_n$. For each frame F_{c_i} ($i = 3, ...n$), the process will firstly try to find the common points with ST_{i-1} which will be called STF_i ($STF_i \subseteq ST_{i-1}$). So we have a certain points STF_i in 3D coordination and its project in the plan of camera: F_{c_i}. Thus the process will try to estimate the orientation and the position of the camera by maximizing the similarity index between the two sets:

$$(R_{c_i}, Pc_i) = \underset{R,P}{\text{argmax}} SIFT(F_{c_i}, T_P(PROJ_R(STF_i)) \tag{3.3}$$

where T_P is the translation operator by vector P and $PROJ_R$ is the projection operator by rotation matrix R.

$$\mathbf{P_{c_i}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3.4}$$

$$\mathbf{R_{c_i}} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} \\ r_{2,1} & r_{2,2} & r_{2,3} \\ r_{3,1} & r_{3,2} & r_{3,3} \end{bmatrix} \tag{3.5}$$

The geometric information of F_{c_i} is represented in position matrix $\mathbf{P_{c_i}}$ and rotation matrix $\mathbf{R_{c_i}}$ (which can be transformed to orientation matrix $\grave{}_\mathbf{i}$)

$$\grave{}_\mathbf{i} = \begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} \tag{3.6}$$

(x,y,z) is the coordination of the center pixel of frame F_{c_i}, $(\theta_x, \theta_y, \theta_z)$ is its orientation respectively in the view of Ox, Oy and Oz.

The process will, then, studies the extra part of frame F_{c_i} which is not exist inside STF_i and try to match it's with some other nearest frame to find additional common points STP_i

$$ST_i = ST_{i-1} \cup STP_i \tag{3.7}$$

The last one (ST_i) becomes the 3D model formed by $\{F_{c_1}, ..., F_{c_i}\}$.

The set ST_n definite the 3D model of all frame. This 3D model form a matrix which represents a cluster of featured points in a Cartesian coordinate system. Notes $\mathbf{M_{ini}}$ the 3D model , it can be represented as:

$$\mathbf{M_{ini}} = \begin{bmatrix} x_1 & x_2 & ... & x_N \\ y_1 & y_2 & ... & y_N \\ z_1 & z_2 & ... & z_N \end{bmatrix} \tag{3.8}$$

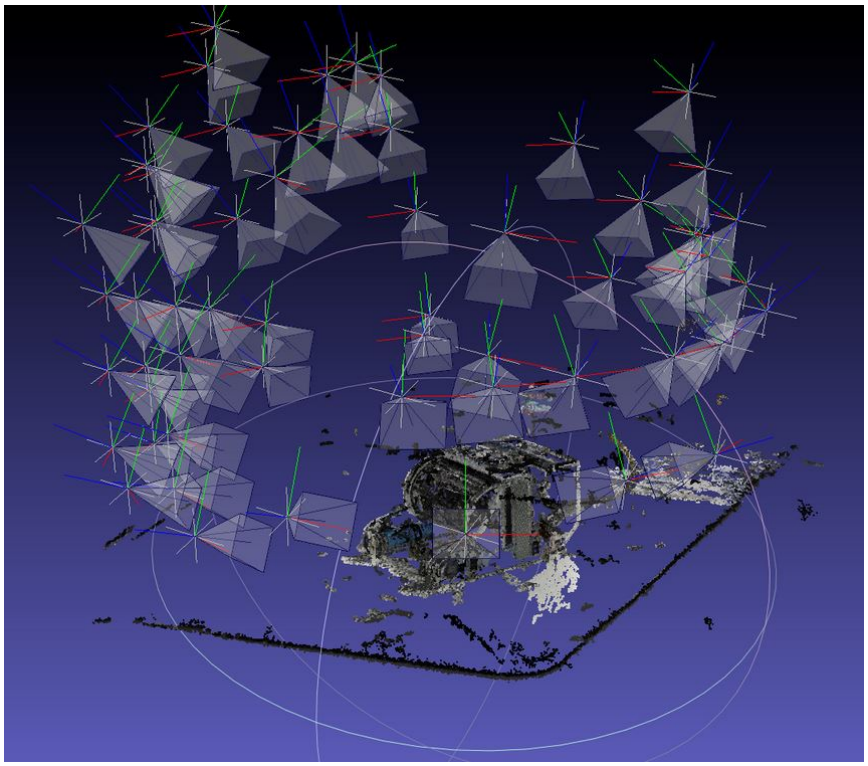


FIGURE 3.10: 3D mesh and camera's position

With color video a RGB matrix $\mathbf{M_c}$ can be associated with $\mathbf{M}$

$$\mathbf{M_c} = \begin{bmatrix} r_1 & r_2 & ... & r_N \\ g_1 & g_2 & ... & g_N \\ b_1 & b_2 & ... & b_N \end{bmatrix} \tag{3.9}$$

where N is the number of feature points and (x_k, y_k, z_k) $(k = 1, ..., N)$ are the coordinates of the k-th point in the space $Oxyz$. (r_k, b_k, g_k) $(k = 1, ..., N)$ represent the pseudo-color (which is computed from intensity) of this point.

The obtained point cloud, which is neither dense nor periodic, must be improved using the Patch-based Multi-view Stereo (PMVS) developed by Yasutaka Furukawa [109] and using the Poisson Surface Reconstruction studied by Michael Kazhdan [110]. By this supplement process, the set of point is transformed into a dense collection of small oriented rectangular. The algorithm of PMVS can be decomposed in 3 steps:

- **Matching**: Pixel-level correspondences of point cloud is computed to enhance a portion of features points. Features found by Harris and difference-of-Gaussians operators are first matched across multiple pictures, yielding a sparse set of patches associated with salient image regions
- **Expande**: Initial matches is spread to nearby pixels to obtain a denser cloud of points.

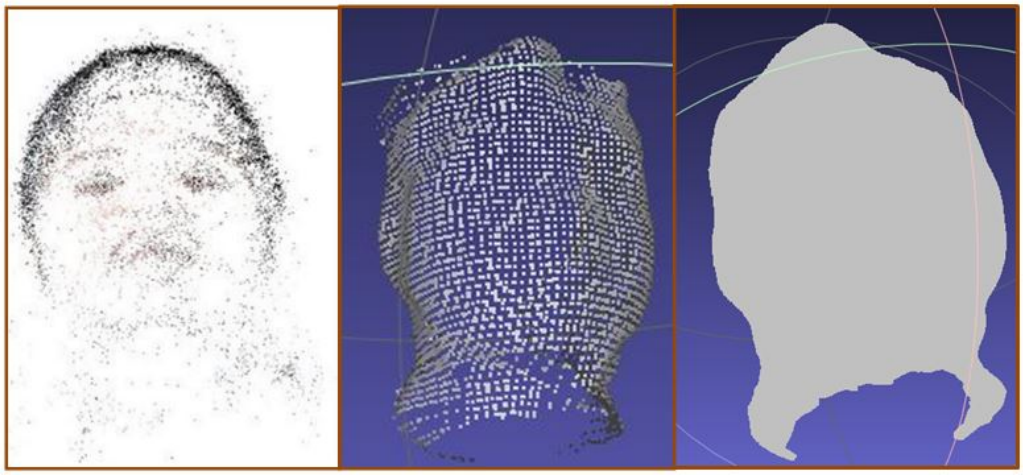


FIGURE 3.11: point cloud (left), dense vertex (middle), surface (right)

- **Filtering**: Intensity constraints (and a weak form of regularization) are used to eliminate incorrect matches.

These 3 steps are repeated for several iterations until the set of points (the mesh) is dense enough. This mesh can be further refined by a mesh-based MVS algorithm that enforces the photometric consistency with regularization constraints like Poisson Surface Reconstruction . The resolution of the mesh model is adaptive, and the size of a triangle depends on the density of the nearby oriented points: The denser the points are, the finer the triangles become. The PSR software outputs a closed mesh model even when patches are only reconstructed for a part of a scene. In order to remove extraneous portions of the mesh, we discard triangles whose average edge length is greater than six times the average edge length of the whole mesh since triangles are large where there are no points.

So, in fact, after the PMVS process, a matrix $\mathbf{M}$ (having the same structure as $\mathbf{M_{ini}}$ but having much more points, here we always use N as the number of points) is obtained. This rebuild 3D model includes a huge number of vertex and triangular surfaces which can be considered as a dense facial surface (Fig.3.11).

This scheme of 3D reconstruction which is applied both in visible imagery and thermal imagery is one of the core features of our study. Geometric data from the 3D model provide robust proof to detect face spoofing attack which is described in the next chapter. On the other hand, depth information extracted from this model is exploited in the fifth chapter to improve the performance of our method of face recognition.

Chapter 4

Face Spoofing Detection Using 3D Model

4.1 Introduction

4.1.1 Problematic and Objectives

Authentication by facial recognition can be exploited as an additional solution to reinforce the security level of our information systems. However, it is proven that this solution is vulnerable. Facial recognition is easily compromised by face spoofing attacks. Therefore, photos and video widely shared on social networks may become a weapon against their owner's security. Attackers have many ways to attack a facial recognition system. They can utilize a photo of a legitimate user printed on a piece of paper or displayed on an LCD screen and present it in front of the camera in operation. They can also replay a video which filmed the victim previously or just use a 3D mask to mislead the face detection process. In 3D-mask attack, attackers have to focus on their target and do firstly manage to construct a 3D mask or maybe a sculpture of the target. If the mask is constructed perfectly, there is less chance to detect it. However, the achievement of this type of attack is quite difficult and expensive. In this paper, the proposed method seeks to detect basically the photo and video-replay attacks.

4.1.2 Contributions

In this study, we construct and investigate a new face-spoofing detection using a 3D model of the head computed from a video captured by the user's smartphone.

- The novel method is designed for one of the most challenging use cases: face recognition using a smartphone which is highly dependent on the user's behavior. However, the study explores some convenient features of smartphone such as the movement capacity and motion sensor to construct an adaptive method.
- The 3D model is super-effective against photo-attack as different in geometric features between a real object and an image is pretty significant.

- The process also compares the prior-motion of the camera and the captured-motion estimated from the input video to justify the credibility of the user. This phase can detect a large portion of video-attack.

4.2 Method Details

In the last few years, we can remark a constant evolution of mobile technology and of the smartphone market. More and more people use smartphones to ease their daily life as well as their professional activities. Myriad mobile applications require or have access to personal or private information of users. Therefore, they need a high level of security. Authentication by facial recognition is proposed as a solution to reinforce the security of mobile systems. However, the problem of face spoofing is always unavoidable. Actual solutions are quite relevant and optimized to settle this problem, but just in some provided cases study. Thus, an efficient solution dedicated to smartphone system is indispensable.

In the case of a smartphone system, which is mobile, images or videos could be captured under different conditions of lighting, under different orientations and with an uncontrollable background. The quality of acquisition could also be affected by the movement of the camera and the movement relative of the acquisition system (e.g. when a user authenticates while he is traveling in a train). In addition, the diversity and the constant evolution of smartphone models, as well as the difficulty in calibrating their cameras, are also among the big barriers for an efficient face spoofing detection solution.

However, the presence of different sensors integrated into a smartphone may be an advantage which allows us to develop a novel dedicated solution to face spoofing detection. Indeed, with the help of the movement sensors and the multitasking ability of smartphones, we can simultaneously capture the device's movement information while filming the user's face by our Android Application. (Notice that all smartphones in our day include at least the gyroscope sensor.) At the end of this phase, the output will include a video of the head and sensors raw data. In the case of legit authentication, information given by movement sensors is a priori coherent with information estimated from the camera's outputs, but it is generally not the same case when a spoofing attack happens. Therefore, it is a good idea to exploit the coherence between these two sources of information as features for face spoofing detection. Our proposed solution relies mainly on this idea.

The proposed solution consists of three major steps. Figure 4.1 shows the detection flowchart of the solution. Firstly, a 3D model of the face is estimated thanks to a 3D reconstruction process. Then, a Photo Attack Detection (PAD) classifier exploiting the 3D shape is employed to retrieve photo attacks (which use a static image of legitimate user, e.g. photo printed on paper or displayed on an LCD screen). The construction of PAD classifier is described in the section 4.2.2. For the ones which pass through the PAD classifier, they will be after that classified thanks to the Video

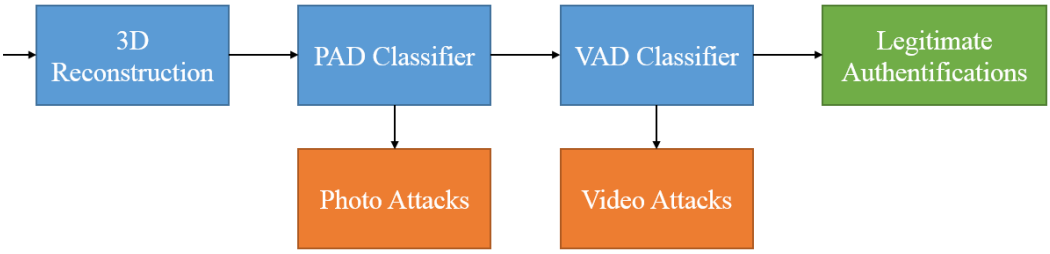


FIGURE 4.1: Flowchart of the whole proposed detection process

Attack Detection (VAD) classifier, described in the section 4.2.3. The VAD classifier permits to detect video-replay attacks. The ones which finally pass through the VAD classifier would be considered as legitimate authentication.

FIGURE 4.2: Camera movements during authentication process

Apart from the necessary movement of smartphone (Figure 4.2) which requires the collaboration of user, all other processes can be automatic. In this study, the video and sensor data collector, 3D reconstructor and classifier are regrouped inside an unique android application. However, the 3D reconstruction is not real-time yet that slows down the detection. In a real scenario, it is recommended to offshore 3D reconstruction and final classification to a dedicated server.

The next section gives details about our solution to face spoofing detection starting with an overview of the method. Step by step, we will explain how our scheme can distinguish photo and video attacks from valid attempts of authentication. The result will be given in the last section where we discuss more about the performance and the perspectives of the method.

4.2.1 Preprocessing: Facial 3D reconstruction

In the process of proposed method, a three-dimensional model of the object (e.g. real face or fake face) is constructed from a video captured during the authentication process. For a better quality of the 3D model, the user is asked to move the phone's camera around their face in such a way as various head poses can be captured in the video. Two simple camera movements are considered in our proposed approach: in

the vertical direction (i.e. upwards or downwards) and in the horizontal direction (i.e. to the left or to the right) (Figure 4.2). These proposed basic movements allow applications to easily communicate with users during authentication. They also permit to simplify the measure of coherence mentioned above.

In our study, 3D reconstruction process is assured by VisualSFM, a 3D reconstruction application developed by C. Wu [108] using Structure From Motion (SFM). The method requires a sequence of images in input. It gives as outputs the 3D reconstruction of the object captured as well as information related to the camera poses. Other recent solutions can replace VisualSFM in this step such as a faster 3D reconstruction proposed by Maninchedda et al . The choice of technology here doesn't affect the final outcome severely but highly depends on the computation capacity of smartphone .

In fact, each frame F_i in the video ($i = 1, ..., n$ where n is the number of frames) is compared to all other frames by method SIFT (scale-invariant feature transform). Two frames (F_{c_1}, F_{c_2}) which maximize the similarity index are chosen to form the base of 3D object. After that, an extraction of common edges and interest points (corners) is applied for all pairs of two frames:(F_i, F_{c_j}) where $i = 1, ..., n$ and $j = 1, 2$ (this extraction based on the derivation of intensity and color of pixels). These features(common edges and points) are tracked from one frame to next so the position and orientation of each frame can be estimated by geometric calculation. Different views of one point which can be obtained from many consecutive frames are extracted to estimate its deep coordination and then its 3D position. The next step is the filter of relevant features points which appears in many frames to form the 3D model.

The 3D model is a matrix which represents a cluster of featured points in a Cartesian coordinate system. Notes **M** the 3D model , it can be represented as:

$$\mathbf{M} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ ... & ... & ... \\ x_N & y_N & z_N \end{bmatrix} \tag{4.1}$$

With color video a RGB matrix $\mathbf{M_c}$ can be associated with **M**

$$\mathbf{M_c} = \begin{bmatrix} r_1 & b_1 & g_1 \\ r_2 & b_2 & g_2 \\ ... & ... & ... \\ r_N & b_N & g_N \end{bmatrix} \tag{4.2}$$

where N is the number of feature points and (x_k, y_k, z_k) $(k = 1, ..., N)$ are the coordinates of the k-th point in the space $Oxyz$. (r_k, b_k, g_k) $(k = 1, ..., N)$ represent the color of this point.

Figure 4.3 gives an example of a real face reconstructed in the form of point cloud.

FIGURE 4.3: Point cloud of a real face 3D model

For each frame F_i $(i = 1, ..., n)$, all the appeared features are saved in a matrix $\mathbf{M_i}$:

$$\mathbf{M_i} = \begin{bmatrix} k_1 & px_1 & py_1 \\ k_2 & px_2 & py_2 \\ ... & ... & ... \\ k_{N_i} & px_{N_i} & py_{N_i} \end{bmatrix} \tag{4.3}$$

where N_i is the number of appeared feature points of frame F_i, (px_j, py_j) $(j = 1, ...N_i)$, is the pixel that represents the j-th point in F_i, k_j is the coordination of this point in $\mathbf{M}$.

The geometric information of F_i is also calculated and represents in position matrix $\mathbf{P_i}$ and rotation matrix $\mathbf{R_i}$ (which can be transformed to orientation matrix $\grave{}_\mathbf{i}$)

$$\mathbf{P_i} = \begin{bmatrix} x & y & z \end{bmatrix} \tag{4.4}$$

$$\mathbf{R_i} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} \\ r_{2,1} & r_{2,2} & r_{2,3} \\ r_{3,1} & r_{3,2} & r_{3,3} \end{bmatrix} \tag{4.5}$$

$$\grave{}_\mathbf{i} = \begin{bmatrix} \theta_x & \theta_y & \theta_z \end{bmatrix} \tag{4.6}$$

(x,y,z) is the coordination of the camera when it captured frame F_i, $(\theta_x, \theta_y, \theta_z)$ is its orientation respectively in the view of Ox, Oy and Oz.

4.2.2 Photo Attack Detection

2D Photo Attack

In the case of photo attacks, the 3D reconstruction given by SFM method is clearly different from the one given in the case of a real face. Figure 4.4 shows different views of the 3D reconstruction of a printed face. It is easy to realize that the form of

the 3D reconstruction is flattering in the case of photo attack. It can be explained by the fact that a real face is a real 3D object which contains much more depth information than a face printed on a piece of paper.

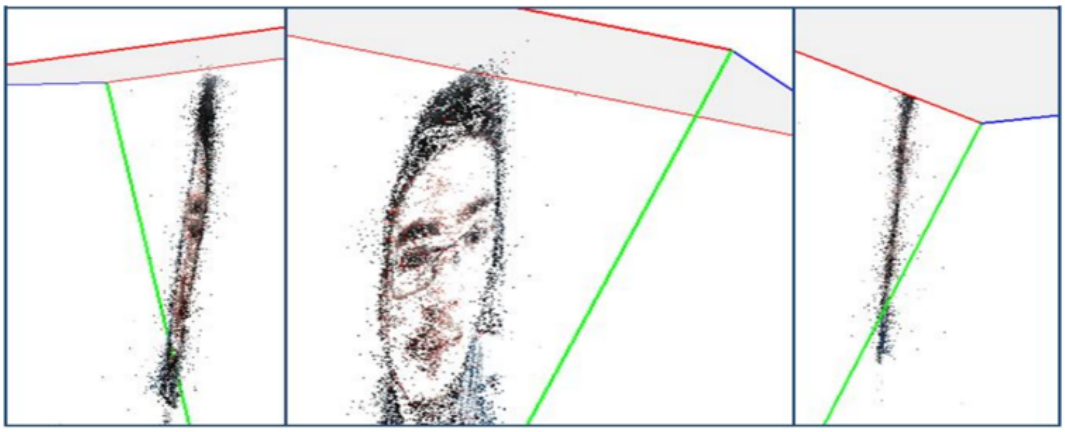


FIGURE 4.4: Different views of a printed face 3D model

Thus, the more a 3D model is flat, the higher possibility of a photo attack. So that we can base on the thickness of the 3D reconstruction to eliminate photo attacks.

The thickness of 3D reconstruction can be relativity estimated using Principal Component Analysis (PCA). The PCA technique permits to transform the 3D reconstruction into a new coordinate system ($w = (w_{(1)}, w_{(2)}, w_{(3)})$), where each coordinate is represented by a principal component. This transformation is defined in such a way that the first principal component ($w_{(1)}$) has the largest possible variance (i.e. it accounts for as much of the variability in the data as possible), and each succeeding component, in turn, has the highest variance possible under the constraint that it is orthogonal to the preceding components. In that way, the variance of point cloud projected in the last component ($w_{(3)}$) is the minimum among all vectors of space that can be used to represent the "thickness".

$$w_{(3)} = \left[\arg\min_{\|w\|=1} \sum_{i=1}^{n} (m_i.w)^2 \right] \tag{4.7}$$

where n is the number of points of the cloud, $m_i = (x_i, y_i, z_i)$ with $i = 1, .., n$ is the coordination vector of i-th point.

For a simplicity of the PCA transformation, the columns of the matrix $\mathbf{M}$ are firstly shifted to have a zero-mean. Without ambiguity, we use the same term $\mathbf{M}$ as the matrix shifted for the following development. The principal component matrix $\mathbf{P}$ is defined as an orthogonal linear transformation of the matrix $\mathbf{M}$:

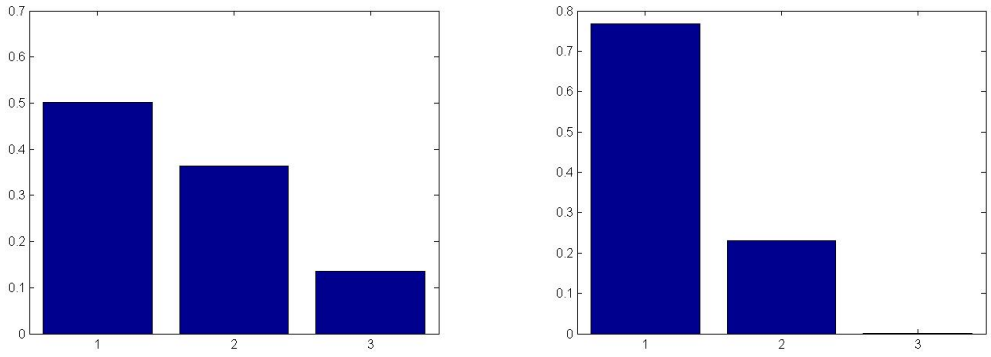
$$\mathbf{P} = \mathbf{MW} \tag{4.8}$$

where the matrix $\mathbf{W}$ is a 3-by-3 matrix whose columns are the eigenvectors of $\mathbf{M}^T\mathbf{M}$.

