

MASTER I - SYSTEMES INTELLIGENTS ET MULTIMEDIA (SIM)

RAPPORT TRAVAIL PERSONNEL ENCADRE
TPE

Surveillance d'examen à l'aide de la vision par ordinateurs et d'apprentissage

Étudiant :
KANA NGUIMFACK KÉVIN

Encadreur
Pr. HO TUONG VINH

Table des matières

1 Résumé	6
2 Introduction	6
2.1 Contexte technologique	6
2.2 Motivation	7
2.3 Objectifs	7
2.4 Résultats attendus	8
2.5 Organisation du travail	8
3 Concepts théoriques et État de l'art	8
3.1 L'apprentissage automatique	9
3.1.1 Le Machine Learning	9
3.1.2 Le Deep Learning	10
3.2 Vision par ordinateur	11
3.3 État de l'art	11
3.3.1 Application basée sur la vision par ordinateur pour la Surveillance des comportements des étudiants en classe	11
3.3.2 La prise de présence basée sur la vidéosurveillance, cas d'étude le collège polytechnique FPT.	12
3.3.3 Apprentissage non supervisé de la structure et de la dynamique des objets à partir de vidéos	13
3.3.4 Prévision du regard dans des vidéos immersives à 360 dégré dynamiques	13
3.3.5 Modélisation conjointe de la pose et de l'expression pour la reconnaissance de l'expression faciale	13
3.3.6 Estimation des poses et actions humaines	14
3.3.7 Reconnaissance de visages	14
3.3.8 Cheating Video Description Based on Sequences of Gestures	15
3.3.9 Résume de la revue littéraire	15
4 Solution proposée et implementation	16
4.1 Spécifications de la solution	16
4.1.1 Spécifications des modèles	16
4.2 Présentation des données	17
4.2.1 Reconnaissance faciale	17
4.2.2 Estimation des poses de la tête	18
4.2.3 Reconnaissancce des expressions faciales (émotions)	19
4.2.4 Reconnaissance des fraudes à l'examen	20
4.3 Implémentéation de la solution	21
4.3.1 Module de reconnaissance faciale	22
4.3.2 Module estimation des poses de la tête	25
4.3.3 Module reconnaissance des expressions faciales	25
4.3.4 Module reconnaissance d'activités de fraudes	26
4.4 Evalution du modèle	27
5 Expérimentation et résultats	28
5.1 Module de reconnaissance faciale	28
5.2 Modèle de pose de la tête (estimation du regard)	29
5.3 Modèle de reconnaissance des expressions faciales	29
5.4 Modèle de reconnaissance de scénariions de tricheries	29

6 Conclusion	32
7 Perspectives	32
8 Références	33

Liste des tableaux

1	Résumé des revues littéraires	16
2	Répartition de la base données pour la reconnaissance faciale	17
3	Détection pose de la tête	18
4	Identification des expressions faciales et leurs labels	19
5	Niveau de difficulté associé aux modules	27
6	Précision module de reconnaissance faciale	28
7	Précision module de reconnaissance faciale	31

Table des figures

1	Les algorithmes du machine learning	9
2	Les algorithmes de deep learning	10
3	Un réseau de neurone convolutif	10
4	Architecture général de la vision par ordinateur	11
5	Architecture du système de vidéosurveillance	12
6	Architecture du système de prise présence	13
7	example images pour reconnaissance faciale	17
8	300W-Lp head pose	18
9	FER_2013 datasets	19
10	analyse des sentiments	20
11	exchange pape	20
12	Looking friend	21
13	Talking	21
14	Using sheet	21
15	No cheating	21
16	Architecture générale d'un système de reconnaissance faciale	22
17	Architecture MTCNN	23
18	Application MTCNN	24
19	Principe fonctionnement FaceNet	24
20	Angles de Eleur.	25
21	réseau convolutionnel de pose faciale.	25
22	Reconnaissance des expressions faciales	26
23	Reconnaissance des expressions faciales	27
24	Précision modèle de reconnaissance et image test	29
25	Performance modèle d'estimation du regard	29
26	image estimation du regard	30
27	Performance modèle de reconnaissance d'expression faciale	30
28	image de reconnaissance d'expression faciale	30
29	image reconnaissance de scénario de triches	31

Nomenclature

O – NET : Output Network

MTCNN : Multi-task Cascaded Convolutional Networks

NME : Normalized Error

ASM : Active Shape Model

AAM : Active Appearance Model

CLM : Constrained Local Model

VGG : Visual Geometry Group

LFW : Database. Labeled Faces in the Wild

AFLW : Database Annotated Facial Landmarks in the Wild

3DDA : 3D Dense Face Alignement

CCTV : Closed-circuit television

FR : Facial regonition

SSH : Single Stage Headless

BNC : Naive Bayes classifier

DTC : Decision Tree classifier

GAN : generative adversarial network

SVM : Support Vector Machine

KNN : k-Nearest Neighbours

CRF : Conditional Random Fields

CNN : Convolutional Neural Networks

VRNN : Varational Recurrent Neural Networks

t – SNE : t-Distributed Stochastic Neighbor

1 Résumé

De nos jours, avec l'évolution permanente des données et des unités de stockage grâce à l'utilisation des matériaux multimédia tels que les téléphones et de visosurveillances tels que les caméras, et de matériels puissants, les domaines que sont la vision par et l'intelligence artificielle sont de plus en plus solliciter à fin d'améliorer notre quotidien. A la lumière des différentes difficultés qu'on rencontre lors de la surveillance des examens aux sains des établissements scolaires et institutions, nous proposons une approche de surveillance automatique de ces évaluations.

Nous insistons sur quatre modèles d'aide à la décision des examens, notamment la reconnaissance faciale pour détecter les usurpateurs d'identité dans les examens, la reconnaissance des expressions faciales et l'estimation du regard à partir de la pose de la tête à fin de déterminer les attitudes des étudiants et un dernier modèle d'identification des scénarios de tricherie au déroulement d'un examen.

Pour ce faire nous utilisons des techniques basées sur les réseaux de neurones convolutifs pour la mise en œuvre de ces modèles. Les modèles FaceNet et le classifieur **SVM** ont été utilisés pour notre modèle de reconnaissance faciale. Les CNN ont été utilisés pour la mise en œuvre des modèles d'estimation de regard et la reconnaissance des expressions faciales. Et le modèle **ResNet50** et le transfert learning ont été utilisés pour la reconnaissance des scénarios de tricherie.

2 Introduction

De nos jours, nous pouvons observer dans notre environnement, en tout lieu et place des installations des dispositifs de surveillance extrêmement puissants que sont "**les caméras**". Nous devons cette vulgarisation technologique grâce à leur coût de plus en plus très abordable. En effet, l'Homme soucieux de la sécurité des personnes et des biens, fait recours à ces dispositifs pour monitorer ses activités et garantir sa sécurité. Les aéroports par excellence regorgent de très grands réseaux de caméras, de même que les réseaux routiers, les supermarchés, les accès dans les bâtiments et aussi l'utilisation des caméras au sein des établissements éducatifs comme excellent support pédagogique.

Nous allons dans notre analyse, situer premièrement le contexte dans lequel s'inscrit notre thème, puis faire état des motivations, ensuite définir les objectifs visés et les résultats attendus, et enfin décrire la démarche à suivre pour atteindre ces objectifs.

2.1 Contexte technologique

Depuis des années, au regard d'importantes évolutions technologiques que connaît le monde d'une part, et de l'explosion des données d'autre part, le stockage de grands volumes de données est désormais possible. Les supercalculateurs voient leurs puissances décuplées, et on assiste également aux fabrications d'équipements électroniques très puissants à l'instar des caméras qui sont constituées de séquences vidéos, et à leur tour constituées d'une longue suite d'images individuelles ou trames. On appelle fréquence d'une caméra le nombre de trames par seconde, elle varie avec la puissance de la caméra, voir même des trillions de trames la seconde (**BLUFFANT baptisé T-CUP**). Dès lors, compte tenue de la vulgarisation des systèmes de vidéosurveillance qui ne cesse de prendre

le large, il devient difficile et voit même impossible pour des opérateurs humains de moniterer de tels systèmes, dû à l'immensité des données produites. Et de plus à la persistance rétinienne car théoriquement **l'oeil ne peut capter environ 20 images par seconde**.

C'est ainsi que les entreprises et institutions adoptent l'automatisation de leur système de vidéosurveillance.

Un système de vidéosurveillance peut être défini, comme étant un système de caméras et de transmissions d'images. Généralement l'analyse des flux d'informations issus de ces systèmes, constitue un excellent support d'aides à la décision.

L'automatisation des systèmes de vidéosurveillance confère non seulement un gain en réactivité, mais aussi en fiabilité. Mais celle-ci demeure un réel challenge notamment pour la vision par ordinateur, mais aussi pour l'apprentissage automatique.

En effet la vision par ordinateur aussi appelée vision artificielle ou vision numérique est une branche de l'intelligence artificielle dont le principal but est l'analyse, le traitement et la compréhension des images prises par **un système d'acquisition** (caméras).

Elle utilise généralement comme support l'apprentissage automatique qui consiste à donner la possibilité à un système informatique d'apprendre par soi-même à partir d'exemples.

Nous sommes donc amenés à concevoir et implémenter un système automatique (**modèles**) d'aide à la surveillance d'examen.

2.2 Motivation

Notre thème s'inscrit dans des domaines de recherche très actif et complexe, que sont la vision par ordinateur et de l'apprentissage automatique.

Le processus de vidéosurveillance automatisé requiert généralement une analyse complète de la scène, pour les détection et la reconnaissance des objets qui constituent un domaine de recherche très active compte tenu de la diversité et la complexité des objets qui nous entourent. Toutefois nous nous interrogeons, sur les différents critères de la scène qui pourront nous permettre de réaliser un système de vidéosurveillance optimale pour monitorer les examens et faciliter la détection des éventuels possibilités de tricheries.

Un tel système permettra en autre de :

- éliminer les efforts fournis pour une surveillance
- gagner en temps
- aider l'enseignant à la décision
- motiver les apprenants (élèves et étudiants) à développer leur aptitudes, pour ne compter sur soit - même durant leurs examens

Pour cela plusieurs objectifs seront définis à fin de dégager une meilleure approche à notre problème.

2.3 Objectifs

L'objectif principale de notre travail, sera de concevoir et d'implémenter un système automatique de vidéosurveillance d'examen (**modèles**), à l'aide de la vision numérique et de l'apprentissage automatique.

