

UNIVERSITÉ NATIONALE DU VIETNAM, HANOÏ
INSTITUT FRANCOPHONE INTERNATIONAL

CIBAMBO Masugentwali Steven

**EMOTION RECOGNITION AND ANTI-SPOOFING IN FACE
RECOGNITION FOR SMART EDUCATION SYSTEM**

**NHẬN DẠNG CẢM XÚC VÀ CHỐNG GIÀ MẠO TRONG NHẬN
DẠNG KHUÔN MẶT CHO CÁC HỆ THỐNG ĐÀO TẠO THÔNG
MINH**

**MÉMOIRE DE FIN D'ÉTUDES DU MASTER
INFORMATIQUE**

Code : Programme pilote

Spécialité : Systèmes Intelligents et Multimédia

HANOÏ - 2021

**UNIVERSITÉ NATIONALE DU VIETNAM, HANOÏ
INSTITUT FRANCOPHONE INTERNATIONAL**

CIBAMBO Masugentwali Steven

**EMOTION RECOGNITION AND ANTI-SPOOFING IN FACE
RECOGNITION FOR SMART EDUCATION SYSTEM**

**NHẬN DẠNG CẢM XÚC VÀ CHỐNG GIÀ MẠO TRONG NHẬN
DẠNG KHUÔN MẶT CHO CÁC HỆ THỐNG ĐÀO TẠO THÔNG
MINH**

**MÉMOIRE DE FIN D'ÉTUDES DU MASTER
INFORMATIQUE**

Spécialité : Systèmes Intelligents et Multimédia
Code : Programme pilote

Sous la direction du Prof. Assoc. :

M. Nguyen Chan Hùng (CEO de VDSmart - Vietnam) et
M. HO Tuong Vinh, Ph.D. (Responsable des Masters UNV/IFI)



M. Nguyen Chan Hùng

M. HO Tuong Vinh

Attestation sur l'honneur

J'atteste sur l'honneur que ce mémoire a été réalisé par moi-même et que les données et les résultats qui y sont présentés sont exacts et n'ont jamais été publiés ailleurs. La source des informations citées dans ce mémoire a bien été précisée.

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất công trình nào khác. Các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

Signature de l'étudiant



CIBAMBO M. Steven

Remerciements

Ce travail de mémoire de Master de recherche est le résultat de l'engagement de plusieurs personnes qui ont décidé de m'accompagner résolument dans cet exaltant parcours.

Je souhaiterais tout d'abord remercier l'équipe de VDSmart pour l'accueil qu'elle m'a réservé, le temps que chacun des ses membres m'a accordé et plus globalement; pour toutes les informations, références bibliographiques, réflexions, corrections, ... que chacun m'a apporté et qui ont nourrit ce travail. Je remercie également cette entreprise de m'avoir fait découvrir et approcher du monde de l'autogestion en m'ouvrant les portes des réseau auxquels elle collabore. Je remercie tout particulièrement le Prof. Ass. Nguyen Chan Hung (respectivement PDG de VDSmart et directeur de ce mémoire) pour son rigueur de travail et le respect de deadline qu'il n'a pas cessé de m'inculquer

Je remercie également le corps d'enseignant de l'Institut Francophone International (IFI) pour la qualité et la méthodologie de leur enseignement au cours de ces deux années passées à l'Université Nationale du Vietnam. Je remercie tout particulièrement M. Nguyen Hong Quang (responsable du Master 1) et M. HO Tuong Vinh (responsable du Master 2) qui m'ont laissé une large part d'autonomie dans les travaux de recherche scientifique tout en m'aiguillant sur des pistes de réflexions riches et porteuses.

Enfin je saisie de cette occasion pour remercier l'entreprise Videia Ed Tech qui m'a apporté son aide pour l'accoplissemement de cette oeuvre au moment où j'en avais plus besoin. Je remercie plus particulièrement M. Huang Chien En, M. Yoo In Seak, M. Nguyen Văn Thành, M. Nguyen Quang Hip, Mlle. Nguyen Th Thuy Tiên and Mme. Nguyen Thi Vân Khánh.

CIBAMBO M. Steven

Résumé

De nos jours, le système de reconnaissance faciale est utilisé dans plusieurs applications principalement pour l'authentification individuelle. Alors que la reconnaissance faciale reste vulnérable de plusieurs types d'attaques ; la détection d'attaques de visage (liveness detection) s'avère une étape cruciale avant de fournir les données faciales au système pour l'identification et/ou l'authentification d'un individu.

Dans ce travail de mémoire, nous nous engageons d'apporter une solution au problème de l'usurpation de visage par l'emploi de Réseau de Neurones à Convolution en utilisant le capteur d'image à double pixel. Le but principal est de parvenir à distinguer un vrai visage du faux dans la mesure la plus possible. Ainsi étant convaincu de l'utilité de l'information contenu dans le depth map d'une image [21] nous avons opté d'en faire usage. La solution proposée pour distinguer un visage réel du faux et basée sur la reconstruction du depth map à partir d'une paire d'images issue de la caméra double pixel et la classification du depth map. Cette solution est enfin destinée à être intégrée dans le système Smart Education et/ou Smart access de l'entreprise VDSmart.

Abstract

Nowadays, the facial recognition system is used in several applications mainly for individual authentication. Whereas facial recognition remain vulnerable to several types of attacks ; Face Anti-Spoofing detectionn is a crucial step before providing facial data to the face recognition system.

In this work, we are committed to providing a solution to the problem of face anti-spoofing attaque through the using of the Convolutional Neural Network with the dual pixel image sensor. The main goal is to be able to distinguish a reel/genuine face to a fake face as much as possible. Thus being convinced of the usefulness of the information contained in the depth map of an image [21] we opted to use it. Thus the proposed solution to distinguish a real face from a fake one is based on the reconstruction of the depth from a pair of images from the dual pixel camera and the classification of the depth map. This solution is finally intended to be integrated into a Smart Education and / or Smartaccess system from the VDSmart company.

Table des matières

Table des figures	i
1 Introduction générale	1
2 Structure d'accueil	2
2.1 Présentation de VDSmart	2
2.2 Ressources humaines	2
2.3 Organigramme	3
2.4 Missions	3
2.5 Recherche et Développement	4
2.6 Projets	4
2.6.1 VDSmart Box	4
2.6.2 VDSmart Access	5
2.6.3 Eye Pro Thermal	6
2.6.4 VDSmart Class	7
2.7 Conclusion	8
3 Analyse du sujet	9
3.1 Contexte de la recherche	9
3.2 Cadre théorique	9
3.2.1 Motivation	9
3.2.2 Définition de quelques termes clés	10
3.2.3 Différents types de présentation d'attaques	13
3.2.4 Les 7 principales émotions	14
3.3 Problématique	18
3.4 Objectifs	18
3.5 Résultats attendus	19
3.6 Conclusion	20
4 État de l'art	21
4.1 Introduction	21

TABLE DES MATIÈRES

4.2	Étude de l'existant	21
4.2.1	Eye Pro Education	21
4.3	Travaux connexes	22
4.3.1	Face anti-spoofing	23
4.3.2	La reconstruction du depth map	27
4.3.3	Reconnaissance automatique d'émotions	29
4.4	Analyse des solutions existantes	32
4.5	Conclusion	33
5	Méthode proposée	34
5.1	Introduction	34
5.2	Caméra à Double Pixel	34
5.3	Génération de Depth Map	35
5.4	Entrainement de depth map par paire	36
5.4.1	La cohérence de la transformation	36
5.4.2	L'étiquetage relative du depth	38
5.4.3	Fonction de perte	38
5.5	Classification de Depth	38
5.6	Conclusion	39
6	Implémentation et analyse des résultats	40
6.1	Introduction	40
6.2	Base de données (dataset)	40
6.3	Architecture réseau	41
6.3.1	Encodage	41
6.3.2	Décodage	42
6.4	La reconstruction du depth map	44
6.5	Classification du depth map	44
6.5.1	Démarche	44
6.5.2	Inférence	46
6.6	Analyse du depth map généré	52
6.7	Performance de la méthode de classification	53
6.7.1	Précision de la classification	53
6.7.2	Entraînement et Validation	54
6.7.3	AUC et la Courbe ROC	54
6.7.4	Matrice de confusion	55
6.7.5	Rapport de classification	56
6.8	Conclusion	57
7	Conclusion générale	58

Table des figures

2.1	Organigramme de la société VDSmart	3
2.2	VDSmart Box	5
2.3	VDSmart Access	6
2.4	Thermal Eye Pro	7
2.5	Salle de classe intelligente standard	8
3.1	4 étapes de la reconnaissance faciale	10
3.2	Liveness detection	11
3.3	Emotion recognition process	13
3.4	Expression de la colère [6]	14
3.5	Expression de la peur [6]	15
3.6	Expression du dégoût [6]	15
3.7	Expression de la joie [6]	16
3.8	Expression de tristesse [6]	16
3.9	Expression de surprise [6]	17
3.10	Expression de mépris [6]	17
4.1	Eye Pro Education [35]	22
4.2	Processus de génération de données synthétiques [16]	23
4.3	Processus de détection de mouvement des yeux [16]	24
4.4	Maillage et déformation d'un objet 3D [12]	25
4.5	Projection du perspective [12]	26
4.6	post-traitement [12]	26
4.7	Multi-modalité PAD [11]	27
4.8	Architecture dual camera based features [22]	28
4.9	Illustration de combinaison d'images [22]	28
4.10	Fusion de CNN basé sur le patch et depth [24]	29
4.11	Extraction de patches et estimation de depth [24]	29
4.12	Détection and filtrage de bord [10]	30
4.13	La procédure de la solution proposée [10]	31
4.14	Etape de pre-traitement [26]	32

TABLE DES FIGURES

5.1 Caméra double pixel	35
5.2 Reconstruction du depth map [23]	36
5.3 Modèle vision stéréo [37]	36
5.4 Classification de depth map	39
6.1 Échantillon de la base de données	41
6.2 Architecture du réseau proposé	43
6.3 Classification du depth map	45
6.4 Architecture Xception pour la classification [7]	45
6.5 Présentation d'attaque - portrait	46
6.6 Présentation d'attaque - paysage	47
6.7 Présentation d'attaque - affichage sur écran	48
6.8 Présentation simultanée	49
6.9 Présentation d'attaque - vrai visage	49
6.10 Présentation d'attaque - photo	50
6.11 Présentation d'attaque - vrai	50
6.12 Reconnaissance d'émotion	51
6.13 Fréquence de dominance d'émotions	51
6.14 La disparité de depth map par différentes méthodes	53
6.15 Courbes de la précision et de la perte	54
6.16 Receiver Operating Characteric	55
6.17 Matrice de confusion	56
6.18 Rapport de classification	56

Liste des sigles et acronymes

AUC	<i>Area Under the Curve</i>
CNN	<i>Convolutional Neural Network</i>
DL	<i>Deep Learning</i>
DTN	<i>Deep Tree Network</i>
EER	<i>Equal Error Rate</i>
FAS	<i>Face Anti-spoofing</i>
FPAD	<i>Face Presentation Attacks Detection</i>
PFH	<i>Fast Point Features Histograms</i>
FPR	<i>False Positif Rate</i>
FR	<i>Face Recognition</i>
FRR	<i>False Reject Ratio</i>
HOG	<i>Histogram of Oriented Gradients</i>
IA	<i>Intelligence Artificielle</i>
ML	<i>Machine Learning</i>
MRF	<i>Markov Random Field</i>
PA	<i>Presentation Attack</i>
PAD	<i>Presentation Attack Detection</i>
PFH	<i>Point Features Histograms</i>
RNN	<i>Recurrent Neural Network</i>
ROC	<i>Receiver Operation Characteristic</i>
SI	<i>System d'Information</i>
SVM	<i>Support Vector Machine</i>
TPR	<i>True Positif Rate</i>
ULBP	<i>Uniform Local Binary Pattern</i>

Introduction générale

Les deux dernières décennies le monde a bénéficié d'une grande disponibilité de données dans tous les secteurs en général, cependant la sécurité s'avère nécessaire pour s'assurer des utilisateurs qui y accèdent. Curieusement la sécurité reste un problème pour la plus part des domaines comme; le e-commerce, e-learning, transport public, la finance, etc.

De nos jours, la Biométrie offre des moyens intéressants pour sécuriser l'accès à un système d'information. Elle utilise les informations physiologique à savoir; l'empreinte digitale, le visage, l'iris et/ou la rétine, la paume de main, etc. pour l'identification individuelle ou l'authentification d'un individu.

La reconnaissance faciale est celui qui se développé rapidement ces dernières années et semble être un bon choix car il ne nécessite pas de contact physique, il est naturel, bien accepté et juste avec un capteur très peu coûteux (webcam) qui est pratiquement disponible sur tous les appareils électroniques d'aujourd'hui et le tout est joué.

Le système basé sur la reconnaissance faciale a été adopter par la plus part des organisations soit pour restreindre l'accès à un endroit spécifique, soit pour la surveillance d'une zone, soit pour se connecter à un système, soit pour la prise de présence, soit pour déverrouiller une porte ou un téléphone, etc. qu'à cela ne tienne, tout ces systèmes restent vulnérable tant qu'ils ne peuvent pas distinguer un vrai visage du faux. Il est ainsi pertinent de s'intéresser aux technique de la détection d'usurpation de visage (Liveness detection en anglais) cherchant accès illégale à un système en faisant une étude comparative et distinctive d'un visage réel et non réel présent devant la caméra. Pour y arriver, nous nous sommes proposés de subdiviser ce travail en cinq sections principales à savoir; d'abord nous commençons par une brève présentation de l'organisme d'accueil (VDSmart) où nous avons passé nos six de recherche (Chapitre 2). Ensuite nous allons également tenter d'analyser la thématique en parlant du contexte dans lequel ce travail a été fait et surtout de la problématique que nous tentons de résoudre (Chapitre 3), la troisième section parle de l'État de l'art où nous faisons une étude de l'existant et des travaux connexes (Chapitre 4). Après avoir présenter la solution proposée (Chapitre 5), le dernier point contient l'implémentation et l'analyse de résultats (Chapitre 6).

Chapitre **2**

Structure d'accueil

2.1 Présentation de VDSmart

VDSmart est une société technologique par action ¹ spécialisée dans la conception d'applications d'Intelligence Artificielle sur l'IoT ² dans des domaines tels que l'éducation (Smart Education), la construction (Smart Building), la finance (banking) et commerce (Smart Retail), etc. La société fait partie de l'écosystème VDS - VIETNAM Digital Spaces, réunissant des experts Vietnamiens du premier plan en IA. La société VD-Smart est également membre fondateur du Vietnam Smart Education group VISEDU qui comprend à son sein 10 sociétés membres et fournit des solutions complètes pour une éducation intelligente.

2.2 Ressources humaines

Pour bien mener les projets de recherche et de développement les ressources humaines comprend :

- Les experts du premier plan en Intelligent Artificielle de l'académie d'IA,
- Les experts en informatique, en automatisation et en robotique avec de nombreuses années d'expérience, auteurs des produits pratiques, dont l'un connus sous le nom de VIEBOT Robot, premiers robots humanoïdes à au Vietnam à partir de 2017.
- Ingénieurs expérimentés dans les domaines de l'intégration de systèmes
- Les programmeurs des langages modernes, tels que Java, Python, C / C ++, AngularJS, capables de développer des applications multiplateformes,
- L'équipe de collaborateurs est composée d'experts en éducation et de conférenciers de premier plan issus d'institutions de formation prestigieuses telles

1. <https://vdsmart.vn>

2. IoT : Internet of Things

que l'Université de l'Education, l'Université de Technologie de Hanoi, l'Académie Vietnamienne d'Agriculture, l'Académie des Postes et des Télécommunications, avec une expérience pédagogique énorme.

2.3 Organigramme

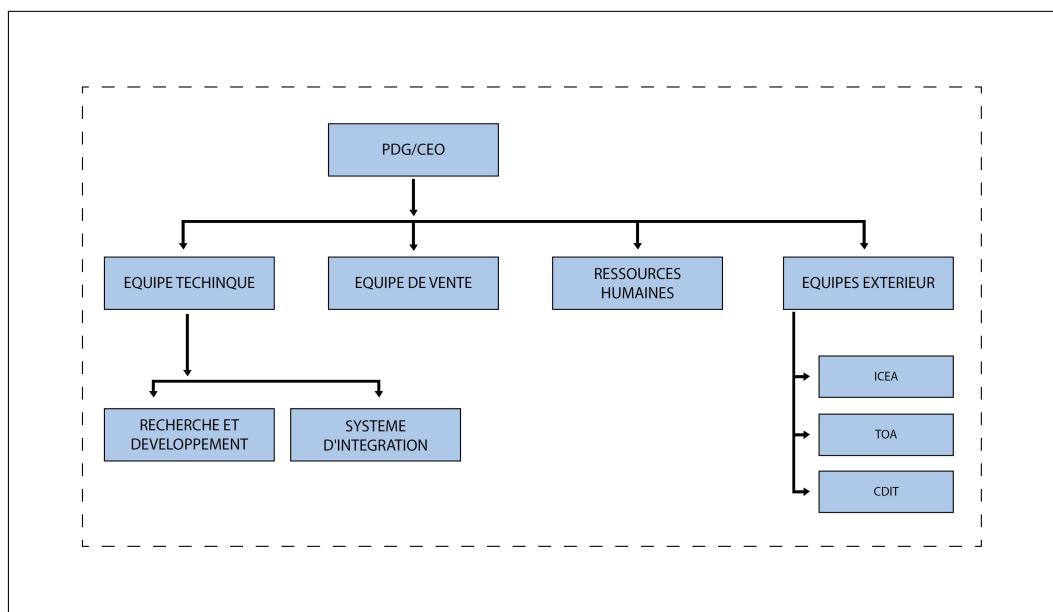


FIGURE 2.1 – Organigramme de la société VDSmart

La figure 2.1 présente l'organigramme de la l'entreprise VDSmart. Et nous entant que stagiaire chercheur nous étions basé dans le département équipe technique plus précisément dans recherche et développement.

2.4 Missions

A son lancement l'entreprise VDSmart s'était fixé deux missions principales à accomplir :

- S'engager à encourager les petites et grandes entreprises à appliquer l'Intelligence Artificielle aux activités commerciales et de production,
- Contribuer à la formation des ressources humaines de haute technologie à la révolution industrielle 4.0 du pays.

2.5 Recherche et Développement

Domaine de l'intelligence artificielle

La recherche sur les applications de la Vision par Ordinateur telles que :

- La reconnaissance faciale (Face recognition)
- La reconnaissance d'émotion (Emotion recognition)
- La reconnaissance d'activité (Activity recognition)
- La reconnaissance d'objets (Objects recognition)

Réseau et systèmes

- Créer des solutions cloud avec les technologies OpenStack, Kubernetes, Docker, Apache Hadoop,
- Créer des solutions pour intégrer les données (Data Lake),
- Systèmes de communication Multimédia comme solution de télévision interne, affichage numérique et télé-conférence,
- Construire des solutions logicielles de gestion intelligentes dans l'éducation,
- Conseil, conception, construction et mise en œuvre de solutions de caméras de surveillance intelligente,
- Conseil pour le système éducatif (école et université) souhaitant une transformation numérique.