Denote v_j the variance of the i-th column of $\mathbf{P}$ ($j = 1, 2, 3$). The order of magnitude of each column, denotes d_i, is given as follows:

$$d_i = \frac{v_i}{v_1 + v_2 + v_3} \tag{4.9}$$

For a face spoofing attack, the points of the 3D reconstruction is in a plan. Therefore, there is almost no information in the third component that makes d_3 very tiny. Meanwhile, for a real face, the thickness play a significant part in the total information. Figure 4.5 (a) and (b) give an illustration of the three principal components obtained respectively from fake and real face 3D models.



(a) (b)

FIGURE 4.5: PCA of real (a) and fake (b) face 3D reconstruction

The different is so net that a simple SVM classifier fed by the order of magnitude d_i can be employed as a PAD classifier without any other processing.

Advanced photo attack

However, in a more sophisticated situation where the attacker uses some deformed photos to create a model which a significant depth dimension, the PCA is not performance enough for the detection. Here, we propose a method to extract the depth image of the face from the 3D model to produce another proof of liveliness.

Plan estimation and coordination In this stage of preprocessing, the depth image can be trivially extracted by fixing a plan which is perpendicular to the normal vector of view and calculating the distance of each vertex to this plan. Ensemble of these distances forms a map which can be called a depth image. (Fig.4.6)

For each pixel in this depth image, one vertex is linked and also its value of vessel intensity. However, 3D reconstruction process is not always stable, it gives relative measure rather than absolute one. This is the reason why depth value must

FIGURE 4.6: Depth image of a face

be adjusted into [0,1].

$$\mathbf{D} = \begin{bmatrix} u_1 & v_1 & d_1 \\ u_2 & v_2 & d_2 \\ ... & ... & ... \\ u_N & v_N & d_N \end{bmatrix} \tag{4.10}$$

where u_i, v_i $(i \in 1, ..., N)$ are the projected coordination of i-th vertex in the plan. d_i is depth value which corresponds to this vertex. However, the point cloud is not uniformly distributed. There are regions that contain much higher density of point than the others. There are also some parts of the face representing almost all the distinguishable features of this user that makes studying other parts is wastes. These problems are fixed in the normalization phase.

Normalization There are two normalizations in this preprocessing: the crop of effective region of the face and the pixelization of depth and vessel image. In the scenario of observation of a front view, the crop of effective region is simply the application of the elliptic mask on the face using the nose tip detection. The location of the nose tip can be easily determined by the depth image and the width of the elliptic mask is calculated by the localization of information region. This crop is applied on both images to eliminate unnecessary points.

The pixelization is, in fact, the transformation of the point cloud to an image (in this case an image of 320×256-pixels). This is very similar to a scaling process apart from the fact that the point cloud is not equi-distributed. Our solution is using an adapted version of bilinear interpolation which can be summarized as follows. For $p_1, p_2 ... p_h \in 1, ..., N$ are the points inside a pixel ($h \in 0, .., N$). Assume that the depth measure follows a linear relation to u-axis and v-axis, it can be represented by a local function $D(u, v)$:

$$D(u, v) = c_0 + c_1 u + c_2 v + c_3 uv \tag{4.11}$$

where c_0, c_1, c_2, c_3 are 4 coefficients to be determined.

In the case of $h \geq 4$, the problem becomes a linear least-square problem estimating $C=(c_0, c_1, c_2, c_3)^T$ that minimizes the sum S (for the case $h = 4$, the equation

becomes a standard bilinear interpolation and the minimized sum must be zero):

$$S = \sum_{j=1}^{h} (d_{p_j} - (c_0 + c_1 u_{p_j} + c_2 v_{p_j} + c_3 u_{p_j} v_{p_j}))^2 \tag{4.12}$$

This equation could be represented in matrix form as follows:

$$S = \sum_{j=1}^{h} (d_{p_j} - D(u_{p_j}, v_{p_j}))^2 \tag{4.13}$$

where

$$\begin{pmatrix} D(u_{p_1}, v_{p_1}) \\ D(u_{p_2}, v_{p_2}) \\ ... \\ D(u_{p_h}, v_{p_h}) \end{pmatrix} = \begin{pmatrix} 1 & u_{p_1} & v_{p_1} & u_{p_1} v_{p_1} \\ 1 & u_{p_2} & v_{p_2} & u_{p_2} v_{p_2} \\ & & .... & \\ 1 & u_{p_h} & v_{p_h} & u_{p_h} v_{p_h} \end{pmatrix} \times \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} \tag{4.14}$$

The result of this minimization problem can be directly obtained by matrix equation:

$$\hat{C} = (Q^T Q)^{-1} Q^T \times d \tag{4.15}$$

where $d = [D(u_{p_1}, v_{p_1}), D(u_{p_2}, v_{p_2}) ... D(u_{p_h}, v_{p_h})]^T$ and

$$Q = \begin{pmatrix} 1 & u_{p_1} & v_{p_1} & u_{p_1} v_{p_1} \\ 1 & u_{p_2} & v_{p_2} & u_{p_2} v_{p_2} \\ & & .... & \\ 1 & u_{p_h} & v_{p_h} & u_{p_h} v_{p_h} \end{pmatrix} \tag{4.16}$$

The depth value of the pixel can be calculated by $D(u_0, v_0)$ with (u_0, v_0) is the center of the pixel. In the rare event when $h < 4$, values of neighbor pixels can be used to feed the bilinear interpolation solution. At the end of this phase, a depth image of 160×128 pixels is obtained .

Gabor transformation In this study, 2D Gabor filters are applied to all depth images in order to characterize each video. The Gabor wavelets contain information of spatial localization, orientation selectivity and spatial frequency selectivity . A lot of robust 2D face recognition algorithms use Gabor wavelet as the principal representation of face which places great emphasis in both spatial frequency and spatial relations. The Gabor kernel can be described as follows:

$$\Psi(Z) = \frac{k_{\mu,\nu}^2}{\sigma^2} exp(\frac{-k_{\mu,\nu}^2 Z^2}{2\sigma^2})[exp(ik_{\mu,\nu} Z) - exp(-\frac{\sigma^2}{2})] \tag{4.17}$$

where μ and ν represent the orientation and scale of the Gabor wavelets. $\Psi(Z)$ is the value of Gabor wavelet at $Z = (t_u, t_v)$. t_u, t_v are the centered coordination of any

point in the plan. The coefficient $k_{\mu,\nu}$ is defined by $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$ with $k_\nu = k_{max}/f^\nu$ and $\phi_\mu = \pi\mu/8$ so that $e^{i\phi_\mu}$ determines the orientation of the kernel and k_ν places it into a scale. In this study, we use five scales $\nu \in \{0, 1, ..., 4\}$ and eight orientations $\mu \in \{0, 1, ..., 7\}$ which make 40 Gabor kernels with the other parameters as following: $\sigma = 2\pi$, $k_{max} = \pi/2$ and $f = \sqrt{2}$.

The representation of an image by Gabor wavelets, so-called the Gabor image, is the convolution of the image with a Gabor kernel. However, the convolution gives each pixel a complex value with two Gabor parts: the real part and the imaginary part. These two parts can be transformed to two types of information: Gabor magnitude features and Gabor phase features. In this study, only Gabor magnitude features are used to describe the face. For 40 Gabor kernel, 40 Gabor image can be computed/ Each image is an ellipse of size 256x320 which includes about 63600 features, so in total 2,544,000 features to feed into the classification algorithm.

Feature Selection and final classifier The richness of Gabor transformation in terms of quantity of features improves significantly the result of classification. However, the complexity of this algorithm increases with the number of features. Therefore, a scheme proposed by Chenghua Xu is applied to divide the whole system into small ones which can work in parallel. This hierarchical selection includes two stages:

LDA sub-sampling: for each Gabor vessel image, the optimal LDA sub-sampling extrudes massively non-efficient or redundant features by minimizing the within-class distance when maximizing the between-class distance.

Unlike the usual sub-sampling method where the sub-windows is uniformly distributed in the image, this optimal method aims for rich-information regions where the features could provide more proof of recognition. Gabor images under different orientation and scale may not share the same sub-sampling pixels. Therefore, 40 sets of sub-sampling positions are constructed correspondents to 40 Gabor depth images. To minimize the within-class distance (explain by scatter matrix S_W) and maximize the between-class distance (S_B), the optimal discriminant vectors constructing the LDA subspace is computed by solving the following criterion in the standard LDA algorithm:

$$W^* = argmax(J(W)) = \frac{W^T S_B W}{W^T S_W W} \tag{4.18}$$

where

$$W = \left\{ \begin{pmatrix} w_{1,1} \\ w_{2,1} \\ ... \\ w_{p,1} \end{pmatrix} \begin{pmatrix} w_{1,2} \\ w_{2,2} \\ ... \\ w_{p,2} \end{pmatrix} \begin{pmatrix} ... \\ ... \\ ... \\ ... \end{pmatrix} \begin{pmatrix} w_{1,v_{max}} \\ w_{2,v_{max}} \\ ... \\ w_{p,v_{max}} \end{pmatrix} \right\} \tag{4.19}$$

Here, $p = 15900$ is the number of pixels in one image and v_{max} is the amount of discrimination vectors (each vector is one column of W). The summation vector V

can be computed as follows:

$$V = \left(\sum_{k=1}^{k=v_{max}} |w_{k,1}|, \sum_{k=1}^{k=v_{max}} |w_{k,2}|, ..., \sum_{k=1}^{k=v_{max}} |w_{k,p}| \right) \tag{4.20}$$

The magnitude of V at a particular position represents the corresponding variations among the training set, which also reflects the corresponding importance in distinguishing the faces. After this stage, only 1278 features in each image are chosen for AdaBoost selection.

AdaBoost learning: a supervisor learning which applies a weak and tiny classifier on each feature of the sample in order to:

- Select the less redundant group of effective features which can discriminate the two hypotheses,
- Construct weak classifiers using these features,
- Build a strong cascaded classifier .

The algorithm of Adaboost learning for feature selection can be introduced as below:

Given example couple images (I_1, J_1, y_1),(I_2, J_2, y_2)...(I_n, J_n, y_n) where y_i=1 when I_i and J_i are all images of genius faces or all images of the same type of attack (photo or video) and y_i=0 in other case (negative examples).

Initialize weights $w_{1,i}$=1/2m or 1/2l, for y_i=0,1, respectively, where m and l are the number of negatives and positives examples.

For t=1,...,T (T is the maximum number of chosen features)

- **1.** Normalize the weights: $w_{t,i} := \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$
- **2.** For each feature, j, train a LDA classifier h_j which using only this feature (which has 2 values). The error is evaluated by: $\epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$.
- **3.** choose the classifier h_j which minimizes the error ϵ_j. j is the feature chosen in this step.
- **4.**Update the weights $w_{t+1,1} = w_{t,i}\beta_t^{1-e_i}$ where e_i = 0 if example x_i is classified correctly, e_i = 1 otherwise and $\beta_t = \epsilon_j/(1-\epsilon_j)$

Each iteration, the algorithm searches for a feature that minimize l'error of classifier which is pondered by the weight of each sample in training set. By this way, each time the algorithm choses a feature as efficient feature, it will update the weight of all the samples so that the incorrectly classified samples become more important for next iterations.

Instead of one image, our examples become any couple of images possible. The database includes 1001 videos of 3 people including sensors data, therein: 451 cases of legitimate authentication, 362 cases of video-replay attack and 188 cases of photo

attack. However, in the training set we used only 100 cases of legitimate authentication, 50 cases of video-replay attacks and 50 cases of photo attack which makes 7400 intra-case couples and 12500 extra-case couples.

In this stage, the AdaBoost selection is used twice as following:

- *Individual learning:* apply AdaBoost method to each Gabor depth image to select the effective feature for each image (about 30-38 per image) and group all these features into one set.
- *Total learning:* apply AdaBoost to this set of features to reduce one more time the number of features (about 127 features in this case).

The final step of training phase is the construction of a cascaded strong classifier from these 127 features. That cascaded classifier contains many layers, each layer is also building by the efficient features in features learning stage. In fact, instead of constructing a big classifier of a lot of features in order to achieve a detection rate D and limit the false positive rate under F, the method aims to build some small independent classifiers that provide a higher detection rate d_l with a huge false positive rate f_l. When these classifiers are used as layers for a bigger one we can choose the layer so that:

$$F = \prod_{l=1}^{l=L} f_l \tag{4.21}$$

$$D = \prod_{l=1}^{l=L} d_l \tag{4.22}$$

where L is the number of layers. In this way, from 12 classifiers with high false positive rate (by 50%), cascaded classifier can be building which limits F at $0.5^{10} \approx 10^{-3}$. The algorithm of Adaboost learning for strong classifier can be introduced as below:

- **1.** Parameters initialization: selecting the value of f (the maximum acceptable false positive rate per layer) and d (the minimum acceptable detection rate per layer). This step depends essentially on the efficiency of features
- **2.**Target determination: selecting the value overall false positive rate (F_{target}). This step depends on the result we aim to
- **3.** P = set of positive examples.
- **4.** N = set of negative examples.
- **5.** Initialization: $F_0 = 1.0$; $D_0 = 1.0$; $l = 0$
- **6.** While $F_l > F_{target}$:
  - $l \leftarrow l + 1$
  - $n_l = 0; F_l = F_{l-1}$

- While $F_l > f \times F_{i-1}$
  - $n_l \leftarrow n_l + 1$
  - Use P and N to train a classifier with n_l features using AdaBoost
  - Evaluate current cascaded classifier on validation set to determine F_l and D_l.
  - Decrease threshold for the l-th classifier until the current cascaded classifier has a detection rate of at least $d \times D_{l-1}$(this also affects F_l)
- $N \leftarrow \emptyset$
- If $F_i > F_{target}$ then evaluate the current cascade detector on the set of non-face images and put any false detections into the set N

The output of our experiment is a cascaded strong classifier with 12 layers and 108 features.

4.2.3 Video Attack Detection

In the scenario of a video attack, a clip of authentic user's head is displayed in a LCD screen in front of the camera. The video is edited so that the head moves in the same way to mimic the process. This attack can pass the PAD classifier since the moving head can provide different views of user's face to construct a genuine 3D model of the head. The form of the 3D reconstruction doesn't give us enough information for spoofing detection. Therefore, we proposed to study, in addition, the camera poses.

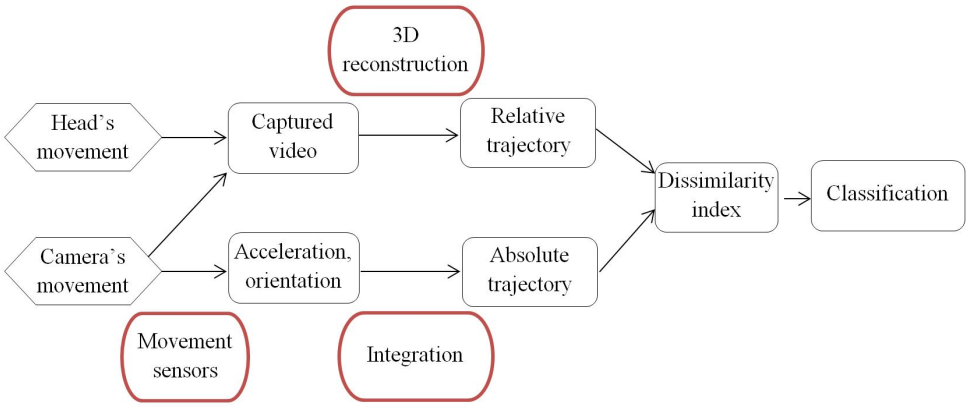


FIGURE 4.7: General schema of video attack detection.

In fact, the movement of the camera can be observed in two ways. The first is the positions and the poses of camera corresponding to each image. These positions which represent the movement of the camera according to the user's face can be located by 3D reconstruction. Figure 4.8 shows an example of the camera's positions estimated from the 3D reconstruction.

The 3D reconstruction describes the movement by position vectors $\mathbf{P_i}$ for the camera's position and by orientation vectors $\mathbf{`_i}$ for the camera poses with $i = 1, .., n$ is the index of the frame. Since the movement is mono-direction (the camera move

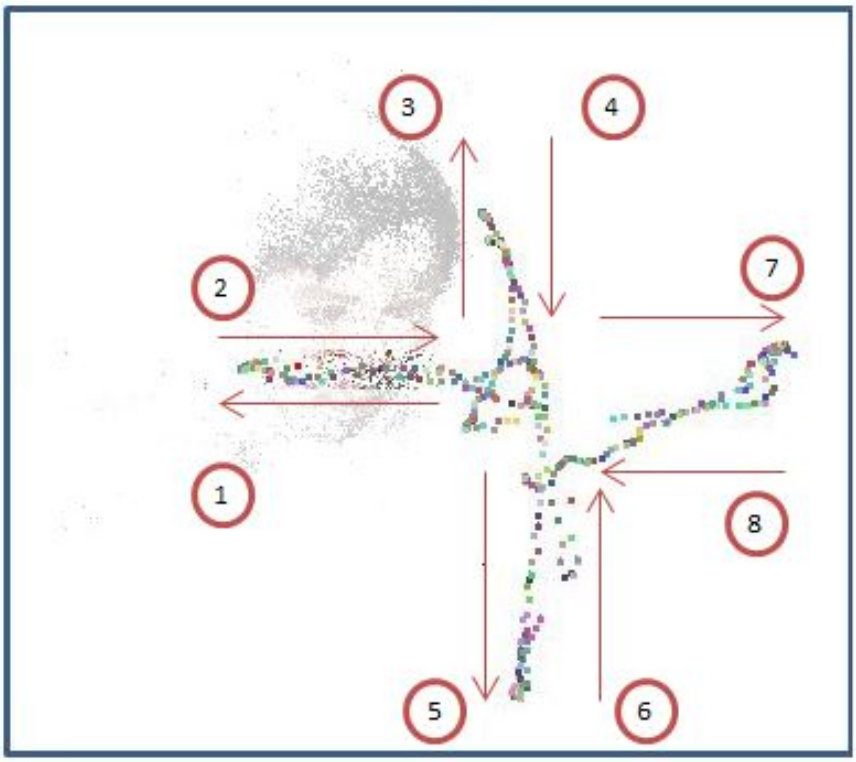


FIGURE 4.8: Example of the camera's positions estimated from the 3D reconstruction. Direction of the camera's move is marked form 1 to 8.

by x-axis, y-axis but not both at the same time), the position and orientation can be represented separately axis by axis in the form of sequence:

$$\mathbf{X_i} = \mathbf{P_{i,x}} \tag{4.23}$$

$$\theta_i^x = \theta_{i,x} \tag{4.24}$$

This trajectory observed by the 3D model is, in fact, the relative movement of the camera in view of the head. These motions come from 2 sources: the absolute movement of this camera and head trajectory.

The second observation is the trajectory of camera captured by movement sensors (e.g. accelerometer, gyroscope). These sensors observe the acceleration of translation and rotation of camera continuously by the time that allows to describe that movement independently. The gyroscope captures rotation acceleration and the accelerometer captures linear acceleration of the device affected by the force of gravity.

$$\mathbf{a}(\mathbf{t}) = \left[a_x, a_y, a_z\right] \tag{4.25}$$

$$\mathbf{a`}(\mathbf{t}) = \left[g_x, g_y, g_z\right] \tag{4.26}$$

Where $t \in \mathbb{R}^+$ is an index of time, g_x, g_y, g_z is gyroscope data. Since the δt (time

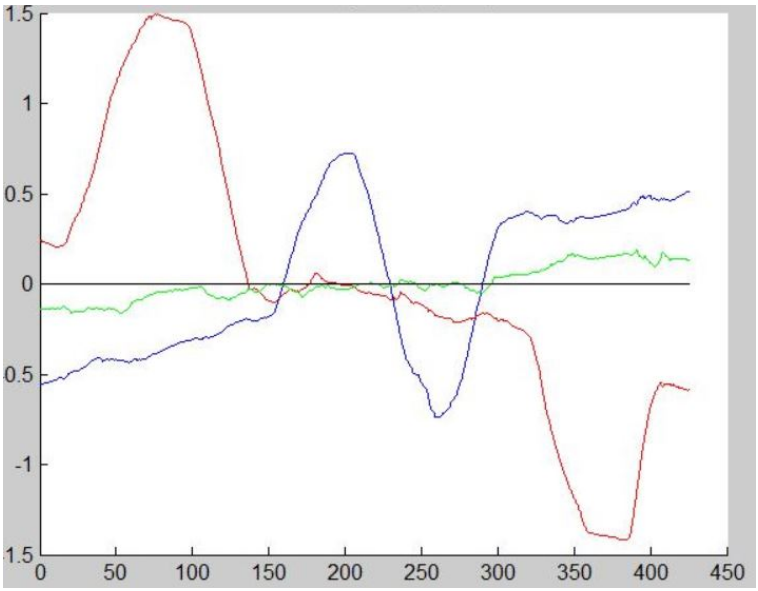


FIGURE 4.9: Orientation of camera described by gyroscope sensor (blue for θ_i^x, red for θ_i^y and green for θ_i^z,

between two consecutive measures) is very small, the acceleration data can be considered as continuous.By the same reason of mono-direction, we can observe the movement only by x-axis or y-axis.

$$\frac{d^2x(t)}{dx^2} = a_x(t) \tag{4.27}$$

$$\frac{d^2\theta_x(t)}{dx^2} = g_x(t) \tag{4.28}$$

By using integration method, the position and orientation of the camera can be calculated from the acceleration as the initial speed and initial position are both zeros. Let's t_{max} is the length of video, on frame is taken each $\Delta t = \frac{t_{max}}{n-1}$ second. Another sequence of the camera's state can be obtained from sensor data:

$$\widehat{\mathbf{X}}_\mathbf{i} = \mathbf{x}(\mathbf{t}_\mathbf{i}) \tag{4.29}$$

$$\widehat{\theta_i^x} = \theta_x(t_i) \tag{4.30}$$

where $i = 1, .., n$, $t_1 = 0$, $t_n = t_{max}$, $t_i = \Delta t.(i-1)$

The camera poses are compared to the information captured from movement sensors (e.g. accelerometer, gyroscope) to obtain some similarity index. For a legit authentication case, the process does not demand any head's movement. The head's motion becomes insignificant that makes a high coherence between absolute and relative trajectory. Meanwhile, when attackers use a video to authenticate, in order to mimic a 3D object, they must display some motion in this video (without motion, it becomes a normal 2D photo). Therefore, the similarity between the two sources would not be assured. Here, the dissimilarity index is, in fact, a factor of head's

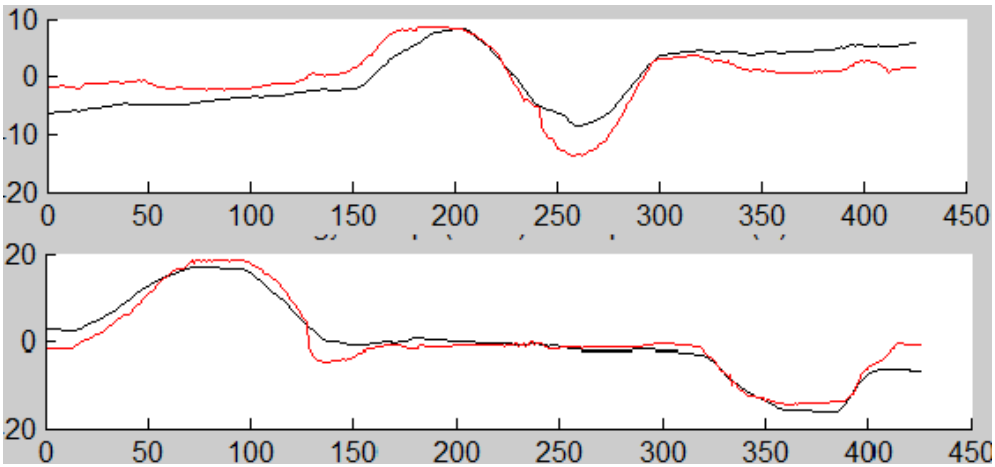


FIGURE 4.10: Correlations between θ_i^x (black)and $\widehat{\theta_i^x}$ (red) and between θ_i^y (black) and $\widehat{\theta_i^y}$ (red).

motion. To estimate this dissimilarity, a simple correlation can be applied for each couple of data as: $(\theta_i^x, \widehat{\theta_i^x})$, $(\theta_i^y, \widehat{\theta_i^y})$, $(X_i, \widehat{X}_i)$ and $(Y_i, \widehat{Y}_i)$. All these features (correlation results) are fed to an SVM classifier to form the VAD classifier mentioned previously.

