



HAL
open science

Contributions à la construction de résumés Vidéos : application à la recherche d'objets génériques et à la reconnaissance faciale

Sahbi Bahroun

► **To cite this version:**

Sahbi Bahroun. Contributions à la construction de résumés Vidéos : application à la recherche d'objets génériques et à la reconnaissance faciale : Recherche à partir d'une base de vidéos Recherche d'objets génériques reconnaissance faciale à partir de vidéos reconnaissance faciale. Informatique [cs]. faculté des sciences de tunis, université de tunis el manar, 2023. tel-04081564

HAL Id: tel-04081564

<https://hal.science/tel-04081564>

Submitted on 7 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



République Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Faculté des Sciences de Tunis, Université de Tunis El Manar



Rapport de recherche présenté en vue de l'obtention de l'Habilitation Universitaire en Informatique

Par

Mohamed Sahbi Bahroun

Docteur en Technologies de l'Information et de la Communication
Maître-Assistant en Informatique à l'Institut Supérieur d'Informatique

Contributions à la construction de résumés Vidéos : application à la recherche d'objets génériques et à la reconnaissance faciale

Soutenu le 11 Février 2023 devant le jury composé de :

Ezzeddine Zagrouba	Professeur à l'Institut Supérieur d'Informatique	Président
Mohamed Hammami	Professeur à la Faculté des Sciences de Sfax	Rapporteur
Faten chaieb	Maître de Conférences à EFREI Paris	Rapporteur
Azza Ouled Zaied	Professeur à l'Institut Supérieur d'Informatique	Examineur
Vincent Charvillat	Professeur à l'Institut National Polytechnique de Toulouse	Examineur

Année Universitaire 2021/2022

Dédicace

à la mémoire de mon père

Remerciements

Je voudrais exprimer toute ma gratitude à Monsieur Ezzeddine Zagrouba, Professeur à l'Institut Supérieur d'Informatique (ISI), pour l'honneur qu'il m'a fait en acceptant de présider le jury de mon habilitation universitaire. Je le remercie également pour toute son aide, ses conseils qu'il n'a cessé de me prodiguer et toute la confiance qu'il m'a accordé.

Je tiens aussi à remercier, Monsieur Mohamed Hammami, Professeur à la Faculté des sciences de sfax (FSS), et Madame Faten Chaieb, Maître de conférences et Responsable du département informatique à EFREI Paris, pour tout l'honneur qu'il m'ont fait en acceptant de rapporter mon habilitation universitaire.

Je suis très sensible à l'honneur que me fait Azza Ouled Zaied, Professeur à l'Institut Supérieur d'Informatique, ainsi que Monsieur Vincent Charvillat, Professeur à l'Institut National Polytechnique de Toulouse, d'avoir accepté de faire partie de mon jury d'habilitation universitaire, manifestant ainsi leur intérêt à mon travail.

Ce manuscrit relate des travaux réalisés au sein du Laboratoire "Limtic" à l'ISI. Je suis très reconnaissant à tous les membres de mon laboratoire et plus particulièrement à mon collègue et ami Professeur Walid Barhoumi pour les discussions enrichissantes que nous avons partagé. Les travaux présentés dans ce manuscrit sont le fruit du travail réalisé par des doctorants que j'ai eu le plaisir de co-encadrer : Hana Gharbi et Rahma Abed. Un grand merci à vous pour le travail que vous avez accompli.

Je suis également très reconnaissant à tous mes collègues de l'Institut Supérieur

d'Informatique pour leurs encouragements, qui par un conseil ou un sourire m'ont soutenu pour mener à terme mes travaux.

Mes remerciements les plus vifs vont enfin à ma famille pour leur soutien indéfectible. Une mention spéciale pour ma femme Meriem Chekir pour tout son soutien principalement tout au long des soirées où elle a dû gérer seule les filles.

Table des matières

Dédicace	i
Remerciement	ii
Table des matières	iv
Table des figures	vii
Liste des tableaux	x
Liste des abreviations	xii
Résumé	xiii
Abstract	xv
Introduction générale	1
1 Mise en correspondance de points d'intérêts par description locale et invariants géométriques	12
Introduction	12
1.1 Approche proposée pour la mise en correspondance de points d'intérêts (MCIG)	13
1.2 Mise en correspondance de points d'intérêts par descripteur local . .	15

1.3	Mise en correspondance par invariants géométriques	17
1.4	Évaluation et résultats expérimentaux	20
	Conclusion	28
2	Construction de résumés vidéos : application à la recherche d'objets génériques	29
	Introduction	29
2.1	Extraction d'images clés basée sur la table de répétabilité (EICCTR) [1]	30
2.2	Extraction d'images clés basée sur les graphes de répétabilité (EICGR) [2]	32
2.2.1	Description générale de la méthode d'extraction des images clés EICGR proposée	33
2.2.2	Sélection de la répétabilité minimale (EICGR-1)	35
2.2.3	Classification des valeurs de répétabilité par maximisation de la modularité (EICGR-2)	37
2.3	Étude comparative des méthodes proposées d'extraction des images clés	41
2.3.1	Évaluation qualitative	43
2.3.2	Évaluation quantitative	46
	Conclusion	48
3	Construction de résumés vidéos par les descripteurs de qualité du visage	50
	Introduction	50
3.1	Extraction des images clés a base de la qualité des images faciales(KS- FQA) [3]	51
3.1.1	Détection Faciale	52
3.1.2	Classification des images	54
3.1.3	Sélection des images candidates	56
3.1.4	Calcul du score de qualité du visage	58

3.1.5	Évaluation de la méthode KS-FQA	60
3.2	Extraction des images clés à base d'un apprentissage profond et la qualité d'image faciale (FQM-CNN) [4]	62
3.2.1	Étape1 : préparation des données d'apprentissage	62
3.2.2	Étape2 : Apprentissage	65
3.2.3	Évaluation de l'architecture CNN proposée	67
3.2.4	Extraction des images clés à base de la qualité d'image faciale	70
	Conclusion	74
4	Construction de résumés vidéos par les descripteurs 2D et 3D du vi- sage : application à la reconnaissance faciale	75
	Introduction	75
4.1	Description des images faciales par la modélisation 3D du visage et la description de la texture (Deep 3D-LBP) [5]	76
4.2	Pré-traitement : Détection faciale et localisation des points de repère	77
4.3	Représentation faciale en 3D et extraction des caractéristiques	78
4.4	Phase d'apprentissage	82
4.5	Évaluation de la méthode Deep 3D-LBP	83
4.5.1	Reconnaissance des visages en présence de la variation à la pose	85
4.5.2	Reconnaissance des visages invariante en fonction de la pose et de l'illumination	87
4.5.3	Reconnaissance de visages avec variation des expressions faciales	88
4.5.4	Reconnaissance faciale sur l'ensemble de données LFW	89
4.5.5	Reconnaissance faciale sur l'ensemble de données YTF	90
	Conclusion	92
	Travaux en cours et perspectives	93
	Bibliographie	98

Table des figures

0.1	Architecture générale du système de recherche de vidéos par le contenu appliquée à la reconnaissance faciale.	10
1.1	Schéma général de la méthode de mise en correspondance de points d'intérêts MCIG [6].	13
1.2	Exemples de codes LBP circulaires identiques	16
1.3	Configurations considérés pour les correspondants candidats de chacune des deux images pour le calcul des invariants géométriques.	18
1.4	Optimisation de la valeur de précision pour la détermination de la valeur du seuil pour différents types de transformation (image "Graffiti" : changement d'angle de vue, images "boats" : couplage rotation+ changement d'échelle et "cars" : changement de luminance).	19
1.5	Image référence "Graffiti" qui subit des transformations de changement d'angle de vue (image 1, ..., image 6)	21
1.6	Résultats comparatifs en termes de nombre d'appariement trouvés lors des transformations de changements d'angle de vue successives pour l'image "Graffiti"	22
1.7	Résultat en termes de précision lors d'un changement progressif d'angle de vue	22
1.8	Image référence " boat" qui subit des transformations en couplage (Rotation+changement d'échelle)	23

1.9 Résultats comparatifs en termes de nombre d'appariements trouvés lors des transformations en couplage (rotation + changement d'échelle) . . .	24
1.10 Résultat en termes de Précision lors des transformations en couplage (rotation + changement d'échelle)	24
1.11 Image référence "cars" qui subit un ensemble de changement de luminance progressive	25
1.12 Résultats comparatifs en termes de nombre d'appariements trouvés lors des transformations de changement de luminance	26
1.13 Résultats en termes de précision lors des transformations de changement de luminance	26
2.1 Processus de classification de la table de répétabilité et d'extraction des images clés	31
2.2 Exemple illustratif d'une représentation graphique de la table de répétabilité	35
2.3 Illustration du principe de partitionnement de graphe en communautés.	37
2.4 Exemple illustratif du principe des approches agglomératives	39
2.5 Exemple Illustratif du principe des approches divisives	39
2.6 Images clés produites pour la vidéo "Foreman.mp4" par les différentes méthodes proposées que (a), (b) et (c) sont respectivement relatives aux méthodes EICCTR, EICGR-1 et EICGR-2	44
2.7 Résultats en termes de temps d'exécution de quelques vidéos choisies pour test des deux bases "OVP" et "YUV" et ce pour méthodes proposée EICCTR, EICGR-1 et EICGR-2	46
2.8 Comparaison de la qualité des résultats obtenus en termes de taux de compression	47
2.9 Comparaison de la qualité des résultats obtenus en termes de taux de PSNR (Rapport signal sur bruit)	48
3.1 Processus général de la méthode KS-FQA [3]	51
3.2 Étapes du détecteur MTCNN [7]	53

3.3	Résultats fournis par le détecteur MTCNN [7]	54
3.4	illustration de la classification par ordre de rang	56
3.5	Sélection des images candidates	57
3.6	Distance ente le centre de région (*) et le centre de masse (+)	59
3.7	Schéma de la méthode FQM-CNN [4]	63
3.8	Exemple des images de références sélectionnées	64
3.9	Module de génération des étiquettes	65
3.10	Architecture du CNN proposé [4]	66
4.1	Schéma de la méthode Deep 3D-LBP	77
4.2	Exemple d'utilisation du détecteur Dlib [8]	78
4.3	Exemple de détection de visages dans des milieux encombrés avec Dlib [8]	79
4.4	Représentation de la texture sous la forme d'un isomap	79
4.5	Un exemple de résultat de l'ajustement de point de repère [9].	80
4.6	Construction de l'image du descripteur.	81
4.7	Architecture du réseau de neurones convolutionnel proposé.	82
4.8	Exemple de de visage de la base Multi-PIE acquis dans des milieux encombrés et des conditions non contrôlées [10]	84
4.9	Exemple de de visage de la base LFW acquis dans des milieux encombrés et des conditions non contrôlées [11]	89
4.10	Exemple de de visage de la base YTF acquis dans des milieux encombrés et des conditions non contrôlées [12]	91

Liste des tableaux

0.2	Récapitulatif des publications pour l'axe 1	5
0.3	Récapitulatif des activités de (co)encadrement pour l'axe 1	5
0.4	Récapitulatif des publications pour l'axe 2	8
0.5	Récapitulatif des activités de (co)encadrement pour l'axe 2	8
1.1	Tableau comparatif des résultats obtenus en termes du temps de calcul (en seconde) entre la méthode proposée MCIG et les méthodes SIFT, SURF et PW-MATCH	27
2.1	Valeurs moyenne en termes de précision et rappel des images clés pro- duites, pour chacune des méthodes, pour toutes les vidéos choisies pour test de la base "OVP"	44
3.1	Estimation de la pose par rapport à la rotation du visage	61
3.2	Taux de précision obtenus en utilisant différentes architectures CNN	69
3.3	Classement des images de la base CVL	71
3.4	Rang du résultat pour la séquence CVL	72
3.5	Comparaison par rapport aux méthodes classiques	73
3.6	Comparaison par rapport aux méthodes profondes	73
4.1	Taux de reconnaissance (%) sous différentes poses avec la base de données Multi-PIE [10]	85

4.2	Taux de reconnaissance Recognition (%) sous différentes poses et illumination avec la base de données Multi-PIE [10]	87
4.3	Taux de reconnaissance (%) en présence des expressions faciales avec la base de données Multi-PIE [10]	88
4.4	Taux de reconnaissance (%) sur le jeu de données LFW [11]	90
4.5	Taux de reconnaissance sur le jeu de données YTF [13]	91

Liste des abréviations

CNN Convolution Neural Network. 62, 65–67, 69

ELU exponential linear unit. 68

FPS Frame Per Second. 56

FQM-CNN Face Quality Measurement using Convolutional Neural Network. 62

HoG Histogramm of Oriented Gradient. 64

LBP Local Binary Pattern. 14, 15, 64

MTCNN Multi-task Cascaded Convolutional Networks. 52

PReLU Parametric Rectified linear unit. 66–68

ReLU Rectified linear unit. 68

SIFT Scale-invariant feature transform. 14

Résumé

La description compacte du contenu d'une vidéo est actuellement une tâche difficile à cause de la très grande quantité de données qu'elle contient. La construction des résumés de vidéos est à l'heure actuelle un domaine de recherche en pleine évolution. C'est dans ce cadre que se s'inscrivent nos travaux de recherche dont l'objectif principal est de donner un aperçu des vidéos de la base sous la forme d'un ensemble d'images clés. Ce résumé sera un point de départ pour la recherche de vidéos par le contenu dans les grandes bases de vidéos. Ainsi, l'extraction de certaines caractéristiques globales et locales s'avère une tâche primordiale. Une des méthodes les plus courantes pour l'extraction d'informations locales s'appuie sur l'utilisation des points d'intérêts représentant une discontinuité. Nos travaux de recherche portent principalement sur deux axes qui sont tous les deux liés au résumé vidéo.

Dans le premier axe de recherche, nous nous intéressons principalement à la recherche d'objets génériques en utilisant une description locale à base de points d'intérêts. Cet axe concerne principalement des applications comme la recherche par le contenu dans une base de vidéos ou le e-commerce. Après avoir extrait les points d'intérêts, ils seront en second lieu décrits par des descripteurs calculés dans le voisinage de ceux ci. Après, vient une étape de mise en correspondance par description locale en ajoutant des invariants géométriques. Les différentes contraintes qu'il faut prendre en considération lors du processus d'extraction de caractéristiques locales et la mise en correspondance sont essentiellement liées avec

le mouvement de la caméra. C'est pour cette raison que les primitives détectées doivent fournir une robustesse aux différentes transformations de l'image (rotations, changements d'échelle, de point de vue, de luminosité, ... etc) pour obtenir des images représentatives de la vidéo.

Notre deuxième axe de recherche porte sur la construction de résumés vidéos dans l'objectif de la reconnaissance faciale. Cet axe peut concerner des applications comme la vidéo-surveillance ou la gestion des archives des chaînes de télévisions. Le visage n'étant pas un objet générique. Les descripteurs globaux ou les descripteurs locaux comme les points d'intérêts sont mal adaptés dans ce cas. les contraintes de son extraction dans les meilleures conditions sont plusieurs. Nous avons alors besoin d'utiliser des descripteurs spécifiques pour bien localiser, décrire et reconnaître le visage. En effet, l'extraction de visages à partir de vidéos dans des milieux encombrés reste un domaine de recherche très compétitif. L'extraction des images clés de visage, sert à définir pour chaque vidéo, l'ensemble des identités qui apparaissent. Cet ensemble devrait être formé des meilleures images de visages. C'est-à-dire, les images frontales, neutres en émotions, de bonne résolution et une quantité de lumière suffisante.

Mots clés : Recherche de vidéos par le contenu, extraction des images clés, points d'intérêts, reconnaissance faciale, apprentissage profond

Abstract

The compact description of the content of a video is currently a difficult task because of the very large amount of data it contains. The construction of video summaries is currently a field of research in full evolution. It is in this context that our research works are inscribed, the main objective of which is to give an overview of the videos of the base in the form of a set of key images. This summary will be a starting point for searching videos by content in major video databases. Thus, the extraction of certain global and local characteristics proves to be a primordial task. One of the most common methods for the extraction of local information is based on the use of points of interest representing a discontinuity. Our research mainly focuses on two axes which are both related to video summarization.