Notre système devra entre autres :

- tracker les apprenants (leur visage ou mains pour la décision)

- détecter les activités suspects (actions de trichéries)
- fournir une robustesse aux différents paramètres de la scène (mouvements rapides, les occlusions, nombres d'entités évolutives)
- détecter les présences et les cas d'usurpation d'identités.
- assister le personnel dans la prise de décision
- capturer les scènes de trichéries

Ces objectifs nous permettent de circonscrire mieux les résultats escomptés par notre système.

2.4 Résultats attendus

Notre système (modèles) devra nous permettre de :

- identifier chaque individu à partir de la reconnaissance faciale
- estimer le regard de chaque individu à partir de la pose de leur tête
- déterminer les expressions faciales des individus
- détecter les cas d'anomalies (triches)

Pour la mise en place de ce système, nous devons nous définir une démarche à suivre.

2.5 Organisation du travail

Notre travail sera divisé en quatre parties.

Dans la première partie nous présentons les **concept théoriques et l'état de l'art**. Les concepts théoriques définissent les différentes technologies qui gravitent autour de notre sujet à savoir la vision par ordinateur et l'apprentissage automatique. Dans L'état de l'art ou encore revue littéraire nous présentons de manière synthétique les travaux existants en rapport avec le nôtre et les différentes algorithmes utilisés.

Dans la deuxième partie il s'agit de la **solution proposée**, elle présente de façon détaillée la démarche à suivre pour concevoir notre solution en nous inspirant des précédents travaux étudiés.

La troisième partie est l'implementation de notre solution proposée et l'interprétation des résultats.

Enfin la dernière partie est l'**évaluation de notre solution**. Dans cette partie nous présentons une étude comparative de notre travail par rapport aux autres travaux, ou par rapport à d'autres algorithmes.

3 Concepts théoriques et État de l'art

La mise en œuvre de notre système automatisé de vidéosurveillance des étudiants dans le cadre des examens en salle de classe, fait intervenir deux domaines principaux, la vision par ordinateur et l'apprentissage. La compréhension de ces domaines demeurent donc indispensables pour cerner le mieux notre sujet et proposer une solution à fin d'atteindre nos objectifs.

3.1 L'apprentissage automatique

Dans le domaine de l'apprentissage automatique, nous distinguons deux branches que sont : le **machine learning** et le **deep learning**.

3.1.1 Le Machine Learning

Arthur Samuel (1959) : « Le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmé ».

Tom Mitchell (1998) : « Un programme informatique se dit d'apprendre de l'expérience E par rapport à une catégorie de tâches T et mesure de la performance P, si sa performance à des tâches T, est telle que mesurée par P, s'améliore avec l'expérience E».

Une vue d'ensemble des algorithmes de machine learning en fonction des méthodes d'apprentissage est présentée ci-dessous.

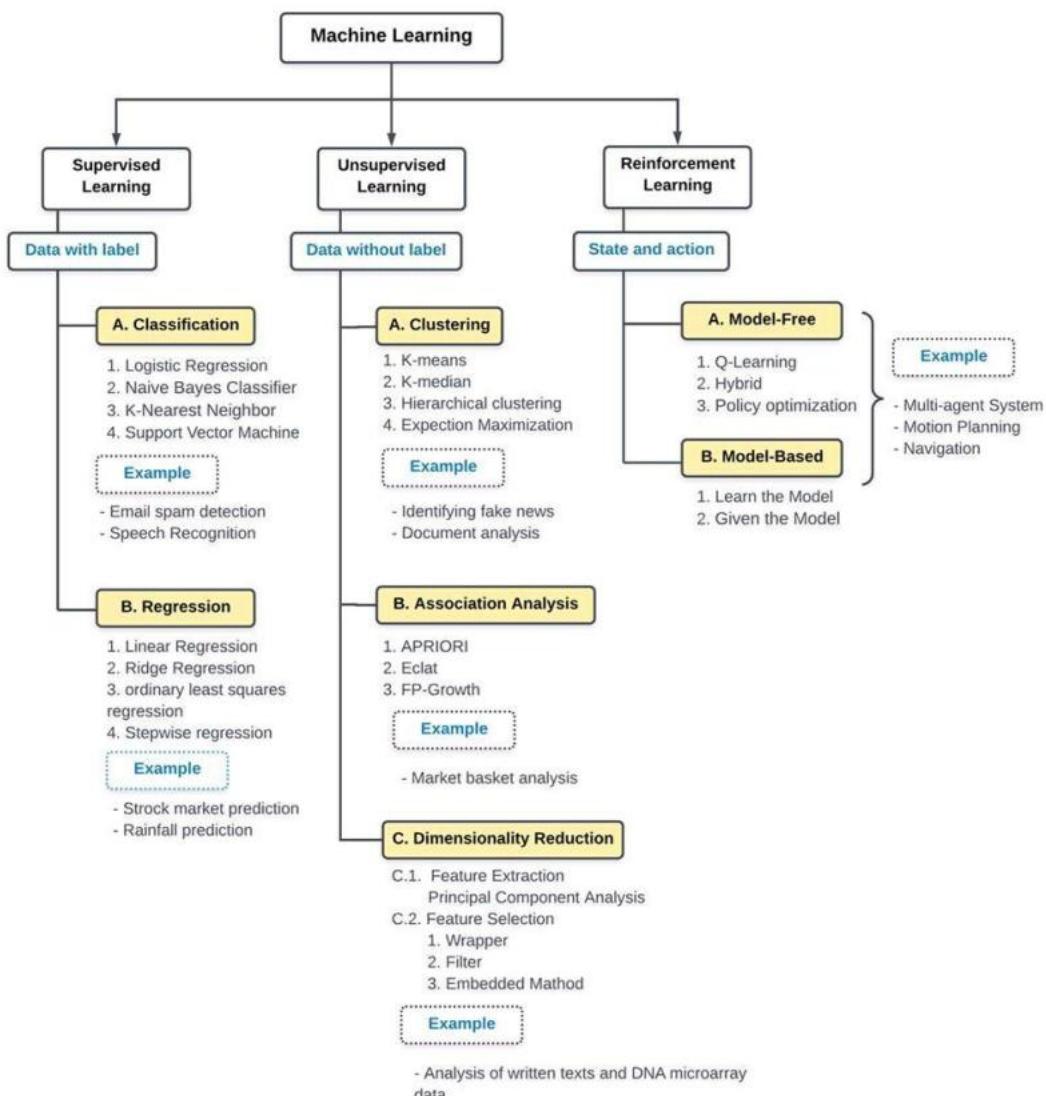


FIGURE 1 – Les algorithmes du machine learning

3.1.2 Le Deep Learning

Le deep learning ou encore apprentissage en profondeur est une sous domaine du machine learning dont l'apprentissage est basé sur une organisation en couches, elle-même basé sur le concept de réseaux neurones artificiels.

Une vue d'ensemble des algorithmes de deep learning est présentée ci-dessous.

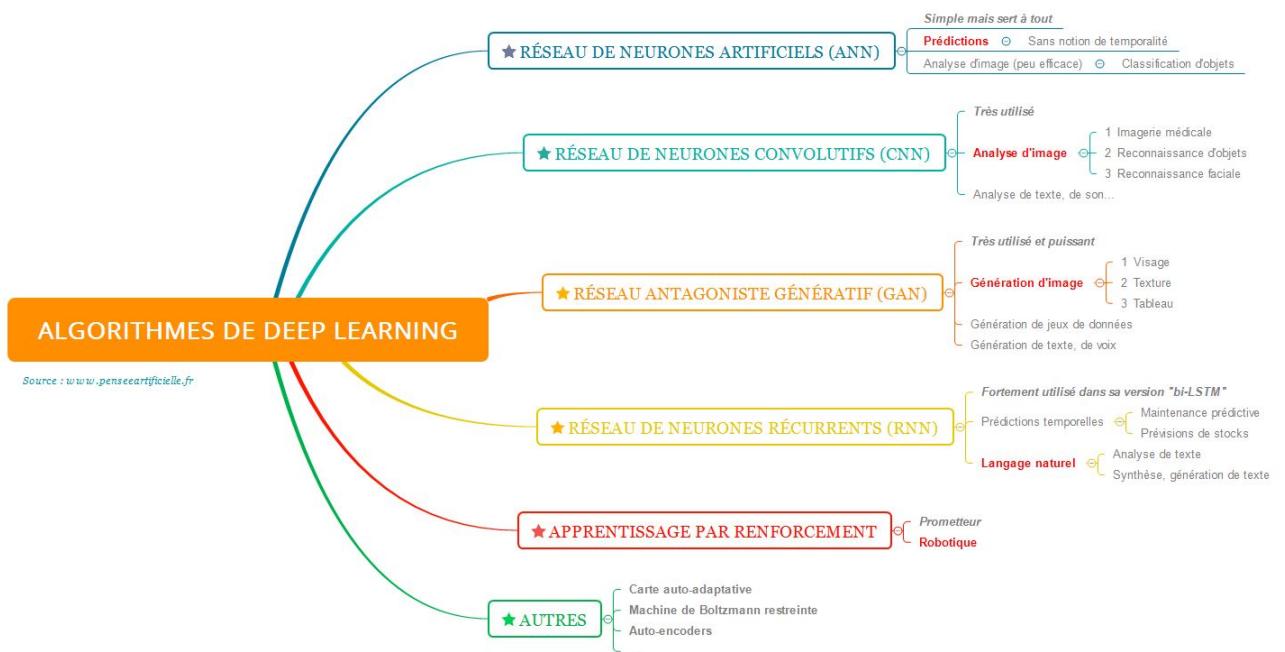


FIGURE 2 – Les algorithmes de deep learning

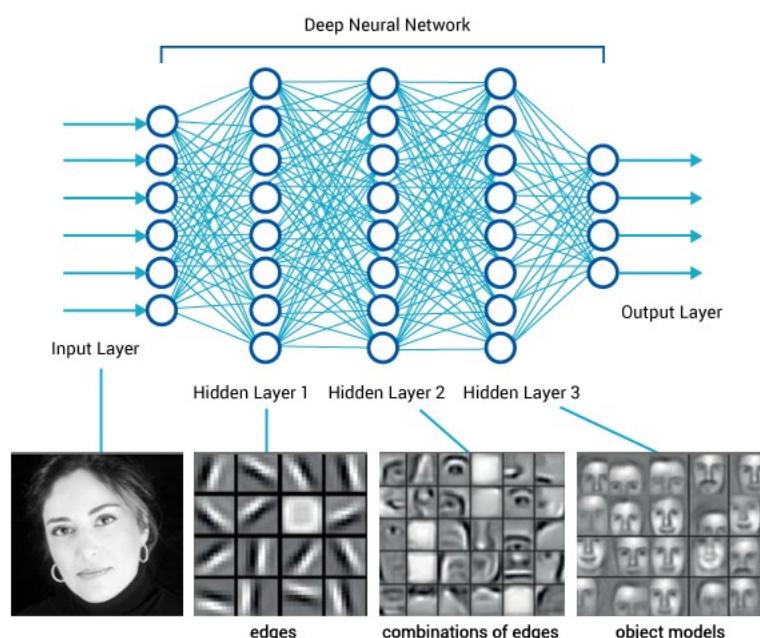


FIGURE 3 – Un réseau de neurone convolutif

3.2 Vision par ordinateur

Partant de l'analyse d'informations pertinentes sur des images en passant par le traitement des images, la vision par ordinateur constitue un excellent outil d'aide à la décision.