L'entreprise VDSmart mène aussi de recherche sur l'automatisation et la mécatronique entre autre; le développement de systèmes embarqués pour les systèmes Internet connectant des objets (IoT) et aussi des solutions de contrôle d'accès avec IA intégrée.

2.6 Projets

2.6.1 VDSmart Box

VDSmart Box est un dispositif d'analyse de vidéos utilisant l'intelligence artificielle pour contrôler les caméras (figure 2.2)



FIGURE 2.2 – VDSmart Box

Un appareil intégré avec des algorithmes d'intelligence artificielle pour le traitement d'image et le contrôle parallèle de plusieurs caméras IP fixes ou de type PTZ³, avec la possibilité d'analyser la reconnaissance faciale, l'émotion du visage, l'actions d'un corps humain. VDSmart AI Box transforme les caméras IP ordinaires en caméras intelligentes sans avoir à investir dans le remplacement de l'infrastructure de la caméra.

2.6.2 VDSmart Access

La solution VDSmart Access est un système biométrique de prise de présence et qui donne l'accès automatique à la porte du bureau; utilisant la technologie de reconnaissance faciale 4.0 la plus moderne avec un noyau d'intelligence artificielle. VDSmart Access permet aux entreprises de toujours comprendre le statut de travail et l'attitude de travail des ses employés. Il y a un avantage remarquable par rapport au système basé sur l'empreinte digitale est d'éviter la fraude, de réduire l'encombrement des portes lors de l'enrôlement de nombreuses personnes. Le système peut également aider avec des alertes de sécurité et de nombreuses autres fonctionnalités de haut niveau pour augmenter la productivité et l'efficacité de la gestion (par exemple, il rappelle automatiquement aux employés quand ils sont en retard au travail et félicite les employés actifs, joyeux anniversaire, etc.) figure 2.3.

3. PTZ : Pan Tilt Zoom, une caméra capable de commander sa direction et de contrôler le zoom

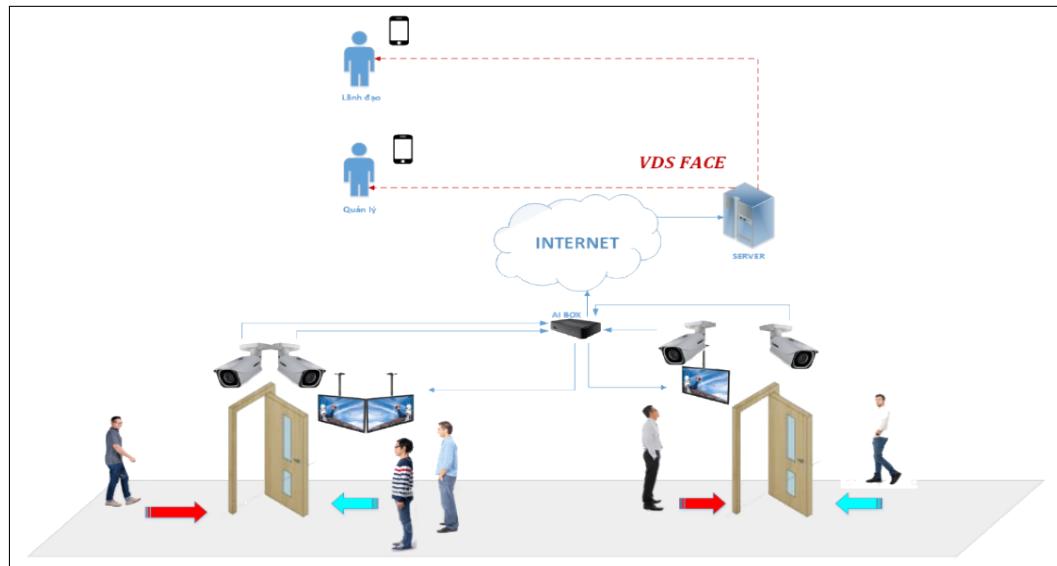


FIGURE 2.3 – VDSmart Access

2.6.3 Eye Pro Thermal

Le système Eye Pro Thermal se compose d'une caméra thermique de haute précision connectée à un ordinateur dédié, d'un écran tactile et d'un logiciel permettant la détection automatique de personnes ou d'objets présentant des températures anormalement élevées à une distance de 1 à 2 mètres. Avec une précision et une sensibilité élevées, le système détecte les symptômes de nombreuses maladies dangereuses telles que le SRAS, le CoVid-19, etc.



FIGURE 2.4 – Thermal Eye Pro

2.6.4 VDSmart Class

La salle de classe intelligente que l'entreprise VDSmart conseille et implémente pour les écoles et universités comme le montre la figure 2.5; offre de possibilités suivantes :

- les apprenants sont assis en petit groupe de 3 à 4 autour d'une table,
- la salle de classe est connectée avec une diffusion TV en direct,
- Signalisation numérique avec une chaîne de télévision interne,
- Technologies de gestion intelligente,
- Vidéo conférence pour l'éducation,
- Solution pour numériser les cours magistraux,
- Bibliothèque numérique intelligente,



FIGURE 2.5 – Salle de classe intelligente standard

2.7 Conclusion

Conventionnellement le stage de fin d'études de Master se déroule dans une entreprise ou un laboratoire suivant les ambitions de l'étudiant et les opportunités qui s'offrent à lui. Pour notre part c'est l'entreprise VDSmart[35] qui nous a accueilli pendant cette période. Cette dernière est spécialisée dans la conception des applications d'Intelligence Artificielle et elle a comme mission principale d'encourager les entreprises d'appliquer de l'IA à leurs activités commerciales et de production. Pour bien mener ce travail et obtenir le résultats attendus d'abord dans le chapitre suivant nous présentons le contexte dans lequel il fait en analysant.

Chapitre 3

Analyse du sujet

3.1 Contexte de la recherche

Depuis la fin des années 1960, l'utilisation de données biométriques pour sécuriser un système d'information ou pour un autre objectif particulier a conduit certaines entreprises à l'intégration de la reconnaissance faciale dans leur système. Il en est ainsi pour l'empreinte digitale, la reconnaissance de l'iris, etc. Cette recherche s'intéresse en particulier à la détection de faux visage pour un système basé sur la reconnaissance faciale (face recognition en anglais) en plus, de la reconnaissance de l'émotion exprimée par le visage de la personne. Il fera partie du projet Eye Pro Education (voir section 4.2.1) et Access VDSmart (voir 2.6.2) tous deux basés sur la reconnaissance faciale. Tous les visages présentés devant la caméra ne sont pas forcément réels (en live ou en direct). Ainsi dire, une photo imprimée d'une personne peut être reconnue par le système même quand l'individu n'est pas présent physiquement. Avec l'intégration de la phase de la détection de faux visage le système est sécurisé contre les différentes attaques. La reconnaissance d'émotion exprimée par le visage d'un apprenant en salle de classe s'avère utile pour l'enseignant afin de pouvoir envisager les possibles améliorations méthodologique et/ou du contenu de la matière.

3.2 Cadre théorique

3.2.1 Motivation

La vision par ordinateur (Computer Vision en anglais) est un thème traité en Master Informatique Systèmes Intelligents et Multimédia qui fait recourt a plusieurs techniques de traitement d'images et plus particulièrement celles de l'apprentissage machine enfin de donner aux machines une compréhension de haut niveau d'une image et/ou une vidéo numérique. La reconnaissance faciale est l'un des cas d'étude traité par la vision par ordinateur, mais la vulnérabilité que continu à subir les systèmes basés sur la reconnaissance faciale plus particulièrement l'accès intelligent (Smart Ac-

cess) a soulevé des questions qui nous ont interpellé et ont piqué notre intérêt face à la conjoncture actuelle de l'évolution de la technologie sur ce thème et surtout de l'industrie 4.0. Dans le même angle d'idée, nous avons jugé bon d'aller plus loin dans notre recherche pour identifier l'émotion exprimé par un visage à temps t.

3.2.2 Définition de quelques termes clés

L'ensemble de ce travail contient quelques termes du domaines qui nécessite de connaissance particulières afin de le comprendre, ci-dessous nous mettons au clair quelques uns;

1. **Reconnaissance faciale :** La reconnaissance faciale est un problème d'identification et de vérification de personnes dans une photographie par leurs visages¹. C'est une tâche qui est exécutée de manière triviale par les humains, même sous une lumière variable et lorsque le visage est modifié par l'âge ou obstrué par des accessoires et des poils du visage. Néanmoins, il est resté un problème de vision par ordinateur difficile pendant des décennies jusqu'à récemment. Souvent on a besoin de reconnaître les personnes dans une photographie pour l'une des raisons suivantes :

- *Identification* : lorsqu'on a besoin d'assigner un nom à un visage
- *Vérification* : pour confirmer que la personne corresponde bien à son ID
- *Authentification* : quand on veut restreindre l'accès à une ressource

Pour arriver l'un des objectifs énumérés ci-haut, la reconnaissance faciale est décrite comme une procédure impliquant quatre étapes principales comme le montre la figure suivante

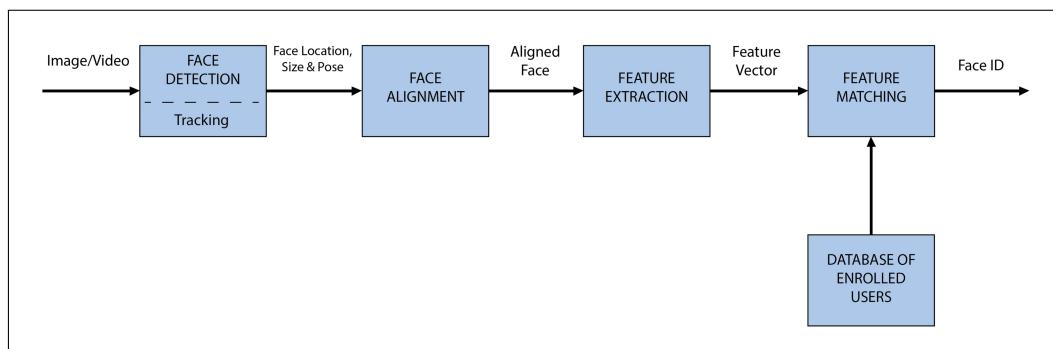


FIGURE 3.1 – 4 étapes de la reconnaissance faciale

Comme le montre la figure 3.1, les quatre étapes clés dans la procédure de la reconnaissance faciale sont ; la détection du visage (face detection), l'alignement du visage (face alignment), l'extraction des caractéristiques (feature extraction) et enfin la reconnaissance faciale (face recognition).

1. <https://machinelearningmastery.com/introduction-to-deep-learning-for-face-recognition/>

- *Détection du visage* : la première étape dans le processus de la reconnaissance faciale est la détection du visage, ceci consiste à localiser un ou plusieurs visage dans une image en marquant une boîte englobante pour chaque face détectée.
 - *Alignement du visage* : l'alignement consiste à la normalisation de la face détectée pour qu'elle soit cohérente avec la base des données.
 - *Extraction de caractéristiques* : cette phase permet d'extraire les caractéristiques du visage qui seront utilisées à la phase de la reconnaissance.
 - Reconnaissance faciale : la reconnaissance fait une mise en correspondance du visage avec un ou plusieurs visages connus dans une base de données préparée.
2. **Liveness detection** : Liveness detection permet de distinguer l'image réelle d'une personne des attaques présentées sous différentes formes (photo imprimée, vidéo ou masques). Le terme scientifique est la détection de présentation d'attaques, qui fait référence à la prévention de la fraude pour la biométrie en général, tandis que le liveness detection est spécifiquement utilisée pour la reconnaissance faciale. Les algorithmes de liveness detection les plus applicables sont indépendants du matériel et ne nécessitent que peu de coopération de l'utilisateur pour une expérience utilisateur optimale².



FIGURE 3.2 – Liveness detection
Source : <https://www.bioid.com/technology>

Sur la figure 3.2, on peut bien voir la différence entre l'image réelle présentée devant la caméra (à droite) et une attaque sous la forme d'une photo imprimée (à gauche).

3. **Machine Learning** : L'apprentissage automatique est une application de l'intelligence artificielle (IA) qui offre aux systèmes la possibilité d'apprendre et de

2. <https://www.bioid.com/liveness-detection/>

s'améliorer automatiquement à partir de l'expérience sans être explicitement programmés. L'apprentissage automatique se concentre sur le développement de programmes informatiques qui peuvent accéder aux données et les utiliser pour apprendre par elles-mêmes³. En général, les méthodes d'apprentissage automatique peuvent être catégoriser en deux : Supervisé et non Supervisé [34].

- *Méthodes supervisées* : ces méthodes se servent de ce qui a été appris dans le passé pour des nouvelles données en utilisant des exemples étiquetés pour prédire les événements futurs. Ces méthodes ont aussi la possibilité de comparer la sortie avec la sortie correcte prévue et trouver l'erreur afin de modifier le modèle en conséquence.
 - *Méthodes non-supervisées* : ces méthodes sont utilisées lorsque les données utilisées pour l'entraînement ne sont ni classifiées ni étiquetées.
4. **Deep Learning** : L'apprentissage profond (également connu sous le nom d'apprentissage structuré profond) fait partie d'une famille plus large de méthodes d'apprentissage automatique basées sur les réseaux de neurones artificiels avec apprentissage par représentation.
 5. **Computer Vision** : La vision par ordinateur (aussi appelée vision artificielle ou vision numérique) est une branche de l'intelligence artificielle dont le principal but est de permettre à une machine d'analyser, traiter et comprendre une ou plusieurs images prises par un système d'acquisition (caméras, etc.)⁴
Quelques applications de Vision par Ordinateur :
 - Détection des défauts (Defect detection)
 - Métrologie
 - Vérification de l'assemblage
 - Lecteur d'écran (Screen reader)
 - Lecteur de Code et de Caractères (OCR)

6. **Reconnaissance d'émotions** : La reconnaissance d'émotion est un processus d'identification d'état d'émotion d'une personne à partir de signaux biologique physique [25]. Le but est de recueillir les données et analyser le sentiment de sujet pour obtenir des réponses possibles appropriées [10]. Les données peuvent provenir de différentes sources physique telle que la voix, le mouvement corporel et autre signes biologique physique. La figure 3.3 montre la chaîne de traitement pour la classification d'émotion.

3. <https://expertsystem.com/machine-learning-definition/>

4. https://fr.wikipedia.org/wiki/Vision_par_ordinateur

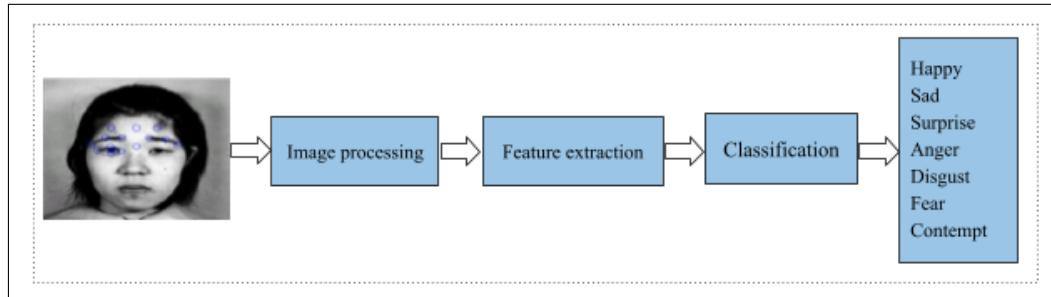


FIGURE 3.3 – Emotion recognition process

3.2.3 Différents types de présentation d'attaques

Techniquement, les attaques contre le système à reconnaissance faciale sont appelées présentation d'attaques (presentation Attacks en anglais) et peuvent être subdivisé deux grandes catégories comme le montre le tableaux ci-dessous;

	Statique	Dynamique
2D	photographies, papier plat, masque en plastique	relecture de la vidéo, diapositive de plusieurs photos
3D	Impression 3D, sculpture, masque	des robots qui reproduisent des expressions, maquillage bien préparé

TABLE 3.1 – Différents type d'attaques [29]

Suivant le cas d'étude, on s'intéresse souvent aux formes d'attaques qui peuvent arriver au système, mais un bon système doit être capable de contrer toutes sortes de présentation d'attaque car, au fur et à mesure que les technologie évolues, les auteurs de présentation d'attaques pensent aussi à améliorer leur façons de faire.

La présentation d'attaque basée sur 3D n'est pas encore un gros problème, la 2D est plus répandue. Cela oblige à détecter et à prévenir la présentation d'attaques. Les exigences sont précises et le produit doit :

- constraint la présentation 2D, statique ou dynamique,
- l'utilisation des images, pas des vidéo,
- travaillé sans l'intervention de l'utilisateur.

L'objectif est d'atteindre une précision maximale en un minimum de temps tout en offrant une expérience conviviale. Un modèle répondant à ces exigences serait facile à intégrer aux systèmes de reconnaissance faciale existants [29].

3.2.4 Les 7 principales émotions

L'expression faciale est l'un des signaux les plus puissants, naturels et universels permettant aux êtres humains de transmettre leurs états émotionnels et leurs intentions [28]. La physiologie humaine est naturellement dotée de ces sept émotions qui peuvent être exprimée de différente manière suivant l'évènement qui se produit. Ci-dessous les sept principales émotions sont listée [6]

- **La Colère** : Selon les chercheurs, l'expression de colère (figure 3.4) fonctionne si bien parce que chaque mouvement du visage donne à une personne une apparence physiquement plus forte. Cette expression permet à la menace de savoir que nous sommes sérieux. C'est l'une des émotions les plus puissantes et cela montre à quel point le visage humain peut être expressif. Cette expression sert d'avertissement, que ce soit simplement pour intimider ou pour montrer qu'un conflit a commencé.



FIGURE 3.4 – Expression de la colère [6]

Présentation du visage : Sourcils tirés vers le bas, paupières supérieures relevées, paupières inférieures relevées, bords des lèvres enroulés, les lèvres peuvent être resserrées.

- **La Peur** : Chaque mouvement du visage basé sur la peur (figure 3.5) nous prépare à une réponse de combat ou de fuite. Cette expression du visage capitalise sur le fonctionnement de notre corps. Élargir nos yeux ouvre notre champ de vision, laisse entrer plus de lumière et permet de voir les menaces qui nous entourent. La même chose peut être dite pour nos voies d'oxygène. Ouvrir nos narines augmente notre apport en oxygène et nous aide à nous préparer à fuir ou à nous battre.



FIGURE 3.5 – Expression de la peur [6]

Présentation du visage : Sourcils relevés et ensemble, paupières supérieures relevées, bouche étirée

- **Le dégoût :** L'expression du dégoût ne montre pas seulement notre répugnance à quelque chose, il fonctionne aussi pour nous protéger. le fait de rider le nez ferme le passage nasal en le protégeant des vapeurs dangereuses et en plissant les yeux les protège des dommages.