In the first line of research, we are mainly interested in the search for general objects using a local description based on points of interest. This axis mainly concerns applications such as search by content in a video database or e-commerce. After extracting the points of interest, they will be secondly described by descriptors calculated in the vicinity of these. Next, comes a matching step by local description by adding geometric invariants. The various constraints that must be taken into consideration during the process of local feature extraction and matching are essentially related to the movement of the camera. It is for this reason that the detected primitives must provide robustness to the various transformations of the image (rotations, changes of scale, point of view, luminosity, etc.) to obtain representative images of the video.

Our second line of research concerns the construction of video summaries for facial recognition. This axis can relate to applications such as video surveillance or the management of the archives of television channels. The face is not a generic object. Global descriptors or local descriptors such as points of interest are poorly suited in this case. The constraints of its extraction in the best conditions are several. We then need to use specific descriptors to properly locate, describe and recognize the face. Indeed, extracting faces from videos in cluttered environments remains a very competitive field of research. The extraction of face keyframes is used to define for each video, the set of identities that appear. This set should consist of the best face images. That is to say, frontal images, neutral in emotion, of good resolution and a sufficient quantity of light.

Keywords : Content based video retrieval, keyframes extraction, interest points, repeatability, matching.

Introduction générale

Contexte et motivations

Dans ce manuscrit d'habilitation universitaire, j'ai essayé de résumer mes principales contributions dans le domaine de la recherche de vidéos par le contenu. Après ma thèse, j'ai intégré l'équipe de recherche SIIVA du laboratoire RIADI de l'École Nationale des Sciences de l'Informatique. Mes activités de recherche se sont élargies à la construction de résumé vidéo. Puis à partir de 2016 j'ai poursuivi mes travaux au sein du laboratoire de recherche Limtic de l'Institut Supérieur d'Informatique (ISI).

Les avancées technologiques dans les domaines liés à internet ont permis aux utilisateurs d'accéder à un grand volume de données Multimédia et principalement de type vidéo. Les applications à base vidéos tels que youtube et les réseaux sociaux sont de nos jours en croissance exponentielle. Ce besoin d'accès rapide et pertinent à ces données a boosté la recherche dans le domaine de l'analyse, synthèse et indexation de vidéos. Ces applications ont pour principal objectif de fournir des techniques de plus en plus efficaces de fouille dans ce type de bases. Les bases contenant des vidéos sont caractérisées par un volume de données très grand et même gigantesque. Généralement, une seule minute d'une séquence d'un film est équivalente à 1500 images (à une fréquence de 25 images par seconde). Lorsqu'un utilisateur a besoin d'une information spécifique, la tâche de recherche manuelle est difficile et même impossible surtout avec la contrainte de temps. Ceci a poussé les chercheurs à

automatiser le processus de recherche et permettre à l'utilisateur l'accès au contenu des vidéos à partir d'une représentation compacte et structurée de ces données. La meilleure représentation était les résumés de vidéos. Ces résumés doivent contenir les informations les plus pertinentes et les plus représentatives sur les vidéos tout en réduisant la redondance. Plus ce résumé est pertinent et fidèle au contenu de la vidéo mieux celle-ci sera décrite. Dans la phase de recherche par la suite, la requête sera comparée au résumé de la vidéo et non la vidéo elle-même pour optimiser le temps de réponse de la requête.

Dans la littérature [14], les méthodes de construction de résumé vidéo sont réparties en deux grandes familles : le résumé statique et le résumé dynamique. Le résumé statique de vidéo se compose d'un ensemble d'images clés extraites des différents plans de celle-ci. Chaque image clé est soigneusement extraite et représente le contenu visuel d'une partie de la vidéo sans redondance. Tandis que le résumé dynamique est considéré plus simple à interpréter que le résumé statique vu qu'il garde l'aspect dynamique de la vidéo ainsi que l'information audio-visuelle. Dans notre travail, nous nous intéressons principalement au résumé statique. Le résultat du résumé statique est un ensemble d'images clés qui sont faciles à visualiser et ils minimisent la complexité du processus d'indexation et de recherche. La requête qui sera introduite par l'utilisateur sera comparée à l'ensemble des images clés extraites de la base de vidéos pour optimiser le temps de recherche.

Problématique et contributions

Notre premier axe de recherche sera consacré au résumé de vidéo contenant des objets génériques. Nous nous intéressons principalement à l'extraction de primitives locales par extraction et description des points d'intérêts. Les points d'intérêts sont des primitives locales définissant une double discontinuité de la fonction d'intensité dans une image. Après avoir extrait les points d'intérêts, ils seront en second lieu décrits par des descripteurs calculés dans le voisinage de ces points. Ceci permet de faciliter l'étape de mise en correspondance des points d'intérêts entre des couples

d'images prises dans des conditions différentes. Les différentes contraintes qu'il faut prendre en considération lors du processus d'extraction de caractéristiques locales sont essentiellement liées au mouvement de la caméra. C'est pour cette raison que les primitives détectées doivent fournir une robustesse aux différentes transformations de l'image (rotations, changements d'échelle, de point de vue, de luminosité, ... etc).

Afin de récupérer les images clés sans redondance, nous allons mettre en correspondance les points d'intérêts. La répétabilité représente le taux d'appariements de points d'intérêts entre deux images. Deux images d'une vidéo qui ont un fort coefficient de répétabilité ont de fortes chances de contenir les mêmes objets. Donc, une parmi ces images sera extraite comme image clé. La majorité des algorithmes de mise en correspondance de points d'intérêts sont principalement basés uniquement sur la description locale autour du point d'intérêts. Pour ces algorithmes tels que SIFT [15] et SURF [16], deux points d'intérêts de deux images correspondent si leurs descripteurs locaux sont similaires. Dans une vidéo, les images peuvent subir différentes transformation dues aux conditions de prise de vue ou au mouvement de la caméra tels que (rotation, translation, échelle, luminosité ..etc). Alors, la description locale autour du point d'intérêt ne suffit plus pour garantir la bonne mise en correspondance des points d'intérêts de deux images. La mise en correspondance par descripteur local peut être erronée à cause de toutes les transformations subites d'une image à une autre dans une vidéo. Nos premières contributions dans cet axe portent sur la proposition d'une nouvelle méthode d'appariement de points d'intérêts à base de descripteurs locaux et ajout d'invariants géométriques pour donner plus de robustesse à cette mise en correspondance.

L'étape suivante consiste à utiliser cet algorithme d'appariement afin de regrouper les images similaires, celles ayant un fort taux de répétabilité et de les représenter par une seule image clé.

Pour ce faire, nous avons présenté une première méthode d'extraction d'images clés. Comme première étape, nous calculons la matrice de répétabilité qui a pour taille le

nombre d'images de chaque plan de la vidéo. C'est une matrice carré, symétrique et ayant comme diagonale l'identité. La valeur (i,j) de cette matrice est la valeur de répétabilité entre les deux images i et j appartenant au même plan. Vu que cette matrice est de très grande taille, nous commençons tout d'abord par réduire la dimension de celle ci par application de l'algorithme d'analyse en composantes principales (ACP). L'ACP est un outil extrêmement puissant de compression et de synthèse de l'information, très utile lorsque l'on est en présence d'une somme importante de données quantitatives à traiter et à interpréter. Une fois la dimension réduite, nous appliquons une classification non supervisée sur cette matrice par la méthode classification ascendante hiérarchique (CAH). La CAH permet d'identifier des groupes homogènes dans une population appelées classes. La CAH présente les avantages suivants : permet le choix de la distance, facile à utiliser et permet de choisir le nombre de groupes à partir du dendrogramme qui donne une visibilité sur les regroupements. Les centres des classes obtenues correspondent aux images clés de ce plan de la vidéo. Les résultats obtenus montrent que nous avons réussi à extraire les images clés sans redondance et sans oubli d'un contenu important de chaque plan.

La première méthode proposée calcule la répétabilité des points d'intérêts entre toutes les images d'un plan de la vidéo. Ceci engendre un temps de calcul très élevé. Dans une seconde étape, dans le but d'améliorer la complexité et le temps de calcul et la pertinence des images clés extraites, nous avons proposé une deuxième méthode d'extraction d'images clés où le traitement s'effectue seulement sur une ensemble d'images sélectionnées par la technique de fenêtrage et non pas sur la totalité des images de la séquence vidéo comme la première méthode proposée. Ensuite, la table de répétabilité est construite uniquement avec les images candidates pour toute la vidéo et non par plan. Une fois la table de répétabilité calculée, elle sera représentée par un graphe où les sommets de ce graphe sont les images et les arcs représentent les valeurs de répétabilité. Enfin, une classification du graphe par le calcul de la modularité est effectuée. Dans le graphe résultant de la classification,

les arcs doivent être groupés en intra classe (les sommets qui appartiennent au même groupe) et inter- classe (les sommets qui appartiennent au groupes différents). Le principe de la classification de graphe est d'enlever certains arcs en fonction de la différence entre les poids de ceux-ci, jusqu'à ce qu'il n'y ait pas une amélioration dans la valeur de modularité du graphe. En effet, la valeur la plus grande de modularité indique une meilleure classification. L'image candidate qui est plus proche du centre de chaque classe est considérée comme image clé. Enfin, les images clés sont organisées dans l'ordre chronologique pour rendre le résumé produit plus compréhensible. Les principales publications relatives à cet axe de recherche sont résumé dans le tableau 0.2 :

TABLE 0.2: Récapitulatif des publications pour l'axe 1

Revue internationale	Conférences internationales
[2] (Gharbi et al., 2019)	[17] (Gharbi et al., 2017) [18] (Messaoudi et al., 2017) [1] (Gharbi et al., 2016) [19] (Gharbi et al, 2016) [6] (Gharbi et al., 2015) [20] (Bahroun et al., 2014)

Cet axe de recherche a fait l'objet des activités de (co)encadrement présentées dans le tableau 0.3 :

TABLE 0.3: Récapitulatif des activités de (co)encadrement pour l'axe 1

Étudiant	Diplôme	Date et lieu de soutenance	titre du mémoire

Hana Gharbi	Thèse	27 Décembre 2018 à la FST	Recherche de vidéos par le contenu basée sur l'extraction des images clés par mise en correspondance des points d'intérêts et classification des valeurs de répétabilité
Imen Ben Youssef	Mastère	02 Juillet 2019 à l'ISI	Extraction des objets saillants pour la recherche des vidéos par le contenu
Hana Gharbi	Mastère	07 Juillet 2012 à l'ISI	Sélection et extraction des descripteurs locaux pour l'analyse des images multi-résolutions

Le deuxième axe de recherche porte sur le résumé vidéo contenant des visages. Le visage n'étant pas un objet générique, les descripteurs globaux de couleur, de texture et de forme ainsi que les descripteurs locaux comme les points d'intérêts sont mal adaptés dans ce cas. Les contraintes de son extraction dans les meilleures conditions sont plusieurs. Nous avons alors besoin d'utiliser des descripteurs spécifiques pour bien décrire et reconnaître le visage. En effet, l'extraction de visages à partir de vidéos dans des milieux encombrés reste un domaine de recherche très compétitif. L'extraction des images clés de visage, sert à définir pour chaque vidéo, l'ensemble des identités qui apparaissent. Cet ensemble devrait être formé par les meilleures images de visages. C'est-à-dire, les images frontales, neutres, de bonne résolution et une quantité de lumière suffisante. Pour cela, nous avons proposé une première méthode d'extraction d'images clés contenant des visages basée sur la qualité de l'image faciale. Afin de s'assurer que la sélection des images clés sera basée uniquement sur des régions faciales, la première étape de ce processus consiste à localiser dans une image la zone où le visage se trouve. Nous avons appliqué le détecteur Multi-task Cascaded Convolutional Networks (MTCNN) [7] qui combine la détection et l'alignement des visages et qui se base sur les CNN afin de renforcer la détection et améliorer son taux de précision. Le MTCNN fournit les

cordonnées de boîte englobante ainsi que l'image du visage détectée avec les cinq points de repère : les deux yeux, le nez et les deux coins de la bouche. Il fournit aussi un score de confiance qui représente la probabilité que le rectangle représente un visage. Ensuite, nous regroupons les images restantes par identité. Nous utilisons pour cela le réseau profond FaceNet [21] et nous extrayons à partir du rectangle englobant un descripteur global de taille 128. Par la suite, pour chaque identité nous sélectionnons les 5 images ayant le score de confiance les plus élevés générés par le MTCNN. L'évaluation de la qualité d'image faciale pour chaque image est effectuée en calculant quatre mesures calculées à partir de la zone extraite du visage et les cinq points caractéristiques qui sont : La pose, la clarté, la luminosité et la résolution. L'image ayant le score le plus élevé sera considérée comme image clé représentante de ce visage.

Nous avons proposé une deuxième méthode nommée FQM-CNN. Cette méthode a pour objectif d'extraire des images clés en se basant sur de la qualité des images faciales mais en utilisant cette fois les descripteurs de visage et les réseaux de neurones convolutionnels (CNN). Dans la première étape, nous nous intéressons à la génération d'une base d'apprentissage à partir d'un certaines bases standards de visage que nous allons étiqueter nous même vu la non disponibilité d'une base de référence. Nous construisons un ensemble de référence Template qui est composé des images représentant chacune une identité. L'image de chaque identité est celle qui a le score de confiance le plus élevé. Le reste des images pour ce visage représentent l'ensemble Probeset. Pour chaque image des deux ensembles Template et Probe, nous extrayons des descripteurs complémentaires qui fournissent une bonne description de la pose, l'expression faciale et l'illumination. Un score de qualité de chaque image de l'ensemble Probe est calculé par rapport à son image référence de l'ensemble Template. Dans la deuxième étape, nous avons proposé pour l'apprentissage un réseau profond CNN qui prend en entrée l'image Probe avec son score de qualité et donne en sortie la classe de la qualité de l'image entrée (10 classes du score de qualité). La phase d'apprentissage dure 500 itérations. L'image

clé de chaque visage est celle qui a la classe du score de qualité le plus élevé. Dans la troisième méthode que nous avons proposé dans cet axe de recherche, intitulée Deep 3D-LBP qui a pour principal objectif la reconnaissance faciale par la génération, pour chaque visage, d'un descripteur robuste face aux problèmes des expressions faciales. Premièrement, nous commençons par la détection des visages et la localisation des 68 points de repère par la méthode Dlib. Par la suite, dans l'étape de description du visage, nous commençons par appliquer une transformation en 3D du visage 2D en utilisant le modèle Surry Face Model qui a pour but principalement d'atténuer la variation de la pose des visages. Ensuite, nous utilisons comme descripteur global sur le visage 3D le mesh-LBP pour l'extraction des caractéristiques. Dans la dernière étape, nous avons proposé pour l'apprentissage un réseau profond CNN qui prend en entrée l'image descripteur générée à travers l'application du mesh-LBP sur la région 3D du visage et donne en sortie l'identité du visage avec un vecteur caractéristique de dimension 4096. C'est la première expérience, où nous effectuons un apprentissage sur des descripteurs de visage, et non des images faciales. Les résultats sont très encourageants et donnent de meilleures performances que des méthodes similaires de la littérature. Les principales publications relatives à cet axe de recherche, sont résumé dans le tableau 0.4 :

TABLE 0.4: Récapitulatif des publications pour l'axe 2

Revue internationale	Conférences internationales
[5] (Bahroun et al., 2021)	[22] (Abed et al., 2021)
[4] (Abed et al., 2020)	[23] (Abed et al., 2020)
[3] (Bahroun et al., 2020)	[18] (Messaoudi et al. 2017)

Cet axe de recherche a fait l'objet des activités de (co)encadrement présentés dans le tableau 0.5 :