4.3 Result and Evaluation

4.3.1 Result

The proposed video-replay attack detection process requires the data of motion sensors integrated within the smartphone. Actually, there are no public face spoofing datasets responding to this requirement. Therefore, we tested the proposed method in a specific database constructed in our laboratory. The database includes 1001 videos of 3 people including sensor data, therein: 451 cases of legitimate authentication, 362 cases of video-replay attack and 188 cases of photo attack. The videos are captured in different light conditions and movement speed by three devices: 2 instances of Samsung Galaxy Alpha and a Samsung Galaxy Tab. The database is divided into 2 sets. Each set has all three types of videos (legitimate, photo-attack and video-replay attack authentication). One set is used for training the PAD and VAD classifiers and the other set is used for testing. From the training set, we used all videos to train the PAD classifier, but we used just legitimately and video-replay attack authentication videos to train the VAD one. We also implemented the multi-scale Local Binary Pattern method, which applies for a single image. Therefore, frames from the videos are used as input to implement this method. Figure 4.11 shows the Receiver Operating Characteristics (ROC) curves of the proposed method and the LBP one. Our method performs much better than the LBP one especially for a small rate of false positive. In fact, a large portion of false alarms links to the bad behavior in capturing phase. A lot of problems such as the blurry frame due to the high speed of movement or the escape of user's face from the input video make

the 3D reconstruction phase unstable. So in a real use case, the process asks naturally for another capturing phase. This option ameliorates the process performance significantly 4.11.

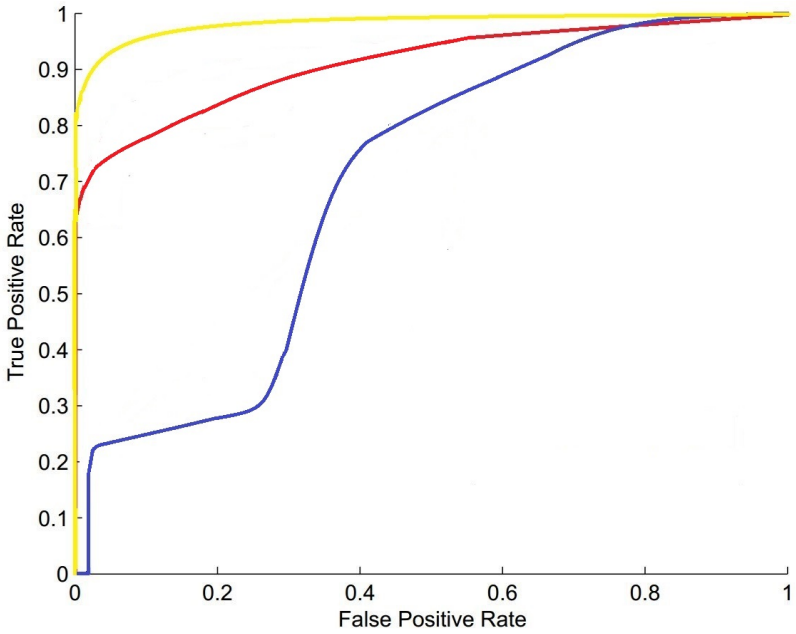


FIGURE 4.11: ROC curve for the proposed detection method (Red) in comparison with the one of LBP method(blue). The proposed method with recapturing option is displayed in yellow.

4.3.2 Evaluation

Our solution for photo attack detection is, in fact, a classification of 3D objects based on their depth images. The different is quite large between the form of a genius face and the form of a material that displays a photo. It is the reason why we can obtain 100% in detection rate. Because of the maximum percentage obtained by different methods of feature extraction, we present only Gabor transformation here to emphasis that Gabor wavelet outperform other methods in case of depth image. An experimental proof of this statement can be found in the next chapter.

In the process of Video attack detection, we assume the worst case of authentication where users have total control of proof capturing phase: authentication using smartphone. We do not have any idea of what they place in the face of the camera, what material they can use to display the video and what type of movement in the attacking video. We do not control which type of smartphone used in the process or any supplemental calibration information. The unstable smartphone's camera (due to vibration of manual manipulation) is the principal source of false detection. The error is generated by the manipulation itself but not by the user which explains the fact that a large percentage of false detection linked to first or second videos of each scenario. In the real application, a demand for "retry" for another capturing phase can reduce significantly the false detection rate.

In other solution where the situation is more controlled, maybe, with a mechanically moving smartphone or a calibrated camera, the performance can be significantly ameliorated. However, 3D mask attack is theoretically immune against our detection because all of our proof is based on the hypotheses that the attacker use a 2D material for authentication. In general, 3D mask attacks are very efficient in case of visible face recognition. A perfect mask can overcome any system using an only normal camera. Nevertheless, this type of attack can be easily detected by other types of sensor such as IR detector. In the next chapter, we will present a novel method of face recognition using an uncooled thermal camera which solves not only face spoofing attack but also illumination problem.

Chapter 5

Face Recognition by Thermal Video Using Vesselness Features in Multiview 3D Projections

5.1 Introduction

In this chapter, we present a novel method for face recognition by a movable thermal uncalibrated camera. The first section mentions some problematic of face recognition in general which leads to the necessity of our method. We will also point out the advantages and limits of our solution in comparing with others. The general process is divided into 2 phases: preprocessing and main-processing which are correspondingly described inside the second and the third section. The last section proposes an ameliorated scheme of the solution which alternates the main-processing with only one frontal pose by a multi-pose method.

5.1.1 Problematic and Objectives

Promoted by impressive growth of computation capacity in recent years, face recognition has conquered a large portion of biometric security domain and has appeared in many applications such as authentication, identification, human tracking and surveillance. Inherit from numerous studies in computer vision over the last there decades, face recognition is now broadly used in a lot of information systems, from giant immigration control in airports to tiny user unlock solution for smartphone, from extremely strict authentication for Internet banking transaction to popular identification for Facebook photo...

However, visible face recognition always suffers from crucial limits due to complex illumination conditions. Another persistent problem in visible face recognition authentication comes from the face spoofing attack using images, video or 3D mask with the face of the victim. These problems have opened a road for thermal imagery to join in the competition. This last one is more robust not only against environmental variations but also against facial expressions. Thermal images can solve many impossible challenges in face recognition like complete dark environment or

monozygotic twin problems. Despite all these advantages, the average accuracy of thermal image classification is always lower than the visible one that requires new solution to improve its performance.

5.1.2 Proposed approach

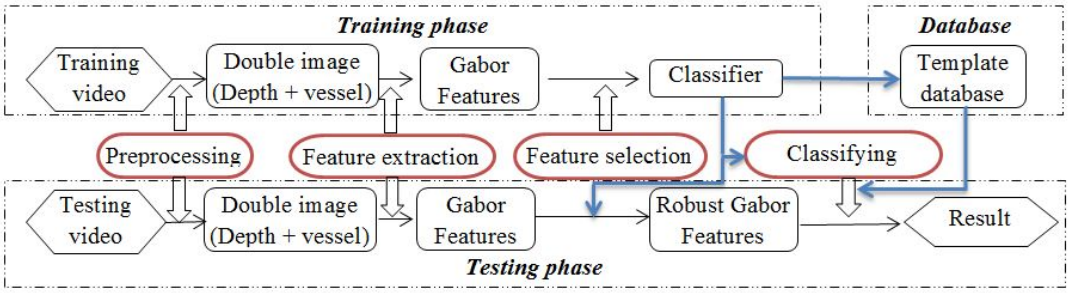


FIGURE 5.1: Framework

In this study, we propose a new solution for face recognition combining 3D information and vesselness features observed from a thermal video. Like other biometric systems, our framework (Fig. 5.1), contains two phases: the training phase and the testing phase. The training phase can be described as follows:

- *Preprocessing:* The 3D model of the head is reconstructed using a video which includes various poses and positions of the head . A blood vessel transformation is applied to each frame of this video to obtain a vessel map . These blood vessel maps are projected on the 3D model in order to obtain a 3D vesselness representation. Thanks to this new model, we can obtain a depth image and a vessel map for each pose. Our first experiment had used only frontal view which had neglected a large portion of information in other poses. Later, the process has been tested for different combinations of views to choose the most relevant result in terms of accuracy and computation.

- *Feature representation:* A lot of transformations widely used in visible imagery (like LBP, Weber, ...) are tested to combine the two informations. We finally chose Gabor filters as the primary transformation in feature representation. These filters with multiple scales and multiple orientations are applied to the depth image and vessel map to extract a lot of Gabor features. Each feature includes two values, one from vessel network, another from depth information.

- *Feature selection:* As described above, many Gabor features are extracted from the depth image and vessel map, but not all of these informations are effective to the recognition system. In this phase, a hierarchical scheme for effective feature selection is proposed using linear discriminant analysis (LDA) and AdaBoost method .

The testing phase uses the same preprocessing scheme as training phase. The feature templates constructed in training phase are now applied to select the feature in

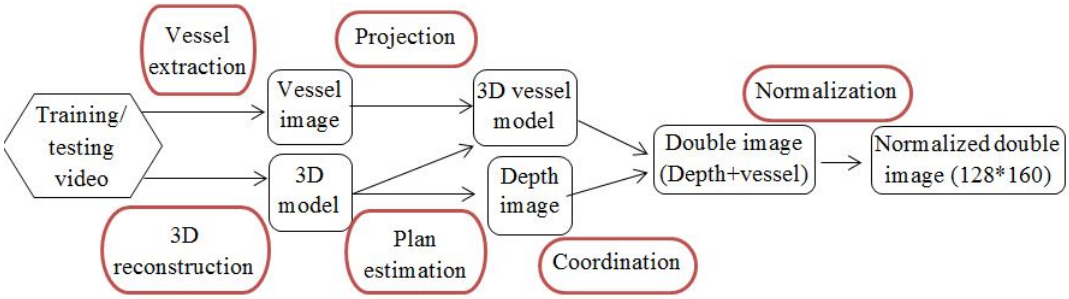


FIGURE 5.2: Preprocessing

testing model. The input is consequently coupled with each training video to compute their similarity using the learned classifier. Finally, an 1-NN(nearest neighbor) algorithm classifies the input video into one category.

5.1.3 Contributions

In this study, we propose and examine a new face recognition solution using a 3D model of the head computed from a thermal video which contains information of vascular network. Its contributions are as follows:

- The reconstruction of a 3D model from a thermal video and the projection of the map of blood vessels on this model give a new 3D representation of vascular network.
- The depth and blood vessels intensity aren't treated like unbound features but are jointed to form a single feature with two values. In this way, the face recognition system bases essentially on the 3D location of vascular network.
- The 3D model is represented in the form of a combination of depth images in different views. This process make possible the feature selection and the classification of a complex object without any loss of information.

5.2 Preprocessing

In this paper, we propose a four-step process (Fig. 5.2): rebuilding of 3D model from every frames of the input thermal video, vessels extraction for each video frame and the projection of these vessels on the 3D model, estimation of the depth image and finally a step of the normalization.

5.2.1 Reconstruction of 3D model

In our study, the input thermal video is supposed to contain many frames from various poses of one head. An algorithm of 3D reconstruction is used to compute a 3D point cloud which describes the head filmed in the video. In this scenario, we use VisualSFM - Structure from motion, developed by Changchang Wu , a robust

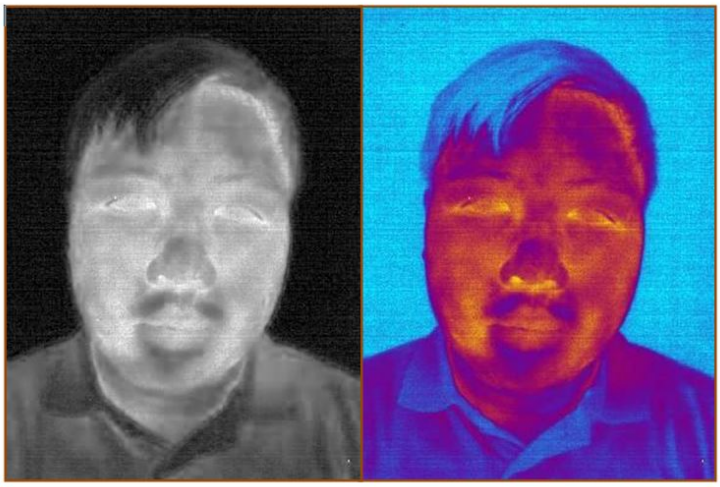


FIGURE 5.3: Gray-scale and iron-palette version

and stable reconstruction algorithm which uses not only the shape but also the color of each pixel to compute the coordination of the face. Therefore, instead of using the gray-scale version of intensity, the experiment observes another color representation of thermal image: the Iron-palette which can be computed from the gray-scale one (Fig.5.3).

These features(common edges and points) are tracked from one frame to next so the position and orientation of each frame can be estimated by geometric calculation. Different views of one point which can be obtained from many consecutive frames are extracted to estimate its deep coordination and then its 3D position.

From the original video, a set of frame F_i ($i = 1, ..., n$ where n is the number of frames) can be extracted. Each frame is an image ($p \times q$ pixels) which contains a view of the face:

$$\mathbf{F_i} = \begin{bmatrix} x^i_{1,1} & x^i_{1,2} & ... & x^i_{1,q} \\ x^i_{2,1} & x^i_{2,2} & ... & x^i_{2,q} \\ ... & ... & ... & ... \\ x^i_{p,1} & x^i_{p,2} & ... & x^i_{p,q} \end{bmatrix} \tag{5.1}$$

In fact, each frame is compared to all other frames by method SIFT (scale-invariant feature transform). Two frames (F_{c_1}, F_{c_2}) which maximize the similarity index are chosen to form the base of 3D object. These common points of F_{c_1} and F_{c_2} will form a first model which is called ST_2 (ST_1 do not exist).

$$(\mathbf{c_1}, \mathbf{c_2}) = \underset{i \neq j}{\text{argmax}} SIFT(F_i, F_j) \tag{5.2}$$

Then, the process sorts all images in the decreased order of $SIFT(F_{c_1}, F_i)$ which makes a complete sequence $c_3, c_4, ..., c_n$. For each frame F_{c_i} ($i = 3, ...n$), the process will firstly try to find the common points with ST_{i-1} which will be called STF_i ($STF_i \subseteq ST_{i-1}$). So we have a certain points STF_i in 3D coordination and its project

in the plan of the camera: F_{c_i}. Thus the process will try to estimate the orientation and the position of the camera by maximizing the similarity index between the two sets:

$$(R_{c_i}, Pc_i) = \underset{R,P}{\text{argmax}} SIFT(F_{c_i}, T_P(PROJ_R(STF_i)) \tag{5.3}$$

where T_P is the translation operator by vector P and $PROJ_R$ is the projection operator by rotation matrix R.

$$\mathbf{P_{c_i}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5.4}$$

$$\mathbf{R_{c_i}} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} \\ r_{2,1} & r_{2,2} & r_{2,3} \\ r_{3,1} & r_{3,2} & r_{3,3} \end{bmatrix} \tag{5.5}$$

The geometric information of F_{c_i} is represented in position matrix $\mathbf{P_{c_i}}$ and rotation matrix $\mathbf{R_{c_i}}$ (which can be transformed to orientation matrix $\grave{}_{\mathbf{i}}$)

$$\grave{}_{\mathbf{i}} = \begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} \tag{5.6}$$

(x,y,z) is the coordination of the center pixel of frame F_{c_i}, $(\theta_x, \theta_y, \theta_z)$ is its orientation respectively in the view of Ox, Oy and Oz.

The process will, then, studies the extra part of frame F_{c_i} which does not exist inside STF_i and try to match its with some other nearest frame to find additional common points STP_i

$$ST_i = ST_{i-1} \cup STP_i \tag{5.7}$$

The last one (ST_i) becomes the 3D model formed by $\{F_{c_1}, ..., F_{c_i}\}$.

The set ST_n definite the 3D model of all frame. This 3D model form a matrix which represents a cluster of featured points in a Cartesian coordinate system. Notes $\mathbf{M_{ini}}$ the 3D model , it can be represented as:

$$\mathbf{M_{ini}} = \begin{bmatrix} x_1 & x_2 & ... & x_N \\ y_1 & y_2 & ... & y_N \\ z_1 & z_2 & ... & z_N \end{bmatrix} \tag{5.8}$$

With color video a RGB matrix $\mathbf{M_c}$ can be associated with $\mathbf{M}$

$$\mathbf{M_c} = \begin{bmatrix} r_1 & r_2 & ... & r_N \\ g_1 & g_2 & ... & g_N \\ b_1 & b_2 & ... & b_N \end{bmatrix} \tag{5.9}$$

where N is the number of feature points and (x_k, y_k, z_k) $(k = 1, ..., N)$ are the

coordinates of the k-th point in the space $Oxyz$. (r_k, b_k, g_k) $(k = 1, ..., N)$ represent the pseudo-color (which is computed from intensity) of this point.

The obtained point cloud, which is neither dense nor periodic, must be improved using the Patch-based Multi-view Stereo (PMVS) developed by Yasutaka Furukawa and using the Poisson Surface Reconstruction studied by Michael Kazhdan . By this supplement process, the set of point is transformed into a dense collection of small oriented rectangular. The algorithm of PMVS can be decomposed in 3 steps:

- **Matching**: Pixel-level correspondences of point cloud is computed to enhance a portion of features points. Features found by Harris and difference-of-Gaussians operators are first matched across multiple pictures, yielding a sparse set of patches associated with salient image regions
- **Expande**: Initial matches is spread to nearby pixels to obtain a denser cloud of points.
- **Filtering**: Intensity constraints (and a weak form of regularization) are used to eliminate incorrect matches.

These 3 steps are repeated for several iterations until the set of points (the mesh) is dense enough. This mesh can be further refined by a mesh-based MVS algorithm that enforces the photometric consistency with regularization constraints like Poisson Surface Reconstruction. The resolution of the mesh model is adaptive, and the size of a triangle depends on the density of the nearby oriented points: The denser the points are, the finer the triangles become. The PSR software outputs a closed mesh model even when patches are only reconstructed for a part of a scene. In order to remove extraneous portions of the mesh, we discard triangles whose average edge length is greater than six times the average edge length of the whole mesh since triangles are large where there are no points.

So, in fact, after the PMVS process, a matrix $\mathbf{M}$ (having the same structure as $\mathbf{M_{ini}}$ but having much more points, here we always use N as the number of points) is obtained. This rebuild 3D model includes a huge number of vertex and triangular surfaces which can be considered as a dense facial surface (Fig.5.4).

5.2.2 Vessel extraction and projection

The vascular network is the product of anatomical observations in thermal imagery. The key idea of this feature is the higher temperature of the blood vessel in comparing with neighbor region. The method is proposed by Buddharaju and is enhanced by Reza Shoja Ghiass. The so-called vascular network is a map of tubular structures extracted from a thermal image.
This type of feature is proven to be a effective transformation in thermal face representation.

For each frame F_i $(i = 1, ..., n)$ consider the two eigenvalues λ_1 and λ_2 of the Hessian matrix computed at a certain image locus and at a particular scale s. Without

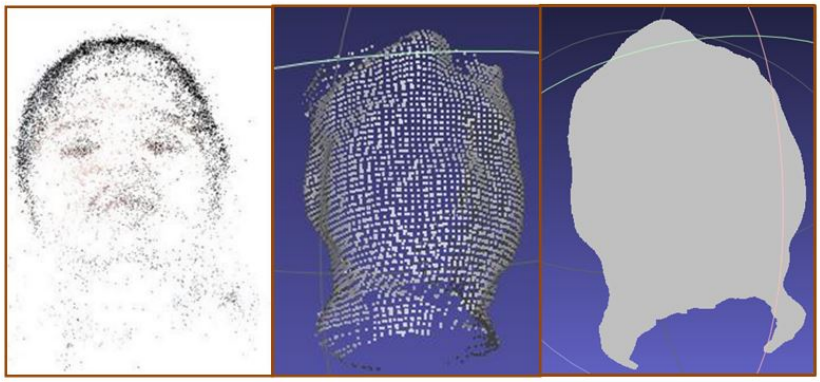


FIGURE 5.4: point cloud (left), dense vertex (middle), surface (right)

loss of generality let us also assume that $|\lambda_1| \leqslant |\lambda_2|$ The two key values used to quantify how tubular the local structure at this scale are $\mathbf{R_A}$ and $\mathbf{S}$:

$$\begin{aligned} \mathbf{R_A} &= \frac{|\lambda_1|}{|\lambda_2|}, \\ \mathbf{S} &= \sqrt{\lambda_1^2 + \lambda_2^2} \end{aligned} \tag{5.10}$$

The former of these measures the degree of local "bloodiness". If the local appearance is blob-like, the Hessian is approximately isotropic and $|\lambda_1| \approx |\lambda_2|$ making $\mathbf{R_A}$ close to 1. For a tubular structure $\mathbf{R_A}$ should be small. On the other hand, $\mathbf{S}$ ensures that there is sufficient local information content at all: in nearly uniform regions, both eigenvalues of the corresponding Hessian will have small values. For a particular scale of image analysis s, the two measures, $\mathbf{R_A}$ and $\mathbf{S}$, are then unified into a single vesselness measure:

$$V(s) = \begin{cases} 0 & if \quad \lambda_2 > 0, \\ (1 - e^{-\frac{R_A}{2\beta^2}}) \times (1 - e^{-\frac{S}{2c^2}}) & otherwise, \end{cases} \tag{5.11}$$

where β and c are the parameters that control the sensitivity of the filter to R_A and S.