FIGURE 3.6 – Expression du dégoût [6]

Présentation du visage : Sourcils tirés vers le bas, nez plissé, lèvre supérieure relevée, lèvres lâches.

- **La joie :** Malgré la connotation amicale, les chercheurs pensent que nos sourires pourraient avoir une origine plus sinistre. De nombreux primates montrent leurs dents pour affirmer leur domination et verrouiller leur statut dans leur structure sociale. Certains chercheurs pensent que c'est ce signe non verbal qui a finalement évolué vers un sourire.



FIGURE 3.7 – Expression de la joie [6]

Présentation du visage : Muscle autour des yeux resserré, rides «pattes d'oie» autour des yeux, joues relevées, coins des lèvres relevés en diagonale.

- **La tristesse :** Selon les chercheurs, la tristesse est une expression difficile à simuler. L'un des signes révélateurs de la tristesse est l'élévation des sourcils, ce que très peu de gens peuvent faire à la demande.



FIGURE 3.8 – Expression de tristesse [6]

Présentation du visage : Coins intérieurs des sourcils relevés, paupières lâches, coins des lèvres abaissés.

- **La surprise :** Bien que l'expression de surprise ne dure qu'une seconde ou deux, la présentation du visage, en particulier les sourcils levés, nous permettent de prendre conscience de notre environnement, de porter notre attention sur un autre événement potentiellement menaçant et de réagir plus rapidement. Que ce soit une bonne ou une mauvaise surprise, la réaction du visage est la même.



FIGURE 3.9 – Expression de surprise [6]

Présentation du visage : Sourcil entier relevé, paupières relevées, bouche ouverte, pupilles dilatées.

- **Le mépris :** Bien que l'émotion de mépris puisse chevaucher la colère et la méfiance, l'expression du visage est unique. C'est la seule expression qui ne se produit que sur un seul côté du visage et qui peut varier en intensité. À son plus fort, un front peut s'abaisser tandis que la paupière inférieure et le coin des lèvres se lèvent du même côté. Dans sa forme la plus secrète, le coin de la lèvre ne peut se lever que brièvement.

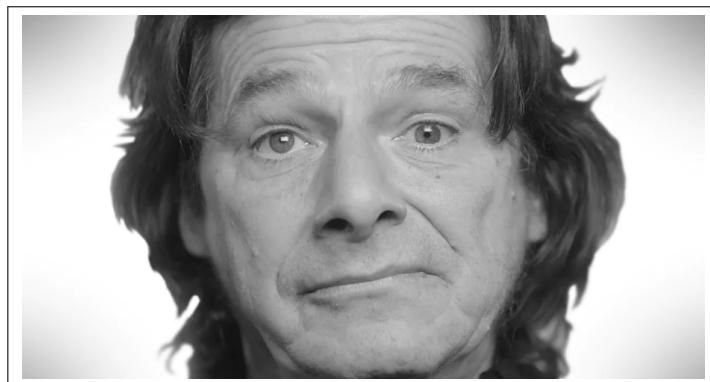


FIGURE 3.10 – Expression de mépris [6]

Présentation du visage : Yeux neutres avec le coin des lèvres tiré vers le haut et en arrière d'un côté.

3.3 Problématique

Problème

Comme nous venons de l'indiquer dans l'introduction, le problème principal dont il est question dans ce travail de recherche est la lutte contre l'usurpation de visage. Ainsi la solution proposée est destinée être intégrée dans le système Eye Pro Education et Smart Access en particulier. De l'autre côté il semble être utile d'évaluer les émotions faciale dominantes parmi celles qui sont détectées.

- Face Anti-spoofing/Liveness detection : un système d'identification, de vérification ou d'authentification basé sur la reconnaissance faciale reste toujours fragile tant qu'il ne peut pas être capable de différencier un vrai visage du faux visage ; c'est-à-dire une image réelle de la personne et non une image pré-imprimée ou une caricature de la personne. Dans le cas où on est capable de détecter un visage (travaux existants) ; pour rendre le système encore plus robuste il faut s'assurer que le visage détecté est un vrai ou faux visage (attaque) avant d'identifier le sujet.
- Reconnaissance d'émotion : le processus de reconnaissance d'émotion faciale reste moins intéressant tant qu'il se limite par le simple fait d'identifier une émotion tout simplement, pour rendre cela encore plus utile dans un environnement éducatif comme celui de Eye Pro Education l'analyse des émotions dominantes est requise enfin d'en tirer le maximum possible.

Question

La première et la principale question à quelle notre recherche tante de répondre est de savoir; si la face présente devant la caméra est-elle réelle ou est une attaque, et la deuxième question est de savoir, parmi les émotions exprimées par les sujets à temps t lesquelles sont dominantes.

Hypothèse

Une mise en place d'un système d'information et d'aide à la décision adapté au mode éducatif (participation au cours, appréciation de la matière ou la méthode de l'enseignant, etc.) contribuera à la concrétisation d'un système éducatif moderne et intelligent.

3.4 Objectifs

Objectif global du projet

Contribuer d'une manière concertée à l'effectivité de la révolution du système éducatif; à travers le développement d'un outil d'aide à la décision adapté au contexte

d'enseignement.

Objectif spécifique du projet

A travers ces quelques objectifs spécifiques nous espérons atteindre l'objectif global de cette recherche ;

- Contribuer à l'évaluation de l'efficacité des méthodes de reconnaissance de visage et aussi à travers le couplage de la détection du vrai/faux visage,
- Contribuer à l'évaluation de la qualité de méthode reconnaissance d'émotion,
- Contribuer à l'évaluation de la vulnérabilité de système de sécurité basé sur la reconnaissance de facial par l'utilisation de méthode de détection du vrai/faux visage,
- Concevoir un modèle prototype d'aide à la décision dans le contexte d'un système éducatif moderne

Notre mission

Nous entant que stagiaire nous avons eu une mission à deux volées sur ce projet enfin d'aider l'entreprise VDSmart d'atteindre ses objectifs. Dans la première rubrique de ce travail de recherche qui est 'La détection d'attaque de visage (Face anti-spoofing detection en anglais)' nous étions chargé de :

- Proposer une base des données performante qui répond mieux au problème de détection d'attaques de visage. Celle-ci peut être soit construite par nous mêmes ou venir de l'extérieur.
- Concevoir et implémenter l'architecture réseau d'entraînement du modèle d'apprentissage automatique pouvant distinguer à une précision près un vrai visage du faux.
- Proposer le plan d'entraînement et entraîner le modèle,
- Évaluer et améliorer le modèle dans le cas échéant.

Dans la seconde rubrique consacrée à la détection d'émotions, signalons que l'entreprise VDSmart avait déjà mis en place un modèle d'apprentissage automatique de détection d'émotions et qui avait déjà été intégré dans le système Eye Pro Education (cfr Chapitre 4), à cet effet il nous avait été demandé de mettre en place un module basé sur la technologie web pour calculer et afficher les fréquences d'émotions des étudiants.

3.5 Résultats attendus

Les résultats attendus pour ce travail de recherche est une mise en place d'un modèle prototype capable de prédire si le visage devant la caméra est un vrai visage ou une attaque à cela s'ajoute un module supplémentaire pour évaluer la tendance d'émotions dominantes des apprenants dans une salle de classe.

3.6 Conclusion

La première étape la plus importante dans un travail de recherche est la compréhension théorique du problème qu'on veut résoudre en le situant dans le temps et l'espace. Pour cela, dans ce deuxième chapitre nous avons tenté de mettre au clair la problématique soulevé par l'entreprise VDSmart pour le système basé sur la reconnaissance faciale en partant par la définition de termes clés, l'objectif du projet jusqu'à l'énoncé des résultats attendus. Notre mission principal étant de proposer une solution au problème de détection d'attaques de visages, dans le chapitre suivant nous nous proposons de revoir les travaux antérieurs que d'autres chercheurs ont réalisé enfin de bien nous orienté.

Chapitre 4

État de l'art

4.1 Introduction

Dans cette section en premier lieu nous parlerons de l'existant qui concerne les produits de l'entreprise VDSmart basés sur la reconnaissance faciale pour lesquels la détection d'attaques de visages s'avère nécessaire, ensuite nous parlerons de travaux connexes en rapport avec la détection d'attaques de visages et de la reconstruction du depth map et enfin nous terminerons par la reconnaissance d'émotions.

4.2 Étude de l'existant

4.2.1 Eye Pro Education

L'idée et la motivation du projet Eye Pro Education sont venues de l'éclosion de l'épidémie de 2019 poussant ainsi l'entreprise VDSmart à penser à une solution pouvant garantir les écoles et même les entreprises à continuer à exercer leur fonction sans avoir besoin à se rendre à leur lieu de travail habituel, cela pour diminuer voir même stopper l'expansion de l'épidémie. La figure 4.1 montre quelques captures du système Eye Pro Education.

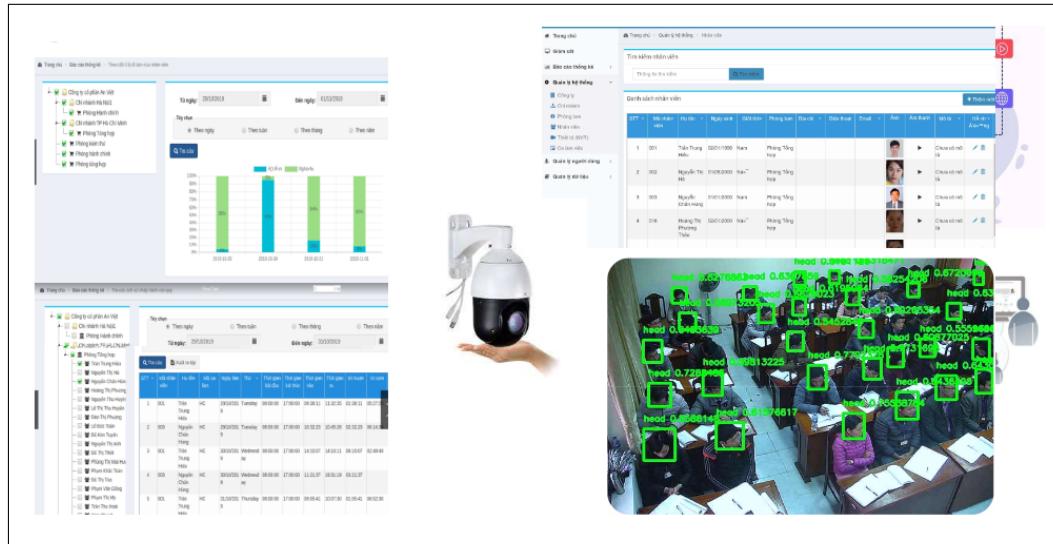


FIGURE 4.1 – Eye Pro Education [35]

Remarquant que ; les réunions traditionnelles doivent être remplacées par des réunions en ligne et la plupart des écoles du pays (Vietnam) se tournent vers l'enseignement en ligne l'entreprise à l'abord évaluer quelques systèmes qui existent et a relevé quelques lacunes à résoudre :

- Certains logiciels libres limités aux langues étrangères et ne garantissent pas la sécurité de l'information,
- Réunions en ligne les participants sont incontrôlables (travailler séparément, rejoindre tardivement, laissez l'ordinateur en ligne et sortir de la réunion, etc.)
- Les étudiants / étudiants en ligne ont de la difficulté à se concentrer plus de 15 minutes. Les enseignants sont absorbés par des conférences incontrôlables, laxisme dans la gestion des étudiants.
- Difficulté de savoir qui est réellement présent (par exemple : faire l'appel)

Après cette étude l'entreprise VDSmart était bien placée enfin de proposer une solution adéquate aux problèmes liés à l'apprentissage en ligne et cela a été effectif. L'expérience utilisateur a fait savoir que le système était toujours vulnérable aux présentations d'attaques basées sur le visage comme la plus part des autres systèmes à reconnaissance faciale et autres fonctions utiles qu'il fallait intégrer, ainsi est l'objet de ce travail.

4.3 Travaux connexes

La présentation d'attaques de visage est le principal aléa pour le système basé sur la reconnaissance faciale. Depuis un certain temps des chercheurs se sont engagés dans la lutte contre ce dernier, et nous avons revu les travaux antérieurs sous deux angles à

savoir; les méthodes basées sur l'utilisation d'image traditionnelle sans aucune information supplémentaire comme le depth map, et celles basées sur la reconstruction ou la génération du depth map.

4.3.1 Face anti-spoofing

Selon le type de caractéristique (features) utilisé, les approches de face anti-spoofing peuvent être catégoriser en deux groupes : les méthodes basées sur le liveness¹ et celles basées sur la texture.

FAS with joint Spoofing Medium Detection and Eye Blinking Analysis :

La méthode proposée par Mikhail Nikitin dans [16] pour le problème de détection d'attaques de visage est basée sur l'analyse et la fusion de deux types de caractéristiques qui sont; la visibilité de spoofing medium devant la caméra et la détection de clignement des yeux.

1. **Détection de Spoofing Medium :** selon l'auteur [16], en plaçant un objet devant un autre il y aura inévitablement une discontinuité de texture visible aux alentours de l'image du premier plan. Par cette affirmation l'auteur propose de créer un algorithme pour détection ce genre de discontinuité sur l'image faciale. Par la quasi inexistance d'une base de données basée sur ce principe, l'auteur part par deux étapes; d'abord la génération des données synthétique (figure 4.2) qui permet d'obtenir les images de la classe attaques basées sur les photos imprimées et vidéos et ensuite la classification binaire de la présence du medium.

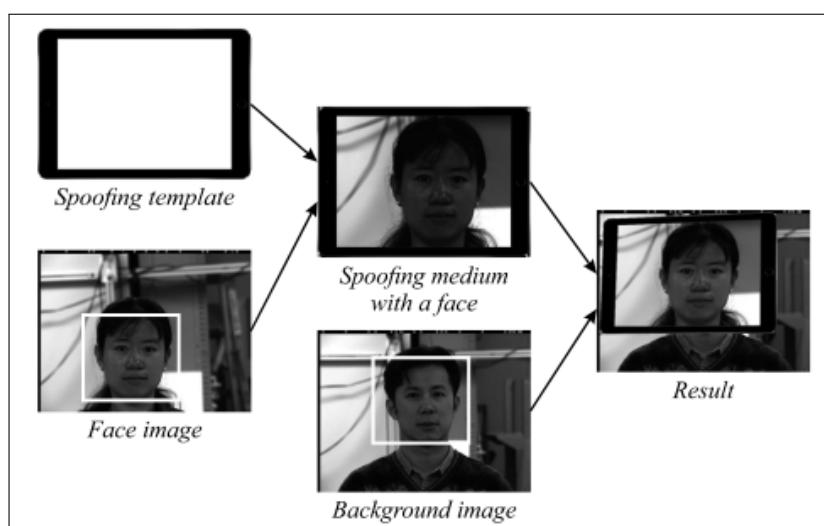


FIGURE 4.2 – Processus de génération de données synthétiques [16]

1. Liveness : tentent de détecter les signes de vie en suivant les mouvements des certaines parties du visage, tels que le clignement des yeux ou les mouvements des lèvres. [16]

2. **Détection de clignement des yeux :** pour être sûr que les yeux clignotent dans une séquence vidéo, l'auteur [16] utilise un modèle de classification d'ouverture des yeux lequel est appliqué à chaque frame de la vidéo. Cette classification donne une probabilité si les yeux sont ouverts sur une image, pour classifier toute la séquence vidéo il analyse la différence entre les probabilités minimum et maximum comme le montre la figure 4.3

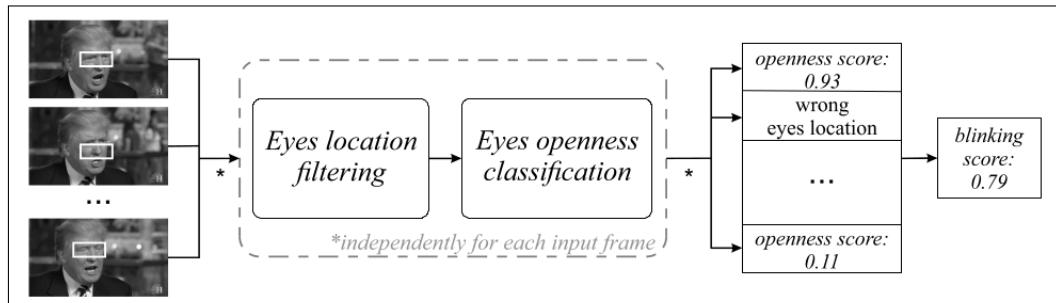


FIGURE 4.3 – Processus de détection de mouvement des yeux [16]

Visiblement la méthode proposée par Mikhail Nikitin dans [16] basée sur la détection de spoofing medium et la détection de clignement des yeux pour la détection d'attaques de visage ne pourra pas détecter une vidéo jouée sur un grand écran rapproché beaucoup plus de la caméra sans laisser voir une autre image. Par là la discontinuité ne sera pas détectée et le clignement des yeux peut être détecté pour de frame contenant un visage.

Improving FAS by 3D Virtual Synthesis :

Dans le même cadre d'idée que l'auteur du [16] pour obtenir les données représentant les différents types d'attaques, l'auteur ici [12] propose aussi une méthode consistant à une synthèse virtuelle des objets 3D pour la détection d'attaques de visages. Pour cela il part par trois étapes à savoir; d'abord le maillage et la déformation des objets 3D puis la projection du perspective et enfin le post-traitement.

1. **Le maillage et la déformation d'objet 3D :** pour parvenir à manipuler un objet ayant une structure 3D sur une surface plane, l'auteur ici propose d'abord de convertir l'image dans un objet 3D enfin de manipuler son apparence. Tout d'abord les quatre coins délimitant la face dans l'image sont étiquetés et la région concernée est recadrée (figure 4.4.a.) Le résultat obtenu de cette première étape est uniformément échantillonné par l'ancre et enfin l'algorithme delaunay² est utilisé pour trianguler ces points et maillé la photo imprimée dans un objet virtuel 3D (figure 4.4.c.)