TABLE 0.5: Récapitulatif des activités de (co)encadrement pour l'axe 2

Étudiant	Diplôme	Date et lieu de soutenance	titre du mémoire
Rahma Abed	Thèse	en cours	Ré-identification de visages dans des milieux encombrés par approche locale et apprentissage profond pour la vidéo surveillance
Nozha Gharnou- gui	Mastère	04 janvier 2020 à l'Eni- Car	Extraction d'images clés pour la reconnaissance faciale dans des vidéos
Rahma Abed	Mastère	14 décembre 2018 à l'ISI	Recherche de vidéos par le contenu pour la vidéo-surveillance basée sur l'extraction des images clés et la qualité de l'image faciale
Mohamed Mes- saoudi	Mastère	08 décembre 2016 à l'Eni- Car	Reconnaissance de visage dans les bases vidéos

Les contributions proposées dans ces travaux interviennent dans le cadre de la recherche de vidéo par le contenu. La Figure 0.1 présente l'architecture générale du système de recherche que nous allons suivre, dans laquelle, nous nous focalisons sur la phase de pré-traitement où à partir d'une vidéo, nous allons extraire les différents objets ou visages qui apparaissent. L'extraction du visage nécessite des traitements différents de celui de l'extraction des objets génériques. Le visage constitue un objet nécessitant des traitements spécifiques. Les résumés générés par toutes les méthodes que nous avons proposé dans les deux axes de recherche que ce soit pour des objets génériques ou pour le visage, seront un point de départ pour la recherche par le contenu dans une base de vidéos. L'utilisateur soumettra en requête une image, la recherche va se faire dans l'ensemble des résumés générés et le système doit retourner l'ensemble de vidéos où l'information cherchée se trouve (visage ou

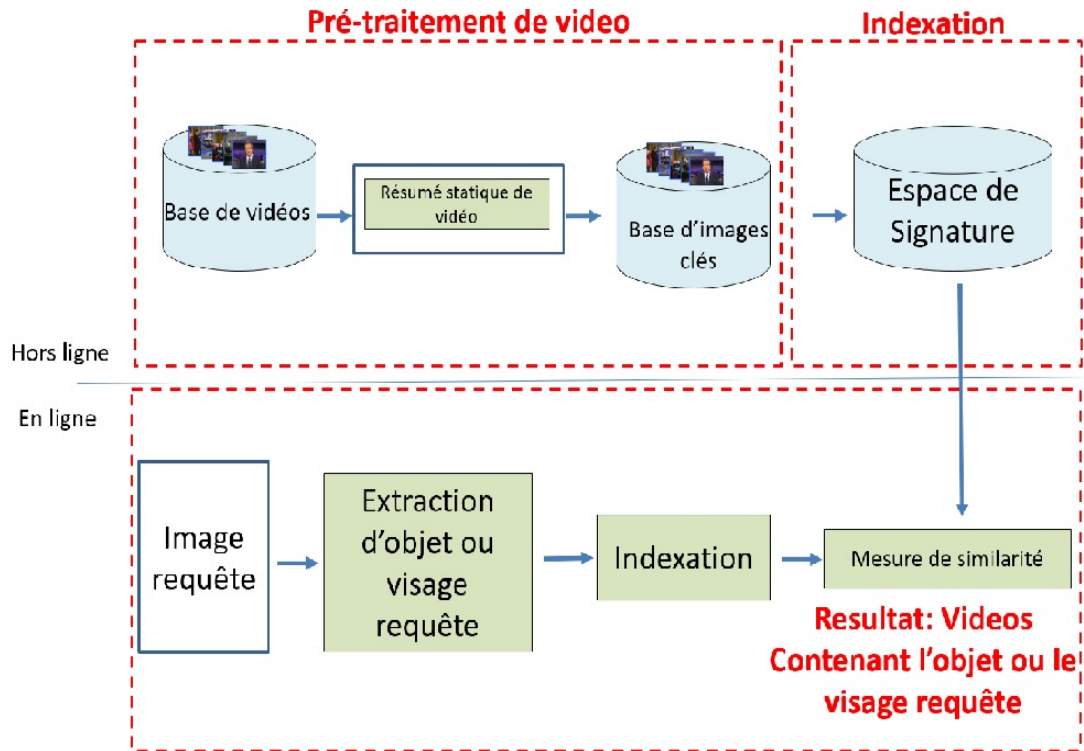


FIGURE 0.1: Architecture générale du système de recherche de vidéos par le contenu appliquée à la reconnaissance faciale.

un objet spécifique par exemple).

Structure du manuscrit

Le présent manuscrit est divisé en quatre chapitres qui sont organisés comme suit :

Le premier chapitre présente la nouvelle méthode de mise en correspondance de points d'intérêts que nous avons introduite pour mesurer la similarité entre deux images d'une vidéo. Cette méthode qui en plus d'utiliser les descripteurs locaux autour des points d'intérêts, ajoute les contraintes spatiales pour donner plus de robustesse au processus d'appariement.

Le deuxième chapitre pose le cadre théorique de nos travaux sur le premier axe de recherche pour la construction de résumés vidéos sous forme d'images clés pour faciliter la recherche d'objets génériques dans des bases de vidéos. Nous présentons

théoriquement les deux approches proposées ainsi qu'une étude expérimentale comparative entre celles ci et les méthodes de l'état de l'art.

Le troisième chapitre présente les différentes approches proposées pour l'extraction d'images clés de visages à partir de vidéos. Ces approches sont basées sur l'extraction de descripteurs de la qualité du visage. Les différentes approches proposées seront détaillées ainsi qu'une évaluation expérimentale pour chacune d'entre elle .

Le quatrième chapitre présente l'approche proposée Deep 3D-LBP pour la reconnaissance de visage à partir de vidéos. Cette méthode combine les avantages de la description de texture et de forme dans l'espace 2D et dans l'espace 3D après transformation de l'image du visage et la neutralisation de celui ci par rapport à la pose, illumination et émotions.

Nous terminerons ce manuscrit par une présentation des travaux en cours et des perspectives.

Chapitre 1

Mise en correspondance de points d'intérêts par description locale et invariants géométriques

Introduction

Notre objectif dans cet axe de recherche est de construire un résumé statique de vidéos contenant des objets génériques qui répond aux défis suivants : être fidèle à la vidéo d'origine et contenant les objets les plus représentatifs de cette vidéo tout en minimisant la redondance.

L'utilisateur introduit comme requête une image (représentant un objet) et il souhaite récupérer les vidéos contenant des images avec des objets similaires à l'objet requête. En effet, le résultat de génération du résumé de vidéo est largement lié uniquement à la robustesse de la méthode de construction du résumé utilisée mais aussi à la qualité de la description utilisée. C'est pour cette raison que nous avons, dans une première partie de cet axe de recherche, proposé une nouvelle méthode de mise en correspondance de points d'intérêts par description locale autour du point d'intérêt et l'ajout des invariants géométriques qui permettra une meilleure robustesse face aux différentes transformations que peut subir une image dans une vidéo à cause du mouvement de la caméra. Ce travail a fait l'objet des publications suivantes : [6] et [20]. Ce travail a fait l'objet de l'encadrement de Hana gharbi en Mastère et en

Thèse.

1.1 Approche proposée pour la mise en correspondance de points d'intérêts (MCIG)

Dans cette partie, nous allons présenter la contribution dans l'étape d'extraction des caractéristiques locales basée sur les points d'intérêts. Le schéma général de la méthode MCIG [6] peut être présenté dans la figure 1.2. Cette méthode est

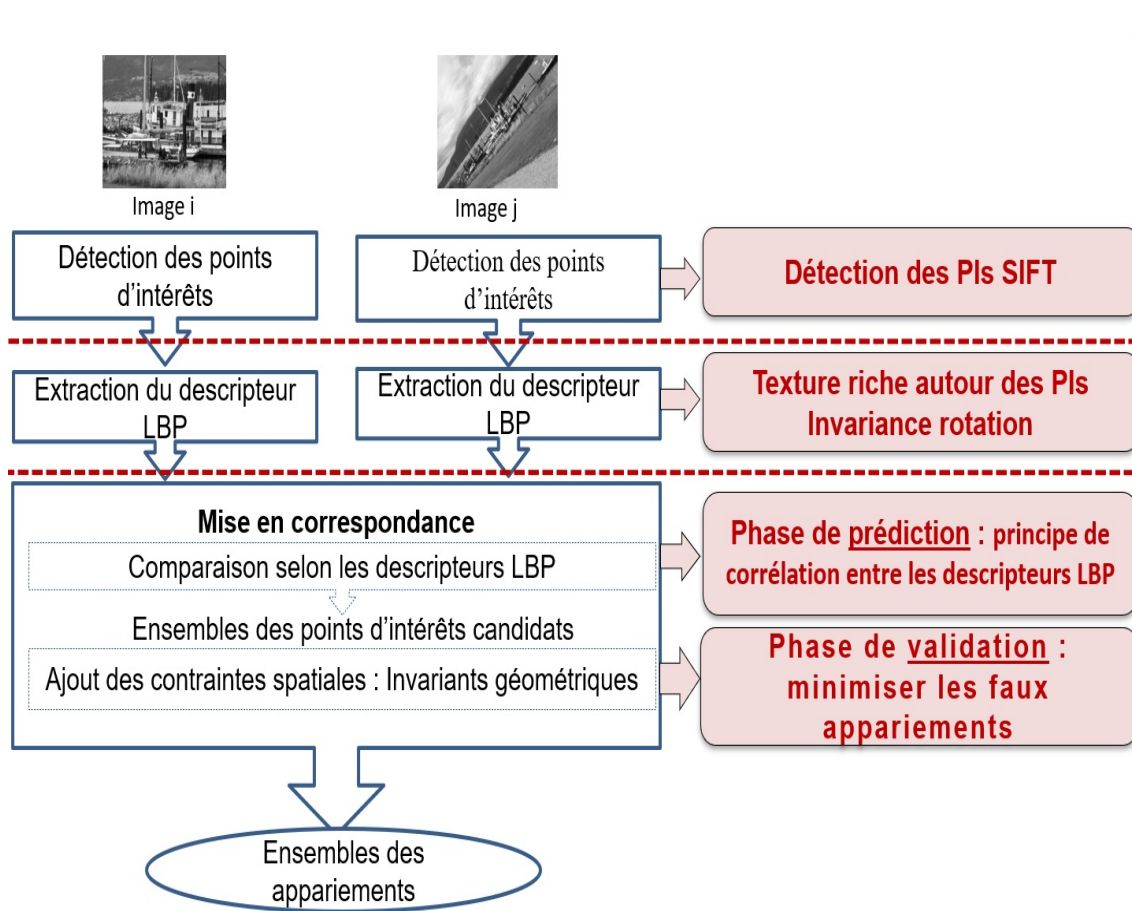


FIGURE 1.1: Schéma général de la méthode de mise en correspondance de points d'intérêts MCIG [6].

composée essentiellement des étapes suivantes :

- Nous commençons par détecter les points d'intérêts dans les deux images que nous voulons appairer afin de calculer leur similarité.

- Un descripteur LBP est calculé autour de chaque point d'intérêt. Pour ce faire, nous avons adapté le descripteur LBP, qui est essentiellement un descripteur de texture [24, 25], au contexte de points d'intérêts. Dans sa forme générique, ce descripteur est connu par sa sensibilité aux rotations [25]. Pour cette raison, nous avons adapté sa représentation pour qu'il soit invariant à la rotation. Nous avons choisi, après plusieurs expérimentations, de prendre 5 rayons de voisinage. Ainsi, le descripteur LBP sera de taille 120. Ceci a permis d'obtenir une meilleure discrimination. Ensuite, vient l'étape de mise en correspondance qui va se faire en deux étapes.
- La première étape consiste à ce que pour chaque point d'intérêt i de la première image 1, on sélectionne un ensemble de points d'intérêts candidats pour l'appariement. Cette sélection est faite sur la base de la similarité de la description locale autour des points d'intérêts dans les deux images à appairer. Un point j de l'image 2 est un candidat pour l'appariement du point d'intérêt i de l'image 1 s'il a le même code LBP que celui de i .
- Dans la deuxième étape, pour l'ensemble des points d'intérêts candidats que nous avons sélectionné dans l'étape précédente, on applique des contraintes géométriques entre chaque point d'intérêt et ses plus proches voisins. Ceci dans l'objectif de donner plus de robustesse au processus d'appariement. Un point de l'image j est correspondant au point d'intérêt de l'image i s'ils ont les mêmes relations de distance et d'angles avec leurs voisins.

Ces deux étapes doivent être précédées par une étape de détection des points d'intérêts. Une multitude de méthodes robustes ont été proposées dans la littérature pour la détection des points d'intérêts. D'après une étude comparative que nous avons effectué [6], nous avons trouvé que le détecteur SIFT (Scale Invariant Feature Transform) s'adapte mieux à notre contexte. Aussi, plusieurs études dans la littérature [26, 27, 28] ont comparé SIFT à différents détecteurs. Cette comparaison est basée essentiellement sur le critère de répétabilité qui permet l'évaluation de la stabilité des points d'intérêts détectés face à différents types de changements. Ainsi,

SIFT montre sa stabilité pour tous les types de transformations et contraintes sauf dans le temps d'exécution qui est relativement lent. Cela est dû essentiellement au grand nombre de points d'intérêts qui sont appariés comparé aux autres méthodes. On trouve qu'un nombre important de faux appariements existent [26, 29]. Ainsi, nous avons utilisé seulement l'étape de détection des points d'intérêts proposée par SIFT. Par la suite, nous avons appliqué la méthode d'extraction des caractéristiques et de mise en correspondance que nous avons proposée dans le but de minimiser, au maximum, le nombre de faux appariements.

Dans ce qui suit, nous allons présenter une description détaillée des deux étapes de description locale et de mise en correspondance des points d'intérêts que nous avons développées.

1.2 Mise en correspondance de points d'intérêts par descripteur local

Nous avons choisi l'opérateur LBP [30] comme descripteur car il décrit bien la texture et il est invariant aux variations des niveaux de gris dus au changement de luminosité dans la scène de la vidéo. En effet, puisque la texture est caractérisée par une grande variation de l'intensité, elle contient donc forcément un grand nombre de points d'intérêts. Par définition, un point d'intérêt est un pixel de l'image qui est caractérisé par une grande variation de l'intensité dans au moins deux directions.

Bien que l'opérateur LBP soit invariant par rapport aux variations des niveaux de gris, les différences sont affectées par l'échelle. Pour atteindre l'invariance par rapport à toute transformation monotone des niveaux de gris due à l'échelle, nous ne considérons que les signes des différences.

Malgré les différents avantages présentés, nous pouvons noter que l'opérateur LBP souffre d'un majeur inconvénient qui est sa sensibilité à la rotation. Dans le cas de texture, la description utilisant le code LBP circulaire [31] a montré des résultats performants en termes de capacité de discrimination. Dans notre cas, une petite taille

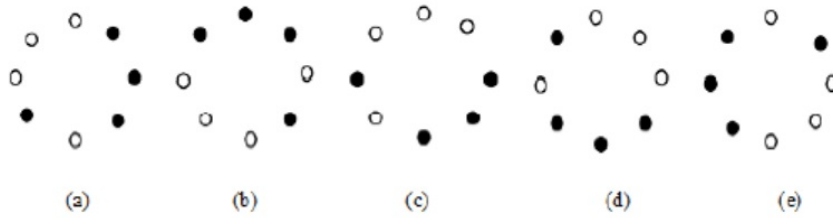


FIGURE 1.2: Exemples de codes LBP circulaires identiques

de voisinage n'est pas assez suffisante pour capter les informations caractérisant la texture locale vu la richesse du voisinage des points d'intérêts.

Dans ce contexte, nous calculons le code LBP pour 5 rayons du voisinage autour du point d'intérêt. Nous calculons le code LBP pour différents rayons de voisinages afin de garantir un descripteur riche en information locale autour des points d'intérêts. Nous avons choisi de prendre 5 rayons de voisinage après plusieurs expérimentations qui a montré un bon compromis entre robustesse de la description et complexité de calcul. Ainsi, le descripteur LBP sera de taille 120.