La figure ci-dessous illustre de manière général les différentes étapes d'un système de vision par ordinateur.

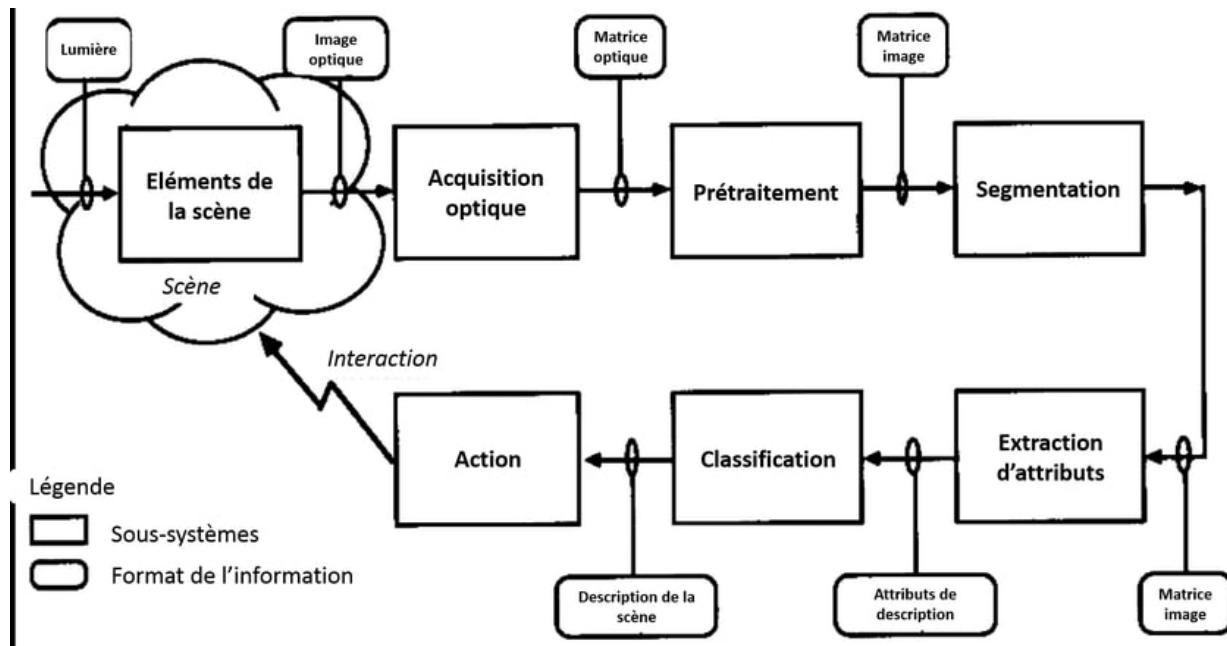


FIGURE 4 – Architecture général de la vision par ordinateur

3.3 État de l'art

Dans cette revue de littérature, nous faisons un inventaire de travaux antérieurs portant sur des thèmes reliés à la surveillance d'examen à l'aide des techniques de vision par ordinateur et d'apprentissage. En effet, la littérature scientifique nous permet de trouver des projets de recherche similaires. Toutefois ces travaux constituent plus pour nous une source d'inspiration.

3.3.1 Application basée sur la vision par ordinateur pour la Surveillance des comportements des étudiants en classe

Bui Ngoc Anh, Ngo Tung Son, Phan Truong Lam et les autres, présentent Une application basée sur la vision par ordinateur pour l'analyse des comportements chez les apprenants, [1]. L'objectif était de mettre sur pied un système capable de suivre l'évolution des apprenants et permettre la prise des décisions à partir des informations obtenues afin d'améliorer les aptitudes des apprenants et les méthodes d'enseignements. Pour ce faire, ils proposent un modèle complexe basé sur plusieurs techniques permettant la détection du visage (SSH), la détection des repères faciales (O-NET et L-NET, partie de MTCNN), l'encastrement du visage (ARcface), la classification des visages (KNN) et l'estimation et la classification du regard (Hopenet et SVM). L'évaluation du modèle est très satisfaisante soit une F1-score de 82.83%.

Les données utilisées sont constituées de 1800 frames soient 10 à 20 étudiants par frame, mais on note plusieurs difficultés notamment les étudiants ne regardant pas la caméra la plupart du temps, les images floues et les visages partiellement visibles.

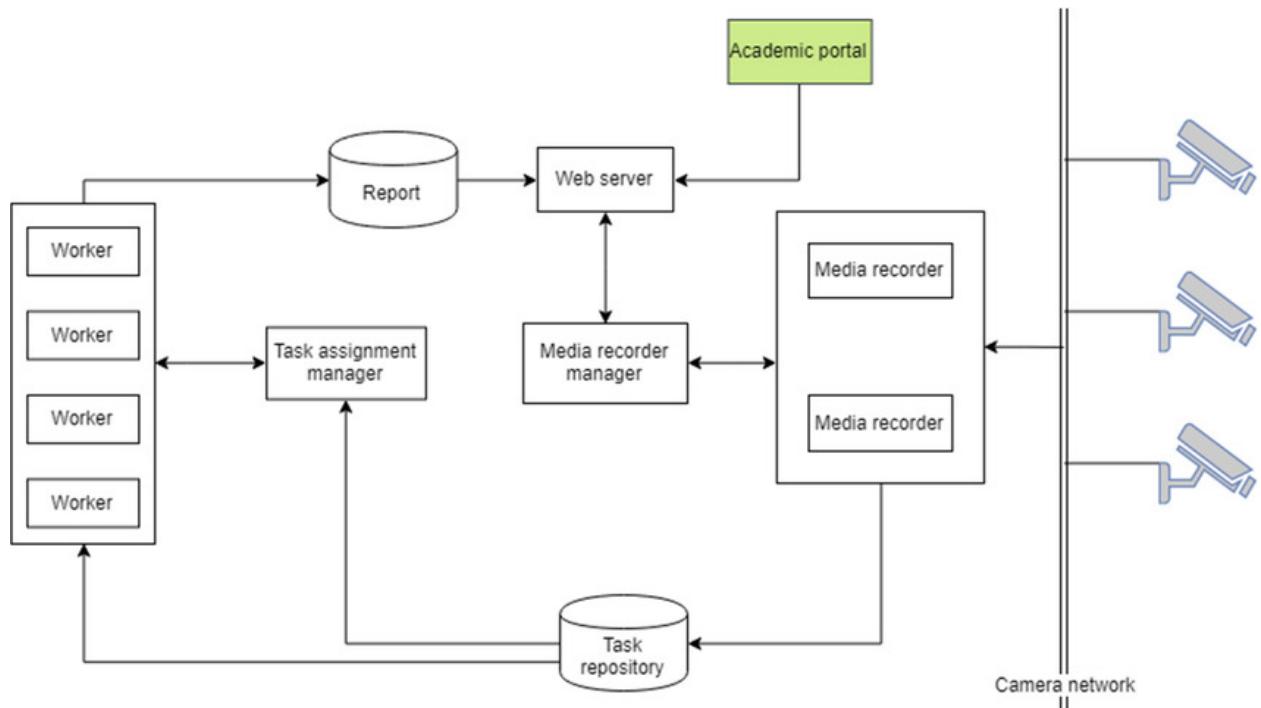


FIGURE 5 – Architecture du système de vidéosurveillance

3.3.2 La prise de présence basée sur la vidéosurveillance, cas d'étude le collège polytechnique FPT.

Ngo Tung Son, Bui Ngoc Anh et Tran Quy Ban, présentent un système ATSS (Attendance Taking Support System) [2] expérimenté au collège polytechnique FPT sur 120 élèves dans cinq classes. Ils ont mis au point une application web de reconnaissances faciale profonde à partir d'un modèle résolvant plusieurs difficultés notamment la luminosité des images, la résolution de la caméra et les images floues. Ce modèle repose sur plusieurs méthodes la détection des régions d'intérêts, la détection du visage et des repères faciaux (MTCNN), l'alignement du visage, l'incorporation du visage (Arcface) et la classification de visages (KNN). L'évaluation du module Arcface à l'aide du t-SNE eut un résultat très satisfaisant soit 98% avec un seuil de 1.04, la matrice de confusion du système et la précision obtenues sont 92.7%.

Les données utilisées ont été obtenues à partir d'une application mobile dans la résolution la plus basse 720 pixels, 30 frames par secondes.

Schedule data

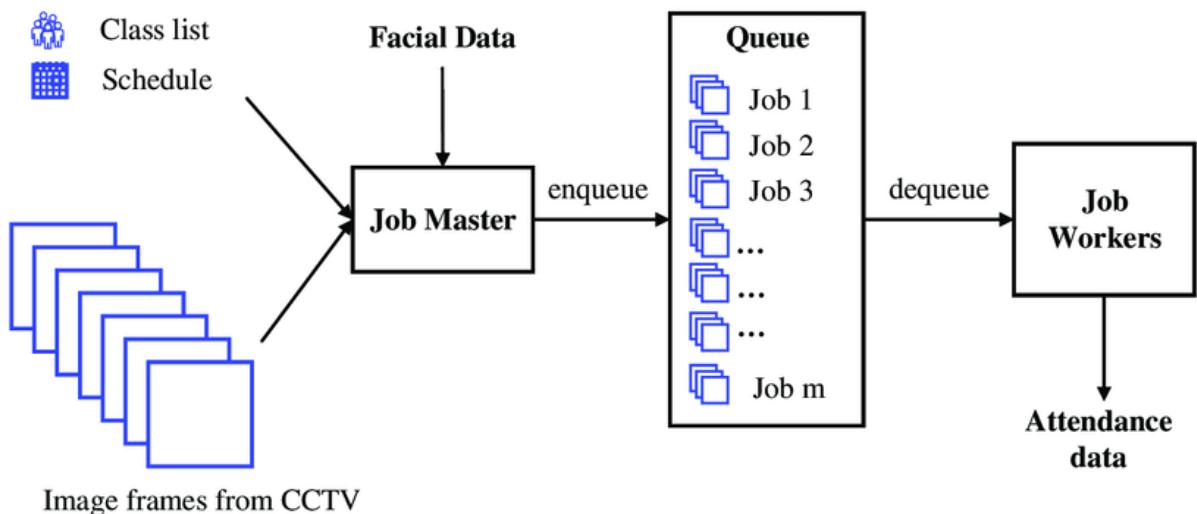


FIGURE 6 – Architecture du système de prise présence

3.3.3 Apprentissage non supervisé de la structure et de la dynamique des objets à partir de vidéos

Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole et Kevin Murphy and Honglak Lee, [13] proposent une méthode de détection de mouvement des objets, qui consiste à extraire et prévoir la structure et la dynamique des objets à partir de vidéos sans la supervision. Cette approche se base sur la détection des points clés des pixels de frames issus des vidéos pour prédire le mouvement de l'object cible (homme). Elle est utilisée un modèle stochastique. Le modèle utilise un réseau neuronal récurrent variationnel (VRNN).