2. https://en.wikipedia.org/wiki/Delaunay_triangulation

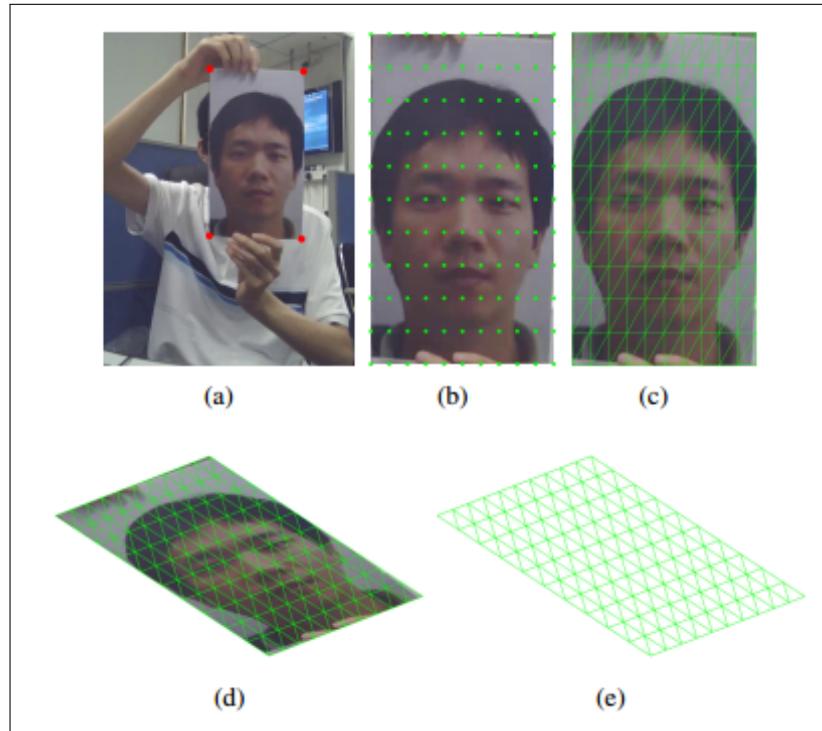


FIGURE 4.4 – Maillage et déformation d'un objet 3D [12]

La figure 4.4.d et 4.4.e représentent la vue 3D du résultat obtenu de cette première phase. Après le maillage, les opérations de transformations 3D comme la rotation ou le fléchissement peuvent être appliquées.

2. **Projection du perspective :** Pour parvenir à projeter l'image issue de la phase précédente, l'auteur ici propose d'abord de se rapprocher de la taille physique de la photo imprimée. Pour cela l'auteur suppose la distance des pixels et la distance réelle entre le centre de deux yeux. Après les différentes projections, l'algorithme Z-buffer³ est utilisé pour l'affichage de résultat.

3. <https://www.geeksforgeeks.org/z-buffer-depth-buffer-method/>

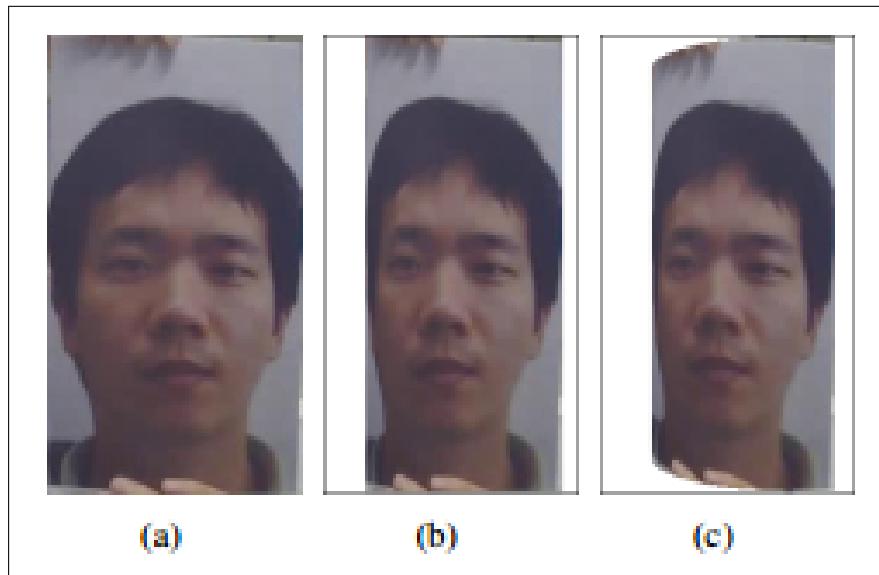


FIGURE 4.5 – Projection du perspective [12]

La figure 4.5.b représente une projection avec une faible perspective, tandis que 4.5.c avec une perspective normale.

3. **Le post-traitement :** L'auteur ici à ce niveau constante qu'après la déformation et la projection du perspective la taille de la photo synthétique est modifiée, pour cela le filtre Gaussien est appliqué pour rendre égale les bordures de la fusion comme on peut le voir sur la figure 4.6.

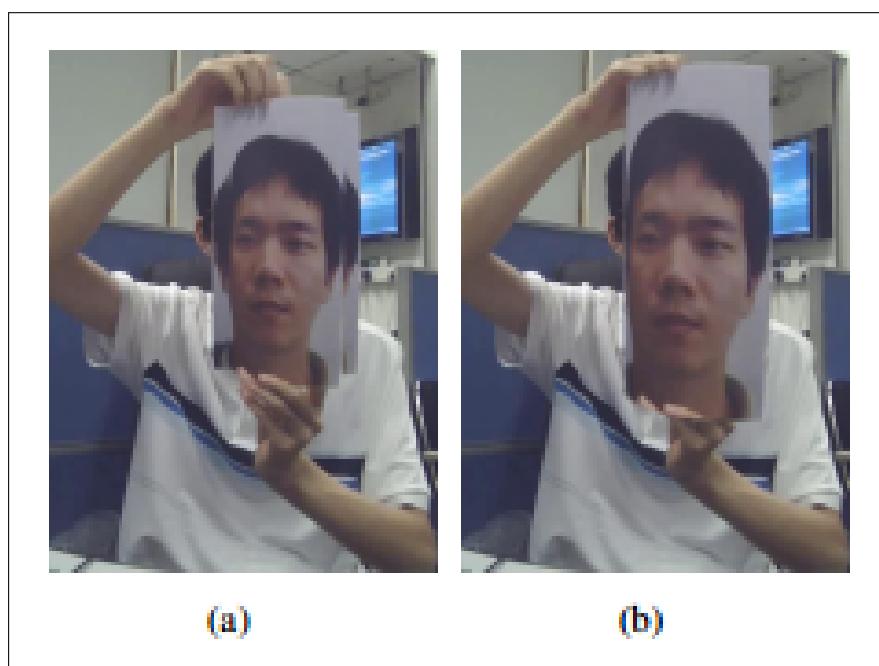


FIGURE 4.6 – post-traitement [12]

Multi-Modal FPAD via Spatial and Channel attentions :

L'auteur de [11] suppose que la plus part des approches traditionnelles pour le problème de détection d'attaques de visage manque les moyens pour de nouveaux types d'application et cela est due à la modalité de données. Pour cela l'auteur propose une approche consistant à fusionner plusieurs modalités étant donné que la base d'expérimentation CASIA-SURF⁴ est caractérisée par la bande de chaînes RGB, Depth and IR modalités.

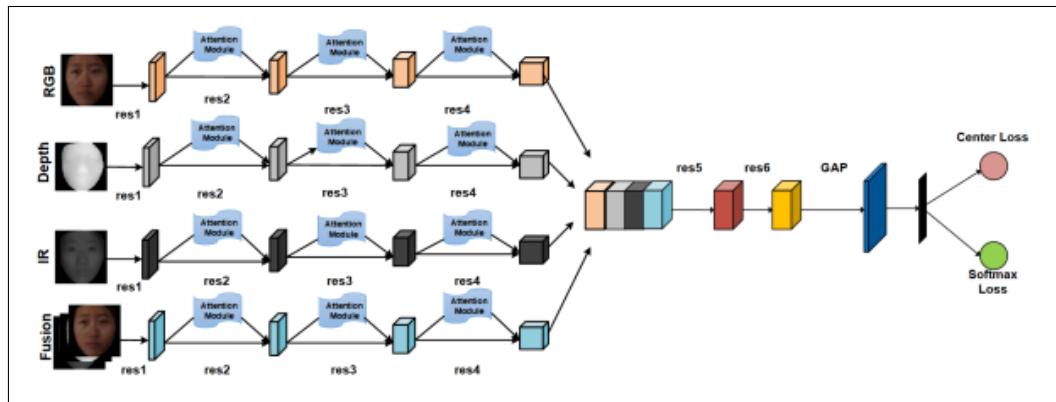


FIGURE 4.7 – Multi-modalité PAD [11]

Comme on peut le voir sur le figure 4.7, l'approche proposée par l'auteur ici est une fusion de quatre branches de modalités telle que la branche de la modalité RGB, Depth, IR et la branche qui met en fusion ces trois dernières. Les caractéristiques extraites de ces quatre banches sont ensuite concaténées et introduites dans une couche partagée pour obtenir le résultat de la classification finale.

4.3.2 La reconstruction du depth map

Dual Camera Based Feature For Face Spoofing Detection

La figure 4.8 montre la vue d'ensemble de la méthode proposée par l'auteur de [22]. Les deux images d'entrée du réseau proviennent d'un système de caméra binoculaire (voir figure 5.1.c). Les deux images capturées au même moment fournissent une texture additionnelle et une structure d'information lesquelles rendent le système de détection des faux visages très robuste et effective.

4. CASIA-SURF : is the largest publicly available dataset for face anti-spoofing in terms of both subjects and modalities.

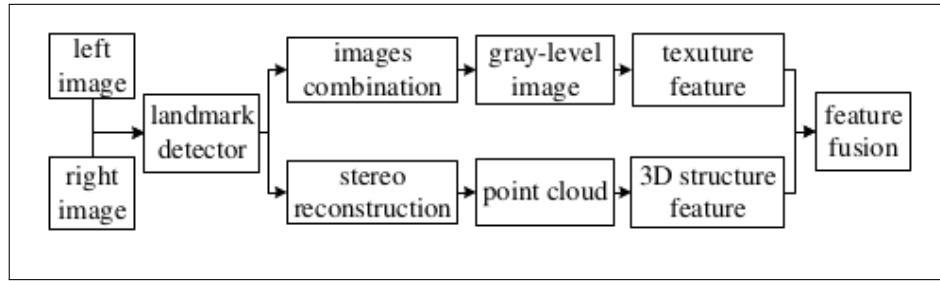


FIGURE 4.8 – Architecture dual camera based features [22]

Pour obtenir le composant 2D, l'auteur de [22] propose un algorithme basé sur la similarité d'image pour combiner la paire d'images d'entrée dans une image de niveau gris à partir de laquelle les caractéristiques Gabor⁵ sont extraites dans les régions d'intérêt. Pour obtenir le depth map qui représente le composant 3D, l'auteur ici se basant sur la méthode de l'Histogramme des Point de Caractéristique (Point Features Histogram) le nuage de points sont obtenus par l'algorithme de reconstruction stéréo. La figure 4.9 illustre bien l'architecture de la méthode proposée par [22]. Enfin les deux composants 2D pour la texture et 3D pour le depth sont normalisés avant d'être concaténés dans un seul vecteur de caractéristiques et utilisé comme la paire d'image d'entrée.

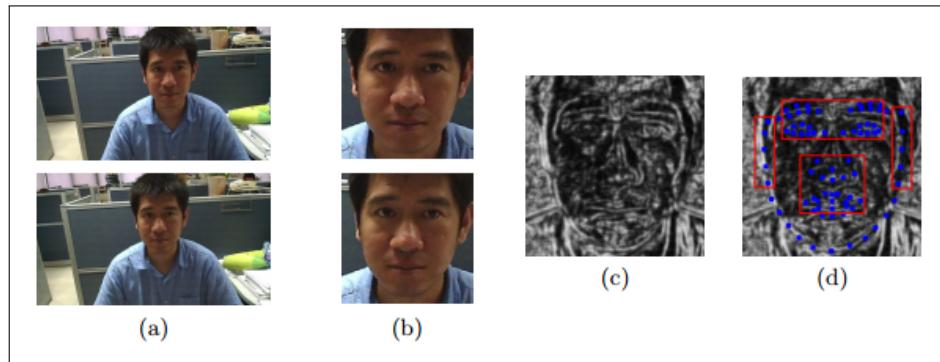


FIGURE 4.9 – Illustration de combinaison d'images [22]

FAS Using Patch and Depth-Based CNN

Pour pallier au problème de FAS, l'auteur de [24] propose l'approche de fusion de deux canaux CNN (figure 4.10) l'un basé sur le Patch de l'image qui est entraîné pour apprendre les caractéristiques d'apparence riche et l'autre sur l'estimation du depth pour estimer le depth map de l'image. Comme [22, 23], l'auteur du [24] suppose aussi que le depth map d'une photo imprimée ou d'une vidéo est plane et que celui d'une photo réelle (en direct) est normal.

5. https://cran.r-project.org/web/packages/OpenImageR/vignettes/Gabor_Feature_Extraction.html

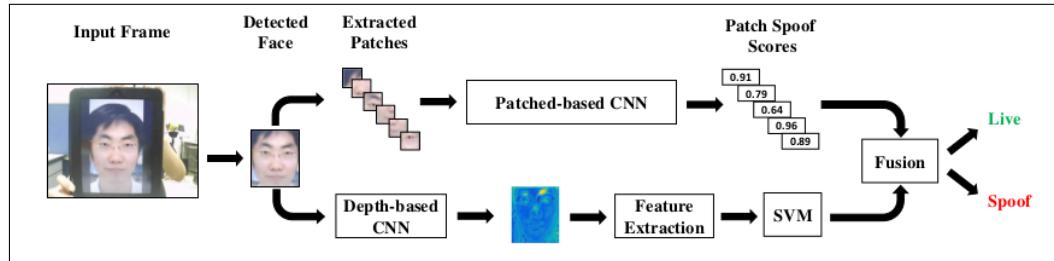


FIGURE 4.10 – Fusion de CNN basé sur le patch et depth [24]

A partir de l'image du visage, l'auteur de [24] choisi aléatoirement 10 patches qui sont capable de distinguer un vrai visage du faux comment on peut le voir sur la figure 4.11

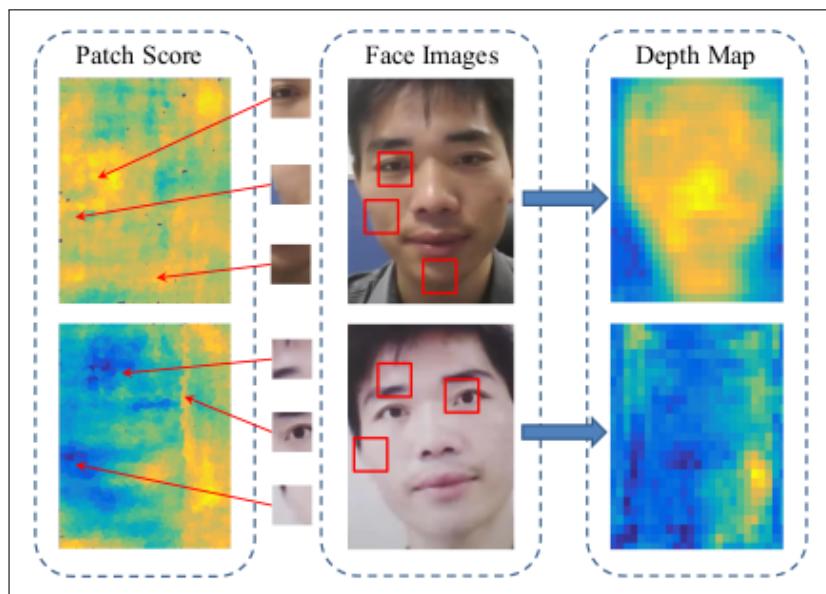


FIGURE 4.11 – Extraction de patches et estimation de depth [24]

Sur la figure 4.11 la colonne de gauche montre le score de sortie de patches locaux pour un visage réel (en haut) et une attaque (en bas), pour [24] les couleurs jaune et bleu représentent respectivement une forte ou faible probabilité d'attaque. La colonne de droite montre aussi la sortie estimée de depth map où les couleurs jaune et bleu représentent respectivement les points les plus proches et plus éloignés.

4.3.3 Reconnaissance automatique d'émotions

Pour le deuxième volé de cette recherche qui est la reconnaissance d'émotions, nous avons aussi revu quelques travaux existants et dont les plus pertinents sont présentés dans cette sous section. Signalons que la démarche générale pour la reconnaissance d'émotions est constituée de trois étapes principales dont l'acquisition de l'image contenant un visage, l'extraction de caractères et la classification (voir figure 3.3)

An ER Model Based on Facial Recognition in Virtual Learning Environment

Les solutions basées sur la sélection des yeux seulement et utiliser les caractéristiques issue de cette région ne donnent pas de bons résultats pour la reconnaissance d'émotions [10]. L'auteur de [10] propose d'ajouter les caractéristique issues de la partie de la bouche en utilisant Haar Cascades. Pour y arriver d'abords il sélectionne et extrait les caractéristiques de la région de la bouche et des yeux et puis il détecte et filtre le bord à l'aide des filtres moyens et médians comme on peut le voir sur figure 4.12.

Sample Images	Proposed Solution (Haar Cascades) Eyes and Mouth				Current Solution (Sobel Edge Detection Eyes)		
	Output	Detected Emotion	Accuracy (%)	Processing Time(sec)	Output	Accuracy (%)	Processing Time(sec)
Group 1							
		Sad	80.5	138		79	142
		Disgust	89.2	113		89	115
		Fear	84.3	158		83	164
		Anger	87	113.5		82	118

FIGURE 4.12 – Détection and filtrage de bord [10]

Sur la figure 3.3 on peut voir la démarche générale proposée par [10] pour le problème de reconnaissance d'émotions. La présentation du visage est le principal ou le seul indicateur l'émotion exprimée par une personne pour cela plusieurs composant du visage interviennent enfin de déterminer avec une précision ressentie par le sujet (voir 3.2.4). Ceci nous pousse à croire que la solution proposé par [10] pourra ne pas reconnaître certaines émotions dans le cas où une émotion est influencée par le nez, sourcil ou une autre partie que la bouche et les yeux.

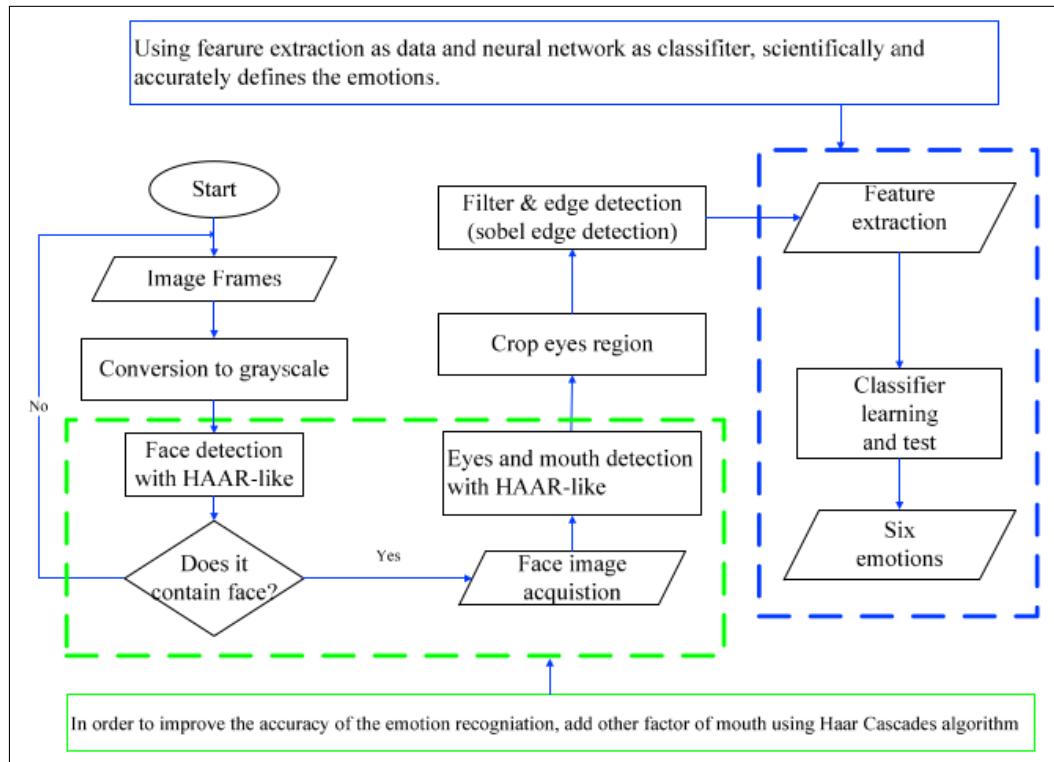


FIGURE 4.13 – La procédure de la solution proposée [10]

Facial Emotions Recognition using CNN

L'illumination, l'occlusion et le changement de la position de la tête sont les facteurs principales qui affectent la qualité du système de reconnaissances d'émotions [33, 18, 19] par conséquent une émotion peut être détecté à la place d'une autre, pour cela l'auteur ici [26] propose sur la phase de pré-traitement la normalisation de l'intensité et l'amélioration du contraste pour permettre au réseau CNN de reconnaître les caractéristiques avec une plus grande précision. Le pré-traitement proposé par [26] est illustré sur la figure 4.14.