Pour la mise en correspondance, le descripteurs LBP de chaque point d'intérêt détecté dans la première image sera comparé aux descripteurs de points d'intérêts de la deuxième image. Cette comparaison sera basée sur leurs descripteurs LBP représentés sous leur forme circulaire, pour chaque point la comparaison se fait pour chaque rayon et son homologue.

Le résultat de cette première étape est que pour chaque point d'intérêt de la première image, nous pouvons avoir un ensemble de points candidats : ceux qui ont des descripteurs LBP similaires. Dans ce qui suit, ces candidats vont subir un autre test basé sur les invariants géométriques pour améliorer le résultat obtenu étant donné qu'une caractérisation selon des paramètres locaux uniquement n'est pas suffisante pour garantir la robustesse de la mise en correspondance.

1.3 Mise en correspondance par invariants géométriques

La sélection des plus proche voisins est une étape préparatrice pour l'étape de filtrage des points candidats correspondants par invariants géométriques. Pour chacune des 2 images à apparier, nous cherchons pour chaque point d'intérêt ses trois voisins les plus proches. Pour ceci, nous nous sommes basés sur le calcul de la distance Euclidienne entre chaque point d'intérêt et le reste des points d'intérêts dans la même image. Après la sélection des plus proches voisins pour chaque point d'intérêt, nous faisons le tri des voisins dans un ordre croissant selon la valeur de la distance euclidienne calculée.

L'ajout des contraintes spatiales, en plus de celles locales, permet d'obtenir des appariements plus robustes [32]. En effet, les points d'intérêts ne sont considérés comme correspondants que s'ils répondent à certaines contraintes spatiales. Ces dernières exigent que les points correspondants candidats entre les deux images soient conformes en termes de structure géométrique. Pour ceci, nous cherchons pour chaque point à apparier des points correspondants candidats tout en sachant que leurs voisins les plus proches respectent les invariants géométriques. C'est à dire, le point à apparier et son correspondant doivent avoir les mêmes relations géométriques avec leurs voisins respectifs. Ceci reste valable quelle que soit la transformation subite par l'image référence. Les invariants géométriques ajoutent plus de robustesse au processus d'appariement.

Les transformations les plus récurrentes que peut subir l'image sont : la rotation, le changement d'échelle et le changement d'illumination. Donc, nous pouvons considérer que le mouvement entre les deux images traitées peut être approximé par une transformation affine [9]. Or, les invariants d'une géométrie affine dans le plan présentent le rapport de longueurs entre des segments colinéaires [33]. D'où vient l'idée de prendre en considération la relation entre chaque point d'intérêt et son voisinage le plus proche des points d'intérêts. Afin de calculer ces invariants,

nous avons besoin pour chaque point d'intérêt de ses trois voisins les plus proches.

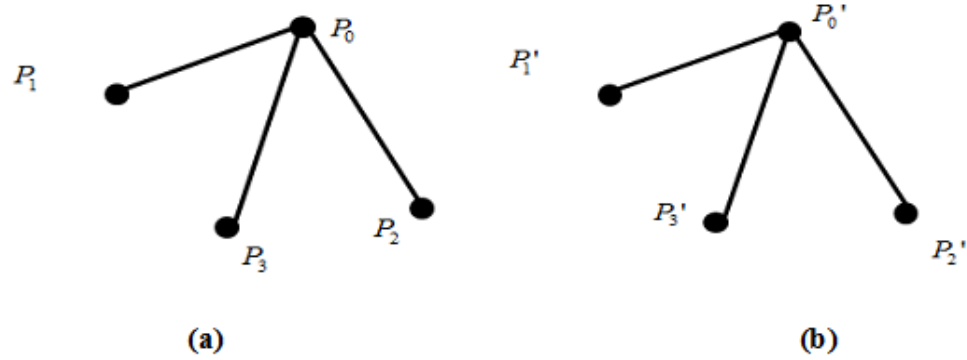


FIGURE 1.3: Configurations considérés pour les correspondants candidats de chacune des deux images pour le calcul des invariants géométriques.

Comme le montre les schémas de la figure 1.3, les invariants sont les coordonnées affines des deux points d'intérêts qu'on souhaite comparer P_0 et P'_0 respectivement à leurs trois voisins les plus proches (P_1, P_2, P_3) et (P'_1, P'_2, P'_3) .

Ces coordonnées sont définies par le système d'équations suivant :

$$\begin{cases} a_1X_1 + a_2X_2 + a_3X_3 = X_0 \\ a_1Y_1 + a_2Y_2 + a_3Y_3 = Y_0 \\ a_1 + a_2 + a_3 = 0 \end{cases} \quad (1.1)$$

Cette configuration est mise en relief de telle sorte que les invariants sont triés selon un ordre croissant. Tel que :

$$a_1 \prec a_2 \prec a_3 \text{ et } b_1 \prec b_2 \prec b_3 \quad (1.2)$$

Par la suite le processus d'appariement ne s'effectue que lors de la satisfaction d'un seuil qui a été choisi après une phase d'expérimentation sur les deux bases génériques graffiti et boats :

$$| a_1 - b_1 | \leq S \quad (1.3)$$

$$| a_2 - b_2 | \leq S \quad (1.4)$$

$$| a_3 - b_3 | \leq S \quad (1.5)$$

Le choix du paramètre S à accorder aux différences entre les invariants est une étape importante qui influence le résultat d'appariement en termes de précision. De ce fait, nous allons, expérimentalement, optimiser le choix de ce paramètre en le variant dans un intervalle de 1 à 2, avec un pas de 0.05. Nous considérons, à la fin la valeur de S qui optimise le taux de précision pour tous les types de transformations que peut subir l'image. Cette façon de choisir la valeur du seuil permet de prouver la pertinence et la fiabilité de la méthode proposée en ne permettant d'apparier que les points d'intérêts qui semblent quasi-ressemblants. Dans ce contexte, nous allons calculer la valeur moyenne de précision (pour chaque type de transformation) en fonction des différentes valeurs du seuil. En se basant sur l'influence de la valeur du seuil, nous choisissons celle qui maximise le résultat.

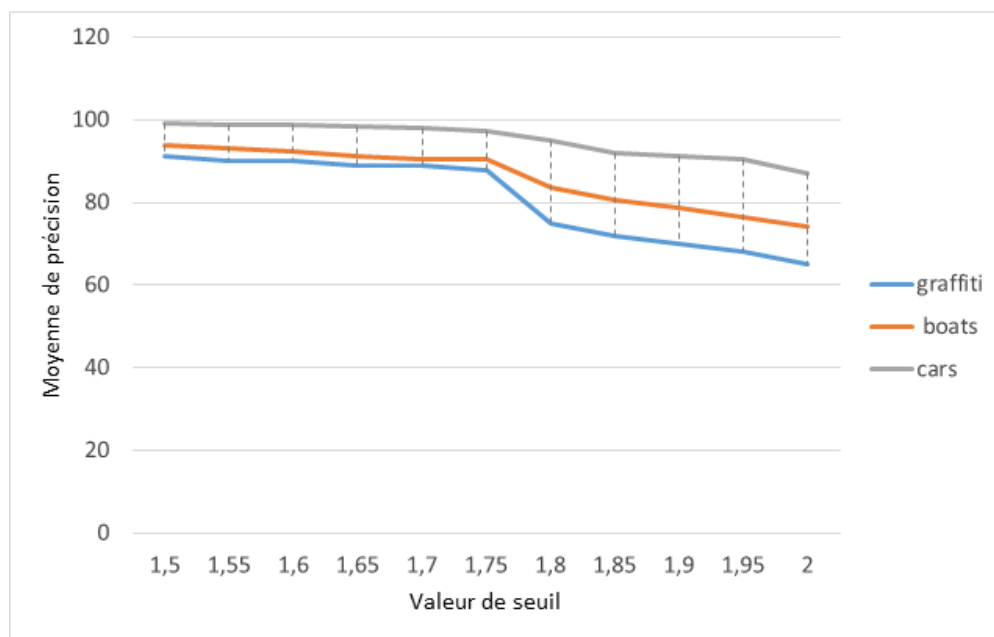


FIGURE 1.4: Optimisation de la valeur de précision pour la détermination de la valeur du seuil pour différents types de transformation (image "Graffiti" : changement d'angle de vue, images "boats" : couplage rotation+ changement d'échelle et "cars" : changement de luminance).

En regardant le graphique de la figure 1.4, nous pouvons remarquer que la précision est stable jusqu'à la valeur 1,75. Au-delà de cette valeur, la moyenne de précision commence à se dégrader. De ce fait, nous pouvons considérer que cette valeur ($S=1,75$) représente un bon compromis entre une stabilité de la précision et un

bon nombre d'appariements. En effet, une accentuation de cette valeur cause une perte de précision et une valeur inférieure de ce seuil impose une grande sélectivité dans le processus de mise en correspondance et par la suite un nombre faible d'appariements. Sachant que la valeur moyenne de précision de SIFT et SURF respectivement : 61.6 et 75.8 pour les images "Graffiti", 84 et 82.2 pour les images "boats" et 95.6 et 96.5 pour les images "cars", ce qui confirme un bon choix de la valeur de S.

1.4 Évaluation et résultats expérimentaux

Nous allons dans cette partie comparer la méthode de mise en correspondance que nous avons proposée aux trois méthodes de mise en correspondance : SIFT [15], SURF [16] et PW-MATCH [32]. Cette comparaison aura pour objectif de montrer l'apport du descripteur local LBP pour la description des points d'intérêts, ainsi que l'apport des contraintes spatiales basées sur les invariants géométriques lors de la phase de mise en correspondance. En effet, nous avons choisi SIFT et SURF car il a été démontré dans la littérature [9, 34], que ces deux méthodes restent des références dans le processus de description locale par points d'intérêts. Pour PW-MATCH aussi il a montré une bonne performance [32], du fait qu'il a été basé sur les contraintes spatiales. Pour ce faire, nous allons étudier pour ces différentes méthodes le taux d'appariement ainsi que la stabilité selon différentes transformations que peut subir une image. Plusieurs types de transformations seront étudiées : le changement de luminosité, le changement de point de vue petit angle et grand angle, le couplage rotation/changement d'échelle.

Nous présentons une évaluation des résultats de la méthode de mise en correspondance proposée sur les différentes séquences de test. Une image de référence est utilisée (image 1 à gauche de chaque séquence), celle-ci sera appariée avec les images qui la suivent respectivement dans chaque séquence. Chacune des séquences d'images est dédiée à étudier un type de transformation.

a) **Changement d'angle de vue** : Dans ce paragraphe, nous présentons les résul-

tats de la méthode d'appariement proposée pour une séquence d'images avec un changement d'angle de vue dégradé de l'angle petit vers le plus grand. Pour ce faire, nous avons pris la séquence de test "Graffiti" de la figure 1.5. Dans cette séquence, l'image initiale (image 1) subit un changement d'angle de vue. Ce changement s'accroît progressivement de l'image 2 vers l'image 6.

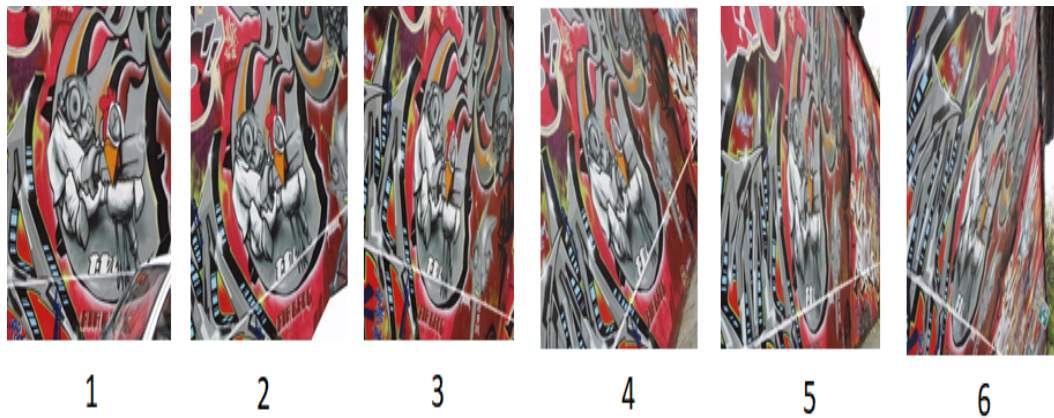


FIGURE 1.5: Image référence "Graffiti" qui subit des transformations de changement d'angle de vue (image 1, ..., image 6)

Nous présenterons respectivement dans les figures 1.6 et 1.7 les résultats du nombre d'appariements et de précision pour des changements d'angle de vue. Ces changements s'accroissent respectivement du changement numéro 1 (entre l'image 1 et l'image 2) vers le changement numéro 5 (entre l'image 1 et l'image 6).

Pour ce type de transformation, la méthode MCIG proposée surpasse nettement les méthodes SIFT, SURF et PW-MATCH en terme de précision pour les changements importants d'angle de vue. De plus, nous pouvons noter d'après la figure 1.7 que la courbe de précision de la méthode MCIG est la plus constante lors des changements : elle décroît moins rapidement d'un changement à un autre plus important. Contrairement aux courbes des autres méthodes qui sont nettement décroissantes.

En ce qui concerne le nombre d'appariements, les résultats de la méthode MCIG proposée décroissent aussi de manière plus constante que les autres

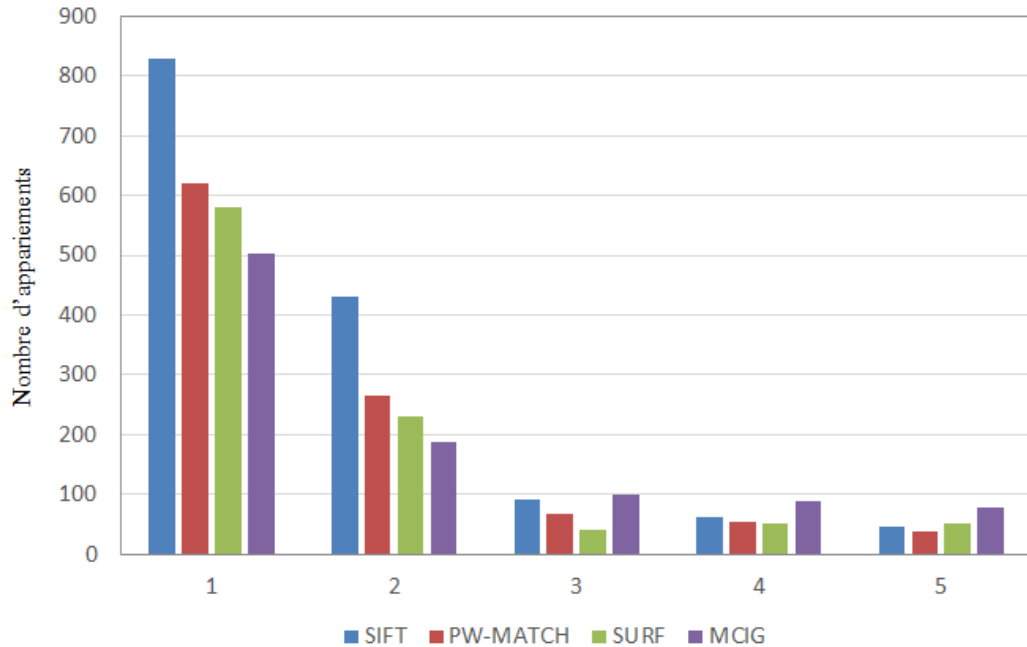


FIGURE 1.6: Résultats comparatifs en termes de nombre d'appariement trouvés lors des transformations de changements d'angle de vue successives pour l'image "Graffiti"

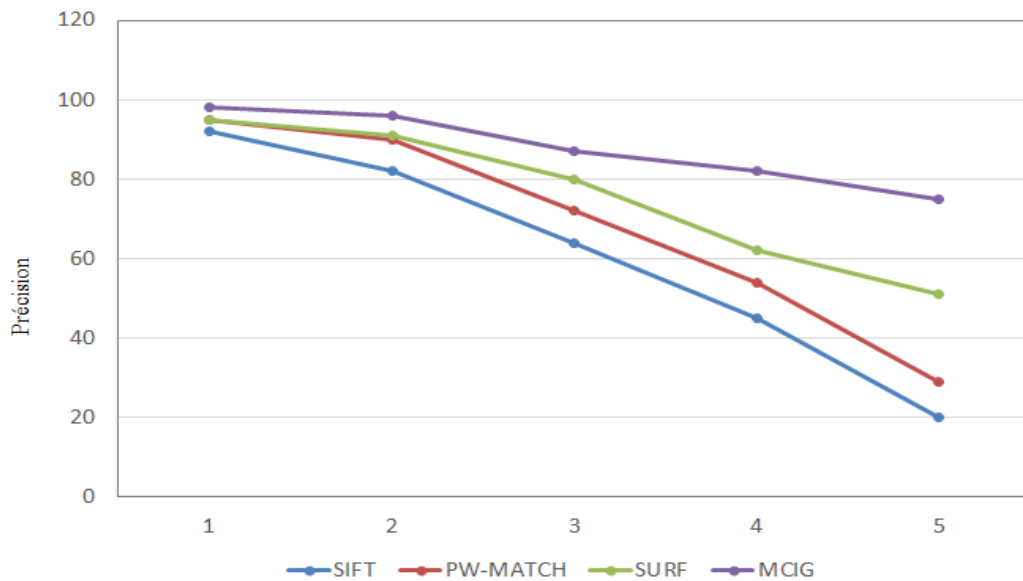


FIGURE 1.7: Résultat en termes de précision lors d'un changement progressif d'angle de vue

méthodes, d'un changement à un autre plus important, bien que ce nombre n'est pas le plus important pour la totalité des transformations.