In fact, the "vessel value" of a pixel is represented by the measure $V(s)$ of the $(6s+1) \times (6s+1)$ block centered at this pixel. Finally, if an image is analyzed across scales from s_{min} to s_{max}, the vesselness of particular image locus can be computed as the maximal vesselness across the range:

$$V_0 = \max_{s_{min} \leq s \leq s_{max}} V(s) \tag{5.12}$$

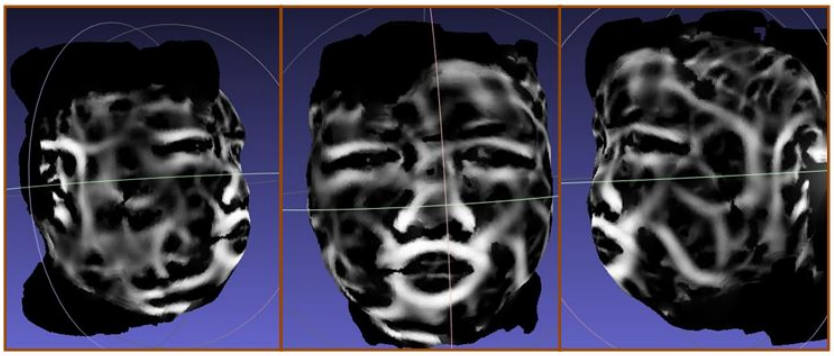


FIGURE 5.5: 3D vascular network model

In the end, each vertex is associated with a value V_0 which presents the vessel probability at this point. Another column V_0 can be added to matrix M:

$$\mathbf{M} = \begin{bmatrix} x_1 & x_2 & ... & x_N \\ y_1 & y_2 & ... & y_N \\ z_1 & z_2 & ... & z_N \\ V_0(1) & V_0(2) & ... & V_0(N) \end{bmatrix} \tag{5.13}$$

For each intensity image, its poses, positions and contributions to the 3D model is computed under a texture map. By using this texture map, these vascular networks can be projected to the 3D model in order to form a 3D vessel model which represents the 3D coordinates of vessel features (Fig.5.5).

5.2.3 Plan estimation and coordination

In this stage of preprocessing, the study aims to obtain a couple of depth and vessel images corresponding to a certain view. The depth image can be trivially extracted by fixing a plan which is perpendicular to the normal vector of view and calculating the distance of each vertex to this plan. Ensemble of these distances forms a map which can be called a depth image. (Fig.5.6)

FIGURE 5.6: Depth image of a face

For each pixel in this depth image, one vertex is linked and also its value of vessel intensity. However, 3D reconstruction process is not always stable, it gives relative measure rather than absolute one. This is the reason why depth value must

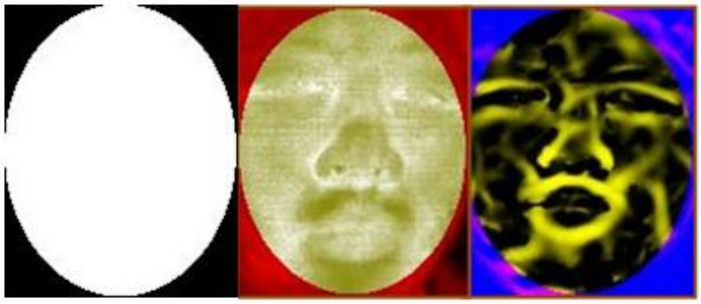


FIGURE 5.7: Elliptic mask (left), cropped intensity image(middle) and cropped vessel image(right)

be adjusted into [0,1]. Base on this information, a vessel image correspondent can be regrouped to the depth information in order to make a double image.

$$\mathbf{D} = \begin{bmatrix} u_1 & u_2 & ... & u_N \\ v_1 & v_2 & ... & v_N \\ d_1 & d_2 & ... & d_N \\ V_0(1) & V_0(2) & ... & V_0(N) \end{bmatrix} \tag{5.14}$$

where u_i, v_i $(i \in 1, ..., N)$ are the projected coordinations of i-th vertex in the plan. d_i and V_i are depth value and vessel measure which corresponds to this vertex. However, the point cloud is not uniformly distributed. There are regions that contain much higher density of points than the others. There are also some parts of the face representing almost all the distinguishable features of this user that makes studying other parts is wastes. These problems are fixed in the normalization phase.

5.2.4 Normalization

There are two normalizations in this preprocessing: the crop of effective region of the face and the pixelization of depth and vessel image. In the scenario of observation of a front view, the crop of effective region is simply the application of the elliptic mask on the face using the nose tip detection. The location of the nose tip can be easily determined by the depth image and the width of elliptic mask is calculated by the localization of information region (Fig.5.7). This crop is applied on both images to eliminate unnecessary points

The pixelization is, in fact, the transformation of the point cloud to an image (in this case an image of 160×128-pixels). This is very similar to a scaling process apart from the fact that the point cloud is not equi-distributed. Our solution is using an adapted version of bilinear interpolation which can be summarized as follows. For $p_1, p_2...p_h \in 1, ..., N$ are the points inside a pixel ($h \in 0, .., N$). Assume that the depth measure follows a linear relation in u-axis and v-axis, it can be represented by a local function $D(u, v)$:

$$D(u, v) = c_0 + c_1 u + c_2 v + c_3 uv \tag{5.15}$$

where c_0, c_1, c_2, c_3 are 4 coefficients to be determined.

In the case of $h \geq 4$, the problem becomes a linear least-square problem estimating $C=(c_0, c_1, c_2, c_3)^T$ that minimizes the sum (for the case $h = 4$, the equation becomes a standard bilinear interpolation and the minimized sum must be zero):

$$S = \sum_{j=1}^{h} (d_{p_j} - (c_0 + c_1 u_{p_j} + c_2 v_{p_j} + c_3 u_{p_j} v_{p_j}))^2 \tag{5.16}$$

This equation could be represented in matrix form as follows:

$$S = \sum_{j=1}^{h} (d_{p_j} - D(u_{p_j}, v_{p_j}))^2 \tag{5.17}$$

where

$$\begin{pmatrix} D(u_{p_1}, v_{p_1}) \\ D(u_{p_2}, v_{p_2}) \\ \dots \\ D(u_{p_h}, v_{p_h}) \end{pmatrix} = \begin{pmatrix} 1 & u_{p_1} & v_{p_1} & u_{p_1} v_{p_1} \\ 1 & u_{p_2} & v_{p_2} & u_{p_2} v_{p_2} \\ & & \dots & \\ 1 & u_{p_h} & v_{p_h} & u_{p_h} v_{p_h} \end{pmatrix} \times \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} \tag{5.18}$$

This equation becomes:

$$d = Q \times C \tag{5.19}$$

where $d = [D(u_{p_1}, v_{p_1}), D(u_{p_2}, v_{p_2}) \dots D(u_{p_h}, v_{p_h})]^T$ and

$$Q = \begin{pmatrix} 1 & u_{p_1} & v_{p_1} & u_{p_1} v_{p_1} \\ 1 & u_{p_2} & v_{p_2} & u_{p_2} v_{p_2} \\ & & \dots & \\ 1 & u_{p_h} & v_{p_h} & u_{p_h} v_{p_h} \end{pmatrix} \tag{5.20}$$

The result of this famous problem can be directly obtained by matrix equation:

$$\hat{C} = (Q^T Q)^{-1} Q^T \times d \tag{5.21}$$

The depth value of the pixel can be calculated by $D(u_0, v_0)$ with (u_0, v_0) is the center of the pixel. This process also works for vessel measure. In the rare event when $h < 4$, values of neighbor pixels can be used to feed the bilinear interpolation solution. At the end of this phase, a double image (I_d and I_V) of 160×128 pixels is obtained where each pixel has a couple of value (d, V) correspond to depth information and vessel measure.

5.3 Feature learning

As mentioned above, **the training phase** (Fig.5.8) includes three steps which are: first, the data representation by Gabor transformation, then the hierarchical feature

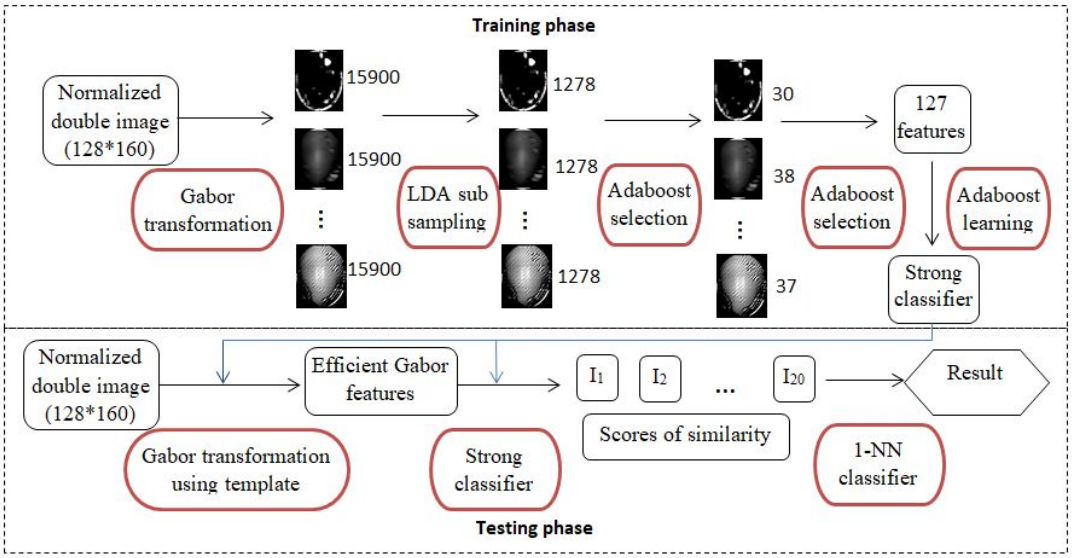


FIGURE 5.8: Learning and testing phase of classifier

selection using LDA sub-sampling and AdaBoost feature learning, and finally the AdaBoost classifier learning .

5.3.1 Gabor Transformation

In this study, 2D Gabor filters are applied to all double images in order to characterize each video. The Gabor wavelets contain information of spatial localization, orientation selectivity and spatial frequency selectivity . A lot of robust 2D face recognition algorithms use Gabor wavelet as the principal representation of face which places great emphasis in both spatial frequency and spatial relations. The Gabor kernel can be described as follows:

$$\Psi(Z) = \frac{k_{\mu,\nu}^2}{\sigma^2} exp(\frac{-k_{\mu,\nu}^2 Z^2}{2\sigma^2})[exp(ik_{\mu,\nu}Z) - exp(-\frac{\sigma^2}{2})] \tag{5.22}$$

where μ and ν represent the orientation and scale of the Gabor wavelets. $\Psi(Z)$ is the value of Gabor wavelet at $Z = (t_u, t_v)$. t_u, t_v are the centered coordination of any point in the plan. The coefficient $k_{\mu,\nu}$ is defined by $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$ with $k_\nu = k_{max}/f^\nu$ and $\phi_\mu = \pi\mu/8$ so that $e^{i\phi_\mu}$ determines the orientation of the kernel and k_ν places it into a scale. In this study, we use five scales $\nu \in \{0, 1, ..., 4\}$ and eight orientations $\mu \in \{0, 1, ..., 7\}$ which make 40 Gabor kernels with the other parameters as following: $\sigma = 2\pi$, $k_{max} = \pi/2$ and $f = \sqrt{2}$.

The representation of an image by Gabor wavelets, so-called the Gabor image, is the convolution of the image with a Gabor kernel. However, the convolution gives each pixel a complex value with two Gabor parts: the real part and the imaginary part. These two parts can be transformed to two types of information: Gabor

magnitude features and Gabor phase features. In this study, only Gabor magnitude features are used to describe the face. For 40 Gabor kernel, 40 Gabor image can be computed/ Each image is an ellipse of size 128x160 which includes about 15900 features, each feature has two dimensions (one dimension for the depth and another for the vessel intensity), so in total 1,272,000 features to feed into classification algorithm.

5.3.2 Feature Selection and final classifier

The richness of Gabor transformation in terms of quantity of features improves significantly the result of classification. However, the complexity of this algorithm increases with the number of features. Therefore, a scheme proposed by Chenghua Xu is applied to divide the whole system into small ones which can work in parallel. This hierarchical selection includes two stages:

LDA sub-sampling: for each Gabor vessel image, the optimal LDA sub-sampling extrudes massively non-efficient or redundant features by minimizing the within-class distance when maximizing the between-class distance .

Unlike the usual sub-sampling method where the sub-windows is uniformly distributed in the image, this optimal method aims for rich-information regions where the features could provide more proof of recognition. Gabor images under different orientation and scale may not share the same sub-sampling pixels. Therefore, 40 sets of sub-sampling positions are constructed correspondents to 40 Gabor double images. To minimize the within-class distance (explain by scatter matrix S_W) and maximize the between-class distance (S_B), The optimal discriminant vectors constructing the LDA subspace are computed by solving the following criterion in the standard LDA algorithm:

$$W^* = argmax(J(W)) = \frac{W^T S_B W}{W^T S_W W} \tag{5.23}$$

where

$$W = \left\{ \begin{pmatrix} w_{1,1} \\ w_{2,1} \\ ... \\ w_{p,1} \end{pmatrix} \begin{pmatrix} w_{1,2} \\ w_{2,2} \\ ... \\ w_{p,2} \end{pmatrix} \begin{pmatrix} ... \\ ... \\ ... \\ ... \end{pmatrix} \begin{pmatrix} w_{1,v_{max}} \\ w_{2,v_{max}} \\ ... \\ w_{p,v_{max}} \end{pmatrix} \right\} \tag{5.24}$$

Here, $p = 15900$ is the number of pixels in one image and v_{max} is the amount of discrimination vectors (each vector is one column of W). The summation vector V can be computed as follows:

$$V = \left(\sum_{k=1}^{k=v_{max}} |w_{k,1}|, \sum_{k=1}^{k=v_{max}} |w_{k,2}|, ..., \sum_{k=1}^{k=v_{max}} |w_{k,p}| \right) \tag{5.25}$$

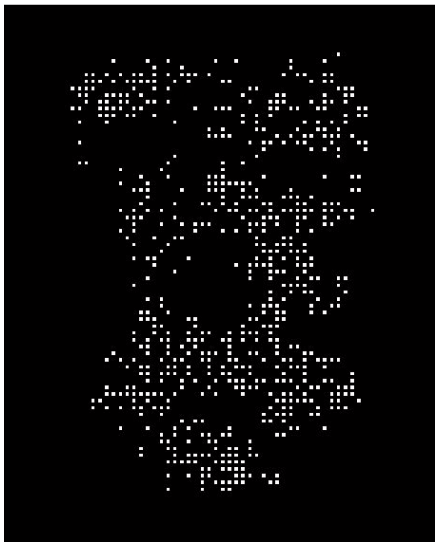


FIGURE 5.9: LDA template for Gabor image with $u = 4$ and $v = 8$

The magnitude of V at a particular position represents the corresponding variations among the training set, which also reflects the corresponding importance in distinguishing the faces. After this stage, only 1278 features in each image are chosen for AdaBoost selection.(Fig.5.9)

AdaBoost learning: a supervisor learning which applies a weak and tiny classifier on each feature of the sample in order to:

- Select the less redundant group of effective features which can discriminate the two hypotheses ,
- Construct weak classifiers using these features,
- Build a strong cascaded classifier .

The algorithm of Adaboost learning for feature selection can be introduced as below:

Given example couple images (I_1, J_1, y_1),(I_2, J_2, y_2)...(I_n, J_n, y_n) where y_i=1 when I_i and J_i are images of a same person (positive examples) and y_i=0 in other case (negative examples).

Initialize weights $w_{1,i}$=1/2m or 1/2l, for y_i=0,1, respectively, where m and l are the number of negatives and positives examples.

For t=1,...,T (T is the maximum number of chosen features)

- **1.** Normalize the weights: $w_{t,i} := \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$
- **2.** For each feature, j, train a LDA classifier h_j which using only this feature (which has 2 values). The error is evaluated by: $\epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$.
- **3.** choose the classifier h_j which minimizes the error ϵ_j. j is the feature chosen in this step.
- **4.**Update the weights $w_{t+1,1} = w_{t,i}\beta_t^{1-e_i}$ where e_i = 0 if example x_i is classified correctly, e_i = 1 otherwise and $\beta_t = \epsilon_j/(1-\epsilon_j)$

Each iteration, the algorithm searches for a feature that minimize l'error of classifier which is pondered by the weight of each sample in training set. By this way, each time the algorithm choses a feature as efficient feature, it will update the weight of all the samples so that the incorrectly classified samples become more important for next iterations.

The AdaBoost learning algorithm essentially solves a two-class classification problem. So, we have to reduce the face recognition multi-class problem to a two-class problem: intra-personal versus extra personal. Instead of one image, our examples become any couple of images possible. We have 4 persons and 5 videos training for each person who makes 40 intra-personal couples and 150 extra personal couples. In our study, since each feature has two values, the LDA classifier (Linear Discriminant Analysis) with two dimensions is chosen as the elemental "weak classifier" used in AdaBoost algorithm in order to consider simultaneously the depth and vessel informations of a feature.

In this stage, the AdaBoost selection is used twice as following:

- *Individual learning:* apply AdaBoost method to each Gabor double image to select the effective feature for each image (about 30-38 per image) and group all these features into one set.
- *Total learning:* apply AdaBoost to this set of features to reduce one more time the number of features (about 127 features in this case).

The final step of training phase is the construction of a cascaded strong classifier from these 127 features. That cascaded classifier contains many layers, each layer is also building by the efficient features in features learning stage. In fact, instead of constructing a big classifier of a lot of features in order to achieve a detection rate D and limit the false positive rate under F, the method aims to build some small independent classifiers that provide a higher detection rate d_l with a huge false positive rate f_l. When these classifiers are use as layers for a bigger one we can choose the layer so that:

$$F = \prod_{l=1}^{l=L} f_l \tag{5.26}$$

$$D = \prod_{l=1}^{l=L} d_l \tag{5.27}$$

where L is the number of layers. In this way, from 10 classifiers with high false positive rate (by 50%), cascaded classifier can be building which limits F at $0.5^{10} \approx 10^{-3}$. The algorithm of Adaboost learning for strong classifier can be introduced as below:

- **1.** Parameters initialization: selecting the value of f (the maximum acceptable false positive rate per layer) and d (the minimum acceptable detection rate per layer). This step depends essentially on the efficiency of features

- **2.**Target determination: selecting the value overall false positive rate (F_{target}). This step depends on the result we aim to
- **3.** P = set of positive examples.
- **4.** N = set of negative examples.
- **5.** Initialization: $F_0 = 1.0; D_0 = 1.0; l = 0$
- **6.** While $F_l > F_{target}$:
  - $l \leftarrow l + 1$
  - $n_l = 0; F_l = F_{l-1}$
  - While $F_l > f \times F_{i-1}$
    - $n_l \leftarrow n_l + 1$
    - Use P and N to train a classifier with n_l features using AdaBoost
    - Evaluate current cascaded classifier on validation set to determine F_l and D_l.
    - Decrease threshold for the l-th classifier until the current cascaded classifier has a detection rate of at least $d \times D_{l-1}$(this also affects F_l)
  - $N \leftarrow \emptyset$
  - If $F_i > F_{target}$ then evaluate the current cascade detector on the set of non-face images and put any false detections into the set N

The output of our experiment is a cascaded strong classifier with 10 layers and 96 features.

5.3.3 Testing phase

In testing phase, the process of preprocessing and feature extraction (Gabor transformation) is the same as in training phase. However, the features are extracted directly by using the selected template from training phase. The testing video is paired with each training video and generate 20 examples of couples. These 20 examples are classified into two classes: intra-personal and extra personal with a score of the similarity ($S(V_0, V_k)$ where $k \in 1, 2, ...20$) for two videos in each couple:

$$S(V_0, V_k) = \sum_{l=1}^{L} \sum_{p=1}^{n_l} \alpha_{l,p} h_{l,p}(V_0, V_k) \tag{5.28}$$

where $L = 10$ is the number of layers and n_l is the number of features in l-th layer. $h_{l,p}$ is a weak classifier which is based on one efficient double-feature and $\alpha_{l,p}$ is its weight. V_0 is the testing video and V_k where $k \in 1, 2, ...20$ are training video.

Among 20 scores of similarity, the $k - NN$ classifier chooses k training video which are the most similar to the input video to decide the result of the test (Fig.5.8). The number of reference sample (k) depends on the number of training samples for

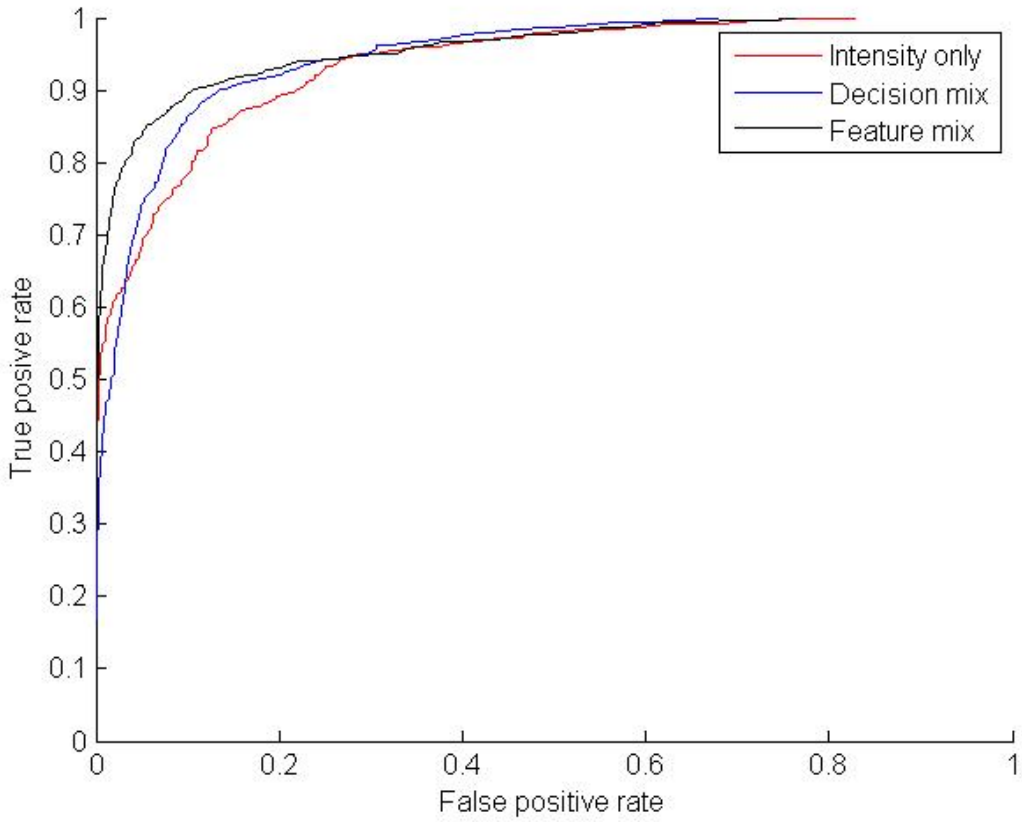


FIGURE 5.10: Roc curve of the strong classifier.

each face. Here in this study, because the database has only 5 samples for each face, we chose $k = 1$ which implies that only one nearest neighbor is enough to identify the face.