3.3.4 Prévision du regard dans des vidéos immersives à 360 degré dynamiques

Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, et Shenghua Gao, explorent la prédiction du regard dans des images dynamiques à 360 degré des vidéos VR. [30]. Pour ce faire deux cartes de saillance sont alimentées et les images correspondantes dans un neurone convolutionnel Réseau (CNN) pour l'extraction des fonctionnalités. Une mémoire à long terme et à court terme (LSTM) est également utilisé pour coder chemin de numérisation historique. Ensuite les fonctionnalités CNN et Fonctions LSTM sont combinées pour la prévision du déplacement du regard entre point de vue sur du temps actuelle et point de vue sur un temps suivant. Des expériences approfondies valident l'efficacité de notre méthode de prédiction du regard dans des scènes VR dynamiques.

3.3.5 Modélisation conjointe de la pose et de l'expression pour la reconnaissance de l'expression faciale

Au regard des nombreux challenges de reconnaissances des expressions faciales dus aux différentes poses possibles, Feifei Zhang, Tianzhu Zhang, Qirong Mao et Changsheng Xu

[29] proposent un modèle d'apprentissage en profondeur de bout en bout en exploitant conjointement différentes poses et expressions synthétisés d'images faciales et la reconnaissance des expressions invariants des poses faciales. Le modèle proposé est basé sur réseau contradictoire génératif (GAN). Dans un premier temps, la structure codeur-décodeur du générateur peut apprendre une représentation générative et discriminante pour les images de visage. Ensuite, la représentation identitaire est explicitement dissociée de l'expression et de la pose à travers les codes d'expression et de pose. Puis, le modèle peut générer automatiquement des images de visage avec différentes expressions sous des poses arbitraires pour agrandir et enrichir l'ensemble de formation pour FER (Facial expression recognition). L'évaluations du modèle démontrer que l'algorithme proposé fonctionne favorablement contre les méthodes de pointe.

3.3.6 Estimation des poses et actions humaines

Xuanyi Dong, Shou-I Yu et les autres [24] proposent un modèle Parsing Induced Learner. Il est composé d'un encodeur d'analyse qui code une image d'entrée en représentation d'analyse de haut niveau, et un adaptateur de paramètres de modèle de pose qui apprend à adapter les paramètres de poser à un modèle en exploitant les représentations d'analyse, qui ensemble apprennent à prédire les paramètres dynamiques du modèle de pose à extraire.

Le modèle a été entraîné sur les bases de données LIP et MPII, fournissant des résultats très satisfaisantes.

Taiki Sekii [26] propose une nouvelle approche pour estimer les poses humaines, appelée réseaux de propositions de pose. Il introduit un état paradigme de détection d'objets à un coup de pointe utilisant des cartes de caractéristiques d'image en grille dans un scénario de détection de pose ascendante. Les propositions de parties du corps, envoyés en tant que propositions de région, et les membres sont détectés directement via un CNN à un coup.

Le benchmark MPII Multi-Person, montre un taux de réussite atteint de 72,8%.

Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong , Zhiru Shi, Jingyi Yu et Sheng-hua Gao [30], proposent un modèle basé sur la structure du graphe HOI (Human Object Interaction) représentée par une matrice d'adjacence. Ce modèle a été évalué à partir des bases de données HICO-DET, V-COCO et CAD-120, donnant des résultats très satisfaisantes.

3.3.7 Reconnaissance de visages

Lisha Chen, Hui Su et Qiang Ji, [12] proposent une méthode de détection de repères faciaux structurés en profondeur basée sur la combinaison d'un réseau convolutionnel profond (CNN) avec un champ aléatoire conditionnel CRF (Conditional Random Fields) qui est un modèle graphique probabiliste. Le modèle obtenu, montre des performances supérieures aux techniques de pointe existantes dans la détection des points de repères faciaux, en particulier une meilleure capacité de généralisation sur des ensembles de données difficiles qui incluent une pose et une occlusion de grande taille.

Xiangyu Zhu, Xiaoming Liu, Member, Zhen Lei, Senior Member, et Stan Z. Li, et Fellow, proposent un modèle appelé **3DDFA**(3D Dense Face alignment) pour détecter l'alignement du visage sur plusieurs poses entre 0 et 90 degrés [23]. Ce modèle est basé sur la combinaison de deux technologies, la regression en cascade et le réseau de neurone en

convolution (CNN). Au regard des système existants notamment ASM, AAM et CLM, le modèle 3DDFA fait preuve d'une efficace supérieur car evaluer sur un ensemble de données difficiles notamment 300W-LP, AFLW et AFLW2000-3D. L' évaluation du modèle donne des résultats satisfaisants sur ces trois ensembles.

Song Zhou et Sheng Xiao[21] proposent une analyse des modèles de reconnaissance faciale 3D existants, corrigéant les lacunes de la reconnaissance en 2D. Cet article traite l'analyse des techniques de reconnaissance faciale sur différentes approches, notamment par la pose, par les expressions et par les occlusions.

Ali Ahmed Elmahmudi et Hassan Ugail, proposent une approche de reconnaissance profonde du visage à l'aide de données faciales imparfaites [8]. Cet article exploite les bases de données EFW et LFW. Le modèle en question est basé sur les réseaux de neurones de convolution (CNN) tels que VGG-Face pour identification des repères faciaux notamment nez, yeux et bouche, et des classificateurs notamment SVM et la similitude cosinus pour évaluer le taux de reconnaissance. Cette approche montre de même que le taux d'apprentissage diminue avec de faibles résolutions d'images.

3.3.8 Cheating Video Description Based on Sequences of Gestures

Ahmad Arinaldi et Mohamad Ivan Fanany utilise le flux vidéo (caméra) pour surveiller les étudiants pendant l'écriture des examens [32]. Ils se basent sur des gestes illégaux tels que :

- donner le codes (geste de la main)
- regarder les amis
- utiliser une feuille de triche
- échange orale
- échange de papiers entre étudiants

Pour cela il utilise deux modèles, un modèle de reconnaissance gestuelle basée sur 3DCNN et XGBOOST et un modèle de génération de langage basé sur un réseau LSTM définissant les activités de deux étudiants et leur interaction. Il obtiennent en moyenne une précision de 81.11% pour 3DCNN + XGBOOST et 95.3% LSTM.

3.3.9 Résumé de la revue littéraire

Nous récapitulons nos différents articles sous forme d'un tableau. Il montre les méthodes et résultats obtenus. Notre étude bibliographique a été très enrichissante tant aux méthodes notamment **3DDA**, **ASM**, **AAM** et **MTCNN** qu'aux données exploitées. Nous relevons plusieurs bases de données notamment **Human3.6M** pour les mouvements humains, **LFW**, **FSA-NET**, **WIDERFACE**, **AFLW**, **Face Net** et **300W-LP** des banques d'images faciales pour la détection des repères faciaux, la reconnaissance et l'alignement du visage.

Révues Littéraires			
Date de Pub.	Références)	Méthodes	Résultats
6 Nov. 2019	[2]	SSH, O-NET, L-NET,...	82.83
21 Fev. 2020	[1]	MTCNN, ARCFACE, KNN	92.7%
Dec. 2019	[12]	CNN, CRF	NME : 2.21
Dec 2019	[13]	VRNN	95%
Nov. 2017	[23], [8]	CNN, REGRESSION CASCADE	
2018	[26]	CNN	72.8%
2018	[27]	GP-NN	80%
2018	[29]	GAN	81.20%
2017	[32]	3DCNN + XBOOST, LSTM	81.20%

TABLE 1 – Résumé des revues littéraires

4 Solution proposée et implementation

Rendu à ce niveau de notre travail, après notre revue littéraire, nous proposons une solution qui va nous permettre de monitorer des examens en salle de classe.

4.1 Spécifications de la solution

Notre solution gravite autour de **quatre modules** principales, la **reconnaissance faciale**, la **reconnaissance des expressions faciales** et l'**identification des poses de la tête** et la **reconnaissance des triches**. La solution proposée devra répondre à un ensemble de spécifications.

4.1.1 Spécifications des modèles

Le système devra enregistrer une session entière et fournit en sortie un rapport. Notre solution devra nous permettre de :

- identifier les étudiants participant à l'examen à fin de détecter les usurpateurs d'identités.
- détecter des mouvements dits suspects de la tête des participants aux examens.
- identifier les différentes expressions des participants aux examens à l'instar de la peur, la colère, la surprise, la tristesse, la joie, la neutralité.
- identifier les activités de triches(Normal, échange de feuille, échange orale, utilisation des cartouches, regarder les amis)

Comme résultats escomptés, nous avons entre autres :

- Classifier les personnes reconnues et non reconnues par le système.
- Classifier les poses corrects et incorrects.
- Classifier les expressions positives et négatives.
- classifier les différentes scènes

La section suivante est réservée à la présentation des données.

4.2 Présentation des données

Plusieurs bases de données existantes seront utilisées pour évaluer la solution proposée. Ces données sont pour la plupart des images, notre webcam pourra être utilisée pour les vidéos.

4.2.1 Reconnaissance faciale

Nous avons utilisé dans ce cadre la base de données **LFW**. Donc 24 classes, nous nous sommes limités à un total de 237 images compte tenu des performances de notre machine.

LFW est une base de données de visage de personnes labelisée et regroupée en classe et excellente pour la reconnaissance faciale.