Ces deux figures prouvent la stabilité de la méthode MCIG face aux change-

ments d'angles de vue et confirment que l'ajout des contraintes spatiales en se basant sur les invariants géométriques minimise considérablement le taux des faux appariements. Certes le nombre d'appariements est inférieur pour les petites transformations mais la précision en termes de bon appariements est toujours plus importante pour la méthode MCIG.

- b) **Couplage rotation et changement d'échelle** : Dans ce paragraphe, nous montrons la robustesse de la méthode MCIG proposée sur un ensemble d'images avec un changement en couplage de rotation et changement d'échelle. Pour ce faire, nous avons pris la séquence de test "boat" présentée dans la figure 1.8. Dans cette séquence, l'image initiale subit un changement progressif de rotation et changement d'échelle (qui s'accroît de l'image 2 vers l'image 6).

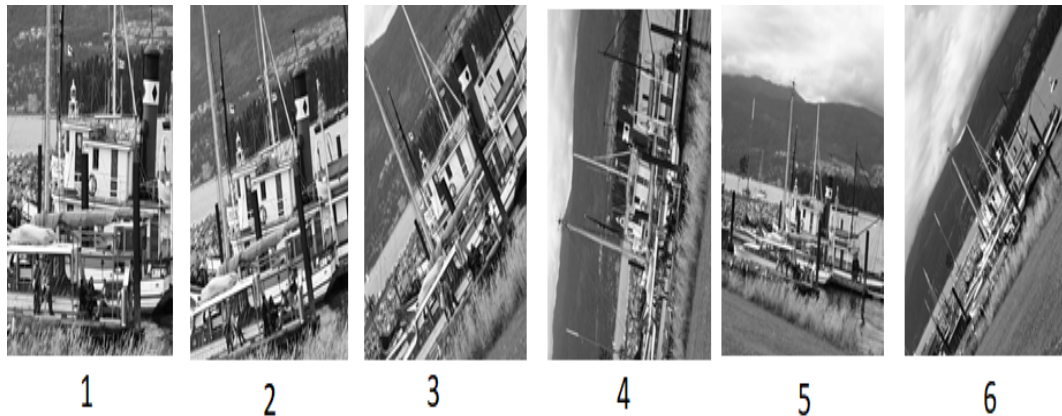


FIGURE 1.8: Image référence " boat" qui subit des transformations en couplage (Rotation+changement d'échelle)

Nous présenterons dans les figures 1.9 et 1.10 les résultats, en termes de nombre d'appariements et de précision, pour des changements en couplage (rotation +changement d'échelle) qui s'accroissent du changement numéro 1 (entre l'image 1 et l'image 2) vers le changement numéro 5 (entre l'image 1 et l'image 6).

D'après la figure 1.10, nous remarquons que la méthode MCIG proposée présente un meilleur résultat en termes de précision que les autres méthodes

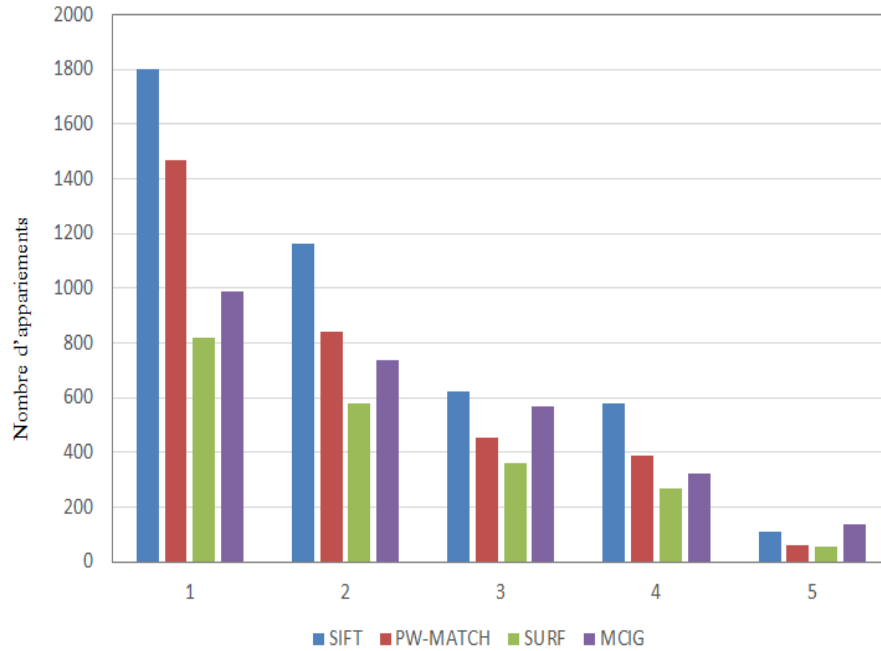


FIGURE 1.9: Résultats comparatifs en termes de nombre d'appariements trouvés lors des transformations en couplage (rotation + changement d'échelle)

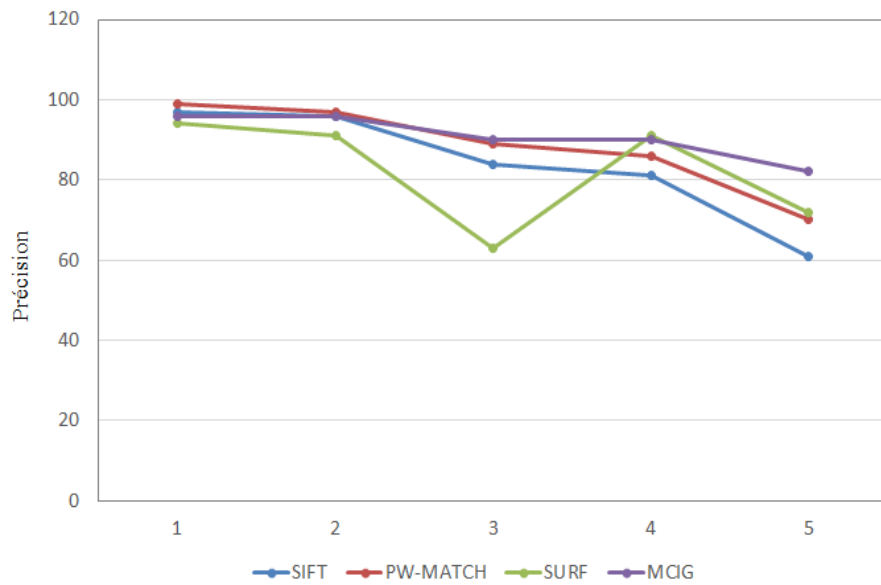


FIGURE 1.10: Résultat en termes de Précision lors des transformations en couplage (rotation + changement d'échelle)

SIFT, SURF et PW-MATCH. Ce résultat est plus clair pour les transformations les plus importantes (dans le cas où la valeur de rotation et de changement d'échelle augmente considérablement). Ceci montre davantage la stabilité des

résultats trouvés indépendamment de la transformation (petite ou grande). En ce qui concerne le résultat en termes du nombre d'appariements, il est inférieur à SIFT, légèrement inférieur que celui de SURF et PW-MATCH pour les petites transformations. Mais, MCIG donne un nombre d'appariements plus important pour les grandes transformations. En termes de précision, la méthode MCIG donne toujours des résultats plus importants que toutes les autres méthodes.

Ces résultats confirment la réussite des deux étapes d'extraction du descripteur LBP et celle de mise en correspondance par invariants géométriques qui aident à minimiser le nombre de faux appariements, tout en montrant une meilleure stabilité face aux changements en couplage (rotation et changement d'échelle).

- c) **Changement de luminosité** : Cette section présente la robustesse de la méthode d'appariement proposée à un changement de luminosité. Pour ce faire, nous avons pris la séquence de test "cars", présentée dans la figure 1.11, dans laquelle l'image initiale subit un changement progressif de luminosité (qui s'accroît de l'image 2 vers l'image 6).



FIGURE 1.11: Image référence "cars" qui subit un ensemble de changement de luminance progressive

Le graphe de la figure 1.12 montre le résultat obtenu en termes de nombre d'appariements lors de la comparaison de la méthode proposée avec les algorithmes SIFT et SURF pour des changements de luminosité qui s'accroissent

du changement numéro 1 (entre l'image 1 et l'image 2) vers le changement numéro 5 (entre l'image 1 et l'image 6). Ce graphe est suivi de la figure 1.13 qui présente une courbe comparative en termes de précision.

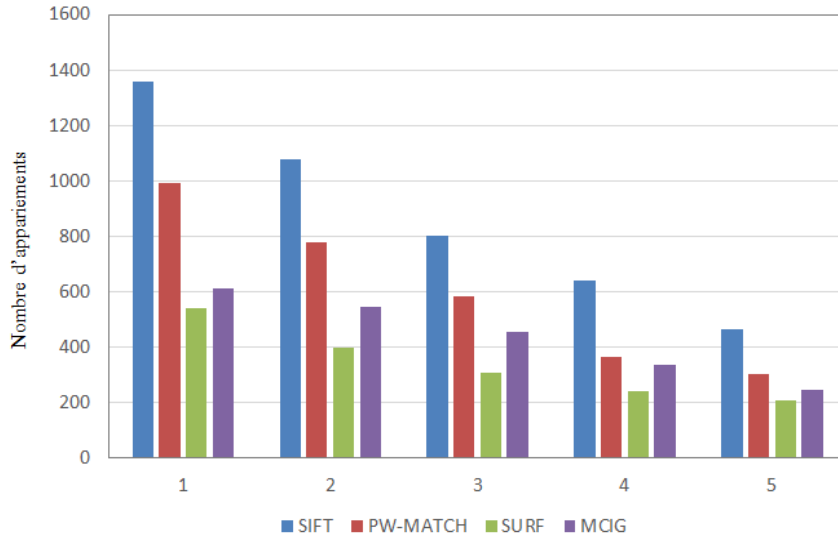


FIGURE 1.12: Résultats comparatifs en termes de nombre d'appariements trouvés lors des transformations de changement de luminance

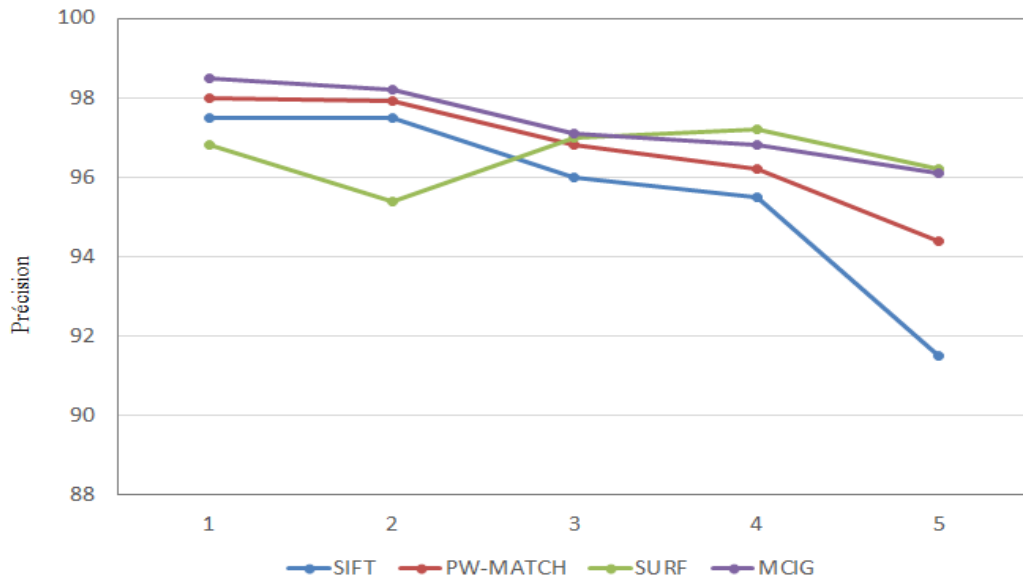


FIGURE 1.13: Résultats en termes de précision lors des transformations de changement de luminance

D'après la figure 1.13, lors d'un changement de luminance, la méthode propo-

sée MCIG présente un meilleur résultat que le SIFT, SURF et PW-MATCH en termes de précision.

Cela est plus visible pour les changements faibles de luminance. Pour les changements de luminance plus importants, le résultat est inférieur à SIFT et PW-MATCH et il est presque égal au SURF.

En ce qui concerne le nombre d'appariements, il est clair que le nombre d'appariements de la méthode MCIG est inférieur que celui des autres méthodes. Mais, il a l'avantage qu'il décroît moins rapidement et donne une valeur de précision plus importante en nombre de bons appariements.

d) Estimation du temps d'exécution

Dans le tableau 1.1, nous évaluons la robustesse de la méthode MCIG proposée en terme du temps d'exécution en comparaison avec les méthodes SIFT, SURF et PW-MATCH. Pour ce faire, nous avons pris un couple d'image ayant une transformation moyenne (image 1 et image 4) de chacune des séquences d'images (Graffiti, Boat, cars) présentées respectivement dans les figures 1.5, 1.8 et 1.11.

TABLE 1.1: Tableau comparatif des résultats obtenus en termes du temps de calcul (en seconde) entre la méthode proposée MCIG et les méthodes SIFT, SURF et PW-MATCH

Méthode	Boat	cars	Graffiti	Moyenne
SIFT [15]	2.5857	2.1471	2.7107	2.4811
SURF [16]	1.9018	2.0251	1.909	1.9453
PW-MATCH [32]	3.2321	2.5367	3.188	2.9856
MCIG [6]	2.2716	2.1105	2.2837	2.2219

A partir de la dernière colonne du tableau 1.1, nous pouvons remarquer que les résultats de la méthode proposée MCIG sont légèrement meilleurs que SIFT et PW-MACTH en termes de temps d'exécution. Ces résultats sont inférieur à SURF. Toutes les expérimentations ont été implémentées sous Microsoft Visual C++ 2010, en utilisant la bibliothèque OpenCV 2.4.3, sur un PC ayant un processeur Intel Core (TM) i5, CPU 2.50 GHZ et de 6 GB de RAM.