As none of the existing thermal database provides video with the movement required in this study, we have decided to test our method in our own thermal video database. This database contains 161 videos of 4 subjects which are captured using a Therm-app camera.

In experiment phase, the database is randomly divided into two subsets: the training set and the testing set. The training set contains 5 videos of each subject (20 in total) and the testing set contains 141 other videos. In order to prove the improvement of this method compared to local matching approach using Gabor transformation, 3 tests are evaluated at the same time: one uses only intensity information (Intensity only), another one mixes the intensity and depth data at the decision level (Decision mix), and the last one mixes the two types of information at feature level (our method: Feature mix).

The experimental results (Table 5.1) prove the advantage of mixing intensity and depth information in feature level. The same concept can be used with any other descriptor like LBP or Weber local descriptors. The different between the performance of three tests is reduced when the number or training videos increases (for training

TABLE 5.1: Average results after repeating 20 times the process of experimentation.

	Accuracy	Precision	Recall
Gabor-Intensity only	79.43%	77.62%	81.76%
Gabor-Decision mix	82.64%	80.12%	85.21%
Gabor-Feature mix	88.43%	87.05%	90.83%
LBP-Intensity only [**?**]	86.11%	85.05%	87.70%
LBP-Mix	87.36%	88.55%	88.79%

set of 40 videos, 10 for each subject, the accuracy of these tests is all around 99%). These results which came from the frontal pose of the head can be enhanced by mixing with other poses from the 3D model to achieve a complete process.
By comparing the proposed method with another process based on LBP [**?**], we observe that LBP is far better than Gabor in description of intensity data. However, by mixing these intensity data with deep information, we obtain a great amelioration in classification using Gabor features (9%) which help this last one surpass the method using LBP-features. This result becomes a solid evidence to our choice of using Gabor-description for geometry data.

5.4 Multi-pose recognition

The result of last experiments proves the advantage of mixing depth and intensity data at feature level. However,by using only frontal view, the process neglects a large source of information in profile views. In fact, the profile views are proven to be more effective than frontal image in automatic face recognition. At this stage, the study aims to get some projected images of 3D model in profile views in order to ameliorate the strong classifier.

5.4.1 Profile views images

The profile views are normally taken at 90°. However, the process of 3D reconstruction is based on a couple of frontal photos at the first step. The farther this process goes from the original couple, the less precise the performance of 3D reconstruction. 90° views are normally the ultimate views this process can get, but they are also the most imprecise. To bias between getting more information and resting accurate, the profile views are projected at 60°. Firstly, the 3D model will be rotated by 60° and -60°.

$$\mathbf{M_{left}} = \theta_{left} \times M \tag{5.29}$$

$$\mathbf{M_{right}} = \theta_{right} \times M \tag{5.30}$$

Where θ_{left} and θ_{right} are respectively the matrix of 60° rotation and -60° rotation. (Notice that the intensity value does not depend on the rotation)

$$\grave{}_{\mathbf{left}} = \begin{bmatrix} -\sin\frac{\pi}{3} & 0 & \cos\frac{\pi}{3} & 0 \\ 0 & 1 & 0 & 0 \\ \cos\frac{\pi}{3} & 0 & \sin\frac{\pi}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.31}$$

$$\grave{}_{\mathbf{right}} = \begin{bmatrix} -\sin-\frac{\pi}{3} & 0 & \cos-\frac{\pi}{3} & 0 \\ 0 & 1 & 0 & 0 \\ \cos-\frac{\pi}{3} & 0 & \sin-\frac{\pi}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.32}$$

The projection of these two models (M_{left} and M_{right}) by z-axis will be processed in the same way of frontal views to make 2 double-images: D_{left} from M_{left} and D_{right} from M_{right}.

$$\mathbf{D_{left}} = \begin{bmatrix} u_1^l & u_2^l & ... & u_N^l \\ v_1^l & v_2^l & ... & v_N^l \\ d_1^l & d_2^l & ... & d_N^l \\ V_0(1) & V_0(2) & ... & V_0(N) \end{bmatrix} \tag{5.33}$$

$$\mathbf{D_{right}} = \begin{bmatrix} u_1^r & u_2^r & ... & u_N^r \\ v_1^r & v_2^r & ... & v_N^r \\ d_1^r & d_2^r & ... & d_N^r \\ V_0(1) & V_0(2) & ... & V_0(N) \end{bmatrix} \tag{5.34}$$

where (u_i^l, v_i^l) and (u_i^r, v_i^r) $(i \in 1, ..., N)$ are respectively the projected coordination of i-th vertex in the left-plan and the right-plan. d_i^l and d_i^r are depth values which corresponds to this vertex.

The normalization phase is now more complicated as the oval model centered on the nose is no longer adapted. The oval model is determined by the nose at the extreme pixel of the left and the chin is at a fixed point of the oval

5.4.2 Experiments and results

The feature extraction of multi-view is similar to the frontal view. The complexity of the training process is estimated as:

$$\mathbf{T} = O(h^2k) + O(k^2) \tag{5.35}$$

FIGURE 5.11: Profile view.

where h is the number of pixels in a double-image and k the number of views using for classification. ($O(h^2k)$ is featured training time and $O(k^2)$ is strong cascaded classifier building time). As the number of features is tripled, the feature training-time is also tripled but the strong cascaded classifier building last 9 times more compared to mono-views process. Because of this link between the number of features and the

FIGURE 5.12: Normalized profile view.

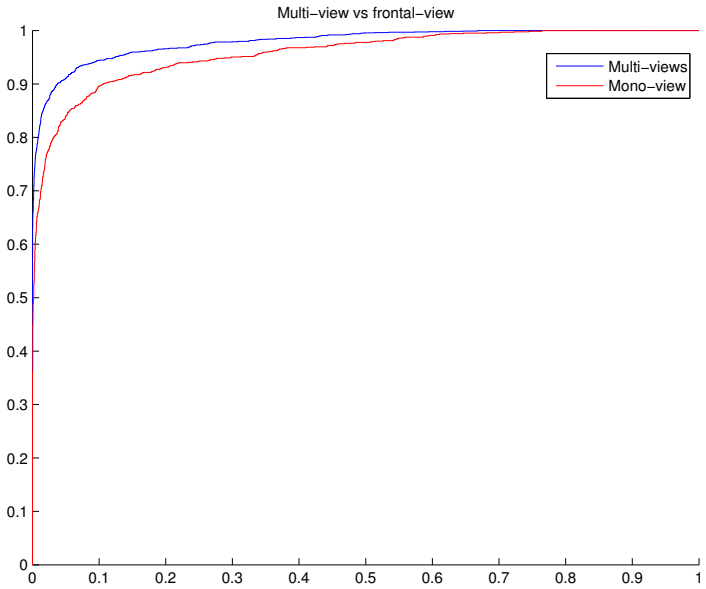


FIGURE 5.13: Roc curve of the strong classifier for mono-view and multi views.

TABLE 5.2: Average results for 20 test sets.

	Accuracy	Precision	Recall
Mono-view	88.43%	87.05%	90.83%
Multi-views	92.27%	89.25%	93.231%

training-time, the study is limited at 3 views which can provide information about most part of the face.

In experiment phase, the training-set contains 20 videos (5 for each subject). The test-set contains 100 videos among the rest. All the sets are chosen randomly. The classification is processed for mono-view and multi-views at the same time in order to provide a comparison between them.

The result ameliorate significantly from using only frontal view of using 3 principal views. This fact proves that a combination of views can represent a 3D model. The result can be ameliorated a little more by augmenting the number of views using in the combination. However, the increase in the computation cost reduces the performance of the method.

Chapter 6

Conclusions and Perpectives

6.1 Conclusion

The development of imaging technology and computing capacity lead us to an era where user's face can be considered as their proof of authentification toward an automatic system. The most convenient and natural method is trying to mimic the human's vision using computer vision. Because of this reason, visible imagery is the first option for every system authentication by facial recognition.

However, visible imagery technology is not robust enough to be used as the only source of identification information. This method has two major limits that make the authentication systems vulnerable. The first limit is its sensitivity to the illumination condition. The very nature of visible image that is captured inside trivial cameras is a reflection of the lighting which hit the object placed in front of the lens. This type of photo depends not only on the color of the object but also on the nature and intensity of light source. The second problem is an active attack where the attacker uses a signature of genius user's face in order to bypass the authentication system. The signature may be a photo, a video found in social network or even a 3D mask of this user.

In the first part of this study, we aim to construct a solution against the face-spoofing attack with minimum equipment required. Our hardest use-case associates to the face recognition method for smartphones with a visible camera. This use-case is complicated since the system is based on a single uncalibrated camera and the scene of authentication depends entirely on users. We adjust our general solution for this case by exploring the capacity of movement and some motion sensors inside the smartphone. From a set of video's frames, the method uses a 3D reconstruction process to build a 3D model of the head which is highly effective against photo-attack as differences in geometric features between a real object and an image is truly large. The video attack can be detected by observing the synchronization between the prior motion of the smartphone (explored by motion sensors) and the captured-motion calculated by the 3D reconstruction process. The limit of our first study is that it can only justify if the object has a real 3D form of user's faces, but it is not able to detect a mask attack.

The delicate mask attack is not readily revealed by visible imagery technology because its imprint is very close to the genius face. However, in thermal imagery where the emission source of the spectrum is human's face, the detection of all types of face-spoofing attack is trivial. The thermal imagery technology can solve the other major problem of visible imagery concerning illumination conditions. Though, in general, thermal images provide less information than visible images. In our second study, we aim to improve the performance of infrared face- recognition method by using a 3D model of the head computed from a thermal video. Vascular network build from the thermal video is now observed in intensity level and geometric feature. The depth information and blood vessels data aren't handled like unchained marks but are associated to form a single feature with two values. By this way, the face recognition method bases mainly on the 3D position of the vascular network.

6.2 Perpectives

For the face-spoofing detection method which is introduced in chapter 4, there is always a type of attack which is not considered in the study: the 3D mask attack. Until this step, our face spoofing detection is independent of the face recognition process. We do not use the same database as authentication system but a database dedicated to our solution. In fact, the geometric information can be used further in face recognition phase which makes 3D mask attacks detectable in this layer. However, this type of solution requires enough images of each user in different positions in the database to compare with the attempt. Another solution is to study the nature of the object's material to distingue the genius face from the attacks. In fact, each type of material generates a special imprint of noise which can be used to detect if the object is made by human's skin.

The thermal face recognition method comes all the way to construct a 3D model of the vascular network. However, we do not make direct use of this 3D model but its projection into a set of depth images. The process with multi-pose depth-images augment our method's performance but cannot achieve an entire comparison of 3D model. In some future works, we aim to construct a method that can study directly these 3D model in order to make uses of all the provided information.

Appendix A

Résumé en français

A.1 Contexte et problématique

Le visage humain est une caractéristique qui permet de distinguer une personne des autres. Reconnaître les parents et la famille est la toute première leçon pour chaque être humain. Depuis le développement des technologies d'imagerie, le cerveau humain est chargé d'une autre tâche : la reconnaissance d'un visage à partir d'une photo. Cette tâche est aujourd'hui la solution d'identification la plus critique dans notre société car notre visage apparaît sur de nombreux papiers tels que des passeports, carte d'identité, permis de conduire, carte d'étudiant... A l'ère du numérique, la reconnaissance faciale est de plus en plus confiée à des systèmes automatiques. Le plus puissant ordinateur est capable d'accomplir de nombreuses tâches complexes, y compris la détection et l'authentification des personnes, le suivi et la prédiction des mouvements, la détection et la classification des maladies.... La reconnaissance faciale est maintenant appliquée dans une large gamme de cas d'utilisation avec différents niveaux et contraintes de sécurité. Il existe des systèmes passifs utilisés par les autorités. Il existe des systèmes actifs calibrés pour le contrôle d'accès dans les aéroports, les entreprises et de nombreuses autres installations. Il existe également de nombreux systèmes distribués non calibrés comme le déverrouillage d'un ordinateur portable et l'identification par un téléphone.

Cependant, la reconnaissance faciale dans le visible est théoriquement vulnérable à l'attaque par usurpation d'identité. Un système d'authentification performant peut être facilement contourné en utilisant une photo du visage d'un utilisateur présentée devant l'appareil photographique. La menace est d'autant plus importante que de nombreuses personnes laissent leurs photos sur Internet, en particulier sur les réseaux sociaux. Afin de renforcer le processus d'authentification, il est possible d'ajouter au système une nouvelle couche de sécurité qui peut réduire cette vulnérabilité. Les solutions existantes sont vastes en termes de technologie, mais presque toutes les méthodes nécessitent un système complexe avec deux ou plusieurs caméras et même d'autres types de capteurs. Plus le système est complexe, moins la solution est difficilement applicable. C'est pourquoi, dans cette étude, nous proposons une nouvelle méthode de détection dédiés aux systèmes mono-caméra comme les smartphones et les tablettes. Cette nouvelle méthode exploite une vidéo

du visage de l'utilisateur dans de nombreuses poses afin de reconstruire le visage 3D qui est ensuite utilisé pour distinguer un visage des tentatives d'attaques.

Un autre problème crucial de la reconnaissance faciale dans le visible est lié au fait que toute la lumière et la couleur que l'on peut observer sur les visages humains n'est qu'un reflet de la lumière venant d'autres sources comme le soleil ou les lampes. L'imagerie dans le visible dépend fortement des conditions d'éclairage. Certaines recherches proposent des méthodes qui peuvent fonctionner correctement à travers le changement d'intensité lumineuse, mais il y a toujours une diminution en termes de précision lorsque la modification est importante. Aucune méthode de reconnaissance faciale dans le visible ne peut être traitée avec un manque important de lumière. Dans ce contexte, l'imagerie infrarouge et particulièrement l'imagerie thermique est devenue une méthode alternative et complémentaire prometteuse pour la reconnaissance faciale. Cependant, jusqu'à ce jour, la reconnaissance du visage avc une caméra thermique n'atteint pas le niveau de maturité requis pour être appliquée à grande échelle. En fait, l'application des images infrarouges dans le processus de reconnaissance faciale est remise en question par l'absence de caractéristiques distinctes sur ces images. Le spectre IR a ses propres problèmes qui peuvent affecter la précision du programme d'identification. Pour faire face à ce problème, nous avons introduit une nouvelle méthode de reconnaissance faciale thermique utilisant un modèle 3D de la tête qui contient des informations sur le réseau vasculaire de la tête. Le processus est dédié au fonctionnement dans différents cas d'utilisation où le seul équipement requis est une caméra thermique.

A.2 Descriptif

Le premier chapitre de la thèse présente, d'une manière générale, le contexte et les problèmes qui ont conduit à cette étude. Il décrit également la structure de cette thèse. Le deuxième chapitre est consacré à présenter un aperçu du domaine de la reconnaissance automatique de visages qui, dans cette étude, est divisé en deux sections principales : la reconnaissance du visage dans le visible et la reconnaissance du visage à partir d'images thermiques. Dans la première section de ce chapitre, la thèse présente l'avantage et la problématique de l'imagerie dans le visible. Il fournit également quelques méthodes représentatives depuis le début de la vision par ordinateur jusqu'à ce jour. La technologie infrarouge est présentée dans la section suivante où nous soulignons comment elle peut contourner la problématique de l'imagerie dans le visible.

Le troisième chapitre présente une nouvelle approche pour modéliser la tête des utilisateurs par des caractéristiques 3D. Le chapitre commence par examiner l'avantage de représenter les données faciales des utilisateurs à l'aide de son modèle 3D. Ensuite, la section suivante donne un aperçu de la méthode de reconstruction

3D à l'aide des différents capteurs. La dernière partie du chapitre présente le processus détaillé de réalisation de ce modèle 3D à partir d'une seule vidéo du visage de l'utilisateur.

Le quatrième chapitre construit et examine une nouvelle méthode pour la détection d'un des problèmes majeurs dans le domaine de la reconnaissance faciale dans le visible : l'attaque par usurpation d'identité. Il commence par un rappel de l'attaque de face-spoofing et comment elle peut affecter le processus d'authentification. La section suivante décrit le schéma détaillé pour détecter cette attaque. Dans la dernière partie, nous fournissons la performance de cette méthode.

Dans le cinquième chapitre, nous proposons une nouvelle solution de reconnaissance faciale utilisant un modèle 3D de la tête obtenu à partir d'une vidéo thermique qui contient des informations sur le réseau vasculaire. De par sa nature, l'imagerie thermique peut évidemment détecter l'attaque par projection du visage et rester invariable aux conditions d'éclairage. Cependant, il y a moins d'informations identifiables dans une image thermique qu'une image visible, ce qui réduit la précision du processus de reconnaissance faciale. La reconstruction 3D fournit des données géométriques du visage qui peuvent être mélangées avec les informations du réseau vasculaire issues de l'imagerie thermique pour améliorer les performances.

A.3 Détection d'une attaque de l'usurpation de visage à l'aide d'un modèle 3D

Ce chapitre construit et examine une nouvelle méthode pour la détection d'un des problèmes majeurs dans le domaine de la reconnaissance faciale dans le visible : l'attaque par usurpation d'identité. Il commence par un rappel de l'attaque de face-spoofing et comment elle peut affecter le processus d'authentification. La section suivante décrit le schéma détaillé pour détecter cette attaque. Dans la dernière partie, nous fournissons la performance de cette méthode.

A.3.1 Introduction

L'authentification par la reconnaissance faciale permet de renforcer un système d'authentification en intégrant un facteur lié à l'identité d'une personne. Cependant, la reconnaissance faciale dans le visible est théoriquement vulnérable à l'attaque par usurpation d'identité. Un système d'authentification performant peut être facilement contourné en utilisant une photo du visage d'un utilisateur présentée devant l'appareil photographique. La menace est d'autant plus importante que de nombreuses personnes laissent leurs photos sur Internet, en particulier sur les réseaux sociaux.

Afin de renforcer le processus d'authentification, il est possible d'ajouter au système une nouvelle couche de sécurité qui peut réduire cette vulnérabilité. Les solutions existantes sont vastes en termes de technologie, mais presque toutes les méthodes nécessitent un système complexe avec deux ou plusieurs caméras et même

d'autres types de capteurs. Plus le système est complexe, moins la solution est difficilement applicable. C'est pourquoi, dans cette étude, nous proposons une nouvelle méthode de détection dédiée aux systèmes mono-caméra comme les smartphones et les tablettes. Cette nouvelle méthode exploite une vidéo du visage de l'utilisateur dans de nombreuses poses afin de reconstruire le visage 3D qui est ensuite utilisé pour distinguer un visage des tentatives d'attaques.

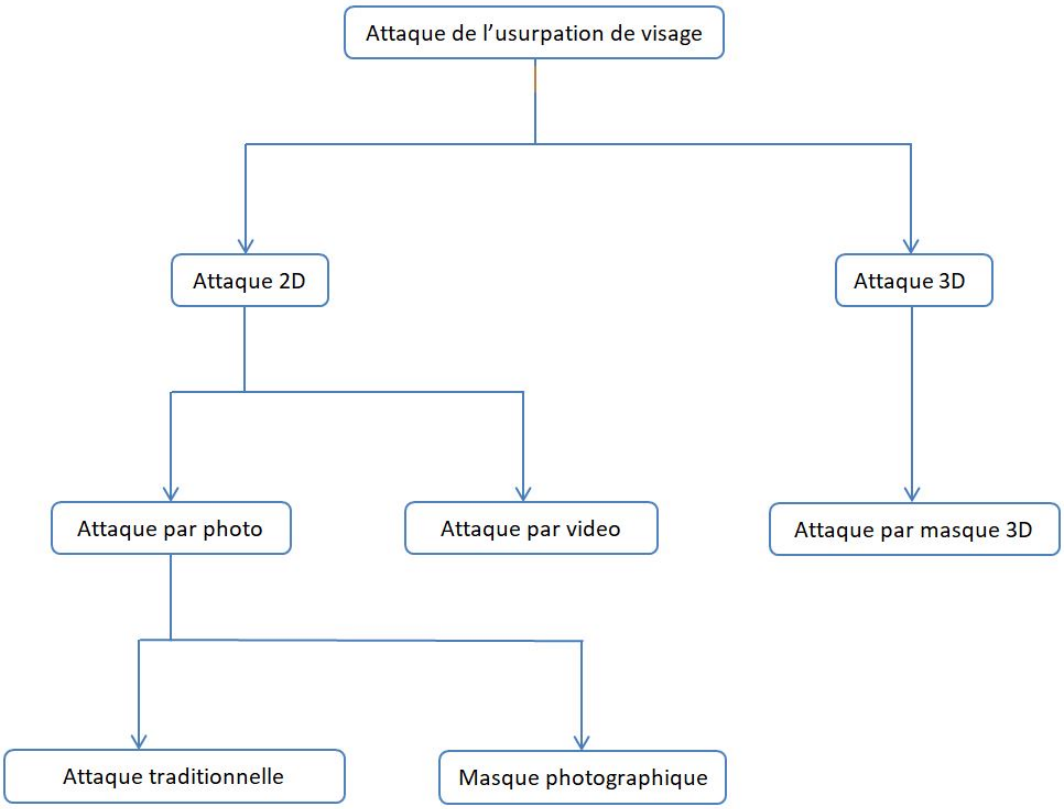


FIGURE A.1: Différentes types de l'attaque de l'usurpation de visage.

A.3.2 Description de la détection de l'attaque

Les principales difficultés de la reconnaissance faciale dans le domaine visible à l'aide de smartphones ou tablette est la variation des conditions d'éclairage, de l'orientation ou encore d'arrière-plans non maîtrisés. De plus le mouvement de l'appareil et le mouvement relatif du système d'acquisition peuvent affecter la qualité de l'acquisition. De même la calibration du système est un frein à l'efficacité de la méthode de détection. Cependant, la présence de différents capteurs dans un smartphone se révèle être un avantage qui nous permet de développer une nouvelle solution opur la détection de Face spoofing. En effet, les capteurs de mouvement et la possibilité des smartphones de gérer plusieurs tâches simultanément permettent de récupérer les informations liées au mouvement pendant la prise de vues de l'utilisateur. a l'issue de cette première étape, nous obtenons une vidéo du visage

de l'utilisateur ainsi que les données brutes des capteurs. Dans le cas d'une authentification légitime, les données liées au mouvement sont cohérentes avec celles de la prise de vue. Ce n'est généralement pas le cas dans le cadre d'une attaque par Face spoofing. La solution que nous proposons repose sur l'exploitation de la cohérence entre ces données.

La méthode proposée se décompose en trois étapes principales présentées sur la Figure A.2.