Image pour détection faciale		
partitions	proportion	description
train_data	80%	80% des données seront utilisées pour entraîner notre modèle
test_data	10%	10% des données seront utilisées pour tester notre modèle

TABLE 2 – Répartition de la base données pour la reconnaissance faciale



FIGURE 7 – example images pour reconnaissance faciale

4.2.2 Estimation des poses de la tête

Nous utiliserons les bases de données **BIWI** et **300W-LP** une base de données avec une diversité de poses. Nous pourrons également explorer une base de données de 6288 features préconçues (taille : 2278) et labelisés(roll, yaw, et pitch : angles de Euler) à partir de ces bases.

Différentes poses de la tête		
pose	orientation	description
vers la gauche	inférieur à zéro dégré	le candidat regarde son voisinage de gauche
centrée	zéro dégré	le candidat a le regard vers l'avant
vers la droite	supérieur à zéro dégré	le candidat regarde son voisinage de droite

TABLE 3 – Détection pose de la tête

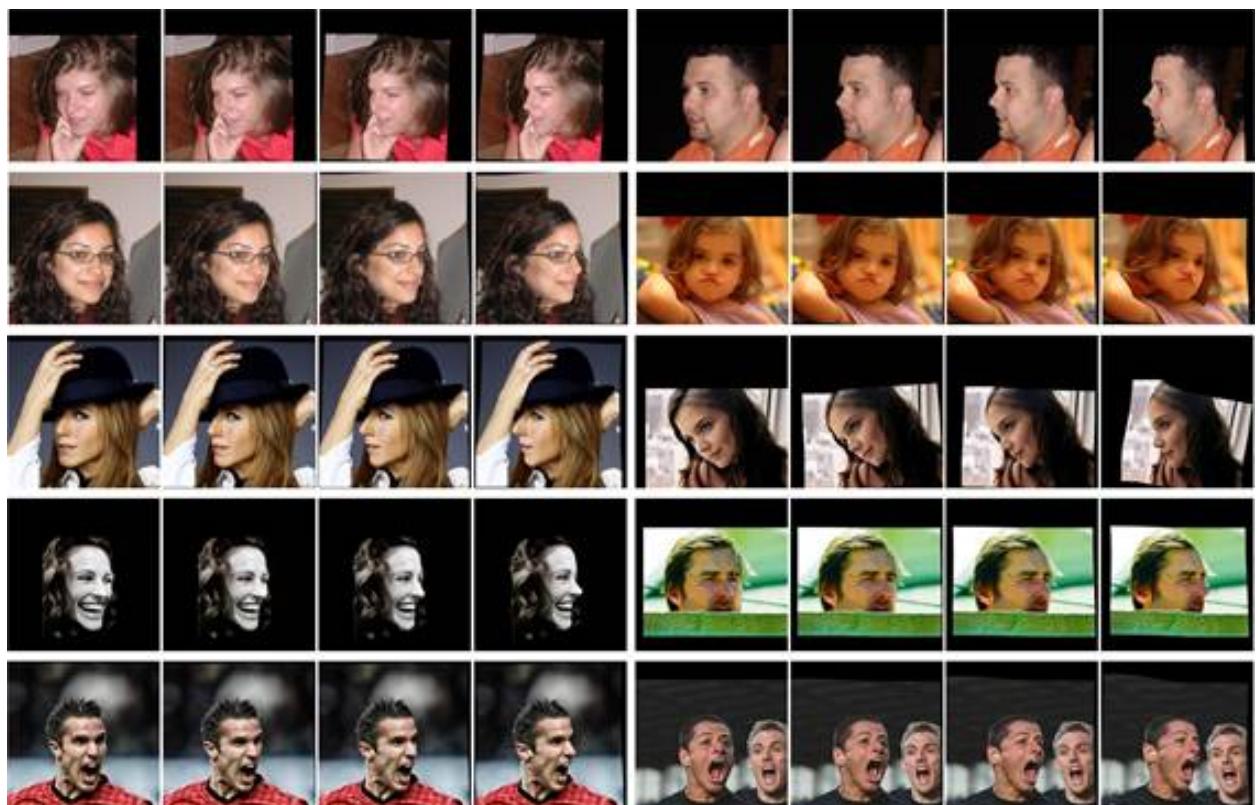


FIGURE 8 – 300W-Lp head pose

4.2.3 Reconnaissance des expressions faciales (émotions)

Les émotions peuvent être un excellent atout pour évaluer la concentration des étudiants, prédire les résultats d'un examen. Nous exploitons la base de données au format vidéo **CMU-MOSI** (Multimodal Opinion Sentiment Intensity) que nous utiliserons et également la base de données FER_2013 de 35887 images labelisées.

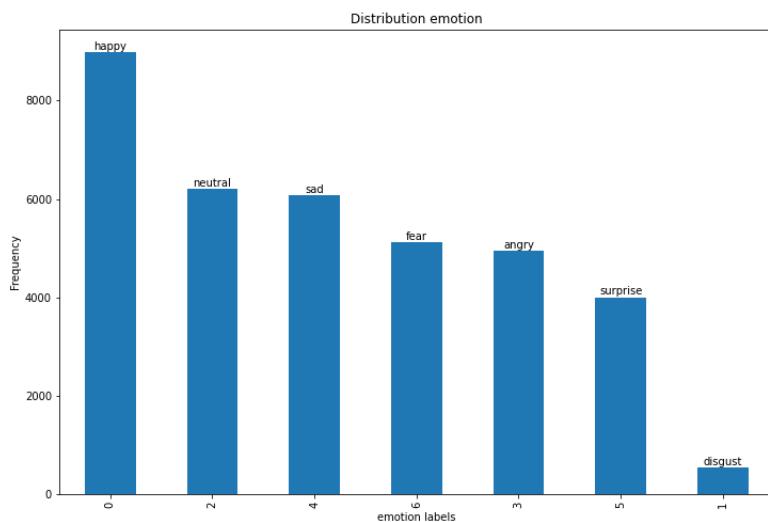


FIGURE 9 – FER_2013 datasets

Nous utiliserons la base **FER_2013** de 35887 images, à cela nous ajoutons une autre source de données obtenues de 13682 images. Soit un total de **49569** images labelisées. Nous partionnons comme suite **80%** de données de test et le reste des **30%** reparties comme suites **80%** données de tests et **20%** données de validations.

Différentes expressions faciales	
expression	description
surprise(5)	l'étudiant est attrapé en situation de fraude
triste (4)	l'étudiant est déboussolé et sera tenté à frauder
colère (0)	l'étudiant est déconcentré par son voisinage
neutre (6)	l'étudiant est concentré sur son travail
joie (3)	l'étudiant trouve satisfaction aux épreuves
dégoût (1)	l'étudiant manifeste un mépris
peur (2)	l'étudiant manifeste la peur d'être pris en flagrant délit

TABLE 4 – Identification des expressions faciales et leurs labels

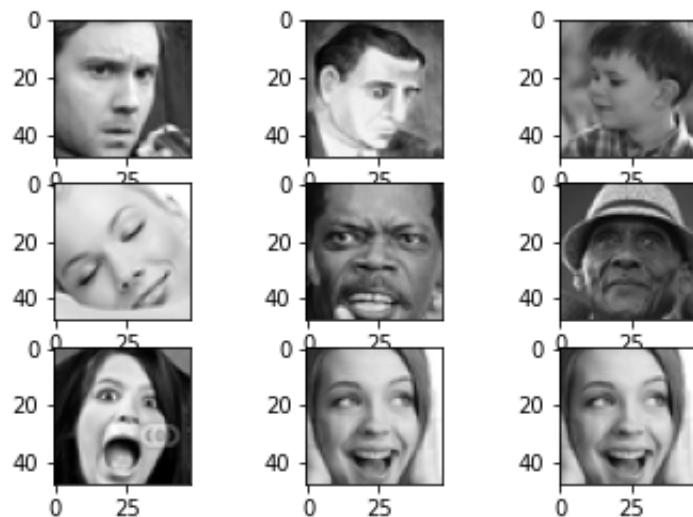


FIGURE 10 – analyse des sentiments

Après cette présentation des données, il est maintenant question de mettre en oeuvre ses différents modules.

4.2.4 Reconnaissance des fraudes à l'examen

Nous utiliserons ici une base de données préconçue par nous même. Elle est constituée de **5 classes** (**0=exchange Paper, 1=looking at friend, 2=Talking Friend, 3=Use Cheat Sheet, 4=No Cheat**).

Nous allons considérés 198 images toutes classes confondues pour l'entraînement, 73 images pour les tests et 19 images pour les validations.

Toutefois comme on peut le constater notre base est limité, que nous améliorons avec le temps.



FIGURE 11 – exchange paper



FIGURE 12 – Looking friend



FIGURE 13 – Talking



FIGURE 14 – Using sheet



FIGURE 15 – No cheating

4.3 Implémentation de la solution

Dans cette section, nous mettons en oeuvre nos modules de reconnaissance faciale, de reconnaissance des expressions et d'estimation de pose de la tête.

Avant d'entrer dans le vif du sujet, il est fort intéressant de rappeler que la mise en oeuvre de ces modules est précédée deux phases indispensables :

- **Phase d'acquisition des images** : les images futurs nécessaires au fonctionnement de notre système proviendront d'un réseau de caméras IP (2 à 3 caméras), préalablement installé dans les salles.
- **Phase de pré-traitement** : Les images provenant du réseaux de caméra peuvent être floues, à cause du mouvement des cibles, des conditions lumineuses et de faibles résolutions des images dépendant du type de caméra et de leur distance aux cibles. La phase de pré-traitement permettra d'améliorer le rendu des images notamment en redimensionnant l'image ,en supprimant le bruit à l'aide des filtre, et également en appliquant les opérations de Segmentation de morphologie (lissage des bords),

ce rendu peut conjointement être obtenu en améliorant le calibre des caméras.

Nous utiliserons des outils et librairies dans le cadre de la mise en oeuvre de notre solution.

- **Langage python** : notre solution sera développée à l'aide de ce langage pour sa puissance et de plus très adapté pour notre tâche.
- **Opencv** : librairie de traitement d'images
- **Tensorflow et keras** : frameworks de deep learning
- **Dlib** : une autre librairie très utilisée pour le deep learning.
- **MTCNN** : un réseau de neurone convolutif pour la detection des images.

4.3.1 Module de reconnaissance faciale

Malgré de nombreux avancés dans ce domaine de recherches, la mise en oeuvre d'un système de vérification et de reconnaissance faciale demeure toujours un réel challenge. La reconnaissance faciale a de très nombreuses applications notamment en **vidéo-surveillance**, biométrie, robotique, commande d'interface homme-machine, photographie, indexation d'images et de vidéos, recherche d'images par le contenu, etc. C'est un système avec plusieurs processus, une architecture générale de ce système se décrit comme suite.