Conclusion

L'objectif principal de la méthode proposée dans cet axe de recherche pour la mise en correspondance de points d'intérêts est qu'elle soit appliquée sur des bases contenant des objets génériques. Ainsi, nous avons proposé une méthode robuste qui permet de donner moins d'appariements que les méthodes classiques tels que SIFT et SURF. Mais, la nouvelle méthode MCIG permet de donner plus de vrais appariements que les deux autres méthodes vu qu'on a ajouté l'étape de sélection par invariants géométriques dans le processus de mise en correspondance et ce en plus de la description locale autour du point d'intérêt. Cette méthode sera utilisée pour l'extraction des images clés à partir d'une vidéo. Dans le chapitre suivant, nous allons discuter les méthodes que nous avons proposé pour la construction de résumés vidéos pour la recherche d'objets génériques.

Chapitre 2

Construction de résumés vidéos : application à la recherche d'objets génériques

Introduction

Le résumé statique forme un aperçu général du contenu de la vidéo sous la forme d'un ensemble d'images clés sélectionnées afin d'éliminer toute redondance tout en gardant le maximum d'information. Dans ce chapitre, après avoir discuté dans le chapitre précédent la méthode de description locale adoptée, nous allons par la suite proposer trois nouvelles méthodes d'extraction des images clés à partir des valeurs de répétabilité entre les différentes images d'une vidéo. Nous allons nous baser sur la méthode MCIG [6] proposée dans le chapitre précédent pour faciliter la recherche d'objets génériques. Ce travail a fait l'objet des publications suivantes : [2],[17],[18], [1] et [19]. Ces contributions ont fait l'objet des encadrements suivants : Hana gharbi (Thèse), Imen Ben youssef (Mastère) et Mohamed Messaoudi (Mastère).

2.1 Extraction d'images clés basée sur la table de répétabilité (EICCTR) [1]

Dans ce paragraphe, nous présentons une première méthode que nous avons proposé pour l'extraction des images clés. Au début, nous calculons les valeurs de répétabilité entre toutes les images d'un plan vidéo. Nous construisons une matrice de répétabilité. C'est une matrice carré, symétrique et ayant comme diagonale l'identité. L'algorithme 1 montre comment nous avons construit la table de répétabilité pour l'ensemble des images candidates de chaque plan.

Data: T : matrice de dimension N x N

N : nombre d'images dans un plan

Result: RM : matrice de répétabilité de taille N*N

initialisation;

for $i \leftarrow 0$ **to** N **do**

for $j \leftarrow i + 1$ **to** N **do**

 // Appliquer l'algorithme d'appariement pour les deux images candidates

 // Calculer la répétabilité entre les images i et j

 T [i][j]= Répétabilité (i,j)

end

end

Fin

Algorithm 1: Algorithme de construction de la table de répétabilité pour chaque plan

La table résultante est sous la forme d'une matrice d'adjacence. C'est une matrice de dimension N*N, où N est le nombre d'images candidates sélectionnées. Les éléments non diagonales, notées r_{ij} , représentent la répétabilité entre les images i et j. Vu que cette matrice est de grande dimension car un plan vidéo contient un nombre très élevé d'images, nous avons opté pour la réduction de la dimension de cette matrice avant de passer à la classification en utilisant l'analyse en composante principale (ACP). Les centres de classes qui résulteront seront les images clés. La figure 2.1 illustre le schéma général de l'approche proposée.

De nombreux algorithmes de classification existent dans la littérature. Ils peuvent être répartis en deux grandes familles : supervisé et non supervisé. Pour les algo-

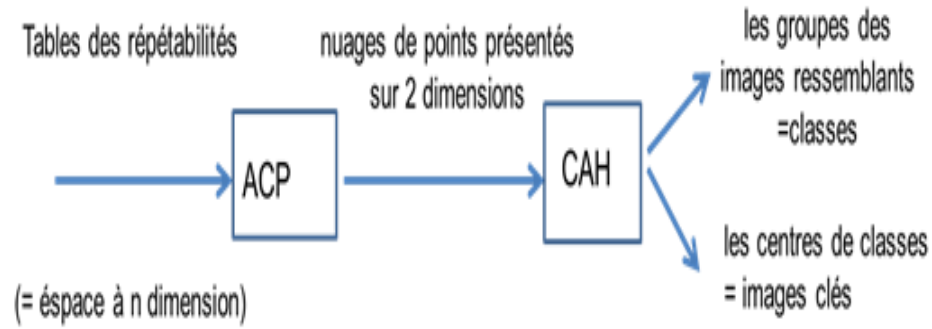


FIGURE 2.1: Processus de classification de la table de répétabilité et d'extraction des images clés

rithmes de classification supervisés, les critères de classification sont connues à l'avance ainsi que le nombre final de classes. Ceci ne répond pas à notre problème. Par contre, les algorithmes de classification non supervisés, où le nombre final de classes est déterminé après la classification ainsi que les critères de classification répond bien à notre problématique.

Parmi les techniques de classification non supervisés, nous pouvons citer la classification ascendante hiérarchique (CAH) [35]. Le but de la CAH consiste à chercher les deux individus les plus similaires et les regrouper dans une même classe d'une manière itérative. La table de répétabilité obtenue vérifie les conditions d'application de cet algorithme. Elle sera l'entrée de la CAH. L'algorithme CAH n'est pas bien adapté aux données d'un grand volume. Étant donné que la table de répétabilité est de grande dimension, cela peut engendrer un coût important, alourdir la complexité, le temps de calcul et donner des résultats pas fiables. En plus, dans les espaces de grande dimension, les intuitions géométriques valables en faible dimension peuvent se révéler fausses. C'est pour cette raison que nous avons décidé d'appliquer un algorithme de réduction de la dimension avant l'étape de classification.

Vu que nous sommes dans le cas de séquences vidéos où les images successives ne changent pas d'une façon significative, il y aura des données très proches qui sont la cause de certaines redondances. La perte d'information que peut causer la réduction de dimension de la table n'aura pas une influence significative sur la qualité des résultats. Au contraire, elle minimise la redondance.

Dans ce contexte, nous avons appliqué l'analyse en composante principale (ACP) [36] sur la matrice de répétabilité des images d'un même plan. Sa principale idée est de réduire la dimension d'un jeu de données tout en gardant un maximum d'informations. Cette technique est capable de convertir un ensemble d'observations de variables éventuellement corrélées en un ensemble de valeurs de variables linéairement dé-corrélés. Le nombre final de variables est beaucoup moins important que le nombre initial ce qui nous permet d'avoir une représentation graphique sous la forme de nuages de points. De plus, la classification dans une faible dimension est moins coûteuse que celle dans une grande dimension. C'est ce qui a motivé l'utilisation de la réduction de dimension en utilisant l'ACP. Ainsi, l'algorithme ACP facilite la visualisation et l'interprétation des données. Cet algorithme permet de présenter le tableau de répétabilité sous forme d'un nuage de points que nous pouvons afficher en 2 dimensions. Cette dimension a été choisie expérimentalement. En effet, après plusieurs tests nous avons remarqué que nous pouvons retenir deux axes étant donné qu'ils représentent presque 86 % de l'inertie totale. Ces nuages de points obtenus seront divisés en classes en utilisant la CAH. Chaque centre de classe sera finalement une image clé.

2.2 Extraction d'images clés basée sur les graphes de répétabilité (EICGR) [2]

Nous présentons dans ce paragraphe, une deuxième méthode d'extraction des images clés que nous avons proposé. Cette méthode est aussi basée sur les mesures de répétabilité. Cependant, cette nouvelle méthode réduit le nombre d'images à traiter de la vidéo pour éviter le passage par l'étape de la réduction de dimension ACP et faciliter la classification. Dans le but de faciliter la sélection des images clés, la représentation des valeurs de répétabilité sera basée sur la notion de graphe. Le schéma général de la méthode ainsi que les différentes étapes seront détaillés dans les sections suivantes.

2.2.1 Description générale de la méthode d'extraction des images clés EICGR proposée

Les graphes, appelés aussi réseaux, sont considérés comme une modélisation naturelle qui peut être associée à un grand nombre de données lors de la résolution des problèmes réels. Les graphes peuvent atteindre des centaines (ou même des milliers) de sommets à étudier. Ceci nous a encouragé à proposer une classification de graphe par maximisation du terme de modularité lors de l'étape de sélection des images clés. Nous allons construire un graphe à partir de la matrice de répétabilité. Puis, nous allons procéder à la classification de ce graphe.

Généralement, la vidéo contient un nombre important d'images. Ces images sont affichées à une fréquence de 25 à 30 images par seconde. Pour les vidéos génériques, les scènes changent normalement lentement. Ceci permet un échantillonnage de la vidéo d'entrée sans avoir un impact significatif sur le résultat du résumé. Dans le but d'éviter la comparaison des images qui sont presque similaires en terme de contenu et afin de minimiser le coût de traitement de ces images, nous avons choisi de sélectionner un certain nombre d'images parmi l'ensemble des images d'une vidéo. L'ensemble de ces images seront appelées ensemble des images candidates (Candidate Set (CS)). La technique utilisée dans la sélection de ces images est celle du fenêtrage. La première image de chaque plan de vidéo est insérée par défaut dans le (CS). Ensuite, en suivant la règle de fenêtrage, le reste des images candidates sera inclut dans l'ensemble (CS). La fenêtre que nous avons définit est de taille F . Puis les images aux positions $F+1$, $2F+1$, $3F+1$ seront extraites pour être analysées ultérieurement. La valeur de F a été fixée expérimentalement pour la valeur de la FPS (frame per second) vu que dans une seule seconde nous ne pouvons pas trouver une variation significative dans le contenu des images consécutives [37].

L'algorithme 2 décrit le processus de sélection des images candidates.

Data: Video $V=f_1, f_2, \dots, f_n$

Result: cs

initialisation;

$\text{fps} := V.\text{getFPS}()$

$i := 1$

while $i < n$ **do**

 cs.add(fi);

$i = i + \text{fps}$;

end

Fin

Algorithm 2: Algorithme de sélection des images candidates

Nous passons par la suite à l'extraction des descripteurs locaux seulement pour les images candidates de chaque plan et non pas pour la totalité des images de ce plan comme la méthode MCIG présentée dans le paragraphe précédent. En effet, cette table de répétabilité pourra être représentée par un réseau $R(X, E, d)$. Où X désigne l'ensemble des sommets de R , E est l'ensemble des arcs reliant les sommets de X et d est l'application distance définie comme suit :

$$d : E \rightarrow R \quad (2.1)$$

$$e \rightarrow d(e) \text{ qui est la distance de l'arc } e$$

Relativement à notre contexte, X désigne l'ensemble des images candidates avec $|X| = N$. Ainsi, chaque image i est un sommet i du graphe $G(X, E)$. E est l'ensemble des arcs reliant ces images et d est la répétabilité entre chaque paire d'images. Chaque sommet i est relié aux sommets $i+1, \dots, N$. Le graphe $G(X, E)$ obtenu est complet, sans circuit, possède une source (l'image 1) qui est également une racine et un puits (l'image N). Notons que les images candidates sont numérotées relativement dans leur ordre chronologique. Nous considérons l'exemple suivant pour 4 images candidates 1, 2, 3 et 4, r_{ij} est la répétabilité associée aux images i et j . Le sens de flèche est relatif à l'ordre chronologique.

Nous construisons ainsi, pour les images candidates, un réseau à partir duquel seront sélectionnées les images qui formeront le résumé de la vidéo. Pour la sélection des

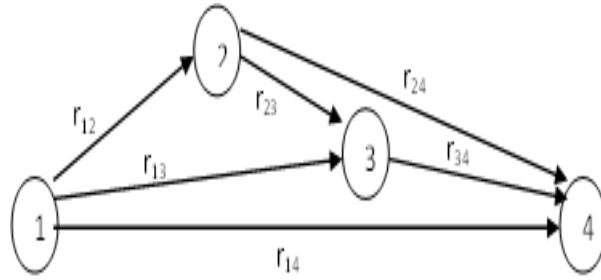


FIGURE 2.2: Exemple illustratif d'une représentation graphique de la table de répétabilité

images clés qui forment les résumés statiques, nous avons proposé deux alternatives :

- La première alternative est basée sur la mesure de répétabilité minimale. Elle est inspirée à partir du principe des algorithmes du plus court chemin. (EICGR-1)
- La deuxième alternative est basée sur la classification du graphe par calcul de modularité. (EICGR-2)

2.2.2 Sélection de la répétabilité minimale (EICGR-1)

Le principe de la méthode proposée est inspiré du principe des algorithmes du plus court chemin. Cela consiste à résoudre le problème en cherchant parmi tous les chemins possibles, vers l'objectif, celui qui donne le plus petit coût. Dans notre cas, le coût est relatif à la valeur de répétabilité. Le graphe est orienté et exige que les sommets consécutifs soient connectés par une arête orientée appropriée puisque les valeurs de répétabilité sont ordonnées dans un sens chronologique. Alors, nous devons commencer par chercher pour chaque graphe la valeur minimale de répétabilité. En effet, la valeur de répétabilité traduit la ressemblance entre les images en termes de contenu. Donc, une valeur minimale de celle-ci traduit la plus faible ressemblance entre les images et inversement.

L'idée initiale consiste à chercher le sommet, dont l'arête sortante possède le coût minimum. Cette arête conduit vers le sommet le moins ressemblant en terme de contenu (répétabilité minimale de la table). Cette valeur minimale doit être à un seuil prédéfini ($=0,2$). Ce seuil a été fixé après plusieurs expérimentations. Son

choix a été relativement strict pour garantir l'extraction des images clés ayant un contenu très hétérogène. Une fois le sommet traité, nous passons au sommet suivant. De ce fait, pas de retour en arrière. Ce qui est bénéfique pour l'élimination de la redondance. Nous présentons dans Algorithme 3 les différentes étapes de cette approche.

Data:

$T[N][N]$; // Matrice de répétabilité de dimension $N \times N$ avec N nombre de CS

$KS = \emptyset$; // Ensemble d'images clés

min ; // Répétabilité minimale de la matrice de répétabilité

$i=j=0$;

$S=0,2$;

Result: KS ;

initialisation;

while $j < N$ **do**

while $i < N$ **do**

if $T[i][j] == min$ **then**

 ajouter i dans KS ;

if ($min < S$)

then

 ajouter j dans KS ;

end

$i=j$;

else

$j++$;

end

end

$i++$;

$j=i$;

end

Fin

Algorithm 3: Sélection d'images clés pour la méthode EICGR-1

Dans la suite, nous présentons la deuxième alternative que nous avons proposé pour l'extraction des images clés à partir de la classification du graphe par maximisation de la valeur de modularité.

2.2.3 Classification des valeurs de répétabilité par maximisation de la modularité (EICGR-2)

Nous avons proposé cette alternative EICGR-2 dans le but d'améliorer la première EICGR-1 présentée dans le paragraphe précédent. Dans cette alternative, pour la sélection des images clés nous allons utiliser la classification automatique qui est considérée comme méthode de classification non supervisée. Elle permet le partitionnement d'un ensemble d'observations sous forme de classes. En effet, la classification automatique conduit à la partition d'une population initiale en un ensemble de groupes disjoints, de sorte que deux individus appartenant à un même groupe auront entre eux un maximum d'affinité et inversement deux individus appartenant à des groupes différents auront un minimum d'affinité. Ceci est effectué selon un critère bien défini selon le contexte. Dans le contexte de description locale par points d'intérêts, la répétabilité est retenue comme critère de similarité en termes de contenu.

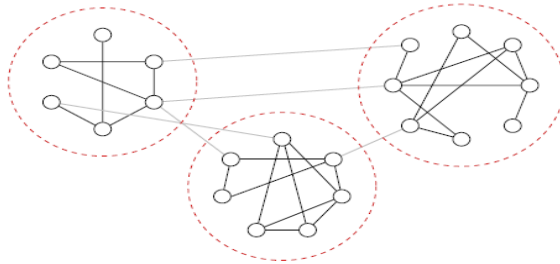


FIGURE 2.3: Illustration du principe de partitionnement de graphe en communautés.

En effet, le principe de la classification de graphes consiste à extraire des groupes de sommets, appelés communautés. Ces sommets sont connectés de façon dense et partagent essentiellement un minimum de caractéristiques avec les sommets appartenant aux restes de communautés [38]. Le but principal de la décomposition est de faciliter l'exploration et la compréhension du réseau. Le fait de se focaliser sur un nombre réduit de classes (groupes d'images) nous permet de mieux extraire les caractéristiques de chacune.