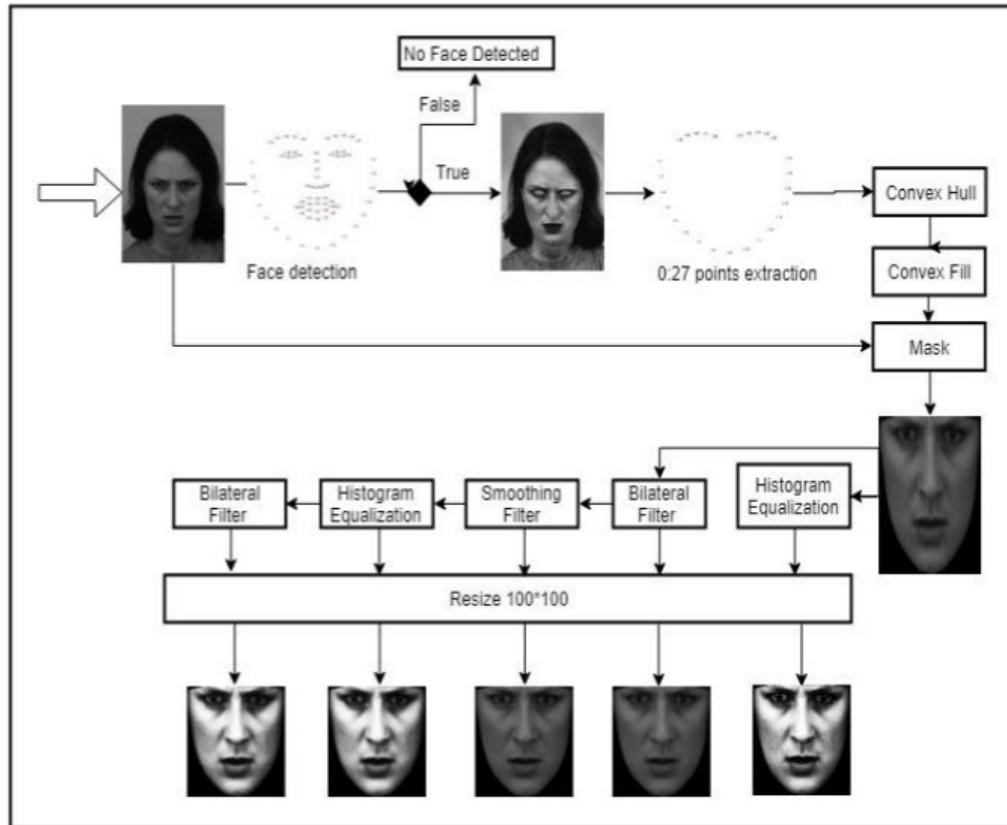


FIGURE 4.14 – Etape de pre-traitement [26]

4.4 Analyse des solutions existantes

Après la révision de quelques travaux sur le problèmes de face anti-spoofing que nous avons menée sous deux axes voici les constantes;

- Liveness detection : les méthodes basées sur la détection de liveness tentent de détecter les signes de vie en suivant le mouvement des certaines parties du visage, tel que le clignement des yeux ou le mouvement des lèvres, par conséquent il a été constater qu'il y a toujours une probabilité qu'un système basé sur le liveness détection soit toujours vulnérable des attaques par vidéo.
- Texture : de même, les solutions basées sur la texture restent aussi vulnérables aux présentations d'attaques par les écrans à haute définition. Actuellement sur le marché électronique nous trouvons les équipements à très haute définition avec un affichage proche du réel qui peut souvent pousser le modèle à confondre une attaque d'une image réelle.
- CNN : Certaines études prouvent que les CNN extraient les caractéristiques de la même façons pour une base d'images et par conséquent une image réelle pourra être confondue à une attaque du fait que la classe des images attaques provient

de la classe des images réelle; pour se faire il faut effectuer une traitement supplémentaire enfin que les caractéristiques de ces deux classes soient distinctes.

Le travail que nous avons revu 4.3.1 basé sur la fusion de la modalité RGB, Depth, IR et enfin une autre modalité combinant les trois premières, utilise une base d'expérimentation qui offre toutes ces modalité, c'est-à-dire lors de la mise en production il sera requit d'utiliser une caméra offrant toutes ces modalités. Voilà pourquoi nous avons opté pour une solution indépendante du matériel et qui et qui promet de distinguer les caractéristiques des deux classes sur base de leur forme.

4.5 Conclusion

Dans ce chapitre dédié à l'étude ciblée, approfondie et critique des travaux antérieurs, nous avons revu ces derniers sous trois axes dont :

- Face anti-spoofing : les solutions basées sur l'utilisation d'images ordinaire sans aucune information supplémentaire comme le depth map. Ces solution se focalisent plus sur la détection de certaines partie du visage comme les yeux, les lèvres, etc.
- La reconstruction du depth map : les solutions basées sur la génération ou la reconstruction du dépth map à partir des images issues d'une caméra à double pixels.
- La reconnaissance automatique d'émotions : en générale la reconnaissance d'émotions passe par trois trois phases respective à savoir; l'acquisition de l'image et le pré-traitement, l'extraction de caractéristiques et enfin la classification.

Avant tout nous avons d'abord présent le système Eye Pro Education sur lequel est prévu être appliquée la solution issue de ce travail de recherche.

Chapitre 5

Méthode proposée

5.1 Introduction

Motivé par le simple fait que ajuster deux caméras ne peut pas être simplement utilisé pour capturer des images et/ou enregistrer des vidéos, la méthode proposée pour le problème d'usurpation de visage (face anti-spoofing en anglais) est basée sur l'utilisation de double caméras; qui peut capturer deux images au même instant pour augmenter la difficulté de l'usurpation de visage. Le modèle est entraîné sur basé de 10.045 images dual caméra rangées en paire incluant 5.167 paires pour la classe réelle (genius) et 4.878 paires pour différents type d'attaques comme la vidéo et photo imprimée. Le réseau étant constitué de deux grandes parties dont l'une pour la prédiction du depth map¹ et l'autre pour la classification du depth map. Étant convaincu que le depth map possède des informations importantes pouvant nous aider à distinguer un visage réel d'une attaque [23, 21], cette section parle en détail de la solution proposée.

5.2 Caméra à Double Pixel

Caméra à double pixel (ou dual pixel camera en anglais) offre deux capteurs de photo au lieu d'un seul comme pour un capteur standard. L'utilisation de dual caméra peut être obtenu en deux possibilités; soit un équipement conçu directement pour ça comme la plus part de smartphone actuels (figure 5.1.a) ou une caméra stéréo (figure 5.1.b) ou soit en montant côté à côté deux caméras standards tout en les orientant dans une même direction de façons que leur vue soit très proche (figure 5.1 c).

1. https://en.wikipedia.org/wiki/Depth_map

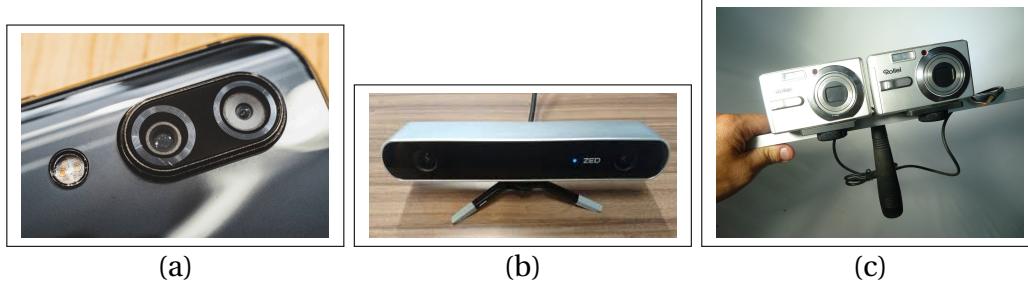


FIGURE 5.1 – Caméra double pixel

La figure 5.1. montre quelques utilisations possibles d'une caméra à double pixel qui au fait joue un rôle important pour l'obtention de depth map d'une image.

5.3 Génération de Depth Map

Le réseau pour générer le depth map d'une paire d'image est constitué de deux blocs essentielles à savoir, l'étape d'encodage et celle de décodage comme le montre la figure 5.2. Dans ce travail le depth map est l'élément clé pour faire la prédiction du visage devant la caméra.

Entrée (Input)

Le double pixel étant basé sur les paires d'images; le réseau prend en entrée deux images qui passent par le même extracteur commun de caractéristiques (features extractors) et la sortie de l'extracteur correspond à 1/4 de la résolution originale de chaque image. Les deux vecteurs de caractéristiques sont en suite concaténés pour former un seul et unique vecteur puis traités par U-Net [17] et la sortie de ce réseau correspond à 1/8 la résolution d'entrée. Pour que notre réseau puisse traiter sans problème les images de différentes tailles nous avons évité d'utiliser les composant fortement connectés.

Sortie (Output)

A la sortie du réseau, nous avons un depth map prédicté correspondant à la paire d'images du dual caméra. Ce résultat sera enfin fourni au réseau de classification de depth pour savoir si le depth vient du vrai visage ou d'une attaque.

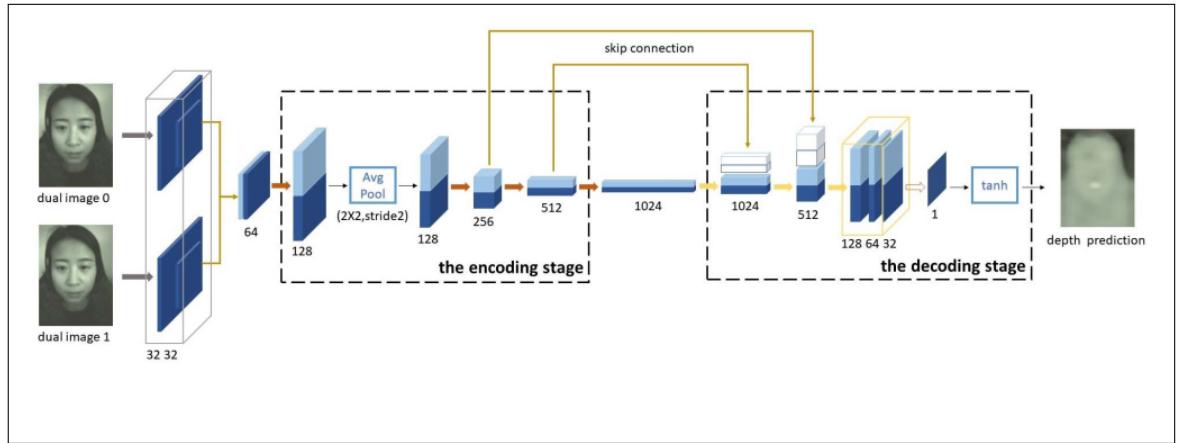


FIGURE 5.2 – Réconstruction du depth map [23]

5.4 Entrainement de depth map par pair

L'écart entre une paire d'images issue de la caméra à double pixel est tellement petit et cela devient difficile d'obtenir la vérité terrain du depth map. Pour y arriver nous avons fait recourt aux deux procédés de supervision à savoir; la cohérence de la transformation et l'étiquetage relative de depth map.

5.4.1 La cohérence de la transformation

Le décalage des pixels entre les deux images est mesuré par leur disparité. Pratiquement si nous déplaçons une image dans la direction de la disparité la valeur de la disparité augmente ou diminue d'une manière ou d'une autre.

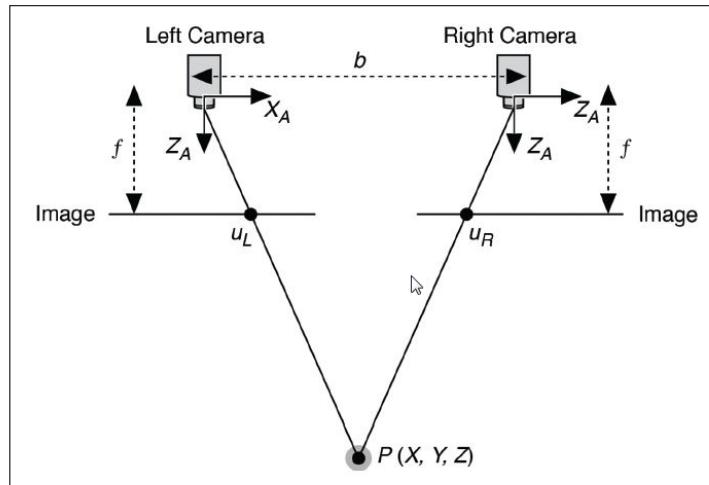


FIGURE 5.3 – Modèle vision stéréo [37]

La figure 5.3, montre le principe de base d'une vision stéréo. Dans ce modèle, la disparité est définie comme la distance entre les points projetés sur l'image plane;

$$u_L - u_R$$

Signalons que le depth et la disparité sont inversement proportionnel et au fur et en mesure que le depth augmente, la disparité augmente aussi exponentiellement [3]. Pour renforcer cette relation nous utilisons une fonction de perte de cohérence où;

$$p = (I_0, I_1)$$

I_0, I_1 représentent la paire d'images double pixel. Nous mettons $I_0^* = I_0$ et déplaçons I_1 par n pixels dans la direction de la disparité pour obtenir I_1^* . Ainsi, la fonction de disparité doit satisfaire les critères suivants :

$$d(x, y) = d^*(x, y) + n \quad (5.1)$$

où x et y représentent les coordonnées d'un point de l'image et d la disparité de la paire p . d^* représente la disparité quand l'image change de position et devient I_0^* et I_1^* . n est une valeur aléatoire comprise entre -16 et 16. Ainsi la perte de la transformation de la cohérence est exprimé comme suit;

$$L_{t1}(p, p^*, n) = L_t((I_0, I_1), n)$$

$$= \frac{1}{hw} \sum_{x=1}^w \sum_{y=1}^h (|d(x, y)| - |d^*(x, y)| - |n|)^2 \quad (5.2)$$

où h, w représentent respectivement la hauteur et la largeur de l'image. Ceci peut être ainsi noté en pratique;

$$\nabla d(x, y) = \nabla d^*(x, y) \quad (5.3)$$

où ∇ représente le gradient au long des axes x et y . En tenant compte de la perte, ceci correspond au formule suivant basé sur le gradient;

$$L_{t2} = (p, p^*, n) = L_t((I_0, I_1), (I_0^*, I_1^*), n)$$

$$= \frac{1}{hw} \sum_{x=1}^w \sum_{y=1}^h (|\nabla d(x, y)| - |\nabla d^*(x, y)|)^2 \quad (5.4)$$

Les deux termes de perte sont pondérés et combinés pour construire la fonction de perte de transformation de cohérence L_t ;

$$L_t = L_{t1} + \lambda L_{t2} \quad (5.5)$$

où λ représente l'hyper paramètre utilisé pour équilibrer les deux termes de perte.

5.4.2 L'étiquetage relative du depth

L'information contenue dans le depth map est très cruciale pour la méthode proposée. En se basant sur ce travail [21] nous effectuons une sélection aléatoire de points dispersés par paire et étiquetons celui qui apparaît en premier. Contrairement à la perte de cohérence, pour la perte relative nous utilisons un ensemble étiqueté de points comme information de supervision. Relative pour la paire d'image d'entraînement p et ses K paire $R = (i_k, j_k, r_k)$, $k = 1, \dots, K$, où i_k représente l'emplacement du premier point dans la paire avec l'index k , j_k représente l'emplacement du deuxième point dans la paire avec l'index k , et $r_k \in \{+1, -1, 0\}$ la relation entre i_k et j_k représente la vérité terrain où

- (+1) plus proche
- (-1) plus loin
- (0) égal

Ici, d représente la disparité du map prédict et d_{ik}, d_{jk} représentent les disparités au point i_k, j_k . Ainsi la fonction de perte de disparité est comme suit;

$$L_p(p, R, d) = \sum_{k=1}^K \psi(p, i_k, j_k, r, d), \quad (5.6)$$

où $\psi_k(I, i_k, j_k, d)$ représente la perte de k -inième paire

$$\psi(p, i_k, j_k, r, d) = \begin{cases} \log(1 + \exp(d_{ik} - d_{jk})), & r_k = +1 \\ \log(1 + \exp(d_{jk} - d_{ik})), & r_k = -1 \\ (d_{ik} - d_{jk})^2, & r_k = 0 \end{cases} \quad (5.7)$$

5.4.3 Fonction de perte

La fonction de perte pour entraîner le réseau de reconstruction du depth map correspond au poids moyen de la perte de transformation de cohérence et de la perte relative.

$$L = L_p + \alpha L_t \quad (5.8)$$

où α met l'équilibre entre les deux pertes.