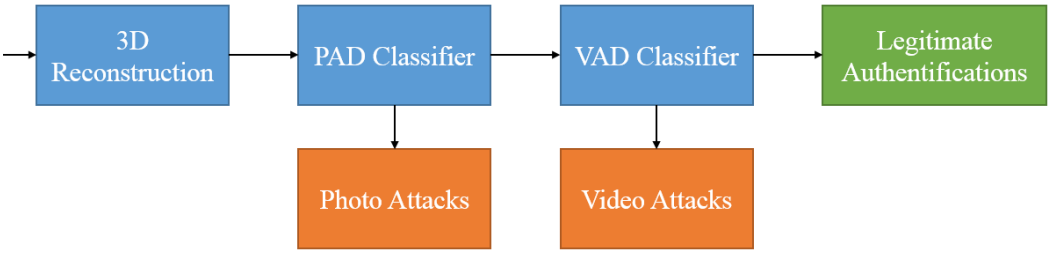


FIGURE A.2: Processus de détection de l'attaque de l'usurpation de visage.

Le première étape consiste à estimer un modèle 3D du visage grâce à un processus de reconstruction 3D. Ensuite un classifieur PAD (Photo Attack Detection) utilise la représentation 3D pour détecter si le visage a été remplacé par une photo (c'est-à-dire si une impression ou une image numérique du visage de l'utilisateur est utilisée pur l'authentification). Si cette étape est validée, un deuxième classifieur VAD (Video Attack Detection) est utilisé pour détecter si l'authentification a été réalisée à partir d'une vidéo. Cela permet de détecter une "video-replay" attaque. Les reconstructions 3D qui satisferaient les deux détections seraient considérées comme authentiques et l'authentification seraient donc validées.

Mis à part les tests de mouvement du téléphone qui sont demandés explicitement à l'utilisateur (de haut en bas et de gauche à droite), le processus est automatisé. Une application Android regroupe la collecte des informations (vidéo et données issues des capteurs), la reconstruction 3D et les classifications. Bien que la reconstruction 3D ne soit pas en temps réel pour le moment, dans un scénario réaliste, il est recommandé d'effectuer la reconstruction 3D et la classification finale sur un serveur dédié.

Les sections suivantes présentent les différentes étapes du processus de détection de Face spoofing.

Reconstruction 3D du visage

Dans la méthode proposée, un modèle 3D de l'objet (visage réel ou falsifié) est construit à partir de la vidéo capturée durant le processus d'authentification. Pour obtenir une bonne qualité du modèle 3D, il est demandé à l'utilisateur de bouger la caméra du téléphone autour du visage de manière à ce que différentes poses soient

capturées. Nous considérons deux mouvements simples : selon la direction verticale et selon la direction horizontale. Ces mouvement permettent notamment de simplifier la mesure de cohérence nécessaire à notre algorithme.

La reconstruction 3D est réalisée avec VisualSFM, une application développée par C. Wu [108] qui utilise le mouvement (Structure From Motion). Cette application nécessite une séquence d'images en entrée et renvoie une reconstruction 3D d'un objet capturé ainsi que des informations relatives aux poses de la l'appareil photo. D'autres solutions plus récentes pourraient être utilisées. Le choix de la solution utilisée n'affecte pas beaucoup le résultat mais dépend essentiellement des capacités de calcul du smartphone utilisé.

Chaque image F_i de la vidéo ($i = 1, ..., n$ où n est le nombre d'images) est comparée aux autres avec la méthode SIFT (Scale-Invariant Feature Transform) ou Transformation de Caractéristiques visuelles Invariante à l'Echelle. Deux images (F_{c_1}, F_{c_1}) maximisant l'indice de similarité sont choisies pour former la base de l'objet 3D. Après cela, une extraction des contours et des points d'intérêt communs est appliquée pour toutes les paires d'images : (F_i, F_{c_j}) où $i = 1, ..., n$ et $j = 1, 2$. Ces caractéristiques (contours et points d'intérêt) sont suivies d'une image à l'autre afin de calculer la position et l'orientation de chaque image. Différentes vues d'un point d'intérêt sont extraites pour estimer sa profondeur et donc ses coordonnées 3D. L'étape suivante est un filtre des points clés apparaissant dans plusieurs images pour réaliser le modèle 3D.

La Figure A.3 présente un exemple de visage réel reconstruit en 3D à l'aide d'un nuage de points.

FIGURE A.3: Nuage de points du modèle 3D d'un visage réel

Détection de l'attaque par photo (PAD)

Dans le cas d'une attaque par photo, la reconstruction 3D donnée par la méthode SFM est très différente de celle obtenue dans le cas d'un visage réel. La Figure A.4

porésente différentes vues d'une reconstruction 3D d'un visage imprimé.

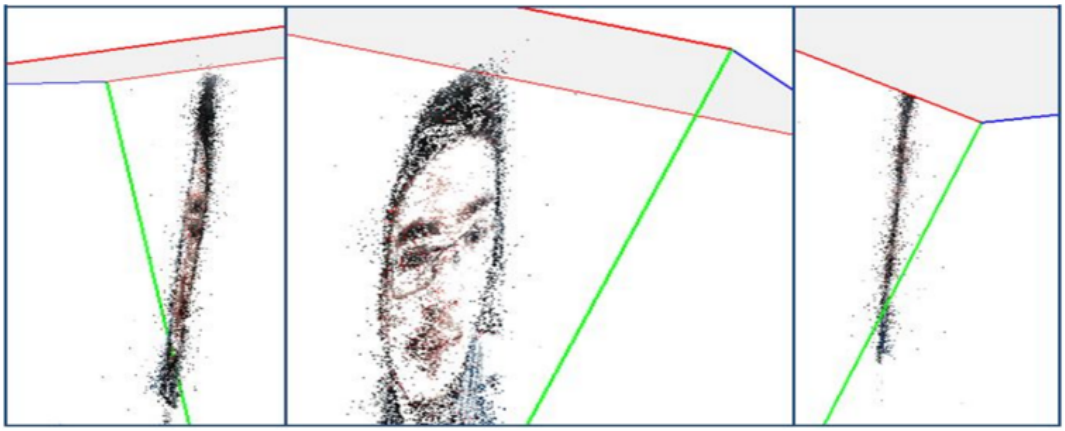


FIGURE A.4: Différentes vues de la modèle 3D d'une attaque par photo.

Détection de l'attaque par vidéo (VAD)

On peut facilement voir que la reconstruction 3D du visage est aplatie dans le cas d'une attaque par photo. Cela s'explique notamment par le fait qu'un visage réel est un vrai objet 3D qui contient davantage d'information sur la profondeur qu'un visage imprimé sur une feuille. Ainsi, plus le modèle 3D est plat, plus grande est la probabilité qu'il s'agisse d'une attaque par photo. Nous avons donc choisi de baser notre détection d'une attaque par photo sur le critère de l'épaisseur de la reconstruction 3D.

L'épaisseur d'une reconstruction 3D peut être estimée au moyen d'une Analyse en Composantes Principales (ACP). Cette méthode permet de transformer la représentation 3D en un nouveau système de coordonnées $(w = (w_{(1)}, w_{(2)}, w_{(3)}))$ où chacune des coordonnées est représentée par une composante principale. Cette transformation est définie de manière à ce que la première composante principale $(w_{(1)})$ ait la variance la plus grande, et que chacune des composantes principales aient la plus grande variable possible sous la contrainte d'être orthogonale aux composantes précédentes. De cette manière, la variance d'un point du nuage projeté sur la dernière composante $(w_{(3)})$ est le minimum de tous les vecteurs de l'espace qui peut être utilisé pour représenté l'épaisseur.

$$w_{(3)} = \left[\arg\min_{\|w\|=1} \sum_{i=1}^{n} (m_i.w)^2 \right]$$

où n est le nombre de points du nuage, $m_i = (x_i, y_i, z_i)$ avec $i = 1, .., n$ est le vecteur de coordonnées du i-ème point.

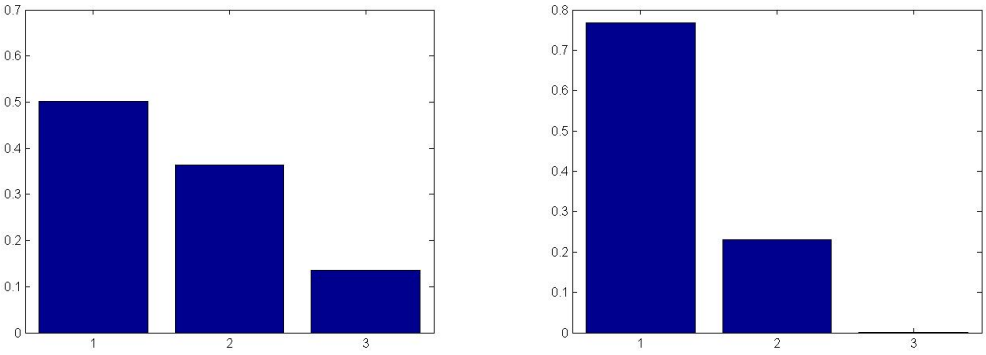
Pour simplifier l'ACP, les colonnes de la matrice $\mathbf{M}$ sont centrées pour avoir une moyenne nulle. La matrice de composantes principales $\mathbf{P}$ est définie comme une transformation linéaire orthogonale de la matrice $\mathbf{M}$: $\mathbf{P} = \mathbf{MW}$.

Soit v_j $(j = 1, 2, 3)$ la variance de la i-ème colonne de $\mathbf{P}$. L'ordre de grandeur de chaque colonne, noté d_i, est donné par :

$$d_i = \frac{v_i}{v_1 + v_2 + v_3}$$

Dans le cas d'une attaque par photo, les points de la reconstruction 3D sont dans un plan. Ainsi, il n'y a quasiment pas d'information dans la troisième composante ce qui rend d_3 très petit. En revanche, pour un visage réel, l'épaisseur joue un rôle important dans l'information totale.

Les figures A.5 (a) et (b) présentent une illustration des trois composantes principales obtenues respectivement pour un modèle 3D de visage réel (a) et falsifié (b).



(a) (b)

FIGURE A.5: PCA de la modèle 3D d'un utilisateur (a) et d'une attaque (b).

La différence est telle qu'un simple classifieur SVM sur les ordres de grandeur d_i peut être utilisé comme classifieur pour la détection d'attaque par photo (PAD).

Attaque par photo avancée

Supposons que l'attaquant utilise une attaque plus évoluée en utilisant des photos déformées qui généreraient un modèle 3D avec une profondeur significative. Dans ce cas, l'ACP n'est pas suffisamment performante pour la détection. Nous proposons donc une méthode pour extraire la profondeur de l'image d'un visage à partir du modèle 3D pour produire une autre preuve d'authenticité.

La première étape consiste à extraire une image de profondeur en fixant un plan perpendiculaire au vecteur normal d'une vue et en calculant la distance de chaque vertex à ce plan. L'ensemble de ses distances forment une carte appelée image de profondeur (voir Figure A.6).

A chacun des pixels de l'image de profondeur est lié un vertex et son intensité. Cependant, le processus de reconstruction 3D n'est pas toujours stable puisqu'il donne une mesure relative et non absolue. C'est pourquoi la profondeur doit être

FIGURE A.6: Image de profondeur d'un visage

ajustée pour être comprise dans $[0, 1]$. Le nuage de points n'est pas uniforme : certaines régions contiennent davantage de points que d'autres régions. Deux normalisations sont donc nécessaires : recadrer la zone du visage avec un masque et une détection du nez et d'autre part, l'échantillonnage de l'image de profondeur en une image de 320×256 pixels.

L'étape suivante consiste à faire une transformation de Gabor. Des filtres de Gabor 2D sont appliqués à toutes les images de profondeur afin de caractériser chacune des vidéos. Les ondelettes de Gabor permettent d'obtenir des informations de localisation spatiale et d'orientation. La représentation d'une image par les ondelettes de Gabor est une convolution de l'image avec un noyau de Gabor. Nous n'utiliserons que les caractéristiques liées à l'intensité pour décrire le visage. Nous obtenons au total 2 544 000 caractéristiques à fournir à l'algorithme de classification.

Après avoir fait la transformation de Gabor, il faut choisir les caractéristiques et construire le classifieur final. Le nombre important de caractéristiques obtenues suite à une transformation de Gabor améliore significativement les résultats de la classification. Cependant, la complexité de l'algorithme augmente également avec le nombre de caractéristiques. Il est proposé de répartir les calculs selon la méthode de Chenghua Xu en divisant le système global en plus petits qui peuvent être traités en parallèle. La sélection des caractéristiques s'effectue en deux étapes : un sous-échantillonage LDA et un apprentissage avec AdaBoost. Pour chacune des images Gabor, le sous-échantillonnage LDA permet d'éliminer les caractéristiques non efficientes ou redondantes en minimisnt la distance au sein d'une même classe et en augmentant la distance entre les classes. L'apprentissage AdaBoost permet quant à lui de sélectionner les ensemble de caractéristiques qui discriminent le mieux les deux hypothèses, de construire un classifieur utilisant ces caractéristiques et de construire un classifieur final en cascade. Nous obtenons au final un classifieur en cascade avec 12 étapes et 108 caractéristiques.

Attaque par vidéo

Dans le scénario d'une attaque par vidéo, une vidéo avec le visage de l'utilisateur est diffusée sur une écran LCD devant le smartphone. La vidéo est faite de telle sorte que le visage bouge de la même manière pour simuler le processus d'authentification. Cette attaque trompe le détecteur PDA puisque le mouvement de la tête permet d'obtenir différentes vues du visage de l'utilisateur et donc de reconstruire un modèle 3D réaliste du visage. Nous nous proposons donc d'utiliser les capteurs du smartphone pour notre nouveau détecteur.

La Figure A.7 présente un exemple des positions de l'appareil photo estimées à partir de la reconstruction 3D.

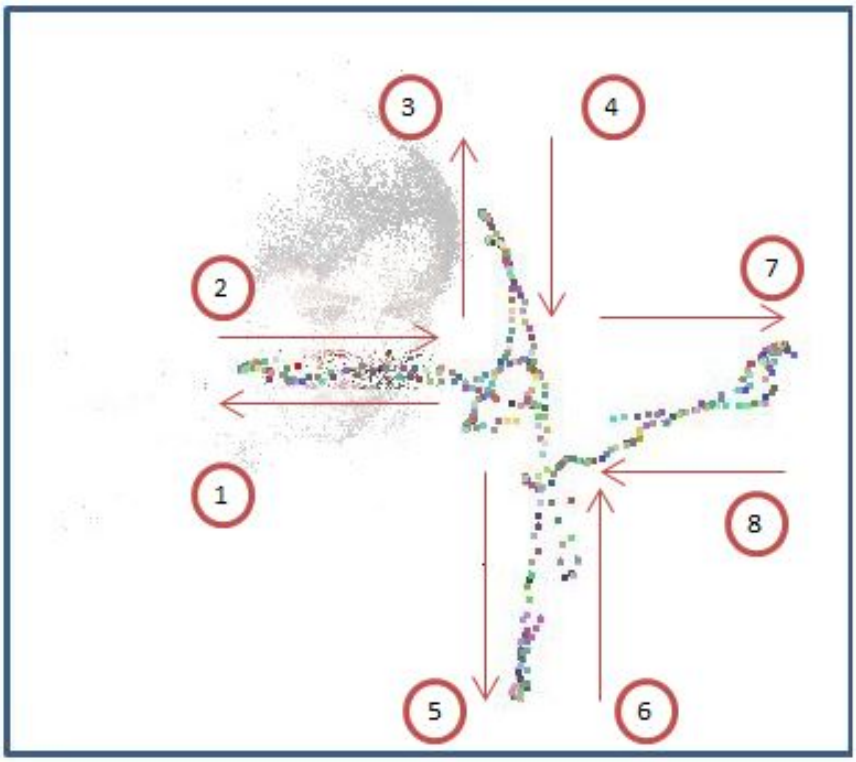


FIGURE A.7: Positions de l'appareil photo estimées à partir de la reconstruction 3D.

Nous pouvons également observer la trajectoire du smartphone enregistrée par les capteurs de mouvement (accéléromètre ou gyroscope). Le capteur gyroscopique mesure la vitesse angulaire alors que l'accéléromètre mesure l'accélération linéaire de l'appareil.

Les poses du smartphone sont comparées aux informations mesurées par les capteurs de mouvement afin d'obtenir un index de similarité. Pour estimer la dissimilarité, une simple corrélation peut être appliqué pour chacun des couples de données : $(X_i,\widehat{X}_i)$ et $(Y_i,\widehat{Y}_i)$, où X_i (resp. Y_i) représente un vecteur de position et d'orientation selon l'axe x (resp. y), $\widehat{X}_i$ (resp. $\widehat{Y}_i$) un vecteur de position et d'orientation calculé à partir des capteurs de mouvement selon l'axe x (resp. y). Ces caractéristiques de

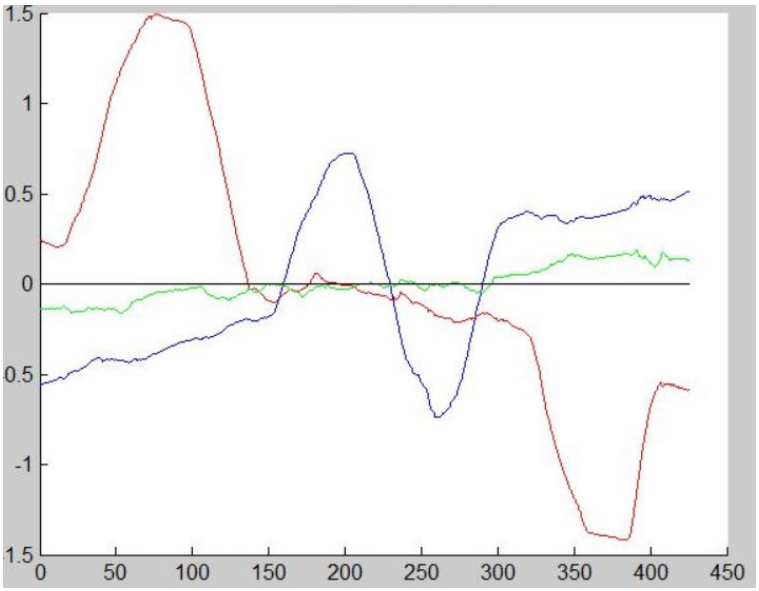


FIGURE A.8: Orientation de l'appareil photo à partir des données du capteur gyroscopique (θ_i^x en bleu, θ_i^y en rouge, θ_i^z en vert)

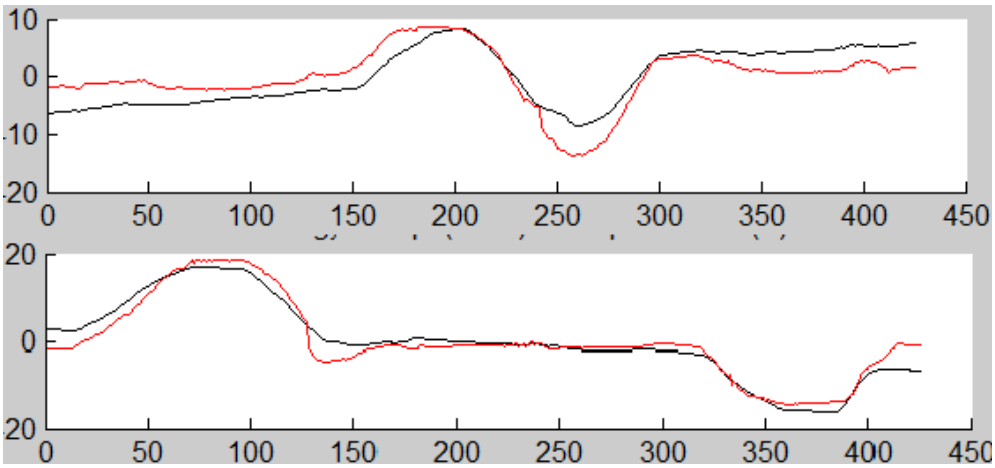


FIGURE A.9: Corrélation entre θ_i^x (noir) et $\widehat{\theta_i^x}$ (rouge) et entre θ_i^y (black) et $\widehat{\theta_i^y}$ (red).

corrélation sont les données discriminantes qui permettent à un classifieur SVM, le VAD de réaliser la détection souhaitée.

A.3.3 Performances de la méthode de détection proposée

Aucune base de données existante ne répond aux contraintes attendues par notre détecteur d'attaque par vidéo (utilisant les données issues des capteurs de mouvement). Nous avons donc effectué nos tests sur une base de données crée dans le laboratoire. Cette base de données contient 1001 vidéos de 3 personnes avec les données des capteurs de mouvement. Ces vidéos contiennent 451 cas d'authentification légitime, 362 cas d'attaque par vidéo et 188 cas d'attaque par photo. Les vidéos sont capturées sous différentes conditions d'illumination et de vitesse de mouvement. 3

smartphones ont été utilisés : 2 Samsung Galaxy Alpha et un Samsung Galaxy Tab. La base de données est séparée en deux ensemble contenant chacun les trois types de vidéos (légitime, attaque par photo et attaque par vidéo). Un ensemble permet de réaliser l'apprentissage, l'autre permet d'effectuer les tests.

Sur la courbe ROC (Figure A.10), nous observons que la méthode proposée présente de meilleurs résultats que la méthode LBP, en particulier pour un taux faible de fausse alarme. Une partie des fausses alarmes provient de la phase de capture. En effet, si l'image est floue , la vitesse trop grande, ou que le visage sort du cadre, la phase de reconstruction 3D n'est pas stable. Dans un scénario réel, une deuxième phase de capture serait effectuée. Nous pouvons mesurer l'amélioration suite à cette deuxième capture sur la courbe ROC (jaune). 4.11.