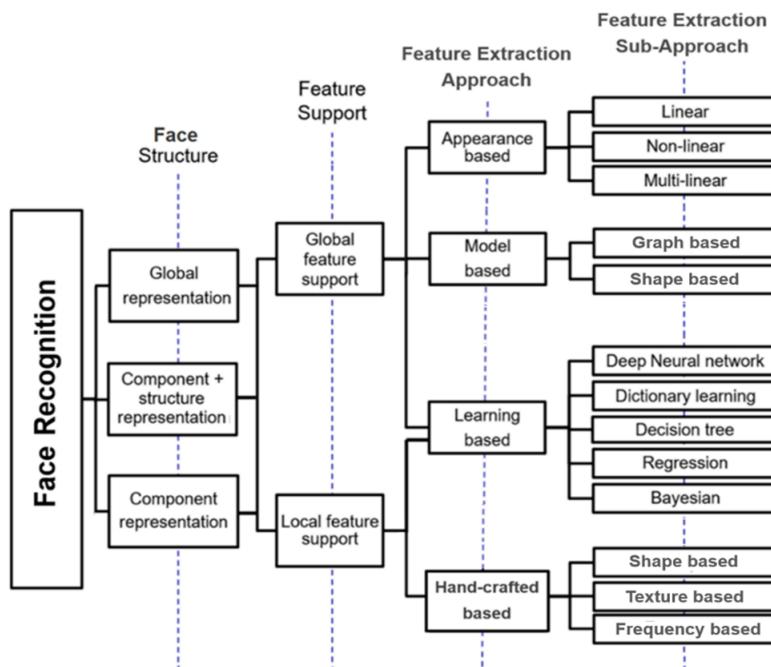


FIGURE 16 – Architecture générale d'un système de reconnaissance faciale

Notre module se décline en deux étapes :

1. Détection et extraction des caractéristiques du visage

Nous utilisons le réseau de neurone **MTCNN**, qui permet conjointement la détection de visage et des repères faciaux (les deux yeux, le nez et les extrémités de la bouche) à partir d'une image en entrée contenant une multitude de personnes, en utilisant les réseaux de neurones convolutifs. Cette méthode est une des plus populaires et très bon outil pour la détection de visages.

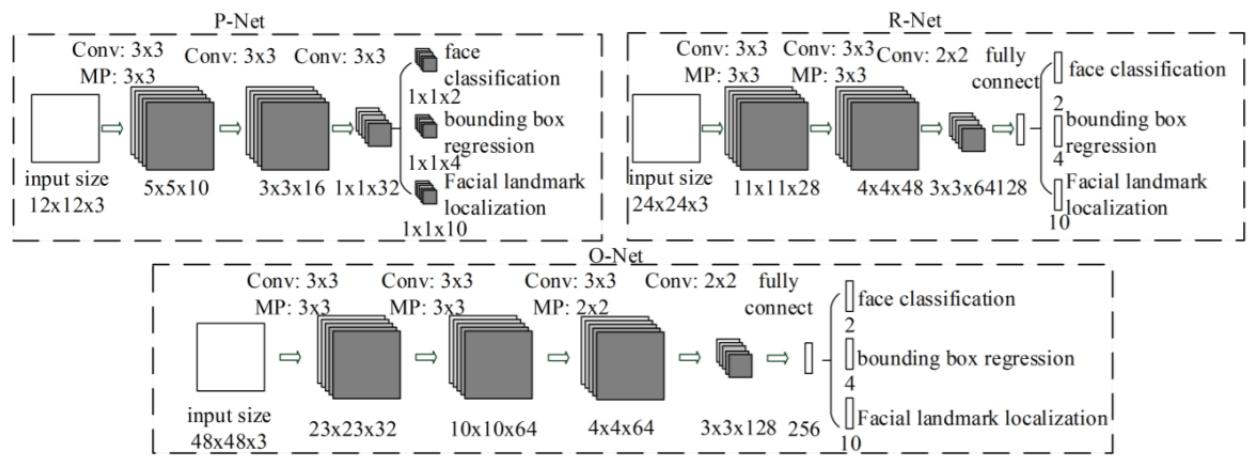


FIGURE 17 – Architecture MTCNN

(a) **P-Net** (Proposal Network)

Dans la première étape, un réseau entièrement convolutif est utilisé pour obtenir les régions cibles et leurs vecteurs de régressions de boîtes englobantes. Ces vecteurs sont utilisés pour calibrer les régions cibles, puis appliquer une suppression non maximale (NMS) pour fusionner des régions fortement chevauchées.

(b) **R-Net** (Refine Network)

Un autre CNN dont le rôle est de rejeter les régions non candidats, puis effectuer une nouvelle calibration et appliquer une suppression non maximale pour fusionner de nouveau les régions fortement chevauchées.

(c) **O-NET** (Output Network)

Similaire au précédent, il donne en sortie **cinq** détails faciaux à savoir l'oeil gauche, l'oeil droit, le nez et l'extrémité gauche et droite de la bouche.

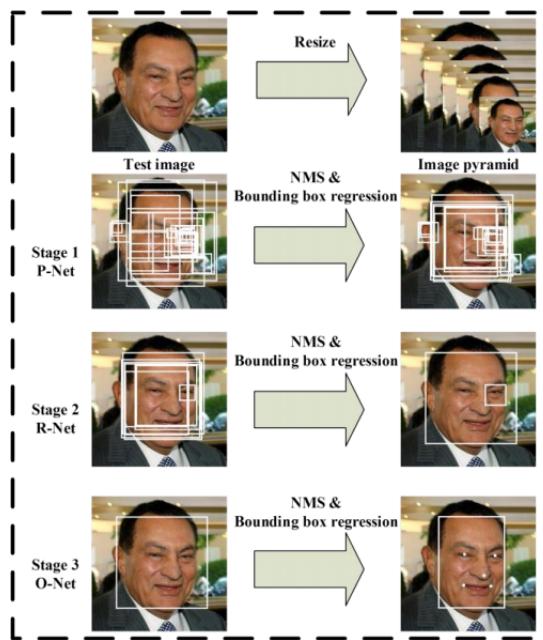


FIGURE 18 – Application MTCNN

2. Reconnaissance faciale

Après obtention des repères faciaux, nous utilisons une technique de reconnaissance faciale, de vérification et de classification par clustering **FaceNet**, nous utilisons ici **InceptionResNetV1** prenant en entrée une image **(160,160,3)**. Il utilise un réseau de neurone convolutif basé sur trois fonctions de perte. FaceNet donne une représentation de visage dans un espace euclidien de telle sorte que les distances dans cet espace correspondent à la similitude du visage, c'est-à-dire qu'une image d'une personne spécifique sera placée plus près de toutes ses autres images par rapport aux images de toute autre personne présente dans la base de données.

A fin d'améliorer la performance du modèle, nous s'y ajoutons des poids pré-entraînés **facenet_weights.h5** pour bénéficier du **transfer learning**.

Pour chaque classe d'images, nous allons extraire les descripteurs.

Ensuite nous utilisons un modèle de classification **SVM** pour les entraînements et les tests sur les ensembles de données d'entraînement et de tests préalablement normaliser et les labels préalablement encodés en utilisant la librairie **preprocessing** du framework **keras** en important les modules **LabelEncoder**, **Normalizer**.

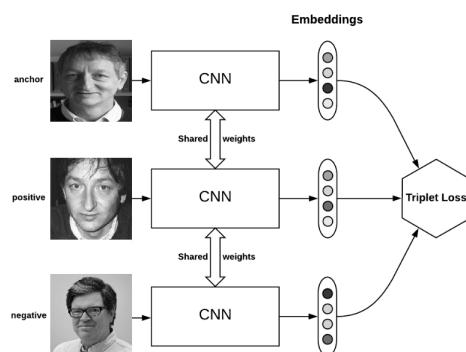


FIGURE 19 – Principe fonctionnement FaceNet

4.3.2 Module estimation des poses de la tête

Pour le module de détection de pose de la tête nous l' implementons à l'aide de la libraries de deep learning **dlib** pour une détection multiple de faces et répères faciaux et lextraction des caractérisques (**68 points de repères faciales**). La pose de la tête sera définie par trois labels :**angles d'Euler** (Pitch, Roll, Yaw) qui représentent la décomposition de la matrice de rotation de la configuration de la tête par rapport au cadre de la caméra.

Nous utilisons la base préconçus de 6288 features (vecteurs de 2278 obtenus en calculant les distances entre chaque pair de points soit $68 \times 67 / 2 = 2278$) avec des labels définis par le triplet (**roll, pitch, yaw**) labelisées à l'aide de **Amazon's Face Detection API..**

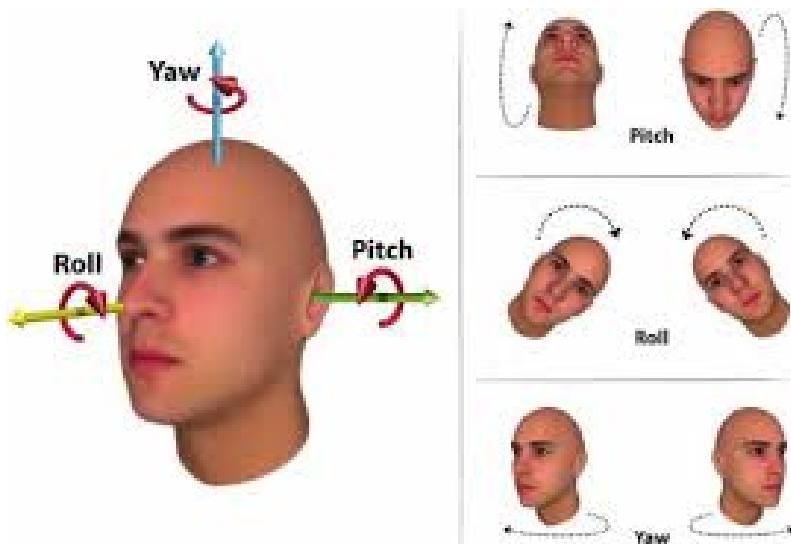


FIGURE 20 – Angles de Euler.

Nous utiliserons des réseaux convolutifs à partir des libraires **keras** et **tensorflow** qui nous permettra d'estimer les poses faciales. Mais notre réseau se limite à 3 couches denses dont la dernière définit les trois classes à prédire.