5.5 Classification de Depth

Détection d'attaque de visage est un problème de classification binaire ; où le résultat attendu peut soit être 0 = fake face (dans le cas d'une attaque) ou 1 = genuine (s'il s'agit d'un visage réel). La classification du depth s'effectue en trois grandes étapes à savoir;

- La prédiction du depth map effectuée via le réseau de reconstruction du depth map ;

- Ensuite le détecteur du visage (face detector) est utilisé pour obtenir les coordonnées landmark dans l'image correspondante;
- De coordonnées landmark trouvées, la région faciale est coupée (face cropped) du depth map (obtenu à l'étape 1)
- Enfin le réseau de classification depth est appelé pour prédire la face

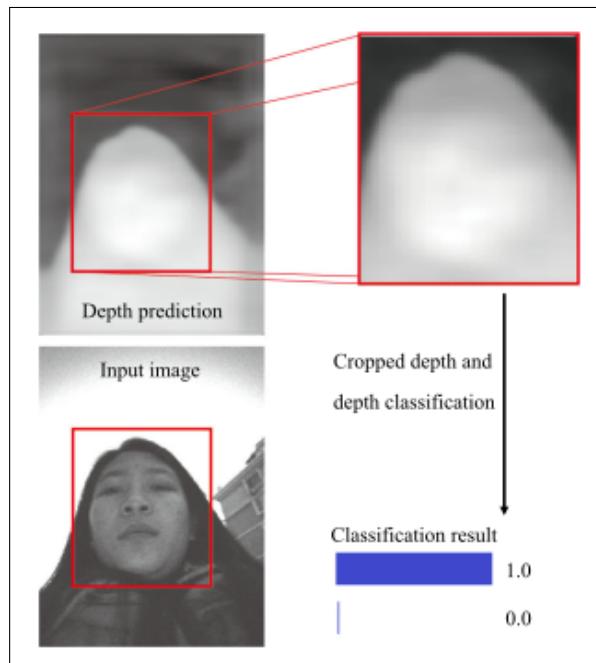


FIGURE 5.4 – Classification de depth map

5.6 Conclusion

La solution que nous avons proposée pour le problème détection d'attaques de visage est basée sur la génération du depth map avec les images issues d'une caméra à double pixels. À l'entrée du réseau une paire d'images est fournie pour en prédire son depth map à la sortie. La classification est effectuée par le modèle Xception network.

Implémentation et analyse des résultats

6.1 Introduction

Ce sixième et dernier chapitre se propose de présenter les outils et techniques auxquels nous avons fait recourt pour la mise en place de la méthode proposée pour le problème de détection d'attaques des visages présentée dans le chapitre précédent. Les résultats obtenus y sont aussi présentés sous une forme analytique enfin de bien comprendre la méthode pour envisager des éventuels améliorations dans le cas échéant.

6.2 Base de données (dataset)

Signalons que, le problème de détection d'attaques des visages diffère grandement de la majorité de problèmes traités en apprentissage automatique plus particulièrement de celui de la reconnaissance faciale (face recognition) par le simple fait que, différencier deux images qui semblent être identiques à la première vue n'est pas chose facile et cela fait à ce que la disponibilité des bases de données fiables pour l'expérimentation devient difficile. A cela on peut ajouter que, l'acquisition d'une base de données pour la détection d'attaque basée sur le visage n'est pas chose simple du fait que les données de la classe attaque (spoof/fake) doivent provenir de la classe réel (real/-live) en les réimprimant et récapturant dans plusieurs vues. Après plusieurs essais avec certaines bases à accès public nous avons enfin opté utilisé cette base [23], elle est constituée de 10.045 images dual pixel rangées par paire dont (5.167 paires réelles et 4.878 paire attaques) pour l'entraînement et 1.546 images dual pixel dont (893 paires réelles et 653 paires attaques) pour le test. Le tableau 6.1 montre comment est réparti les données dans les deux classes à savoir real et fake classe. La classe attaque est constituée de deux types d'attaques (photo imprimée et vidéo) et la rapport de ces deux types d'attaques est approximativement de 1 :1. La figure 6.1 montre quelques échantillons de la base de données. A gauche on a la classe live/real qui représente la personne en direct devant la caméra, et à droite on a quelques échantillons de la classe

fake/spoof photo imprimée et vidéo pré-enregistrée.



FIGURE 6.1 – Échantillon de la base de données

Classes	Entraînement	Test	Total
real	5.167 paires	893 paires	6.060 paires
fake	4.878 paires	653 paires	5531 paires
Total	10.045 paires	1.546 paires	11.591 paires

TABLE 6.1 – Répartition de la base de données

6.3 Architecture réseau

Le réseau pour la reconstruction du depth map est constitué de deux blocs principaux à savoir; la phase d'encodage et celle de décodage. Les deux étant séparés par un saut de connexion (skip connection) voir figure 5.2.

6.3.1 Encodage

Comme le montre la figure 5.2, l'étape d'encodage prend en entrée une paire d'images de niveau gris issue d'une double caméra avec une résolution de 2016 x 756 pixels. Dans la première partie de l'encodage, deux couches de convolution partageant le même poids sont utilisées et les vecteurs de caractéristiques obtenus de ces deux couches sont concaténés. La deuxième partie de l'encodage est constituée de trois couches de convolution avec une couche de mise en commun (pooling layer). Au niveau de l'interconnexion entre les deux blocs, les vecteurs de caractéristiques sont extraits sur différents échelles, ceux ci seront plus tard utilisés au niveau du décodage.

6.3.2 Décodage

Lors de la phase du décodage, nous utilisons des échantillons de caractéristiques grâce à la méthode du plus proche voisin (Nearest Neighbor) puis nous les concaténons avec leur correspondant extraits à l'interconnexion figure 5.2. Après deux modules d'échantillonage et concaténation une couche de convolution est utilisée pour produire le map final de disparité avec une dimension de 252 x 96. La figure 6.2 illustre l'architecture du réseau de construction du depth map.

CHAPITRE 6. IMPLÉMENTATION ET ANALYSE DES RÉSULTATS

Layer (type:depth-idx)	Output Shape	Param #
Sequential: 1-1	[-1, 2048, 7, 7]	--
└ Conv2d: 2-1	[-1, 32, 111, 111]	864
└ BatchNorm2d: 2-2	[-1, 32, 111, 111]	64
└ ReLU: 2-3	[-1, 32, 111, 111]	--
└ Conv2d: 2-4	[-1, 64, 109, 109]	18,432
└ BatchNorm2d: 2-5	[-1, 64, 109, 109]	128
└ Block: 2-6	[-1, 128, 55, 55]	--
└ Sequential: 3-1	[-1, 128, 55, 55]	26,816
└ ReLU: 3-2	[-1, 128, 109, 109]	--
└ Conv2d: 3-3	[-1, 128, 55, 55]	8,192
└ BatchNorm2d: 3-4	[-1, 128, 55, 55]	256
└ Block: 2-7	[-1, 256, 28, 28]	--
└ Sequential: 3-5	[-1, 256, 28, 28]	102,784
└ ReLU: 3-6	[-1, 256, 55, 55]	--
└ Conv2d: 3-7	[-1, 256, 28, 28]	32,768
└ BatchNorm2d: 3-8	[-1, 256, 28, 28]	512
└ Block: 2-8	[-1, 728, 14, 14]	--
└ Sequential: 3-9	[-1, 728, 14, 14]	728,120
└ ReLU: 3-10	[-1, 728, 28, 28]	--
└ Conv2d: 3-11	[-1, 728, 14, 14]	186,368
└ BatchNorm2d: 3-12	[-1, 728, 14, 14]	1,456
└ Block: 2-9	[-1, 728, 14, 14]	--
└ Sequential: 3-13	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-14	[-1, 728, 14, 14]	--
└ ReLU: 3-15	[-1, 728, 14, 14]	--
└ Block: 2-10	[-1, 728, 14, 14]	--
└ Sequential: 3-16	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-17	[-1, 728, 14, 14]	--
└ ReLU: 3-18	[-1, 728, 14, 14]	--
└ Block: 2-11	[-1, 728, 14, 14]	--
└ Sequential: 3-19	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-20	[-1, 728, 14, 14]	--
└ ReLU: 3-21	[-1, 728, 14, 14]	--
└ Block: 2-12	[-1, 728, 14, 14]	--
└ Sequential: 3-22	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-23	[-1, 728, 14, 14]	--
└ ReLU: 3-24	[-1, 728, 14, 14]	--
└ Block: 2-13	[-1, 728, 14, 14]	--
└ Sequential: 3-25	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-26	[-1, 728, 14, 14]	--
└ ReLU: 3-27	[-1, 728, 14, 14]	--
└ Block: 2-14	[-1, 728, 14, 14]	--
└ Sequential: 3-28	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-29	[-1, 728, 14, 14]	--
└ ReLU: 3-30	[-1, 728, 14, 14]	--
└ Block: 2-15	[-1, 728, 14, 14]	--
└ Sequential: 3-31	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-32	[-1, 728, 14, 14]	--
└ ReLU: 3-33	[-1, 728, 14, 14]	--
└ Block: 2-16	[-1, 728, 14, 14]	--
└ Sequential: 3-34	[-1, 728, 14, 14]	1,613,976
└ ReLU: 3-35	[-1, 728, 14, 14]	--
└ ReLU: 3-36	[-1, 728, 14, 14]	--
└ Block: 2-17	[-1, 1024, 7, 7]	--
└ Sequential: 3-37	[-1, 1024, 7, 7]	1,292,064
└ ReLU: 3-38	[-1, 728, 14, 14]	--
└ Conv2d: 3-39	[-1, 1024, 7, 7]	745,472
└ BatchNorm2d: 3-40	[-1, 1024, 7, 7]	2,048
└ SeparableConv2d: 2-18	[-1, 1536, 7, 7]	--
└ Conv2d: 3-41	[-1, 1024, 7, 7]	9,216
└ Conv2d: 3-42	[-1, 1536, 7, 7]	1,572,864
└ BatchNorm2d: 2-19	[-1, 1536, 7, 7]	3,072
└ SeparableConv2d: 2-20	[-1, 2048, 7, 7]	--
└ Conv2d: 3-43	[-1, 1536, 7, 7]	13,824
└ Conv2d: 3-44	[-1, 2048, 7, 7]	3,145,728
└ BatchNorm2d: 2-21	[-1, 2048, 7, 7]	4,096
└ Conv2d: 1-2	[-1, 512, 7, 7]	1,049,088
└ ReLU: 1-3	[-1, 512, 7, 7]	--
└ Dropout2d: 1-4	[-1, 512, 7, 7]	--
└ AdaptiveAvgPool2d: 1-5	[-1, 512, 1, 1]	--
└ Linear: 1-6	[-1, 2]	1,026

Total params: 21,857,066
Trainable params: 21,857,066
Non-trainable params: 0
Total mult-adds (M): 708.51

=====
Input size (MB): 0.57
Forward/backward pass size (MB): 105.84
Params size (MB): 83.38
Estimated Total Size (MB): 189.79
=====

FIGURE 6.2 – Architecture du réseau proposé

6.4 La reconstruction du depth map

Le réseau décrit ci-haut est entraîné sur une base de portraits (dual images, paire annotée), laquelle contient de petits portraits de la classe réelle (real/live) avec une dense annotation, larges portraits avec une annotation dispersée, et larges portraits de la classe attaque (spoof/fake) avec une dense annotation aussi.

Comme le montre la figure 5.2, l'entrée du réseau correspond à une paire d'images issue de la caméra à double pixel et la sortie correspond à la disparité de l'image de gauche. Le réseau est initialisé aléatoirement et le taux d'apprentissage (learning rate) commence à partir de 0.001. L'optimiseur Adam est utilisé pour entraîner le réseau avec approximativement 200.000 pas(steps) / epochs. Dans la transformation de la perte de cohérence, le paramètre de transformation correspond à une valeur aléatoire comprise entre -16 et +16. La fonction `tf.contrib.image.transform`¹ de Tensorflow est utilisée pour produire l'image transformée, en suite l'image originale et la transformée passent par la propagation directe et à la sortie du réseau on a deux maps de disparité et la fonction de perte de cohérence est calculée en utilisant ces deux maps.

6.5 Classification du depth map

6.5.1 Démarche

La résolution de l'image d'entrée du réseau de classification est de 256 x 256 pixels comme le montre la figure 6.3. Pour découper la région du visage à partir du depth map original, 68 points de repères sont détectés. La moyenne de coordonnées et les valeurs standards de ces points de repères sont calculées et la fonction OpenCV `getAffineTransform`² est utilisée pour calculer la matrice affine à partir du depth map au canvas.

1. http://tensorflow.biotecan.com/python/Python_1.8/tensorflow.google.cn/api_docs/python/tf/contrib/image/transform.html

2. https://docs.opencv.org/3.4/d4/d61/tutorial_warp_affine.html

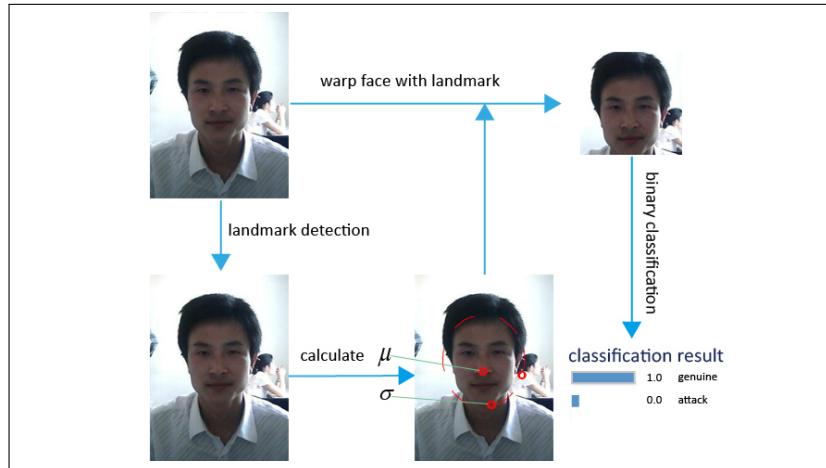


FIGURE 6.3 – Classification du depth map

Le point central représente les coordonnées moyennes, et ceux d'en bas et de droit (figure 6.3) sont calculés en utilisant les valeurs standards. Pour la classification, nous avons utilisé le réseau Xception[7] avec l'emploi du module Squeeze-and-Excitation[13] comme le montre la figure 6.4. Le poids initial est entraîné sur la base de données ImageNet et nous augmentons aléatoirement l'échantillon d'entraînement et enfin converti ce dernier en entraînement standard de classification d'images.

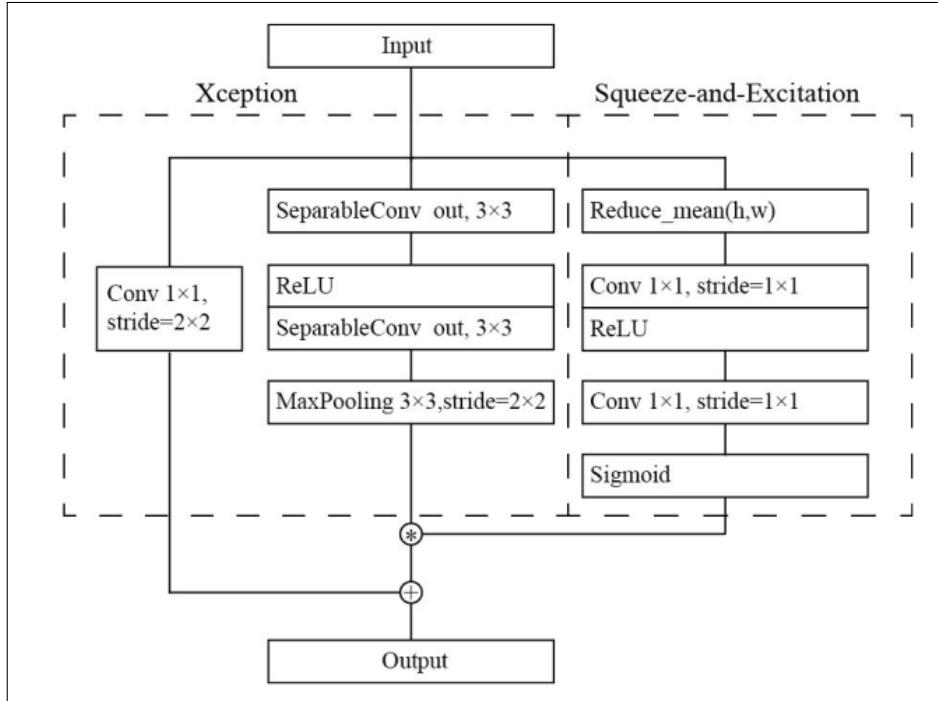


FIGURE 6.4 – Architecture Xception pour la classification [7]

6.5.2 Inférence

A part le test que nous avons effectué lors de l'entraînement du modèle montrant la performance de ce dernier à pouvoir distinguer un visage réel du faux, lors de l'inférence nous avons aussi pu constater que les différents types d'attaques sont bien détectés aussi, comme on peut le voir sur la figure 6.5, et 6.6.

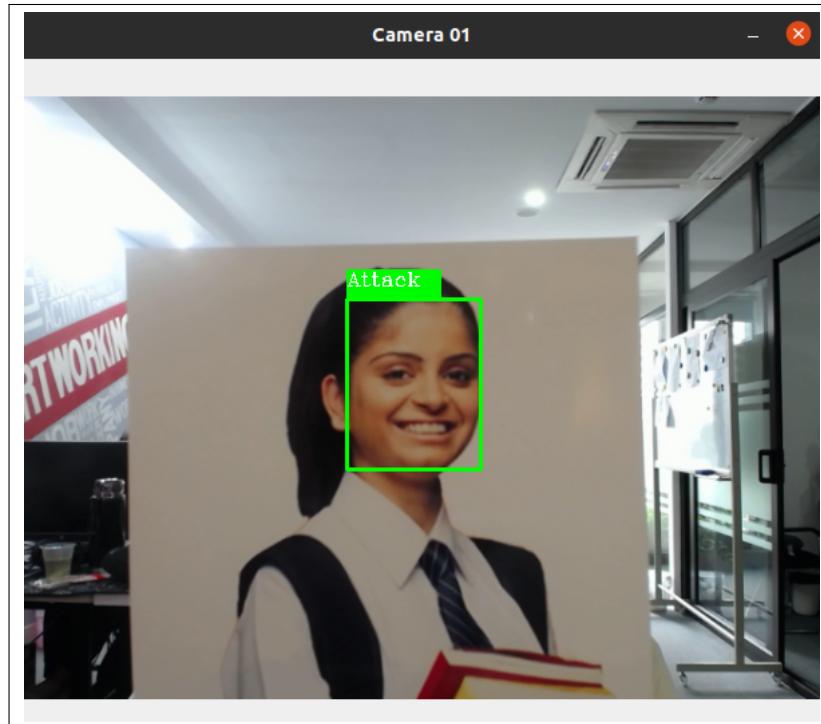


FIGURE 6.5 – Présentation d'attaque - portrait

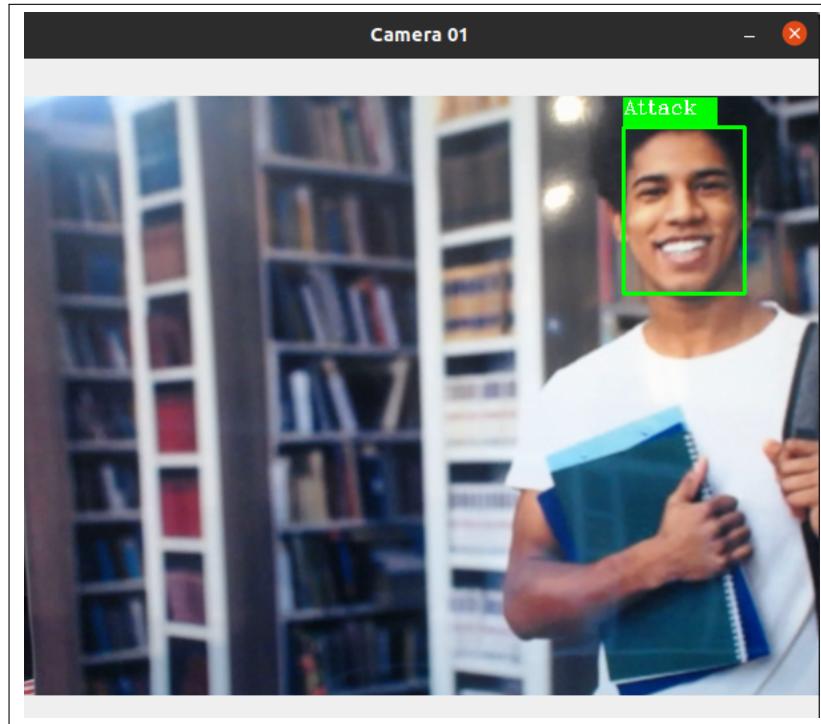


FIGURE 6.6 – Présentation d'attaque - paysage

L'attaque par affichage sur écran est celle la plus rependue allant d'écrans ordinaires jusqu'aux écrans à haute définition, image et vidéo comprises. Comme le montre la figure 6.7, de plusieurs vue que nous avons braqué l'écran affichant un visage ou jouant une vidéo devant la caméra, ceci a été détecté comme une attaque. La figure 6.8 nous montre un exemple d'un visage en direct qui est aussi détecté comme réel.