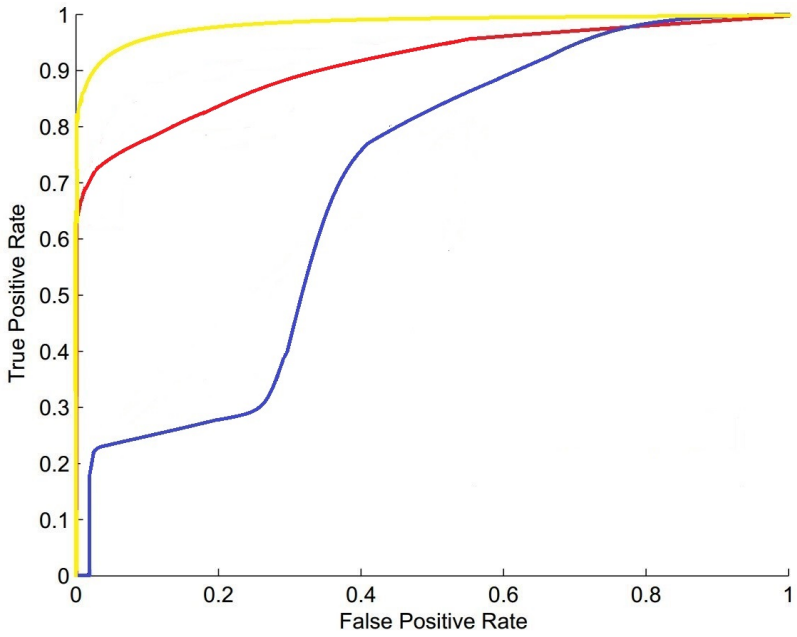


FIGURE A.10: Courbe ROC de la méthode proposée (rouge) en comparaison avec la méthode LBP (bleu). La méthode proposée avec une seconde prise (jaune)

Dans le cas de l'attaque par photo, la classification est réalisée essentiellement à partir de la profondeur de la reconstruction 3D. LA différence entre une image falsifiée et une image réelle est suffisamment importante pour que le taux de détection avoisine les 100%. Dans le cas de l'attaque par vidéo, nous considérons le pire cas, c'est-à-dire que l'utilisateur a un contrôle total sur la phase de capture.Nous ne savons pas ce qu'il filme, quel mouvement il effectue ou quel smartphone il utilise. Une erreur peut être générée par la manipulation du smartphone, c'est pourquoi la possibilité d'une deuxième prise réduit significativement le taux de non détection. Une attaque avec un masque parfait serait efficace contre nos détecteurs. C'est pourquoi dans le chapitre suivant, nous présentons une méthode de reconnaissance faciale utilisant un capteur IR qui permet détecter cette attaque.

A.4 Reconnaissance du visage par imagerie thermique

La reconnaissance du visage par imagerie thermique est la deuxième application étudiée dans cette thèse. Nous proposons d'étudier une nouvelle méthode d'authentification à partir d'images thermiques. Dans cette partie, nous présentons une nouvelle méthode de reconnaissance faciale à partir d'une caméra thermique mobile non étalonnée.

A.4.1 Problématique

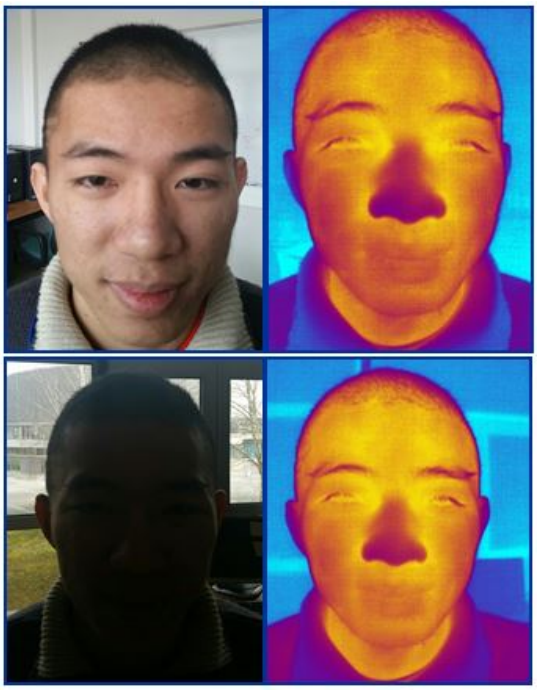


FIGURE A.11: Les images thermiques ne sont pas sensibles à la condition d'illumination

La reconnaissance du visage est une méthode d'authentification biométrique non-intrusive. L'utilisateur n'a pas besoin de présenter des preuves supplémentaires au système d'authentification, seul son visage est suffisant. Cette méthode est utilisée dans plusieurs systèmes de sécurité et de surveillance. Les photopgraphies numériques dans le domaine visible sont fortement sensibles aux changements d'illumination. C'est la raison pour laquelle, nous avons proposé d'utiliser des caméras thermiques comme source secondaire pour la reconnaissance faciale.

En effet, une caméra thermique capture une partie du spectre infrarouge (3-8μm et/ou 8-15μm) pour estimer la température des objets placés devant la lentille. Elle permet de visualiser pixel par pixel une image thermique de la scène qui est robuste aux changements des conditions d'acquisition (figure A.11).

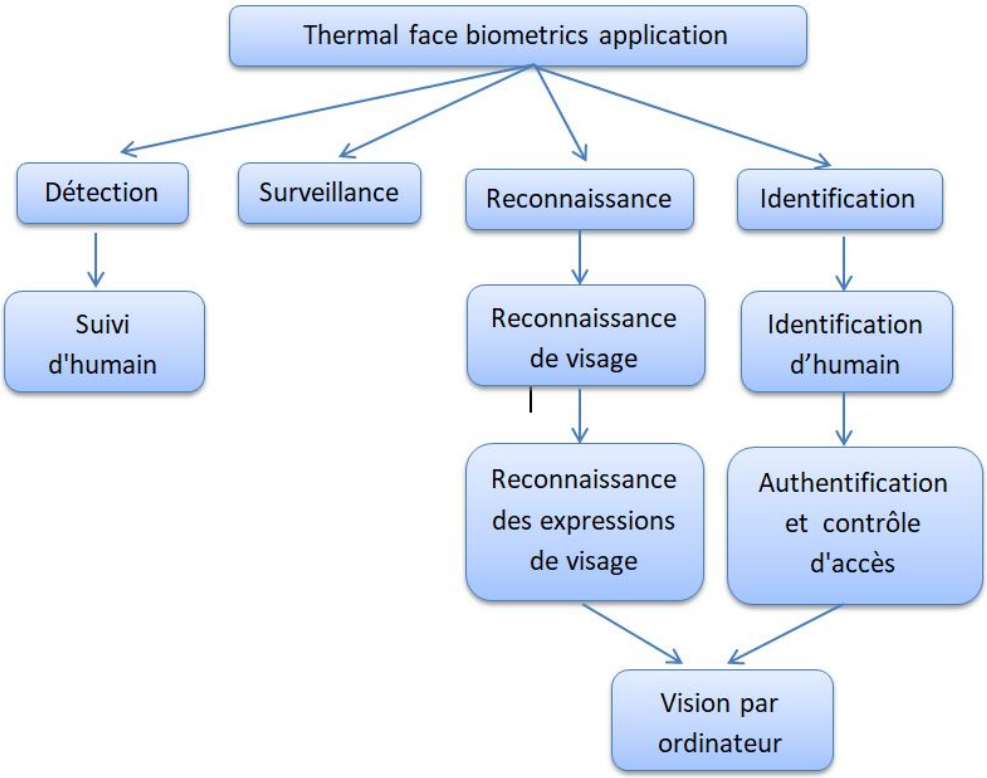


FIGURE A.12: Différentes applications.

A.4.2 État de l'art

Des méthodes de reconnaissance faciale dans l'infrarouge peuvent se regrouper en 4 groupes:

- *Apparence globale :* utiliser la totalité de l'image infrarouge de l'apparence d'un visage pour la reconnaissance.
- *Caractéristique:* utiliser des caractéristiques extraites de l'image infrarouge tels que la géométrie du visage, son réseau vasculaire ou la figure de perfusion sanguine.
- *Multi-spectral:* modéliser le processus de formation d'une image infrarouge pour décomposer des images de visages. Certaines approches utilisent directement les données de capteurs d'imagerie multi-spectrale ou hyperspectrale pour obtenir des images faciales à travers différentes sous-bandes de fréquences.
- *Multi-modal:* combiner des informations contenues dans les images infrarouges avec celles obtenues avec d'autres types de modalités, telles que les données du spectre visible, afin d'exploiter leurs complémentarités.

A.4.3 Solution proposée

Notre méthode de reconnaissance du visage s'effectue en deux étapes: l'extraction d'informations caractéristiques et leur classification pour identifier la personne.

Extraction d'informations caractéristiques: La comparaison directe de deux photographies numériques est fortement sensible aux petits changements des conditions d'acquisition. C'est pourquoi il est nécessaire d'extraire des informations utiles qui permettent de différencier les images. En raison de leur nature différente, les caractéristiques extraites dans chacune de modalités d'imagerie (visible ou infrarouge) sont différentes. Par exemple, des détails du visage peuvent ne pas apparaitre dans l'image thermique si la température de surface reste la même.

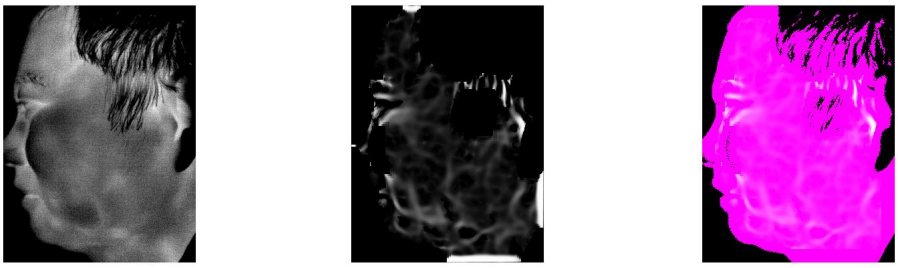


FIGURE A.13: Réseau vasculaire d'une image thermique

C'est la raison pour laquelle, des nouvelles caractéristiques adaptées à l'imagerie thermique ont été étudiées. Dans ce contexte, Ghiass et al [6] ont proposé d'utiliser le réseau vasculaire qui est situé sous la peau pour réaliser une classification. Le réseau vasculaire est un élément caractéristique (c'est à dire, il peut différencier des personnes) qui est normalement plus chaud que d'autres parties du visage. Grâce à ce petite décalage, on peut extraire ce réseau veineux de l'image thermique (figure A.13). Pourtant, cette méthode présente des difficultés pour identifier une personne dans le cas d'une pose différente à celle stockée dans la base de données. L'image thermique est aussi très instable au changement de température extérieur qui peut causer des imperfections dans l'extraction du réseau vasculaire.

Pour améliorer cette méthode et pour s'affranchir de l'impact du changement de pose du visage, **une reconstruction 3D** de la tête par une vidéo/un ensemble d'images consécutives est proposée. En utilisation ce modèle de visage, nous cherchons à **localiser la position 3D du réseau vasculaire** qui permet de mieux caractériser le visage et donc de mieux identifier les personnes.

Classification: Après avoir extrait les informations caractéristiques et construit un modèle, la tâche suivante concerne la comparaison entre le visage capturé et ceux

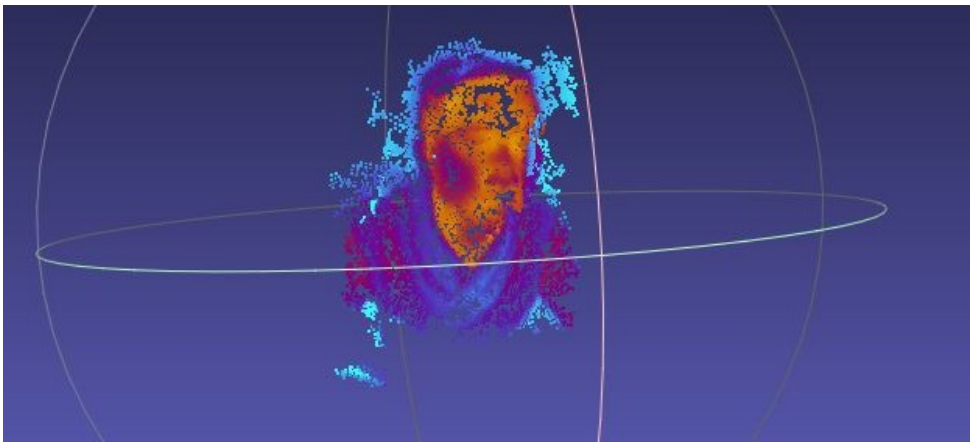


FIGURE A.14: La reconstruction 3D de visage en utilisant des images thermiques.

sauvegardés dans la base de données. Ce processus de comparaison est nommé classification. Dans le cas de la reconnaissance du visage, on utilise surtout **la classification k-NN (k-nearest neighbor:k plus proches voisins)**, c'est à dire, le résultat va être déduit à partir de **k** modèles dans la base de données les plus similaires au visage capturé.

A.4.4 Résultats

Après avoir reconstruit la forme 3D du visage et obtenu le réseau vasculaire pour chaque image, nous avons réussi à localiser ce réseau dans la forme 3D (figure A.15 , A.16).

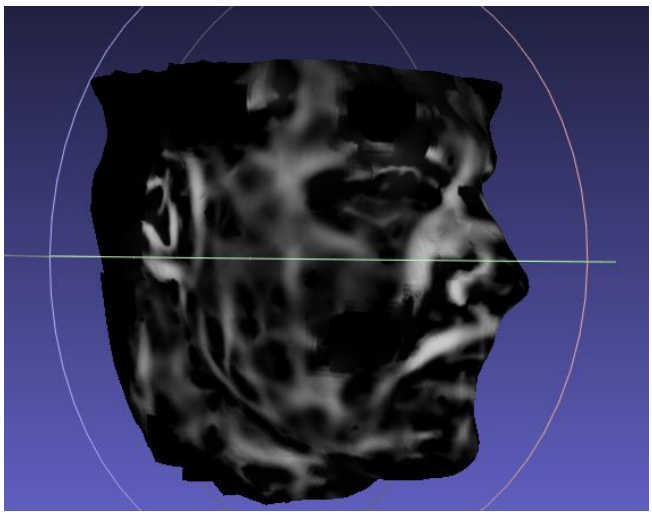


FIGURE A.15: Projection du réseau vasculaire dans le modèle 3D pour personne A.

Ces résultats confirment l'hypothèse que le réseau vasculaire en 3D peut être utilisé comme preuve d'identification des personnes. La qualité de la reconstruction peut encore être amélioré. Nous envisageons deux pistes d'améliorations:

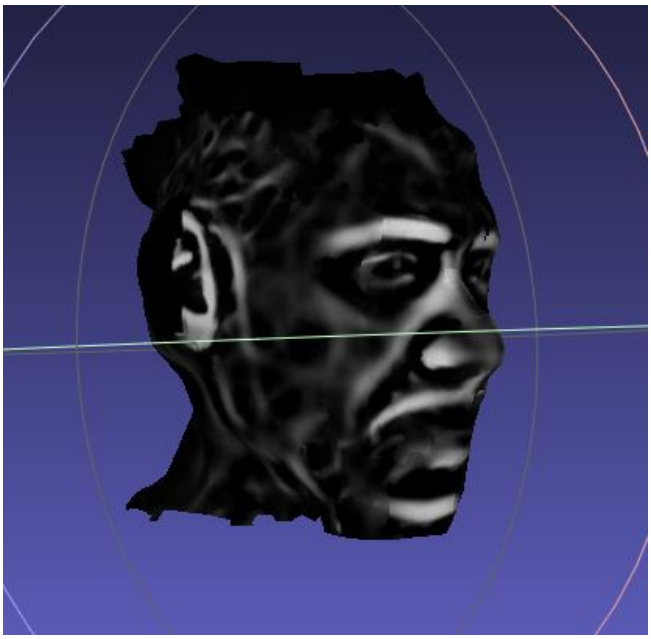


FIGURE A.16: Projection du réseau vasculaire dans le modèle 3D pour personne B.

- *l'optimisation de la méthode de reconstruction 3D en temps et en performances,*
- *l'amélioration de la méthode d'extraction du réseau vasculaire.*

Les capteurs utilisés pour reconstruire ces modèles 3D sont des ThermApps (capteurs à petits prix 1000-2000 euros qui transmettent des données directement vers un téléphone pour visualiser et traiter des images). Ces capteurs sont particulièrement bien adaptés à notre projet puisqu'il est facile d'obtenir des images provenant de différents points de vue grâce à leur petite taille.(figure A.18)

A.5 Conclusion

Le développement des technologies de l'imagerie et de la capacité de calcul nous a menés à une époque où le visage de l'utilisateur peut être considéré comme une preuve d'authentification. La méthode la plus simple et la plus naturelle est d'essayer d'imiter la vision de l'homme en utilisant la vision par ordinateur. Pour cette raison, l'imagerie dans le visible est la première option pour un système d'authentification par reconnaissance faciale.

Cependant, la technologie de l'imagerie dans le visible n'est pas assez robuste pour être utilisée comme seule source d'information pour l'identification. Cette méthode a deux limites majeures qui rendent les systèmes d'authentification vulnérables. La première limite est sa sensibilité aux conditions d'éclairage. L'image dépend non seulement de la couleur de l'objet mais aussi de la nature et de l'intensité de la source lumineuse. Le deuxième problème est qu'il existe des attaques où

TABLE A.1: Résultats moyens après ayant répété 20 fois le processus d'expérimentation.

	Accuracy	Précision	Rappel
Uniquement Gabor-Intensité	79.43%	77.62%	81.76%
Mixage Gabor-décision	82.64%	80.12%	85.21%
Mixage Gabor-caractéristique	88.43%	87.05%	90.83%
Uniquement LBP-Intensité [**?**]	86.11%	85.05%	87.70%
Mixage-LBP	87.36%	88.55%	88.79%

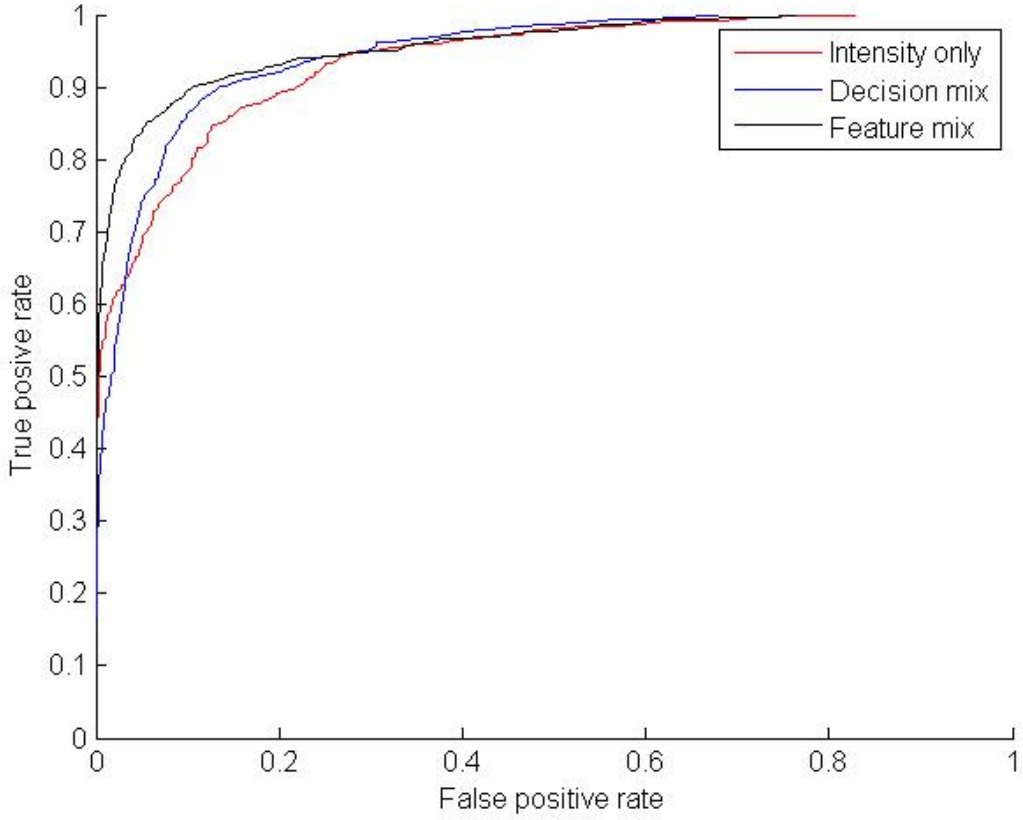


FIGURE A.17: La courbe de Roc du classificateur final.

l'attaquant utilise une signature de visage de l'utilisateur afin de contourner le système d'authentification. La signature peut être une photo, une vidéo trouvée sur un réseau social ou même un masque 3D de l'utilisateur.

Dans la première partie de cette étude, nous avons construis une solution contre les attaques avec un minimum d'équipement requis. Notre cas d'utilisation une méthode de reconnaissance faciale pour les smartphones à caméra dans le visible. Ce

FIGURE A.18: Caméra ThermApp couplée avec un smartphone

cas d'utilisation est compliqué car le système est basé sur une seule caméra non calibrée et la scène d'authentification dépend entièrement des utilisateurs. Nous avons ajusté notre solution pour ce cas en explorant le mouvement de certains capteurs à l'intérieur du smartphone. A partir d'un ensemble d'images vidéo, la méthode utilise un processus de reconstruction 3D pour construire un modèle 3D de la tête qui est très efficace contre la photo-attaque car les différences de caractéristiques géométriques entre un objet réel et une image sont très importantes. L'attaque vidéo peut être détectée en observant la synchronisation entre le mouvement préalable du smartphone et le mouvement calculé par le processus de reconstruction 3D. La limite de notre première étude est qu'elle n'est pas capable de détecter une attaque par masque 3D.

L'attaque délicate du masque n'est pas facilement détectable par une technologie d'imagerie dans le visible car son empreinte est très proche du visage. Cependant, en imagerie thermique où la source d'émission du spectre est le visage humain, la détection de tous les types d'attaques est plus aisée. La technologie d'imagerie thermique peut résoudre l'autre problème majeur de l'imagerie visible concernant les conditions d'éclairage. Dans notre deuxième étude, nous avons amélioré les performances de la méthode de reconnaissance faciale infrarouge en utilisant un modèle 3D de la tête calculé à partir d'une vidéo thermique. La construction du réseau vasculaire à partir de la vidéo thermique est maintenant observée au niveau de l'intensité et des caractéristiques géométriques. Les informations de profondeur et les données sur les vaisseaux sanguins sont associées pour former une seule caractéristique. Ainsi, la méthode de reconnaissance faciale se base principalement sur la position 3D du réseau vasculaire.

A.6 Perspectives

Pour la méthode de détection du spoofing qui est présentée au chapitre 4, il y a un type d'attaque qui n'a pas été pris en compte dans l'étude : l'attaque avec un masque 3D. Jusqu'à cette étape, notre détection d'usurpation d'identité faciale est indépendante du processus de reconnaissance faciale. Nous n'utilisons pas la même base de données que le système d'authentification mais une base de données dédiée à notre solution. En fait, les informations géométriques peuvent être utilisées dans la

phase de reconnaissance faciale, ce qui rend les attaques avec un masque détectables. Cependant, ce type de solution nécessite suffisamment d'images de chaque utilisateur dans différentes positions de la base de données pour pouvoir comparer avec la tentative. Une autre solution est d'étudier la nature du matériau de l'objet pour distinguer le visage des attaques. En fait, chaque type de matériau génère une empreinte spéciale qui peut être utilisée pour détecter si l'objet est fait de peau humaine.

La méthode de reconnaissance faciale thermique permet de construire un modèle 3D du réseau vasculaire. Cependant, nous n'utilisons pas directement ce modèle 3D mais sa projection dans un ensemble d'images de profondeur. Le processus avec des images de profondeur augmente les performances de notre méthode mais ne permet pas d'obtenir une comparaison complète du modèle 3D. Dans nos travaux futurs, nous avons l'intention de construire une méthode permettant d'étudier directement le modèle 3D afin d'utiliser toutes les informations fournies.