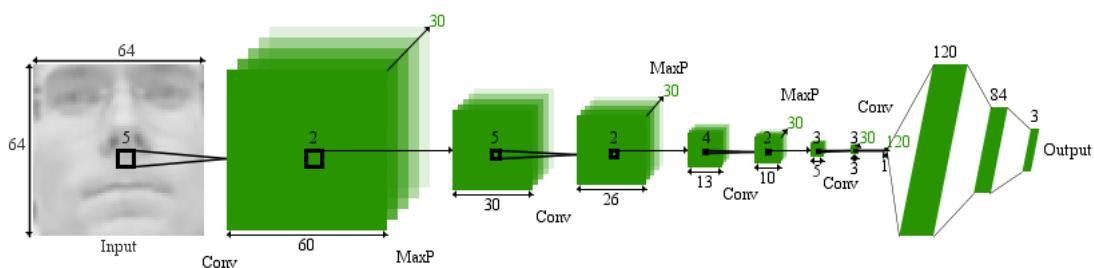


FIGURE 21 – réseau convolutionnel de pose faciale.

4.3.3 Module reconnaissance des expressions faciales

Dans ce module nous utiliserons **7 classes** (**0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral**) à fin de classer les différentes expressions faciales des apprenants durant des examens.

Les images ajoutées à la bases FER_2013 ont été traités et ramenées à la dimension **48 X 48** niveau de gris à l'aide du module **image** de la librairie **preprocessing** de keras tensorflow.

Nous utiliserons les réseaux de neurones convolutifs pour accomplir cette tâche.

Le réseau neuronal recevra en entrée une image en niveaux de gris 48x48, puis sortira la confiance de chaque expression.

L'architecture du réseau comprend 5 couches convolutives, 3 couches de sous-échantillonnage et une couche entièrement connectée.

- La première couche du CNN est une couche de convolution, qui applique un noyau de convolution de 3 x 3 et produit 64 images de 48 x 48 pixels.
- Cette couche est suivie d'une couche de sous-échantillonnage qui utilise la mise en pool maximale (avec une taille de noyau 3 x 3) pour réduire l'image au tiers de sa taille.
- Les deuxième et troisième couches convolutives, qui produisent 64 images de 16 x 16 pixels, suivies d'une couche de sous-échantillonnage avec une taille de noyau 2 x 2.
- Les quatrième et cinquième couches convolutives produisent 128 images de taille 8 x 8 pixels et utilisent la mise en commun maximale avec le noyau 2 x 2.
- Les sorties sont données à une couche cachée entièrement connectée qui a 1024 neurones. Enfin, le réseau a six ou sept nœuds de sortie (un pour chaque expression qui produit leur niveau de confiance) qui sont entièrement connectés à la couche précédente.

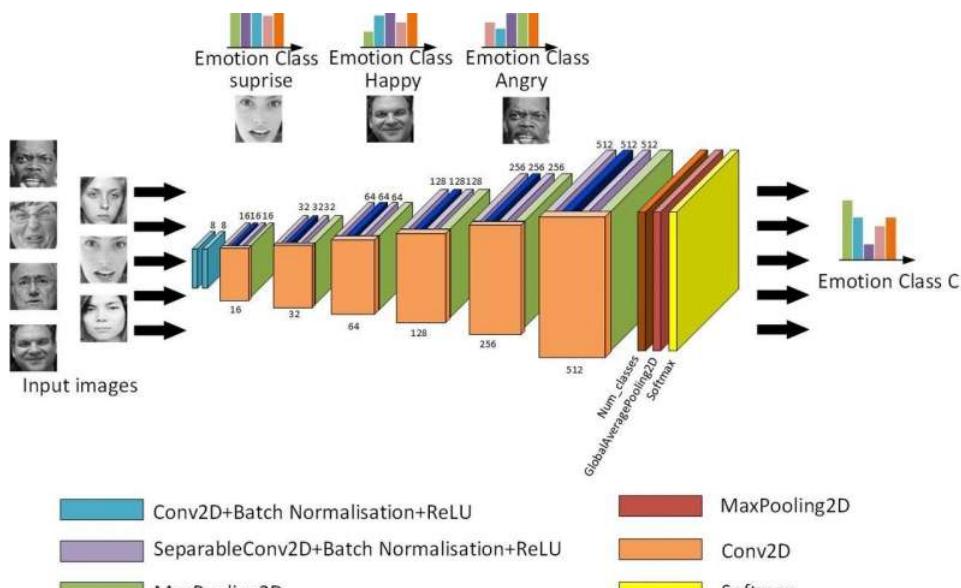


FIGURE 22 – Reconnaissance des expressions faciales

4.3.4 Module reconnaissance d'activités de fraudes

Dans ce module, nous mettrons des réseaux de neurones convolutifs à fin de classifier les scènes d'examen en **5 classes**.

Les données de notre base sont traités et ramenées à la dimension **224 X 224** à l'aide du module **image** de la librairie **preprocessing** de keras tensorflow.

Pour la mise en oeuvre de notre modèles nous utilisons le modèle **ResNet50** intégrer à **keras**, à fin de bénéficier du **transfer Learning** en utilisant les poids pré-entraînés **imagenet** sur le dataset ImageNet. Nous notons que Les réseaux pré-entraînés de pointe inclus dans la bibliothèque principale de Keras représentent certains des réseaux de neurones convolutionnels les plus performants sur le défi ImageNet au cours des dernières années. Ces réseaux démontrent également une forte capacité à généraliser à des images en dehors de l'ensemble de données ImageNet via l'apprentissage par transfert, comme l'extraction de caractéristiques.

Nous remplaçons les couches denses par les nôtres, en définissant deux couches denses.

- Les sorties sont données à une couche cachée entièrement connectée qui a 512 neurones.
- Enfin une dernière couche **full connected** donc la sortie a cinq noeuds (un pour chaque activité qui produit leur niveau de confiance) qui sont entièrement connectés à la couche précédente précédé par une couche **dropout** à fin d'éviter l'overfitting.

Les labels ont été préalablement encodé par un encodage **one-hot encoding**

Difficultés de mise en oeuvre des modules	
modules	niveau difficulté
Réconnaissance faciales	très difficiles (domaine recherche très actif)
détection pose de la tête	difficiles (les cibles mobiles)
Reconnaissance des expressions faciales	difficile (cibles mobiles)
Reconnaissance des activités frauduleuses	difficile (de plus très peu de données)

TABLE 5 – Niveau de difficulté associé aux modules

4.4 Evalution du modèle

Pour évaluer nos modèles, nous utiliserons la précision (accuracy) et la matrice de confusion.

Predicted			
Actual	Negative	Negative	Positive
	Positive	FN	TP

TN = True Negative
 FP = False Positive
 FN = False Negative
 TP = True Positive

FIGURE 23 – Reconnaissance des expressions faciales

- **TN** : le nombre d'échantillons de classe négative devrait être de classe négative.
- **FP** : le nombre d'échantillons dont la classe négative est prédite comme classe positive.

- FN : le nombre d'échantillons de classe positive devrait être de classe négative.
- TP : le nombre d'échantillons pour la prédiction de la classe positive comme classe positive.

La mesure de la performance qui détermine le pourcentage d'échantillons correctement classifié par le classificateur, se définit comme suite :

$$\text{performance} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{erreur} = \frac{FN + FP}{TP + FN + FP + TN}$$

La solution étant définie et implementée, nous devrons analyser.

5 Expérimentation et résultats

Dans cette section, nous analysons les résultats de nos quatres modules, solution à notre problème.

5.1 Module de reconnaissance faciale

Nous présentons un tableau récapitulatif des précisions obtenues pour nos quatres modèles de classification :

- SVM(SVC) : Supports Vectors Machine
- KNN : K-Nearest Neighbors
- DTC : Decision Tree Classifier
- BNC : Bayes Naive Classifier

Comparaison des modèles de classification		
Modèles	Précision Train)	Précision Test
SVM	98%	96%
KNN	89%	96%
DTC	100%	46%
BNC	100%	58%

TABLE 6 – Précision module de reconnaissance faciale

OBSERVATION : Nous constatons que bien les modèles DTC et BNC donnent une performance égale à l'unité, leur précision sur les données demeure médiocres.

Ainsi le modèle KNN est celui qui rivalise le mieux avec notre modèle SVM, alors nous pouvons utiliser soit l'un ou l'autre pour notre module.

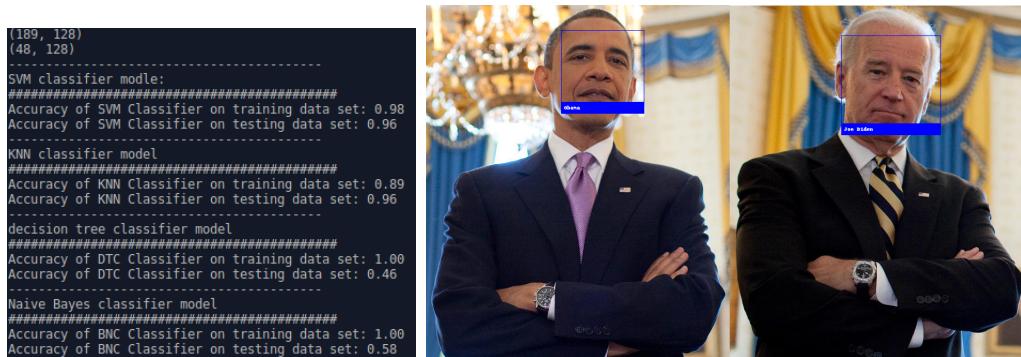


FIGURE 24 – Précision modèle de reconnaissance et image test

5.2 Modèle de pose de la tête (estimation du regard)

Nous présentons ici un ensemble de graphique extraire de l'entraînement de notre modèle à fin d'évaluer la précision de notre modèle.

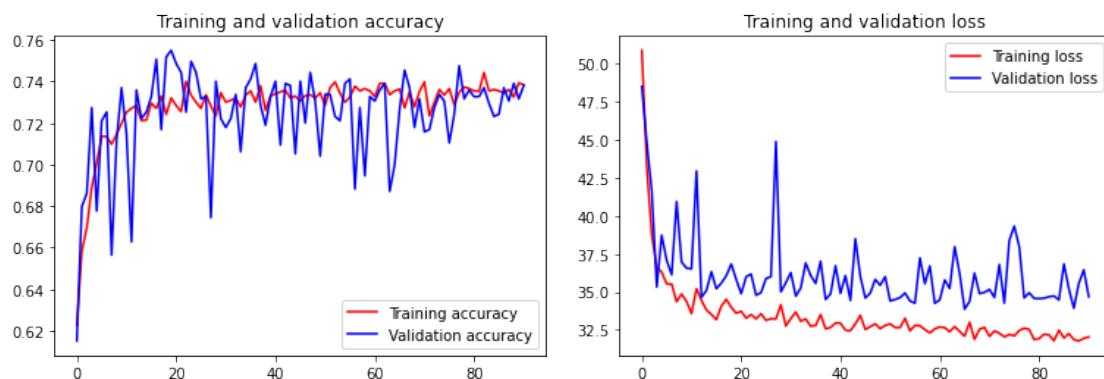


FIGURE 25 – Performance modèle d'estimation du regard

Nous constatons que la performance de notre modèle augmente progressivement au cours de l'entraînement et la courbe de validation lui est fidèle, donc ajuste le mieux les paramètres d'entraînement. La courbe de minimisation d'erreur est conséquent de celle de la performance, l'erreur diminue lorsque la performance augmente.