Dans la littérature, plusieurs méthodes ont été développées pour la classification

de sommets d'un graphe [39, 38]. Elles sont généralement basées sur la définition d'une mesure de similarité reliant les sommets (peut s'appuyer sur la projection du graphe dans son espace euclidien). La classification de sommets d'un graphe peut aussi se faire à travers des méthodes génératives supposant que le graphe peut être généré en se basant sur un modèle aléatoire où les densités inter-communautés et intra-communautés sont différents comme dans [40, 41]. Elles peuvent être basées sur une optimisation d'un critère de qualité de la classification, tel que l'exemple populaire de la modularité qui a été introduit dans [42]. Plusieurs revues telles que dans ([8, 43]) ont donné un panorama complet des différentes méthodes de classification de sommets composants un graphe. La mesure de modularité a été introduite pour la classification de graphes et a montré des résultats performants [44, 45].

En effet, la modularité permet de guider la recherche de la partition P . Plus spécifiquement, la modularité permet de mesurer pour chaque partition P possible une valeur $M(P)$ de modularité. Celle-ci fournit un indice sur la qualité de la partition générée. La maximisation de cette fonction M permet l'identification de la meilleure structure de communautés dans le réseau donné.

Deux catégories d'approches sont largement étudiées :

- Les approches agglomératives : appelées aussi ascendantes, comme montré dans la figure 2.4, selon lesquelles nous partons de la partition atomique (ensemble des singletons), et nous fusionnons deux communautés à chaque itération. Les communautés à fusionner sont celles qui promettent une modularité maximale. Un exemple de cette catégorie est donné dans [42].
- Les approches divisives : appelées aussi descendantes, comme montré dans la figure 2.5, selon lesquelles nous partons d'un graphe entier. À chaque itération, nous cherchons à scinder une communauté parmi celles existantes en deux de sorte à maximiser la fonction de modularité. Un exemple de cette catégorie est donné dans [45].

Notre travail s'inscrit dans le contexte des approches divisives. Ainsi, nous avons

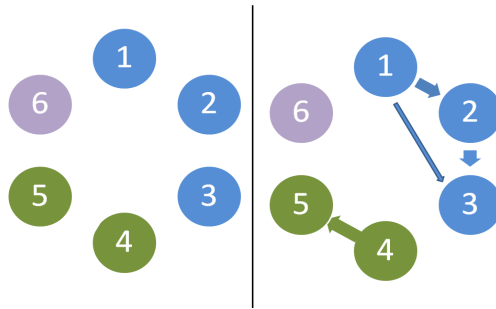


FIGURE 2.4: Exemple illustratif du principe des approches agglomératives

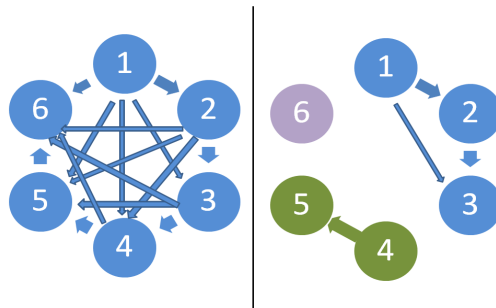


FIGURE 2.5: Exemple Illustratif du principe des approches divisives

adapté le graphe de répétabilité à ce contexte et ce pour trouver la partition optimale maximisant le critère de modularité en relation avec la répétabilité qui est le critère essentiel reliant les sommets du graphe (qui représentent les frames candidates). Pour cela, nous avons étendu le problème de maximisation de modularité pour faciliter la sélection des images clés.

Dans le graphe résultant de la classification, les arcs doivent être groupés en intra-classe (les sommets qui appartiennent au même groupe) et inter-classe (les sommets qui appartiennent au groupes différents). Le principe de la classification d'un graphe de similarité G en se basant sur la maximisation de modularité est de préserver les arcs intra-classe et supprimer les arcs inter-classe. Le principe est d'enlever certains arcs en fonction de la différence entre les poids d'arcs, jusqu'à ce qu'il n'y ait pas une amélioration dans la valeur de modularité du graphe. En effet, la valeur la plus grande de modularité indique une meilleure classification [44, 43]. Dans ce contexte, nous avons modifié la fonction de poids pour qu'elle s'adapte mieux au principe de répétabilité (1-répétabilité).

La modularité $M(c_1, c_2, \dots, c_k)$ pour la classification de graphe pour un nombre de k classes c_1, c_2, \dots, c_k est défini comme suit :

$$M(c_1, c_2, \dots, c_k) = \sum_{i=1}^k \delta_{i,i} - \sum_{i \neq j} \delta_{i,j}, \quad (2.2)$$

$$\text{Etant } \delta_{i,j} = \sum_{\{u,v\} \in E, v \in c_i, u \in c_j} w(v, u)$$

Notons que chaque arête $v, u \in E$ n'est incluse qu'au plus une fois dans le calcul. Plus la valeur de répétabilité est élevée, plus la classification est meilleure.

Data:

$G, E, W, T [N][N]$

Result: Clusters c_1, c_2, \dots, c_k

initialization;

repeat

 Sélectionner les arcs de plus grande valeur;

 Éliminer ces arcs du G ;

 Trouvez les composants connectés du G ;

 Calculer la modularité (M);

until

Aucune amélioration de la modularité sur deux itérations successives;

Obtenir les classes individuelles à partir du G final représentés par des sous graphes disjoints;

Fin

Algorithm 4: Sélection d'images clés pour la méthode EICGR-2

Les autres composants connectés du graphe final après la fin de l'élimination des arêtes représentent les classes individuelles. Sachant que T est la table de répétabilité avec N nombre de (CS). L'algorithme 4 résume les différentes étapes du processus de classification du graphe en utilisant la maximisation de la valeur de modularité. L'image candidate qui est plus proche du centre de chaque classe est considérée comme image clé. Enfin, les images clés sont organisées dans l'ordre chronologique pour rendre le résumé produit plus compréhensible.

2.3 Étude comparative des méthodes proposées d'extraction des images clés

Dans ce qui suit, Nous évaluons la qualité des résumés construits pour chacune des méthodes proposées EICCTR, EICGR-1 et EICGR-2 afin de montrer la qualité des images clés extraites. Cette évaluation sera basée sur une combinaison de critères subjectifs (la qualité) et objectifs (la quantité). Dans un premier temps, nous avons établi une évaluation subjective qui consiste à juger si le résumé généré contient des segments importants en comparaison avec le contenu de la vidéo originale. La notion de subjectivité vient en partie d'une comparaison des segments générés automatiquement en appliquant les méthodes proposées et ceux de la vérité terrain fournie par la base. Dans une seconde étape, nous passons à une évaluation objective. Cette dernière consiste à mesurer les performances des systèmes proposés tel que le temps pris pour la génération des résumés, le taux de compression et le rapport signal/bruit. Dans ce contexte, dans la comparaison subjective, nous allons comparer les résultats obtenus avec les résumés fournis par la base OVP (The Open Video Project (2016) - <http://www.open-video.org>) ainsi que quatre méthodes importantes dans la littérature qui ont utilisé la base OVP dans leurs processus expérimentaux :

- DT [46] : utilise l'algorithme de Triangulation de Delaunay pour classifier les images des vidéos puis le centre de chaque classe sera inséré dans l'ensemble des images clés.
- STIMO (STill and MOving Video Storyboards) [47] : cette méthode génère un résumé statique en utilisant l'histogramme de couleur HSV. Il utilise l'algorithme de variation moyenne FPF (Farthest Point-First).
- VSUMM (Video summarization) [48] : algorithme simple d'extraction de résumé basé sur la classification k-moyenne et l'histogramme HSV pour caractériser la couleur.
- VISCOM (Video Summarization using Colorco-Occurrence Matrices [49] :

Une méthode qui utilise les matrices de cooccurrences des couleurs dans le processus de sélection des images clés.

Pour la comparaison objective, sachant que les quatre méthodes utilisées dans la comparaison subjective se basent sur la description globale, nous allons comparer les méthodes proposées avec la méthode (parmi les quatre méthodes de la littérature citées précédemment) qui a donné un meilleur résultat dans la comparaison subjective avec 2 autres méthodes utilisant la description locale pour la génération des résumés statiques :

- Méthode proposée par [50] : Cette méthode est basée sur la distance de x^2 des histogrammes de couleur HSV et le descripteur SIFT.
- Méthode proposée par [18] : Cette méthode est basée sur la description locale utilisant le détecteur SURF et la méthode FLANN pour la sélection des images clés.

Cette comparaison va nous permettre de mettre en valeur des méthodes proposées (qui sont basées sur la description locale) par rapport à ceux de la littérature basées aussi sur la description locale. L'ensemble des vidéos que nous avons utilisé pour nos tests comprend des séquences des deux bases suivantes qui contiennent des vidéos caractérisées par un contenu diversifié (documentaire, pédagogique, conférence, dessins animés et historique). Chaque vidéo a été divisée en plans à l'aide de la méthode basée sur la distance chi carré des histogrammes [51].

- "YUV" (YUV Video Sequences - <http://trace.eas.asu.edu/yuv/>) : Cette base a été choisie pour la richesse des vidéos en termes de résolutions et contenu diversifié.
- "OVP" (The Open Video Project (2016) - <http://www.open-video.org>) : Nous avons choisi cette base vu qu'elle fournit au préalable des résumés pour chaque vidéo. Ces résumés sont considérés comme une vérité terrain "OVP summaries".

- Comparaison par rapport aux résumés d'utilisateurs (CUS) métrique [48]. Dans cette méthodologie proposée par Sandra et al., [48] puis améliorée par Vinicius et al., [49], le résumé automatique produit sera comparé avec des résumés construits par des utilisateurs (ensemble de 5 résumés produits manuellement par 5 différents utilisateurs). Les résumés des utilisateurs, dans cette méthodologie, sont considérés comme des références : vérité terrain.

2.3.1 Évaluation qualitative

Il est important de pouvoir évaluer la qualité des résumés générés automatiquement. Cependant, l'évaluation de la qualité des résumés générés est une tâche délicate. C'est l'une des parties les plus difficiles à mettre en place dans le processus de développement de méthodes de création des résumés vidéos. Il est extrêmement difficile de donner une définition formelle de ce qui est un bon résumé. Dans un premier lieu, nous allons montrer quelques exemples de résultats des méthodes d'extraction des images clé EICCTR, EICGR-1 et EICGR-2 proposées, et ce pour les bases "YUV" et "OVP".

Vu que cette dernière dispose d'une vérité terrain, ces résultats seront suivis par différentes mesures de métriques utilisés tels que le rappel, la précision et la F1-mesure. Ceci pour les résumés trouvés automatiquement à travers les différentes méthodes proposées en comparaison avec d'autres méthodes de l'état de l'art. Nous allons monter par la suite pour quelques vidéos appartenant aux deux bases de tests les résultats en termes de temps d'exécution normalisé (en seconde). Tous ces résultats nous permettrons d'établir une étude comparative subjective entre les différentes méthodes proposées. Nous présentons dans la figure 2.6 les images clés de la vidéo "foreman".

Dans la figure 2.6, nous aurons pour cette vidéo : 5, 4 et 3 images clés en appliquant respectivement les méthodes proposée EICCTR, EICGR-1 et EICGR-2. Le tableau 2.1 montre les différentes moyennes en termes de précision, rappel pour quelques vidéos choisies poest de la base "OVP", et ceux pour les méthodes



FIGURE 2.6: Images clés produites pour la vidéo "Foreman.mp4" par les différentes méthodes proposées que (a), (b) et (c) sont respectivement relatives aux méthodes EICCTR, EICGR-1 et EICGR-2

proposées EICCTR, EICGR-1 et EICGR-2 en comparaison avec d'autres méthodes de la littérature.

Comme première lecture du tableau 2.1 ainsi que la figure 2.6, nous pouvons

	Précision (%)	Rappel (%)
OVP	58.4	65.7
VSUMM	72.1	64.1
DT	54.7	43.3
STIMO	51.9	62.1
VISCOM	64,9	81,1
EICCTR	68.6	72,1
EICGR-1	62.7	58.9
EICGR-2	66.1	79.8

TABLE 2.1: Valeurs moyenne en termes de précision et rappel des images clés produites, pour chacune des méthodes, pour toutes les vidéos choisies pour test de la base "OVP"

remarquer que les valeurs obtenues pour les méthodes proposées pour l'extraction

des images clés, sont en général bonnes en comparaison avec le reste des valeurs résultantes des autres méthodes existantes dans la littérature. En effet, il est très clair que les résultats des méthodes proposées EICCTR, EICGR-2 surmontent les méthodes existantes en termes de précision et de rappel. Les résultats de la méthode EICGR-1 sont bonnes aussi par rapport à la littérature (meilleures que DT et STIMO et presque similaires à OVP et VSUMM) mais légèrement inférieures à la méthode existante VISCOM et aux deux autres méthodes proposées EICCTR, EICGR-2.

En ce qui concerne le temps de calcul, la figure 2.7 montre que la méthode EICGR-2 est la moins coûteuse en termes de temps d'exécution. En second lieu, arrive la méthode EICGR-1 puis EICCTR. Ainsi, nous pouvons noter que cela est dû au fait que le traitement dans la méthode EICCTR s'effectue sur toutes les images de la vidéo en plus du passage par le découpage de la vidéo en plans et le passage par la réduction de dimension. De même pour la méthode EICGR-1 qui vient en deuxième place, cela peut être dû au passage par l'étape de segmentation de la vidéo en plans en plus de celle de la recherche de la valeur minimale pour chaque table relative aux différents plans. Les résultats de la méthode EICGR-2 prouvent davantage que le traitement de toute la vidéo sans passage par décomposition par plans et par la suite la génération des images candidates pour toute la vidéo, en plus de la classification utilisant le critère de modularité ont minimisé considérablement le temps de calcul sans affecter la qualité des résultats.

En conclusion, d'après le tableau 2.1 et la figure 2.7, il est clair que la méthode EICGR-2 présente un bon compromis entre la qualité et le temps d'exécution. La méthode EICCTR donne aussi des bons résultats mais avec un coût un peu plus élevé en termes de temps d'exécution et de complexité. Pour remédier à ce problème, nous avons proposé EICGR-1 qui a minimisé ces deux contraintes mais nous avons perdu en termes de qualité. Ainsi, EICGR-2 a été proposée pour améliorer à la fois le temps de calcul, la complexité ainsi que la qualité.

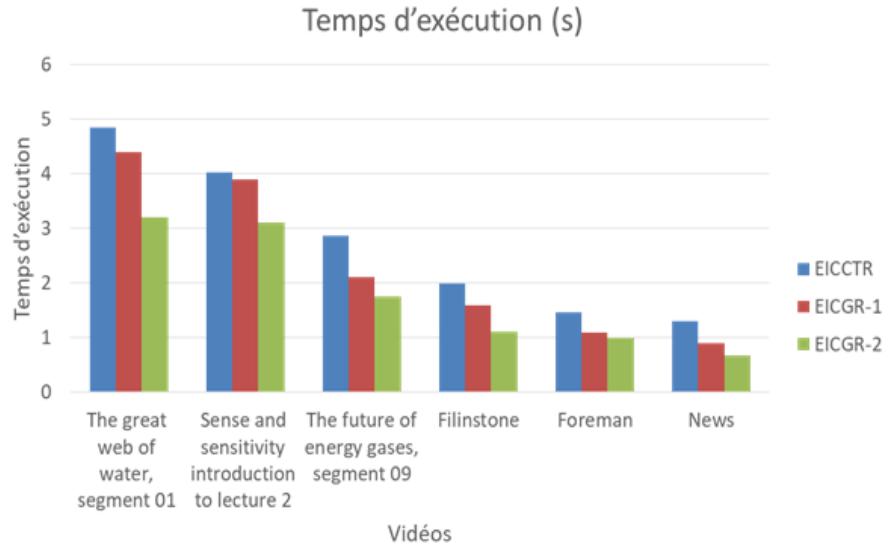


FIGURE 2.7: Résultats en termes de temps d'exécution de quelques vidéos choisies pour test des deux bases "OVP" et "YUV" et ce pour méthodes proposées EICCTR, EICGR-1 et EICGR-2

2.3.2 Évaluation quantitative

Afin d'évaluer quantitativement les résultats des résumés statiques produits par les trois variantes des méthodes proposées, nous déterminons les métriques d'évaluation de taux de compression (CR%) et de rapport signal bruit (PSNR).