Bibliography

[1] R. Singh and H. Om, "An overview of face recognition in an unconstrained environment," in *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, Dec. 2013, pp. 672–677.

[2] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Proceedings*, Jun. 1991, pp. 586–591.

[3] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[4] L. Sirovich and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces," *Journal of the Optical Society of America. A, Optics and image science*, vol. 4, pp. 519–24, Apr. 1987.

[5] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," in *Audio- and Video-based Biometric Person Authentication*, ser. Lecture Notes in Computer Science, J. Bigün, G. Chollet, and G. Borgefors, Eds. Springer Berlin Heidelberg, 1997, pp. 125–142.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] T. J. Stonham, "Practical Face Recognition and Verification with Wisard," in *Aspects of Face Processing*, ser. NATO ASI Series, H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, Eds. Dordrecht: Springer Netherlands, 1986, pp. 426–441. [Online]. Available: https://doi.org/10.1007/978-94-009-4420-6_44

[8] S.-H. Lin, S.-Y. Kung, and L.-J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 114–132, Jan. 1997.

[9] S. Y. Kung and J. S. Taur, "Decision-based neural networks with signal/image classification applications," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 170–181, Jan. 1995.

[10] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, "Identification of human faces," *Proceedings of the IEEE*, vol. 59, no. 5, pp. 748–760, May 1971.

[11] Y. Kaya and K. Kobayashi, "A basic study on human face recognition," Dec. 1972, pp. 265–289.

[12] I. J. Cox, J. Ghosn, and P. N. Yianilos, "Feature-based face recognition using mixture-distance," in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 1996, pp. 209–216.

[13] B. S. Manjunath, R. Chellappa, and C. v. d. Malsburg, "A feature based approach to face recognition," in *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 1992, pp. 373–378.

[14] R. J. Baron, "Mechanisms of human facial recognition," *International Journal of Man-Machine Studies*, vol. 15, no. 2, pp. 137–178, Aug. 1981. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0020737381800016

[15] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042–1052, Oct. 1993.

[16] X. Chen, P. J. Flynn, and K. W. Bowyer, "IR and visible light face recognition," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 332–358, Sep. 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1077314205000226

[17] D. A. Socolinsky, A. Selinger, and J. D. Neuheisel, "Face recognition with visible and thermal infrared imagery," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 72–114, Jul. 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1077314203000754

[18] D. A. Socolinsky and A. Selinger, "A comparative analysis of face recognition performance with visible and thermal infrared imagery," in *Object recognition supported by user interaction for service robots*, vol. 4, Aug. 2002, pp. 217–222 vol.4.

[19] R. Chellappa and W. Zhao, "Robust Face Recognition Using Symmetric Shape-from-Shading," Nov. 1999.

[20] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang, "Total variation models for variable lighting face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1519–1524, Sep. 2006.

[21] B. Klare, A. A. Paulino, and A. K. Jain, "Analysis of facial features in identical twins," in *2011 International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1–8.

[22] R. Singh, M. Vatsa, and A. Noore, "Face recognition with disguise and single gallery images," *Image and Vision Computing*, vol. 27, no. 3, pp. 245–257, Feb.

2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885607001060

[23] N. Ramanathan, R. Chellappa, and A. K. R. Chowdhury, "Facial similarity across age, disguise, illumination and pose," in *2004 International Conference on Image Processing, 2004. ICIP '04.*, vol. 3, Oct. 2004, pp. 1999–2002 Vol. 3.

[24] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, Jun. 2002.

[25] K. Kollreider, H. Fronthaler, and J. Bigun, "Evaluating liveness by face images and the structure tensor," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, Oct. 2005, pp. 75–80.

[26] J. Galbally, S. Marcel, and J. Fierrez, "Biometric Antispoofing Methods: A Survey in Face Recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[27] M. Bagga and B. Singh, "Spoofing detection in face recognition: A review," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 2037–2042.

[28] K. Patel, H. Han, and A. K. Jain, "Secure Face Unlock: Spoof Detection on Smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, Oct. 2016.

[29] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, vol. 5, no. 2, pp. 83–91, 2016.

[30] A. Pinto, W. R. Schwartz, H. Pedrini, and A. d. R. Rocha, "Using Visual Rhythms for Detecting Video-Based Facial Spoof Attacks," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 1025–1038, May 2015.

[31] S. Kumar, S. Singh, and J. Kumar, "A comparative study on face spoofing attacks," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, May 2017, pp. 1104–1108.

[32] I. Chingovska and A. R. d. Anjos, "On the Use of Client Identity Information for Face Antispoofing," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 787–796, Apr. 2015.

[33] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "An Investigation of Local Descriptors for Biometric Spoofing Detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 849–863, Apr. 2015.

[34] J. Yang, Z. Lei, D. Yi, and S. Z. Li, "Person-Specific Face Antispoofing With Subject Domain Adaptation," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 797–809, Apr. 2015.

[35] A. Anjos, M. M. Chakka, and S. Marcel, "Motion-based counter-measures to photo attacks in face recognition," *IET Biometrics*, vol. 3, no. 3, pp. 147–158, Sep. 2014.

[36] D. Wen, H. Han, and A. K. Jain, "Face Spoof Detection With Image Distortion Analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.

[37] N. Poh, C. H. Chan, J. Kittler, S. Marcel, C. M. Cool, E. A. Rua, J. L. A. Castro, M. Villegas, R. Paredes, V. Struc, N. Pavesic, A. A. Salah, H. Fang, and N. Costen, "An Evaluation of Video-to-Video Face Verification," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 781–801, Dec. 2010.

[38] N. Evans, S. Z. Li, S. Marcel, and A. Ross, "Guest Editorial: Special Issue on Biometric Spoofing and Countermeasures," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 699–702, Apr. 2015.

[39] L. Sun, Z. Wu, S. Lao, and G. Pan, "Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera," in *2007 11th IEEE International Conference on Computer Vision(ICCV)*, 2007, pp. 1–8. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICCV.2007.4409068

[40] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics Systems Under Spoofing Attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, Sep. 2015.

[41] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[42] J. Maatta, A. Hadid, and M. Pietikainen, "Face spoofing detection from single images using micro-texture analysis," in *2011 International Joint Conference on Biometrics (IJCB)*, Oct. 2011, pp. 1–7.

[43] G. Kim, S. Eum, J. Suhr, D. Kim, K. Park, and J. Kim, "Face liveness detection based on texture and frequency analyses," in *2012 5th IAPR International Conference on Biometrics (ICB)*, Mar. 2012, pp. 67–72.

[44] H. Bashier, S. Lau, P. Han, L. Ping, and C. Li, "Face Spoofing Detection Using Local Graph Structure," in *Proceedings of the 2014 International Conference on Computer, Communications and Information Technology*, 2014.

[45] H. P. Nguyen, F. Retrainty, F. Morain-Nicolier, and A. Delahaies, "Face spoofing attack detection based on the behavior of noises," in *2016 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2016, December 7–9, Greater Washington, D.C., USA, 2016*, 2016, pp. 1–5. [Online]. Available: http://dx.doi.org/10.1109/GlobalSIP.2015.7416924

[46] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Face and Gesture 2011*, Mar. 2011, pp. 436–441.

[47] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face Spoofing Detection Using Colour Texture Analysis," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1818–1830, Aug. 2016.

[48] I. Chingovska, A. Anjos, and S. Marcel, "On the Effectiveness of Local Binary Patterns in Face Anti-spoofing," 2012. [Online]. Available: https://infoscience.epfl.ch/record/192369

[49] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2013, pp. 1–8.

[50] N. Erdogmus and S. Marcel, "Spoofing Face Recognition With 3d Masks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1084–1097, Jul. 2014.

[51] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of Face Spoofing Using Visual Dynamics," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 762–777, Apr. 2015.

[52] S. R. Arashloo, J. Kittler, and W. Christmas, "Face Spoofing Detection Based on Multiple Descriptor Fusion Using Multiscale Dynamic Binarized Statistical Image Features," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2396–2407, Nov. 2015.

[53] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face Spoofing Detection Through Visual Codebooks of Spectral Temporal Cubes," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4726–4740, Dec. 2015.

[54] A. Agarwal, R. Singh, and M. Vatsa, "Face anti-spoofing using Haralick features," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sep. 2016, pp. 1–6.

[55] K. Kollreider, H. Fronthaler, and J. Bigun, "Non-intrusive liveness detection by face images," *Image and Vision Computing*, vol. 27, no. 3, pp. 233–244, Feb. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885607000893

[56] J. Hyung-Keun, J. Sung-Uk, and Y. Jang-Hee, "Liveness Detection for Embedded Face Recognition System," *International Journal of Biological and Medical Sciences*, 2006.

[57] L. Sun, G. Pan, Z. Wu, and S. Lao, "Blinking-based Live Face Detection Using Conditional Random Fields," in *Proceedings of the 2007 International Conference*

on Advances in Biometrics, ser. ICB'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 252–260. [Online]. Available: http://dl.acm.org/citation.cfm?id=2391659.2391688

[58] M. Gavrilescu, "Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks," *IET Biometrics*, vol. 5, no. 3, pp. 236–242, Sep. 2016. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2015.0078

[59] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee, "Face liveness detection using variable focusing," in *2013 International Conference on Biometrics (ICB)*, Jun. 2013, pp. 1–6.

[60] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-Time Face Detection and Motion Analysis With Application in "Liveness" Assessment," *Trans. Info. For. Sec.*, vol. 2, no. 3, pp. 548–558, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1109/TIFS.2007.902037

[61] X. Maldague, *Theory and practice of infrared technology for nondestructive testing*. Wiley, Apr. 2001, google-Books-ID: ts9RAAAAMAAJ.

[62] G. Friedrich and Y. Yeshurun, "Seeing People in the Dark: Face Recognition in Infrared Images," in *Biologically Motivated Computer Vision*, ser. Lecture Notes in Computer Science, H. H. Bülthoff, C. Wallraven, S.-W. Lee, and T. A. Poggio, Eds. Springer Berlin Heidelberg, 2002, pp. 348–359.

[63] H. Chang, Y. Yao, A. Koschan, B. Abidi, and M. Abidi, "Improving Face Recognition via Narrowband Spectral Range Selection Using Jeffrey Divergence," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 1, pp. 111–122, Mar. 2009.

[64] F. Nicolo and N. A. Schmid, "A Method for Robust Multispectral Face Recognition," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Kamel and A. Campilho, Eds. Springer Berlin Heidelberg, 2011, pp. 180–190.

[65] A. Seal, D. Bhattacharjee, M. Nasipuri, and D. K. Basu, "Minutiae based thermal face recognition using blood perfusion data," in *2011 International Conference on Image Information Processing*, Nov. 2011, pp. 1–4.

[66] E. A. Jaeger and W. Tasman, *Duane's Ophthalmology 2013*. Lippincott Williams & Wilkins, Oct. 2012, google-Books-ID: CkzeMQEACAAJ.

[67] O. Arandjelovic, R. Hammoud, and R. Cipolla, "Thermal and Reflectance Based Personal Identification Methodology under Variable Illumination," *Pattern Recognition*, vol. 43, pp. 1801–1813, May 2010.

[68] R. S. Ghiass, O. Arandjelovic, H. Bendada, and X. Maldague, "Illumination-invariant face recognition from a single image across extreme pose using a dual dimension AAM ensemble in the thermal infrared spectrum," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–10.

[69] F. J. Prokoski, R. B. Riedel, and J. S. Coffin, "Identification of individuals by means of facial thermography," in *Proceedings 1992 International Carnahan Conference on Security Technology: Crime Countermeasures*, Oct. 1992, pp. 120–125.

[70] L. B. Wolff, D. Socolinsky, and C. K. Eveland, "Quantitative measurement of illumination invariance for face recognition using thermal infrared imagery," Jan. 2003, pp. 140–151.

[71] F. Prokoski and R. B. Riedel, "Infrared Identification of Faces and Body Parts," Apr. 2006, pp. 191–212.

[72] L. Wolf, A. Shashua, C. H. A. Il, and C. H. A. Il, "Learning over Sets using Kernel Principal Angles," p. 19.

[73] R. S. Ghiass, O. Arandjelović, H. Bendada, and X. Maldague, "Infrared face recognition: A literature review," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Aug. 2013, pp. 1–10.

[74] R. Ghiass, O. Arandjelovic, H. Bendada, and X. Maldague, "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognition*, Jan. 2014.

[75] D. A. Socolinsky, L. B. Wolff, J. D. Neuheisel, and C. K. Eveland, "Illumination invariant face recognition using thermal infrared imagery," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–527–I–534 vol.1.

[76] G. Hermosilla, J. Ruiz-del Solar, and R. Verschae, "An enhanced representation of thermal faces for improving local appearance-based face recognition," *Intelligent Automation & Soft Computing*, vol. 23, no. 1, pp. 1–12, Jan. 2017.

[77] S. M. Desa and S. Hati, "IR and visible face recognition using fusion of kernel based features," in *2008 19th International Conference on Pattern Recognition*, Dec. 2008, pp. 1–4.

[78] R. Cutler, "Face Recognition Using Infrared Images and Eigenfaces," Mar. 1999.

[79] M. A. Akhloufi and A. Bendada, "Infrared face recognition using texture descriptors," in *Thermosense XXXII*, vol. 7661. International Society for Optics and Photonics, May 2010, p. 766109. [Online]. Available: https://

www.spiedigitallibrary.org/conference-proceedings-of-spie/7661/766109/Infrared-face-recognition-using-texture-descriptors/10.1117/12.849764.short

[80] Z. Wu, M. Peng, and T. Chen, "Thermal face recognition using convolutional neural network," in *2016 International Conference on Optoelectronics and Image Processing (ICOIP)*, Jun. 2016, pp. 6–9.

[81] M. O. Simón, C. Corneanu, K. Nasrollahi, O. Nikisins, S. Escalera, Y. Sun, H. Li, Z. Sun, T. B. Moeslund, and M. Greitans, "Improved RGB-D-T based face recognition," *IET Biometrics*, vol. 5, no. 4, pp. 297–303, Apr. 2016. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2015.0057

[82] M. Peng, C. Wang, T. Chen, and G. Liu, "NIRFaceNet: A Convolutional Neural Network for Near-Infrared Face Identification," *Information*, vol. 7, no. 4, p. 61, Dec. 2016. [Online]. Available: https://www.mdpi.com/2078-2489/7/4/61

[83] C. Orji, E. Hurwitz, and A. Hasan, "THERMAL IMAGING USING CNN AND KNN CLASSIFIERS WITH FWT, PCA AND LDA ALGORITHMS," *Computer Science*, p. 11.

[84] A. Kwaśniewska, J. Rumiński, and P. Rad, "Deep features class activation map for thermal face detection and tracking," in *2017 10th International Conference on Human System Interactions (HSI)*, Jul. 2017, pp. 41–47.

[85] M. S. Sarfraz and R. Stiefelhagen, "Deep Perceptual Mapping for Thermal to Visible Face Recognition," *arXiv:1507.02879 [cs]*, Jul. 2015, arXiv: 1507.02879. [Online]. Available: http://arxiv.org/abs/1507.02879

[86] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination Invariant Face Recognition Using Near-Infrared Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627–639, Apr. 2007.

[87] H. Méndez, C. S. Martín, J. Kittler, Y. Plasencia, and E. García-Reyes, "Face Recognition with LWIR Imagery Using Local Binary Patterns," in *Advances in Biometrics*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Jun. 2009, pp. 327–336.

[88] Z. Xie and Z. Wang, "Joint Encoding of Multi-scale LBP for Infrared Face Recognition," in *Genetic and Evolutionary Computing*, ser. Advances in Intelligent Systems and Computing. Springer, Cham, 2015, pp. 269–276.

[89] N. Wang, Q. Li, A. A. A. El-Latif, J. Peng, and X. Niu, "An enhanced thermal face recognition method based on multiscale complex fusion for Gabor coefficients," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2339–2358, Oct. 2014.

[90] G. Majumder and M. K. Bhowmik, "Gabor-Fast ICA Feature Extraction for Thermal Face Recognition Using Linear Kernel Support Vector Machine," in *2015 International Conference on Computational Intelligence and Networks*, Jan. 2015, pp. 21–25.

[91] A. Seal, S. Ganguly, D. Bhattacharjee, M. Nasipuri, and D. K. Basu, "Automated Thermal Face recognition based on Minutiae Extraction," *arXiv:1309.1000 [cs]*, Sep. 2013, arXiv: 1309.1000. [Online]. Available: http://arxiv.org/abs/1309.1000

[92] Z. Xie and G. Liu, "Blood perfusion construction for infrared face recognition based on bio-heat transfer," *Bio-Medical Materials and Engineering*, vol. 24, no. 6, pp. 2733–2742, 2014.

[93] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos, "Physiology-Based Face Recognition in the Thermal Infrared Spectrum," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 613–626, Apr. 2007.

[94] T. Bourlai and B. Cukic, "Multi-spectral face recognition: Identification of people in difficult environments," in *2012 IEEE International Conference on Intelligence and Security Informatics*. Washington, DC, USA: IEEE, Jun. 2012, pp. 196–201. [Online]. Available: http://ieeexplore.ieee.org/document/6284307/

[95] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition," in *2018 International Conference on Biometrics (ICB)*, Feb. 2018, pp. 174–181.

[96] S. Saxena and J. Verbeek, "Heterogeneous Face Recognition with CNNs," in *Computer Vision – ECCV 2016 Workshops*, ser. Lecture Notes in Computer Science, G. Hua and H. Jégou, Eds. Springer International Publishing, 2016, pp. 483–491.

[97] J. Kim, S. Yu, I.-J. Kim, and S. Lee, "3d Multi-Spectrum Sensor System with Face Recognition," *Sensors (Basel, Switzerland)*, vol. 13, no. 10, pp. 12 804–12 829, Sep. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3859038/

[98] G. J. Iddan and G. Yahav, "3d IMAGING IN THE STUDIO (AND ELSEWHERE...)," vol. 4298, p. 8.

[99] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2008, pp. 1–7.

[100] F. Blais, M. Picard, and G. Godin, "Accurate 3d acquisition of freely moving objects," in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, Sep. 2004, pp. 422–429.

[101] S. Goel and B. Lohani, "A Motion Correction Technique for Laser Scanning of Moving Objects," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 225–228, Jan. 2014.

[102] K. H. Strobl, E. Mair, T. Bodenmüller, S. Kielhöfer, W. Sepp, M. Suppa, D. Burschka, and G. Hirzinger, "The self-referenced DLR 3d-modeler," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2009, pp. 21–28.

[103] K. H. Strobl, E. Mair, and G. Hirzinger, "Image-based pose estimation for 3-D modeling in rapid, hand-held motion," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 2593–2600.

[104] B. K. P. Horn, "Shape from Shading," B. K. P. Horn and M. J. Brooks, Eds. Cambridge, MA, USA: MIT Press, 1989, pp. 123–171. [Online]. Available: http://dl.acm.org/citation.cfm?id=93871.93877

[105] A. R. Bruss, "Shape from Shading," B. K. P. Horn and M. J. Brooks, Eds. Cambridge, MA, USA: MIT Press, 1989, pp. 69–87. [Online]. Available: http://dl.acm.org/citation.cfm?id=93871.93875

[106] M. J. Brooks, "Two Results Concerning Ambiguity in Shape From Shading," in *AAAI*, 1983.

[107] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas-relief ambiguity," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 1997, pp. 1060–1066.

[108] C. Wu, "Towards Linear-Time Incremental Structure from Motion," in *Proceedings of the 2013 International Conference on 3D Vision*, ser. 3DV '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 127–134. [Online]. Available: http://dx.doi.org/10.1109/3DV.2013.25

[109] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[110] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," p. 10.

Kim Trong NGUYEN

Doctorat : Optimisation et Sûreté des Systèmes

Année 2019

Reconnaissance faciale et détection de l'usurpation d'identité par l'utilisation d'un modèle 3D

L'amélioration des technologies d'imagerie nous conduit à une ère où le visage de l'utilisateur peut être reconnu comme une preuve d'authentification. L'imagerie dans le visible est naturellement la première option pour tout système de reconnaissance faciale. Cependant, cette technique présente deux inconvénients majeurs qui rendent le système d'identification vulnérable : sa dépendance à l'égard de la source lumineuse et la difficulté à détecter la projection d'un visage. La solution de reconnaissance faciale pour smartphones est le cas d'utilisation choisi. À partir d'un ensemble d'images vidéo, la méthode construit un modèle 3D de la tête en utilisant un schéma de reconstruction dédié. Cette méthode est très efficace contre les attaques par photographie, car les différences entre un visage et une image sont importantes. L'attaque par vidéo peut être détectée en déterminant une désynchronisation entre le mouvement du smartphone et le mouvement capturé par le processus de reconstruction 3D. En imagerie thermique où la source d'émission du spectre est le visage humain, la détection de tous les types d'attaques par projection du visage est facile et les conditions d'éclairage n'affectent pas les images thermiques. Dans notre deuxième étude, nous visons à améliorer la performance de la méthode de reconnaissance faciale thermique à l'aide d'un modèle 3D du réseau vasculaire calculé à partir d'une vidéo infrarouge. De nombreuses expérimentations sur des données réelles ont souligné la pertinence de l'approche proposée.

Mots clés : identification automatique – imagerie tridimensionnelle – perception des visages – imagerie infrarouge.

Face Recognition and Face Spoofing Detection Using 3D Model

The improvement of imaging technology leads us to an era in which user's faces can be acknowledged as a biometric proof of authentication toward an automatic system. Visible imagery is naturally the first option for every facial recognition system. However, visible imagery has two major drawbacks that make the identification systems vulnerable: its dependency on the light source and its incompetence toward face-spoofing attacks. The first part of this study aims to construct a solution against the face-spoofing attack with minimum equipment required. The face recognition solution for smartphones is our hardest use-case because of the uncalibrated camera and unpredictable behaviors of users. From a set of video's frames, the method builds a 3D model of the head using a dedicated reconstruction scheme. This model is highly effective against photo-attack as differences between a real object and an image is truly large. The video attack can be detected by examining the synchronization between the prior motion of the smartphone (explored by motion sensors) and the captured-motion calculated by the 3D reconstruction process. In thermal imagery where the emission source of the spectrum is human's face, the detection of all types of face-spoofing attack is trivial, and the illumination conditions do not affect thermal images. Though, in general, thermal images present less information than visible images. In our second study, we aim to improve the performance of thermal face- recognition method using a 3D model of the vascular network computed from an infrared video.

Keywords: face-recognition – three-dimensional imaging – face-spoofing attack – infrared imaging – vascular network.

Thèse réalisée en partenariat entre :

EPF ÉCOLE D'INGÉNIEUR-E-S
Sceaux I Troyes I Montpellier

Grand Troyes Communauté d'Agglomération

utt UNIVERSITÉ DE TECHNOLOGIE TROYES

Ecole Doctorale "Sciences pour l'Ingénieur"