Notre modèle a atteint une performance de 73.85%.

5.3 Modèle de reconnaissance des expressions faciales

Nous présentons ici un ensemble de graphique extraire de l'entraînement de notre modèle à fin d'évaluer la précision de notre modèle.

Nous avons obtenu une performance en moyenne de 98%.

5.4 Modèle de reconnaissance de scénarions de tricheries

Nous présentons un tableau récapitulatif des expériences menées lors de l'entraînement du modèle.



FIGURE 26 – image estimation du regard

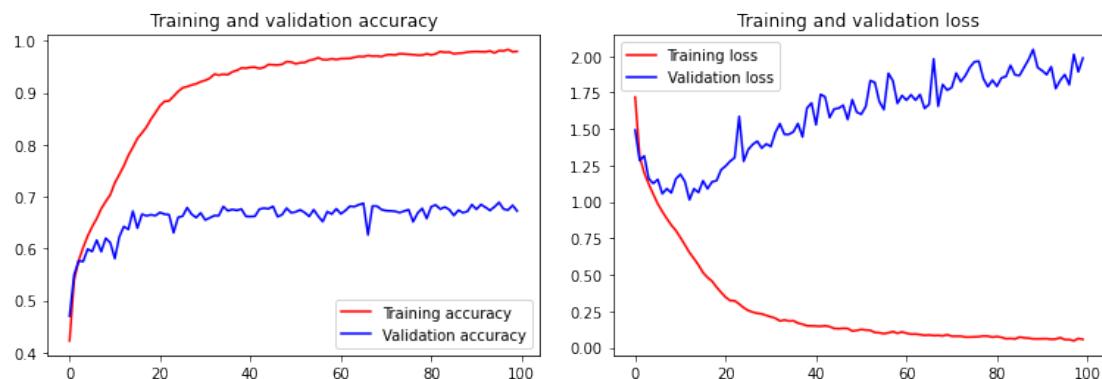


FIGURE 27 – Performance modèle de reconnaissance d'expression faciale



FIGURE 28 – image de reconnaissance d'expression faciale

experimentations du modèle		
epoch	Batch)	Performance
50	32%	90,2%
150	32	96%
200	32%	97,97%
500	16%	99,9%

TABLE 7 – Précision module de reconnaissance faciale

La carence en données n'a pas influencé négativement les performances du modèles. Mais toutefois un enrichissement de notre dataset pourra mieux améliorer nos prédictions.

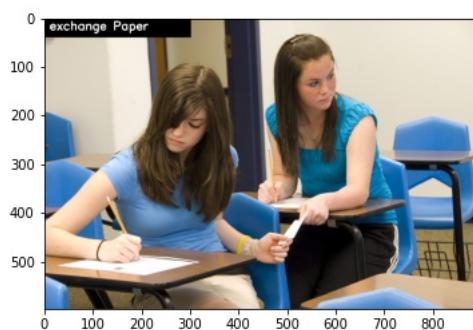


FIGURE 29 – image reconnaissance de scénario de triches

6 Conclusion

Rendu au terme de notre travail, où il était question pour nous de mettre en place un système de surveillances des examens à l'aide de la vision par ordinateur et de l'apprentissage, nous resortons de ce labeur en retenant quatre modèles à savoir : le modèle de reconnaissance faciale, le modèle d'estimation du regard à partir de la pose de la tête, le modèle de reconnaissance des expressions faciales et le modèle de reconnaissance d'activités de tricherie aux examens. Les modèles obtenus peuvent avoir une pléthora d'utilisation. Nous travail nous à permis mieux d'appréhender le monde de l'intelligence artificielle et du computer vision et les différentes challenges autour de ces domaines. Comme tous travaux scientifiques, nous restons ouverts à toutes remarques et suggestions pouvant améliorer ce travail.

7 Perspectives

Plusieurs perspectives sont envisagées :

- Entrainer encore plus et améliorer chaque modèle
- Mettre en oeuvre un système informatique de monitoring local ou à distance des examens basé sur ses modèles.
- Aller encore plus loin vers un système d'examen en ligne
- améliorer encore plus la robustesse de ces modèles pour les applications futurs

8 Références

1. Ngo Tung Son, Bui Ngoc Anh, Tran Quy Ban, Le Phuong Chi, Bui Dinh Chien, Duong Xuan Hoa, Le Van Thanh, Tran Quang Huy, Le Dinh Duy and Muhammad Hassan Raza Khan. Implementing CCTV-Based Attendance Taking Support System Using Deep Face Recognition : A Case Study at FPT Polytechnic College, 21 Feb. 2020.
2. Bui Ngoc Anh, Ngo Tung Son, Phan Truong Lam, Le Phuong Chi, Nguyen Huu Tuan, Nguyen Cong Dat, Nguyen Huu Trung, Muhammad Umar Aftab and Tran Van Dinh. A Computer-Vision Based Application for Student Behavior Monitoring in Classroom, 6 Nov. 2019.
3. Hansung Lee, So-Hee Park, Jang-Hee Yoo, Se-Hoon Jung and Jun-Ho Huh. Face Recognition at a Distance for a Stand-Alone Access Control System, 31 Jan. 2020.
4. Jirapong Manit, Luise Preuß, Achim Schweikard, Floris Ernst. Human Forehead Recognition : A Novel Biometric Modality based on NIR Laser Backscattering Feature Image using Deep Transfer Learning, Sep. 2019.
5. Loris Nanni, Sheryl Brahnam and Alessandra Lumini. Face Detection Ensemble with Methods Using Depth Information to Filter False Positives, 28 Nov. 2019
6. Elena Carlotta Olivetti, Maria Grazia Violante, Enrico Vezzetti, Federica Marcolin and Benoit Eynard. Engagement Evaluation in a Virtual Learning Environment via Facial Expression Recognition and Self-Reports : A Preliminary Approach, Dec. 2019
7. Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, Ali Etemad. ATTENTION-AWARE FUSION FOR RGB-D FACE RECOGNITION, Feb. 2020
8. Ali Elmahmudi and Hassan Ugail. Deep face recognition using imperfect facial data April 2019.
9. Zhijian Liu, Haotian Tang, Yujun Lin and Song Han. Point-Voxel CNN for Efficient 3D Deep Learning, Jul. 2019
10. Alireza Sepas-Moghaddam, Fernando Pereira and Paulo Lobato Correia. Face Recognition : A Novel Multi-Level Taxonomy based Survey, Jan. 2019
11. Dong Hoon Shin, Kyungyong Chung and Roy C. Park. Detection of Emotion Using Multi-Block Deep Learning in a Self-Management Interview App, 11 Nov. 2019
12. Lisha Chen, Hui Su and Qiang Ji. Deep Structured Prediction for Facial Landmark Detection, NIPS 2019.
13. Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin Murphy and Honglak Lee. Unsupervised Learning of Object Structure and Dynamics from Videos, NIPS 2019.
14. Zhou, L., Li, W., Du, Y., Lei, B., Liang, S. Adaptive illumination-invariant face recognition via local nonlinear multi-layer contrast feature. J. Vis. Commun. Image Represent, 2019.
15. Ran, H., Xiang, W., Zhenan, S., Tieniu, T. Wasserstein CNN : Learning Invariant Features for NIR-VIS Face Recognition. IEEE Trans. Pattern Anal. Mach. Intell, 2019.

16. Cai, Y., Lei, Y., Yang, M., You, Z., Shan, S. A fast and robust 3D face recognition approach based on deeply learned face representation. *Neurocomputing*, 2019.
17. Shakeel, M.S. ; Lam, K.-M. ; Lai, S.-C. Learning sparse discriminant low-rank features for low-resolution face recognition. *J. Vis. Commun. Image Represent*, 2019.
18. Mei Wang and Weihong Deng. Deep Face Recognition : A Survey, Jul. 2018
19. Glass, A.L. ; Kang, M. Dividing attention in the classroom reduces exam performance. *Educ. Psychol.* Jul. 2018.
20. Oscar Karnalim, Setia Budi, Sulaeman Santoso, Erico D. Handoyo and Hapnes Toba, Huyen Nguyen, Vishv Malhotra. FACE - Face At Classroom Environment : Dataset and Exploration, Nov. 2018.
21. Song Zhou and Sheng Xia. 3D face recognition : a survey, Nov. 2018
22. Setia Budi, Oscar Karnalim, Erico D. Handoyo, Sulaeman Santoso, Hapnes Toba, Huyen Nguyen and Vishv Malhotra. IBAtS - Image Based Attendance System : A Low Cost Solution to Record Student Attendance in a Classroom, Dec. 2018.
23. Xiangyu Zhu, Xiaoming Liu, Member, Zhen Lei, Senior Member, and Stan Z. Li, and Fellow. Face Alignment in Full Pose Range : A 3D Total Solution, Dec. 2017.
24. Xuanyi Dong, Shou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang and Yaser Sheikh. Supervision-by-Registration : An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors 2018.
25. Xuecheng Nie, Jiashi Feng, Yiming Zuo and Shuicheng Yan. Human Pose Estimation with Parsing Induced Learner. 2018.
26. Taiki Sekii. Pose Proposal Networks 2018.
27. Siyuan Qi, Wenguan Wang, Baoxiong Jia , Jianbing Shen, and Song-Chun Zhu. Learning Human-Object Interactions by Graph Parsing Neural Networks 2018.
28. Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. Pose Partition Networks for Multi-Person Pose Estimation 2018.
29. Feifei Zhang, Tianzhu Zhang, Qirong Mao and Changsheng Xu. Joint Pose and Expression Modeling for Facial Expression Recognition 2018.
30. Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze Prediction in Dynamic 360 degre Immersive Videos 2018.
31. Karim Iskakov, Egor Burkov, Victor Lempitsky, Yury Malkov. Learnable Triangulation of Human Pose 2019.
32. Ahmad Arinaldi and Mohamad Ivan Fanany.Cheating Video Description Based on Sequences of Gestures May 2017