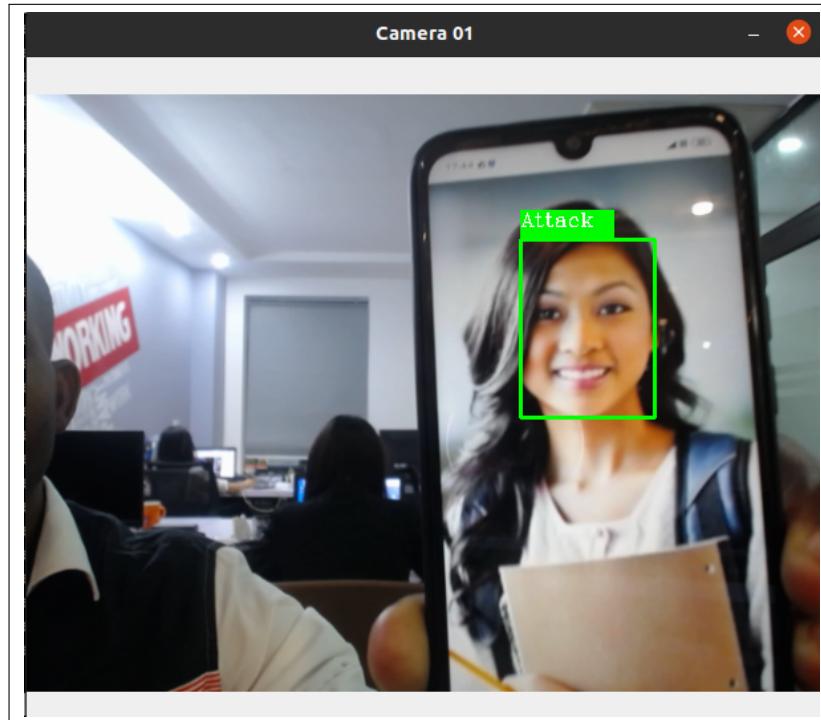


FIGURE 6.7 – Présentation d’attaque - affichage sur écran

Nous avons aussi testé au même moment un visage en direct et une attaque basé sur l'affichage d'écran figure 6.8 et les deux on été respectivement détectés comme vrai et faux. La figure 6.9 montre un visage en direct de la caméra et qui est détecté vrai. La figure 6.10 montre une attaque basé sur une photo imprimée contenant plusieurs visage en salle de classe et tous les visage détectés sont belle et bien reconnus comme faux.

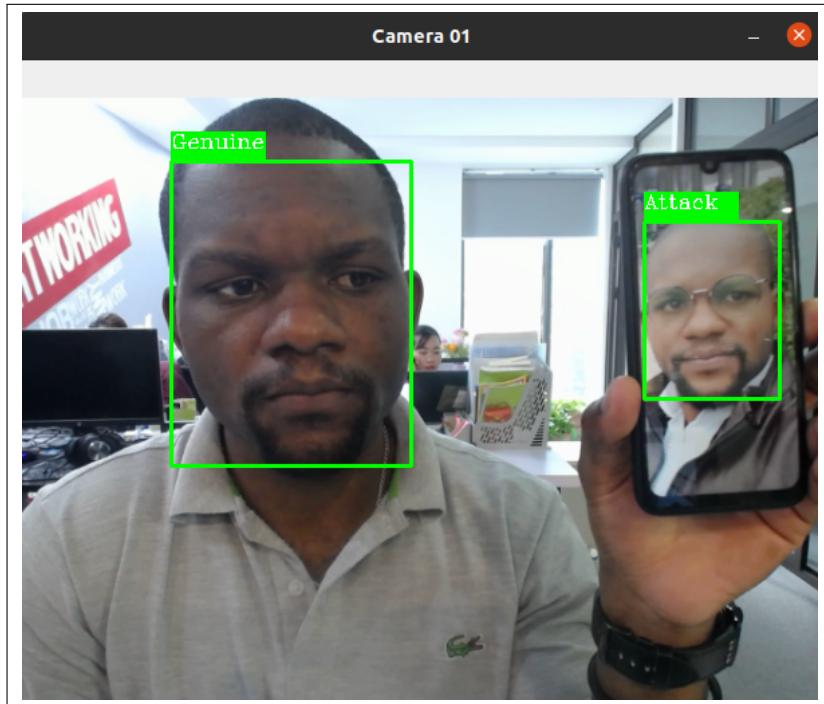


FIGURE 6.8 – Présentation simultanée

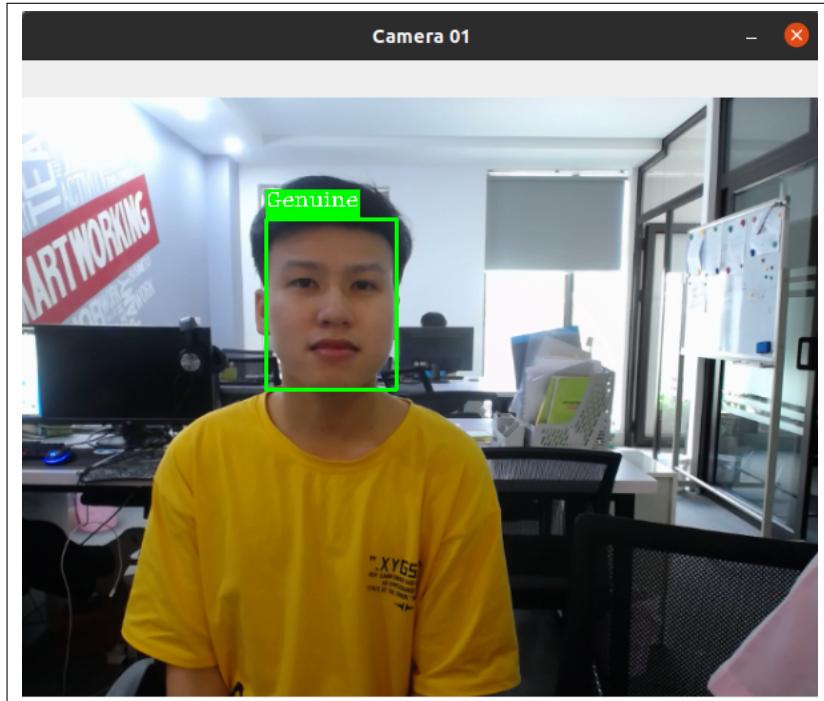


FIGURE 6.9 – Présentation d'attaque - vrai visage

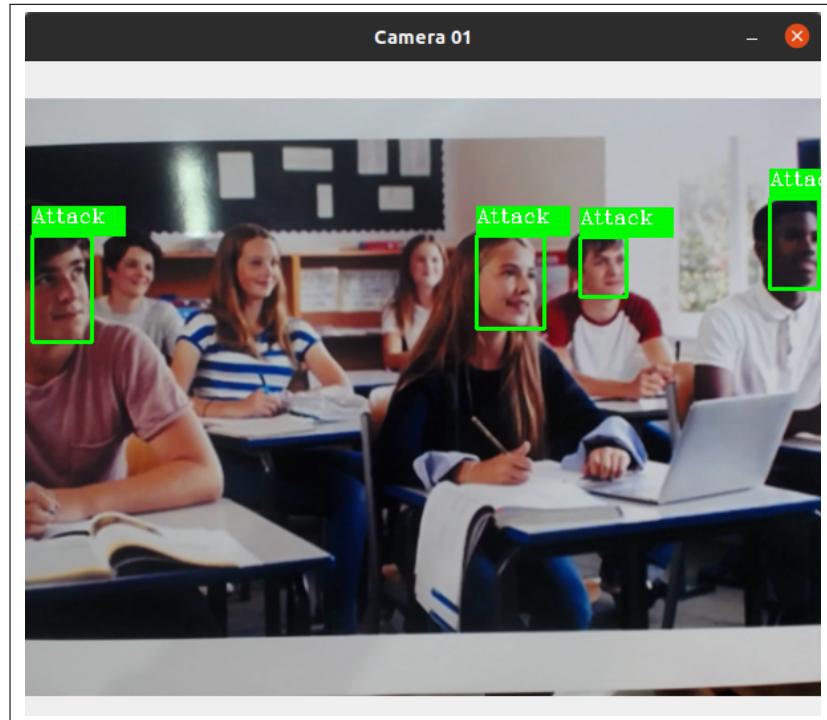


FIGURE 6.10 – Présentation d’attaque - photo

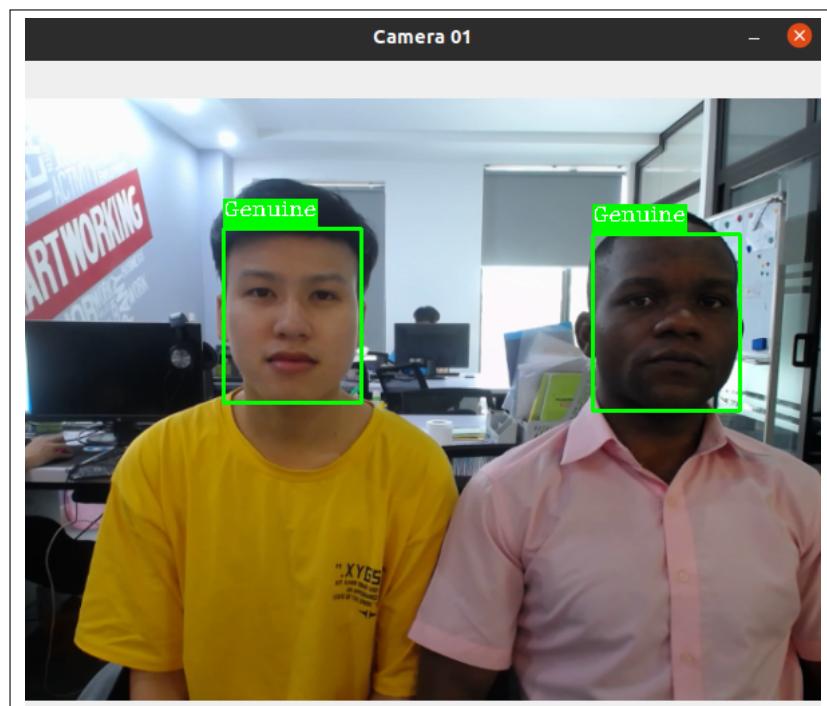


FIGURE 6.11 – Présentation d’attaque - vrai

CHAPITRE 6. IMPLÉMENTATION ET ANALYSE DES RÉSULTATS

A ce niveau, le modèle pour la détection d'émotions mis en place l'entreprise VDSmart était déjà fonctionnel au sein du système Eye Pro Education et pour notre part il fallait mettre place un module basé sur les technologies Web moderne pour visualiser les fréquences des émotions détectés comme le montre la figure 6.12 les quatre émotions choisies sont celles les plus fréquentes dans une salle de classe et qui sont affichées sur un moniteur de contrôle et leur fréquence est mesurée et affichée aussi comme on peut le voir sur la figure 6.13.



FIGURE 6.12 – Reconnaissance d’émotion

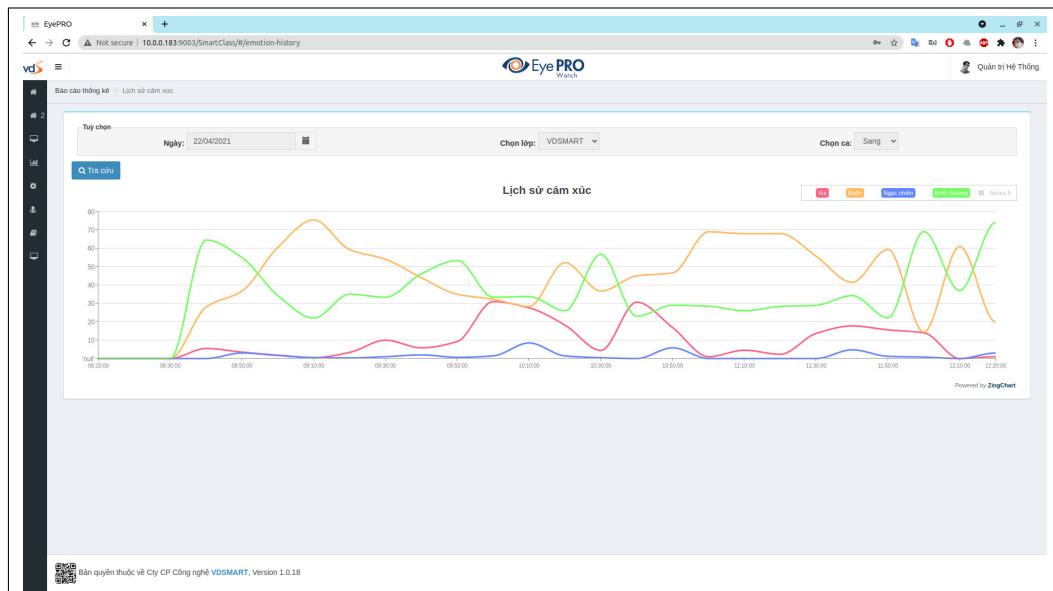


FIGURE 6.13 – Fréquence de dominance d’émotions

6.6 Analyse du depth map généré

Le coeur de tout le travail étant basé sur la génération du depth map pour parvenir à différencier un visage réel du faux, nous avons aussi tenté de visualiser quelques résultats obtenus par d'autres méthodes, comme le montre la figure 6.14 pour les deux classes respectives. La première colonne (a) de chaque classe représente l'entrée images double pixels pour chaque méthode. La deuxième colonne (b) correspond à la méthode MCCNN[14] qui produit un depth map trop lisse dû à l'étape d'agrégation, elle produit un taux élevé de bruit avec une faible disparité, il est difficile pour cette méthode d'établir l'échange entre la robustesse et la préservation de détails[23]. La méthode CVF[15] troisième colonne (c) et POC[31] colonne (d) fournissent une depth map plus net; Mais dans l'entre temps, leur résistance contre le bruit ou de régions sans texture n'est pas satisfaisante. En revanche la méthode proposée colonne (e) produit le résultat convaincant pour tous les échantillons de deux classes.

Méthode	AUC	EER	TPR(1.0)	TPR(0.5)
CVF [15]	0.975699	7.5000%	0.6867	0.5911
MCCNN [14]	0.975689	9.1825%	0.7504	0.4962
POC [31]	0.984355	6.6667 %	0.7657	0.5038
Proposée	0.999690	0.7839 %	0.9939	0.9923

TABLE 6.2 – Comparaison de différentes méthodes de génération de depth map

Dans certains cas particuliers, le taux de vrais positifs (TPR) est un clé indicateur en pratique. Dans le tableau 6.2 (données d'évaluation) la qualité du depth map de dual pixel caméra aide à la classification du depth.

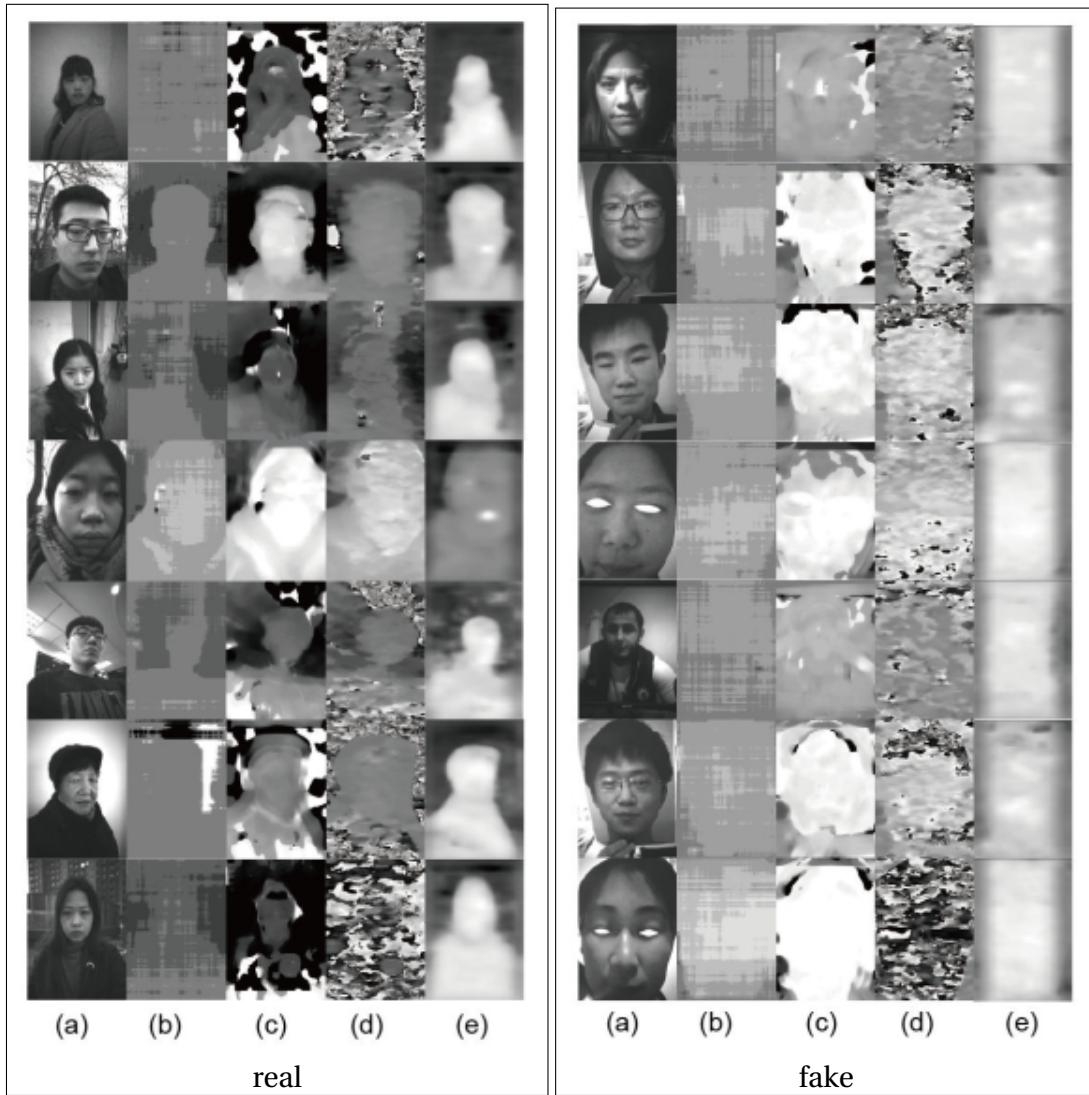


FIGURE 6.14 – La disparité de depth map par différentes méthodes

6.7 Performance de la méthode de classification

6.7.1 Précision de la classification

Comment l'auteur ici [36] l'estime, un système de recherche documentaire parfait fournira des réponses dont la précision et le rappel sont égaux à 1, cela veut dire que l'algorithme trouve la totalité des documents pertinents -rappel et ne fait aucune erreur -précision. Sur la figure 6.18 du rapport de la classification de notre modèle, la valeur de la précision et du rappel pour les deux classes sont aussi très élevées pour dire que il est capable de distinguer les deux classe jusqu'à une précision prêt.

6.7.2 Entrainement et Validation

Pour visualiser l'évolution de la précision et de la perte lors de la phase d'entraînement et de validation, nous nous sommes servi de l'outil TensorBoard³ comme on peut le voir sur la figure 6.15. La figure ci-dessous montre le résultat que nous avons obtenu pour les 10 epochs sur lesquels nous avons lancé notre modèle. L'évolution des courbes laisse dire de la stabilité et de l'habileté que sera ce modèle. Pour la précision nous avons obtenu 0.9909 de le premier epoch et 0.9997 au dixième lors de la phase d'entraînement tandis que lors de la validation de le premier epoch nous avons obtenu 0.9088 et 0.9861 au dixième. Pour la perte d'erreur de le premier epoch on avons obtenu 2.9664e-5, 0.01525 et 1.9079e-5, 0.0531 au dixième pour les deux phases respectives (entraînement et validation).

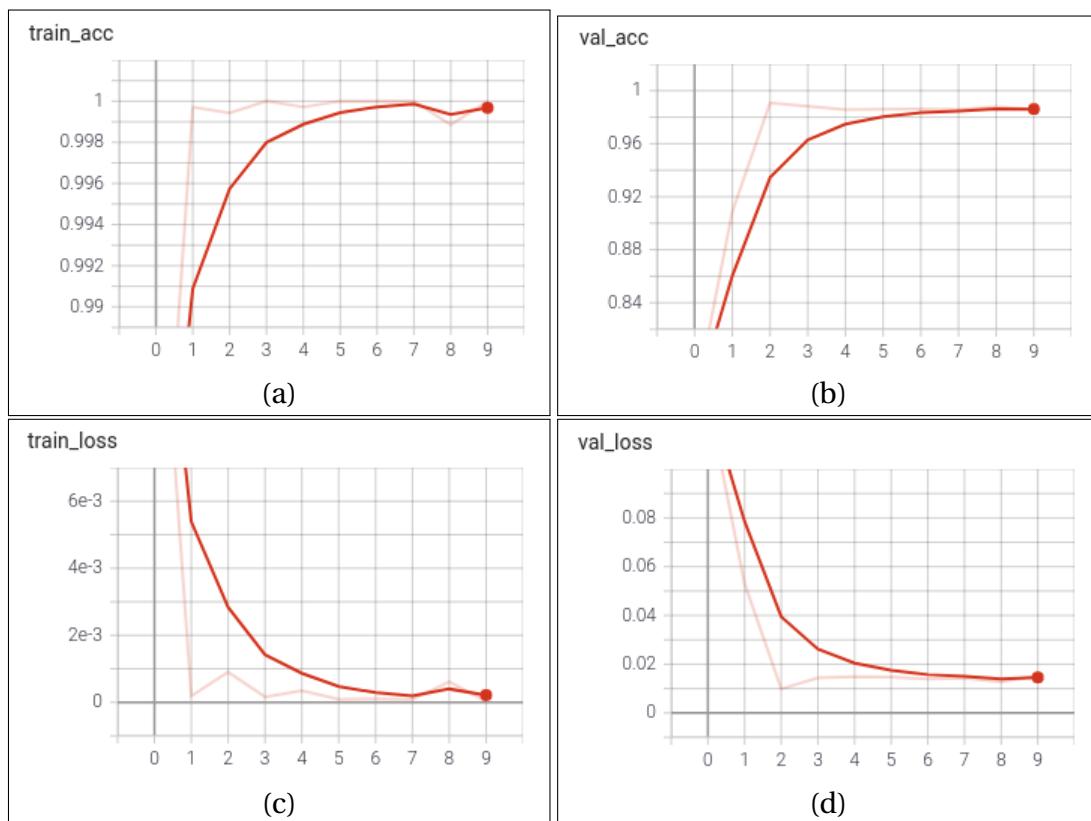


FIGURE 6.15 – Courbes de la précision et de la perte

6.7.3 AUC et la Courbe ROC

Notre problème étant basé sur une classification binaire avec les données reparties équitablement, nous avons choisi d'utiliser la courbe ROC (Receiver Operating Characteric) pour évaluer les performances de classification du modèle proposé. En

3. https://pytorch.org/tutorials/intermediate/tensorboard_tutorial.html

théorie, il est dit que ; un modèle habile est représenté par une courbe qui s'incline en haut à gauche du tracé de la courbe ROC, et le modèle présenté dans ce travail présente cette caractéristique comme on peut le voir sur la figure 6.16.

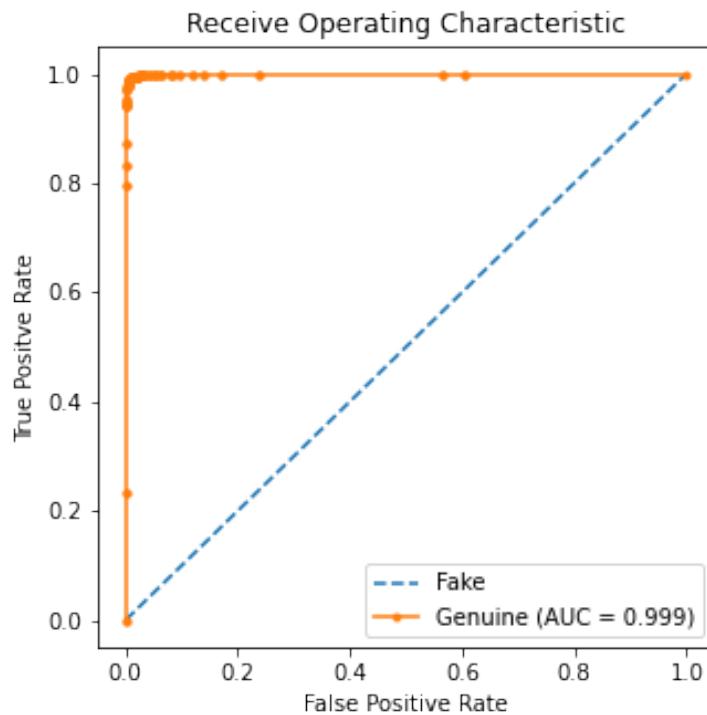


FIGURE 6.16 – Receiver Operating Characteric

L'aire en dessous de la courbe (AUC) du modèle est de 0.997, cela peut être confirmé par la valeur très élevée de vrais positifs prédits par le modèle. Sur l'axe des abscisses du graphe présenté sur la figure 6.16 la probabilité des faux positifs est très petite par rapport aux vrais négatifs ce qui est bon pour un modèle; même chose pour pour l'axe des ordonnées où nous avons la probabilité de vrais positifs est très grande et celle des faux négatifs est petite. Ajouter à ceci le résultat obtenu lors de l'inférence ça prouve comment le modèle peut bien classifier un vrai visage du faux.

6.7.4 Matrice de confusion

Un autre outils auquel nous avons fait recourt pour évaluer la performance de notre modèle est la matrice de confusion. Celle nous a permis aussi de vérifier aussi à quelle fréquence nos prédictions sont correctes par rapport à la réalité pour le problème de détection d'attaque de visage. L'organisation de la matrice de confusion prévoit que, plus le diagonale contient les valeurs élevés mieux est le modèle. Sur la figure 6.17 on peut voir que, la classe attaque (fake) est prédite avec une précision prêt comme le

montre la matrice normalisé. Mais en général le modèle distingue bien les éléments de la classe Genuine et la classe Fake.

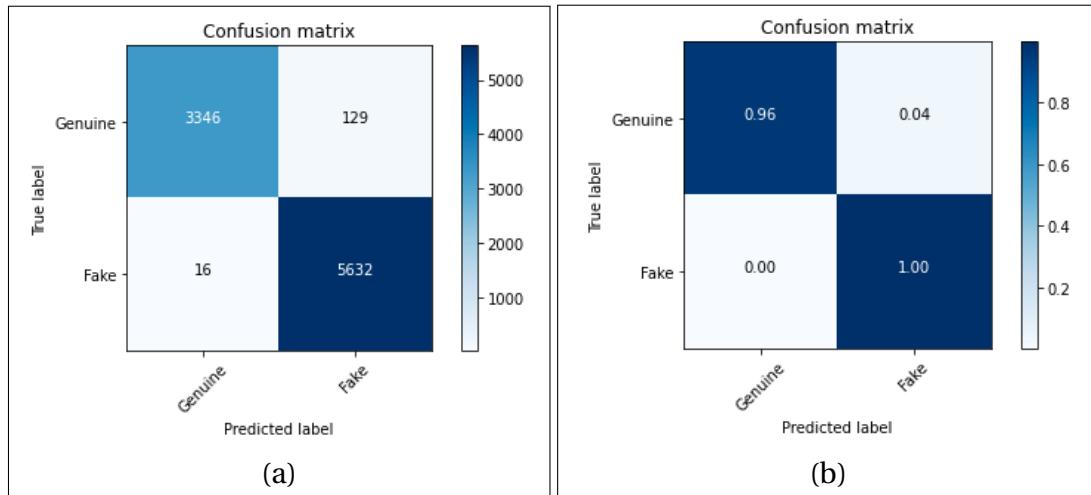


FIGURE 6.17 – Matrice de confusion

6.7.5 Rapport de classification

La précision obtenue dans le rapport de classification 6.18 montre quelle proportion d'identifications est effectivement correcte, d'où 0 représente la classe Genuine et 1 Fake (voir matrice de confusion 6.17).

	precision	recall	f1-score	support
0	0.96	1.00	0.98	3362
1	1.00	0.98	0.99	5761
accuracy			0.98	9123
macro avg	0.98	0.99	0.98	9123
weighted avg	0.98	0.98	0.98	9123

FIGURE 6.18 – Rapport de classification

Le rappel (recall) obtenu dans le rapport de classification, montre toutes les instances positives identifier par le classifier. Pour chaque classe on peut voir que le rapport est à une précision prêt.

6.8 Conclusion

Au terme de ce chapitre, nous pouvons confirmer que la mission qui nous a été confié est belle et bien accomplie voir l'analyse des résultats présentés ci-haut. Le modèle entraîné sur cette base [23] que nous proposons pour distinguer un vrai visage du faux, prédit un depth map de l'image donnée à partir duquel le visage est localisé et coupé enfin d'être fourni au réseau de classification pré-entraîné pour savoir si c'est une attaque qui est présenté ou non.

Conclusion générale

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre de notre projet de fin d'étude de Master de recherche en Systèmes intelligent et Multimédia intitulé "Emotion recognition and face anti-spoofing in face recognition for smart education system". Ce travail avait pour ambition une mise en place d'un modèle d'apprentissage automatique pour la détection d'attaques basée sur le visage à la contre des systèmes d'informations utilisant la reconnaissance faciale pour l'authentification des usagés. La visée de ce travail était donc de s'assurer que toute donnée représentant l'image du visage vient d'une personne en direct de la caméra et non d'une photo pré-imprimée ou un rejoué d'une vidéo à partir d'un équipement électronique.

Après une conception initiale d'un système d'informations intelligent pour le système éducatif, comprenant la gestion des fréquentations, l'assiduité des étudiants et autres fonctions. Une étude a prouvé qu'une image d'une personne non présente en classe ou surplace peut être reconnue et interprétée comme cette personne et cela pourrait entraîner de conséquences graves à tout le système. A cet effet la phase de la détection d'attaque de visage s'avérait très importante.

Il a fallu dans un premier temps avoir une compréhension très profonde sur la différence entre une image en direct et une image pré-imprimée (attaque) présentée devant la caméra, pour cela nous sommes passés par la lecture de quelques travaux déjà réalisés et qui ont obtenu de résultats intéressants [21, 17, 16, 12, 11, 22, 7, 13, 14]. Ainsi c'est à travers cette étude itérative qu'on a pu approcher une solution optimale d'une mise en place d'un modèle pouvant détecter une attaque vis à vis d'un système de RF. La méthode proposée dans ce travail étant basée sur l'estimation du depth map d'une image, cela nous a demandé l'utilisation de la caméra à double pixel pour l'obtention d'images organisées en paire, et d'autres techniques comme la génération de depth map, la cohérence de la transformation, l'étiquetage relative de depth map et enfin la fonction de perte.

A cet effet, nous conseillons d'effectuer une expérimentation de notre méthode en adoptant une autre manière d'utiliser le depth map, à titre d'exemple une fusion du depth map avec un autre type de caractéristique comme celles issues de CNN ou LBP.

CHAPITRE 7. CONCLUSION GÉNÉRALE

Cependant, il ne faut pas négliger l'aspect de la nature ou la source de la base d'expérimentation qui est un facteur majeur dans un tel cas d'étude et qui malheureusement n'est pas très disponible dans nos le secteur.

Bibliographie

- [1] ACNUSA. Acnusa, october 2019. <https://www.acnusa.fr/fr/la-pollution-de-lair/sources/18>.
- [2] M. Barbu. *La belle thèse*. PhD thesis, Université, 2002.
- [3] C. W. T. Blog. Stereo vision basics, March 2014. <http://chriswalkertechblog.blogspot.com/2014/03/stereo-vision-basics.html>.
- [4] J. Brownlee. A gentle introduction to deep learning for face recognition, March 2020. <https://machinelearningmastery.com/introduction-to-deep-learning-for-face-recognition/>.
- [5] J. Brownlee. A gentle introduction to deep learning for face recognition, March 2020. <https://machinelearningmastery.com/introduction-to-deep-learning-for-face-recognition/>.
- [6] CBC. The nature of thing, March 2020. https://www.cbc.ca/natureofthings/m_features/the-seven-universal-emotions-we-wear-on-our-face.
- [7] F. Chollet. Xception : Deep learning with depthwise separable convolutions. *Google, Inc.*, April 2017.
- [8] F. Dupont. *Les choux farcis*. Un gros éditeur, 2004.
- [9] N. Dupont. *Réparer son vaisseau*. L'Alliance, 2009.
- [10] D. Y. et al. An emotion recognition model based on facial recognition in virtual learning environment. *ScienceDirect*, October 2018.
- [11] G. W. et al. Multi-modal face presentation attack detection via spatial and channel attention. *ResearchGate*, June 2019.
- [12] J. G. et al. Improving face anti-spoofing by 3d virtual synthesis. *arXiv*, page 1, April 2019.
- [13] J. H. et al. Squeeze-and-excitation networks. *cs.CV*, May 2019.
- [14] J. Z. et al. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, May 2016.

- [15] M. G. et al. Fast cost-volume filtering for visual correspondence and beyond. *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, June 2011.
- [16] M. N. et al. Face anti-spoofing with joint spoofing medium detection and eye blinking analysis. *ResearchGate*, page 1, June 2019.
- [17] O. R. et al. U-net convolutional networks for biomedical image segmentation. *Internal Conferene on Medical Image Computing and Computer Assisted Intervention*, page 1, May 2015.
- [18] P. T. et al. Emotion recognition using facial expressions. *Procedia computer science*, May 2017.
- [19] P. U. et al. A study on facial expression recognition in assessing teaching skills : Datasets and methods. *Procedia computer science*, May 2019.
- [20] S. C. et al. An overview of face liveness detection. *IJIT*, October 2014.
- [21] X. S. et al. Dual camera based feature for face spoofing detection. *Institute of Auttomation Chinese Academique of Science*, pages 1,2,3, september 2016.
- [22] X. S. et al. Dual camera based feature for face spoofing detection. *ResearchGate*, June 2016.
- [23] X. W. et al. Single-shot face anti-spoofing for dual pixel camera. *IEEE*, page 7, march 2020.
- [24] Y. L. et al. Face anti-spoofing using patch and depth-based cnn. *ResearchGate*, October 2017.
- [25] K. Feng and T. Chaspari. A review of generalizable transfer learning in automatic emotion recognition, February 2020. <https://www.frontiersin.org/articles/10.3389/fcomp.2020.00009/full>.
- [26] F. Ghaffar. Facial emotions recognition using convolutional neural net. *Institute of Informatique Science, Taiwan*, June 2018.
- [27] A. INNOVATION. 7 applications of computer vision, August 2020. <https://www.atriainnovation.com/en/7-applications-of-computer-vision/>.
- [28] S. Li and W. Deng. Deep facial expression recognition : A survey. *IEEE*, October 2018.
- [29] S. Maksymenko. Anti-spoofing techniques in face recognition, February 2021. <https://mobidev.biz/blog/face-anti-spoofing-prevent-fake-biometric-detection>.
- [30] M. Mauvais. Mon roman inachevé. il est chouette mon roman, feb 2000.
- [31] M. A. Muquit and T. Shibahara. A high-accuracy 3d measurement system using phase-based image matching. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, March 2006.
- [32] D. D. Sarkar. A comprehensive hands-on guide to transfer learning, March 2020. <https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a>.

- [33] P. N. Sra. A wide ranging view of face emotion recognition system. *International Journal of Control and Automation*, May 2020.
- [34] E. S. Team. What is machine learning? a definition, May 2020. <https://www.expert.ai/blog/machine-learning-definition/>.
- [35] VDSmart. Vdsmart, December 2020. <https://vdsmart.vn>.
- [36] Wikipedia. Précision et rappel, April 2021. [https://fr.wikipedia.org/wiki/Pr](https://fr.wikipedia.org/wiki/Précision_(apprentissage_automatique))
- [37] Zone.ni.com. Parts of a stereo vision system, December 2020. https://zone.ni.com/reference/en-XX/help/372916T-01/nivisionconcepts/stereo_parts_of_a_stereo_vision_system/.