Le résultat de l'extraction d'images clés doit être compact afin d'éviter la redondance. Dans ce contexte, nous avons utilisé le taux de compression pour vérifier ce critère. Nous avons calculé le PSNR pour chaque couple (F_u, F_v) d'images clés de taille $(M*N)$ extraites. Ensuite, nous considérons la moyenne des PSNRs pour chaque vidéo. Plus que les images clés F_u et F_v sont similaires, plus la valeur du PSNR est élevée. Les valeurs importantes du PSNR reflètent une redondance des images clés extraites et les valeurs réduites indiquent leur diversité. Nous comparons ainsi les valeurs obtenues avec trois autres méthodes : celle fournie par la méthode de VISCOM [49] (qui contribue avec des résultats importants dans le domaine d'extraction d'images clés et qui a donné un meilleur résultat en termes de qualité par rapport aux différentes méthodes testées dans la littérature) et celles proposées par Tapu et al., [50] et Messaoudi et al., [1] qui sont basées sur la description locale. Les résultats obtenus sont reportés dans les figures 2.8 et 2.9.

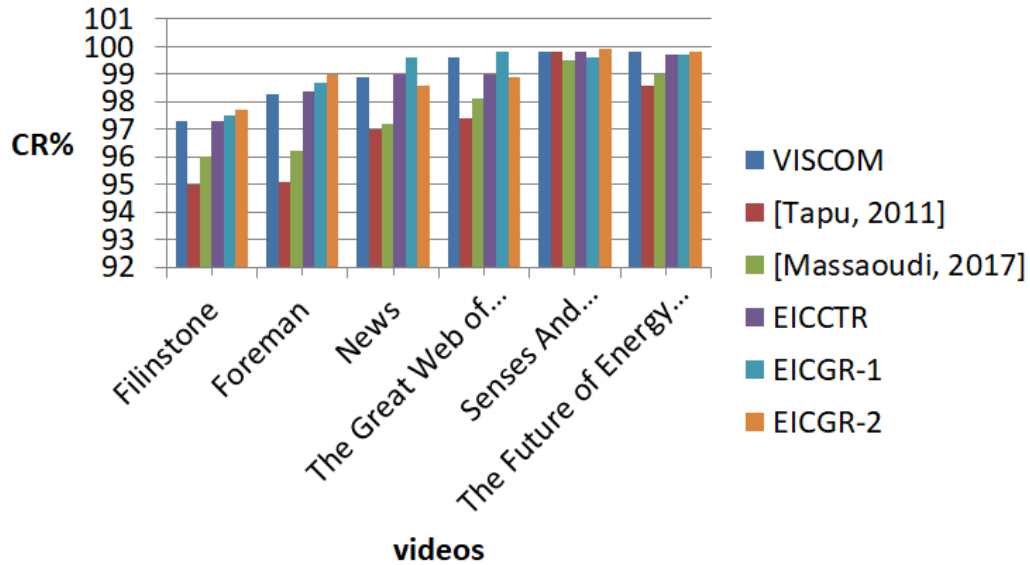


FIGURE 2.8: Comparaison de la qualité des résultats obtenus en termes de taux de compression

A partir de la figure 2.8, nous constatons que les méthodes que nous avons proposées fournissent des taux de compression élevés (toujours supérieurs à 97.3%). Ces taux sont pour la majorité de vidéos supérieures (parfois égaux) au taux trouvé pour VISCOM qui se base essentiellement sur la description globale. Cependant, ils sont largement supérieurs aux méthodes proposées par Tapu et al. et Massaoudi et al. qui sont basées sur la description locale. En effet, une valeur du CR% élevée indique que les images clés produites sont différentes et par la suite une réduction considérable dans la redondance des images clés extraites.

Ces résultats confirment davantage notre supposition initiale qui considère que la description locale par points d'intérêts est une bonne solution pour l'extraction des images clés, vu la robustesse du processus d'extraction des points d'intérêts face à différentes transformations.

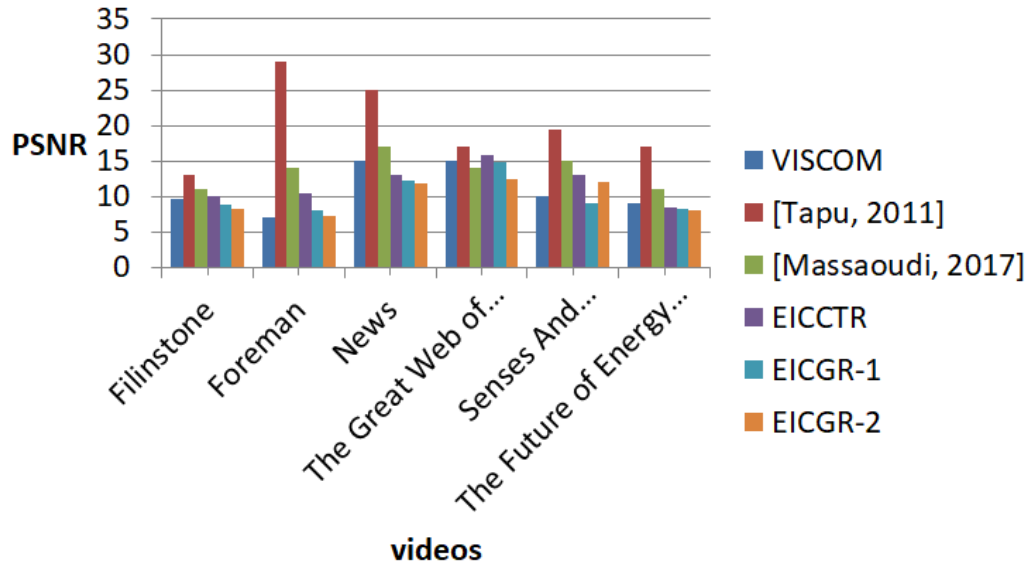


FIGURE 2.9: Comparaison de la qualité des résultats obtenus en termes de taux de PSNR (Rapport signal sur bruit)

Dans la figure 2.9, nous montrons une comparaison des méthodes proposées avec celles de la littérature en termes de PSNR. Une première lecture du graphique, permet de remarquer la réussite du processus d'extraction. En effet, les images clés sont considérées similaires, si la valeur du PSNR est élevée. Inversement, des valeurs réduites indiquent une diversité de ces images clés en termes de contenu. Il est clair que les valeurs enregistrées de PSNR sont faibles. Ceci est confirmé en comparant les valeurs relatives aux méthodes que nous avons proposées avec ceux des autres méthodes de la littérature. Ces résultats confirment que notre méthode EICGR-2 extrait les images clés les plus significatives et pertinentes. Ceci favorise la minimisation de la redondance.

Conclusion

L'objectif principal des méthodes proposée dans cet axe de recherche pour l'extraction des images clés, est qu'elles soient appliquées sur des bases contenant des objets génériques. Ainsi, nous avons atteint notre objectif initial, qui consiste à définir un ensemble d'images clés, qui seront considérées comme les informations

les plus importantes contenues dans une vidéo. Ceci a été réalisé en tirant profit des avantages de la description locale et la mise en correspondance de points d'intérêts. Les images clés extraites faciliteront la recherche de ces vidéos par la suite. Dans le chapitre suivant, nous allons discuter le deuxième axe de recherche qui concerne la construction de résumés vidéos pour la reconnaissance faciale.

Chapitre 3

Construction de résumés vidéos par les descripteurs de qualité du visage

Introduction

Notre objectif dans ce deuxième axe de recherche est de concevoir un mécanisme pour représenter chaque vidéo par l'ensemble des images faciales des identités qui y apparaissent principalement dans des milieux encombrés où on ne contrôle pas la pose du visage, les conditions d'illuminations et les émotions. Dans une vidéo, le visage peut subir plusieurs transformations dus au mouvement de la caméra qui seront un frein à son identification. Ces transformations peuvent être la variation la pose du visage, changement des conditions d'illumination, l'illumination, les occultations et surtout les émotions. Pour ce faire, tout au long de cet axe de recherche, nous allons proposer de reconnaître des visages à partir de vidéos en essayant de répondre aux différents défis dues aux transformations que peut subir le visage dans une vidéo. C'est à dire, parmi l'ensemble des visages d'une identité dans une vidéo, quelle est l'image de ce visage avec la meilleure position frontale, dans de bonnes conditions d'illumination et neutre en émotions. Nous allons, dans un premier temps, construire un résumé de vidéos avec tous les visages qui apparaissent. Ce résumé servira, dans une deuxième étape pour la phase de recherche. A partir d'une base de vidéos, nous allons essayer de récupérer toutes les vidéos où une

personne mise en requête apparaît. Au cours de chapitre, nous allons proposer deux nouvelles méthodes d'extraction des images clés en se basant sur les descripteurs de la qualité du visage. Ce travail a fait l'objet des publications suivantes : [4], [3] et [23]. Ce travail correspond à l'encadrement en thèse et en mastère de Rahmed Abed. Ainsi que l'encadrement en mastère de Nozha Gharnougui et Mohamed Messaoudi.

3.1 Extraction des images clés a base de la qualité des images faciales(KS-FQA) [3]

Dans les méthodes proposées dans cet axe de recherche nous tirons profit des avantages des descripteurs de la qualité d'image faciale. Le processus général de la méthode proposée KS-FQA [3] est présenté dans la Figure 3.1.

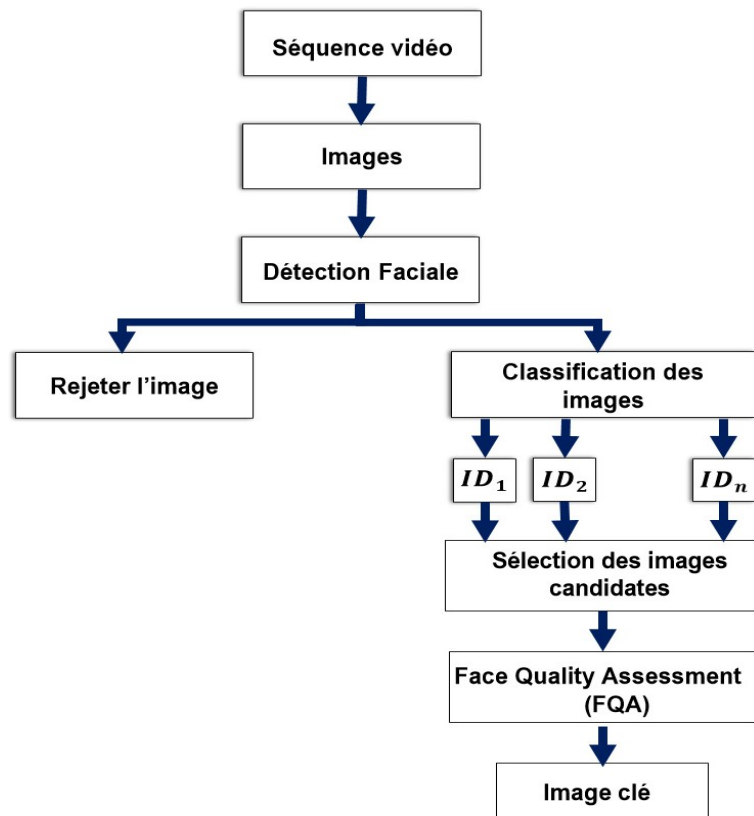


FIGURE 3.1: Processus général de la méthode KS-FQA [3]

Nous présentons les différentes étapes de la méthode KS-FQA [3] pour l'extraction

des images clés de visage :

- La détection faciale est une étape primordiale dans tout système d'analyse et traitement des images faciales. L'objectif est de limiter les régions des visages dans une image donnée [52].
- La qualité de l'image est considérée comme un facteur important qui affecte la performance des algorithmes de traitement d'image [53]. Par conséquent, l'évaluation de la qualité de l'image (Image Quality Assessment (IQA)) a fait l'objet de plusieurs recherches dans divers domaines de traitement d'image et de la vision par ordinateur, comme la détection/ reconnaissance faciale [54], la détection des objets [55], la détection d'événements [56] et le suivi d'objets dans les vidéos [57].
- La sélection de l'image clé est basée sur les scores de qualité (Face Quality Assessment (FQA)) en spécifiant une règle de sélection précise (score minimal, maximal, etc).

A la fin de ce processus, nous obtenons, pour chaque identité, une seule image qui le représente. Cette image possède le meilleur score de qualité parmi toutes les images où ce visage apparaît.

3.1.1 Détection Faciale

Nous utilisons le Multi-task Cascaded Convolutional Networks MTCNN[7] comme détecteur de visage dans une image. Ce détecteur combine la détection et l'alignement des visages afin de renforcer la détection et améliorer son taux de précision [58]. Le MTCNN se base entièrement sur les techniques d'apprentissage profond, particulièrement les CNN pour la localisation de la région du visage. Le principe de ce détecteur est illustré dans la figure 3.2.

Trois sorties sont fournies par P-Net : classification faciale, coordonnées de boîte englobante et les coordonnées des points de repères.

- **Classification faciale** : Le but de la classification est de calculer la probabilité pour que l'objet encadré soit un visage.

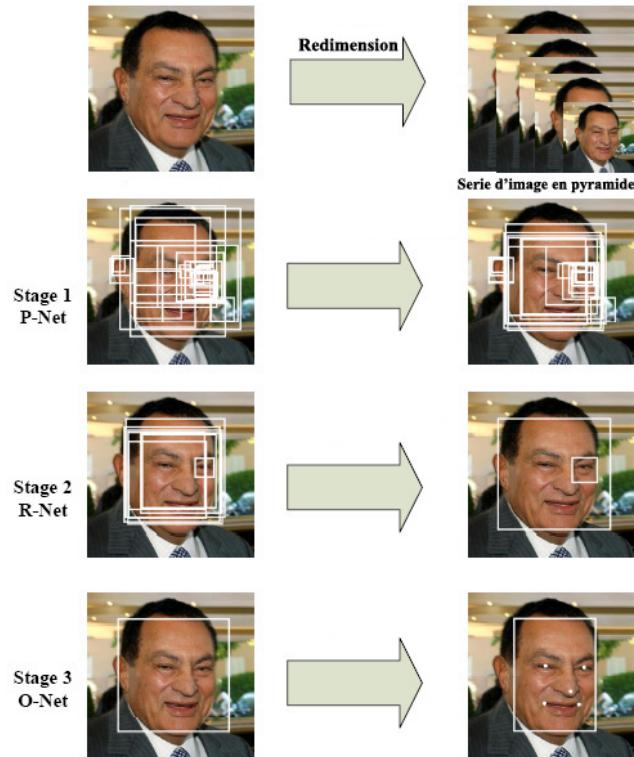


FIGURE 3.2: Étapes du détecteur MTCNN [7]

— **Boîte englobante** : Pour chaque fenêtre candidate, l’algorithme calcule la différence entre la fenêtre déterminée par P-Net et celle de la vérité de terrain la plus proche. Chaque boîte est représentée par quatre valeurs :

[coordonnées de point supérieur gauche (x,y), hauteur, largeur]

— **Localisation des points de repère** : Cette sortie fournit les coordonnées (en abscisse et ordonnées) des cinq points de repère (les deux yeux, le nez et les deux coins de la bouche).

Nous avons choisi d’utiliser ce détecteur pour plusieurs raisons. Premièrement, il s’agit d’un détecteur à base d’un apprentissage profond qui permet, non seulement la détection des visages, mais aussi l’alignement de ceux ci. Deuxièmement, le MTCNN fournit plusieurs informations utiles à utiliser par la suite, à savoir le score de confiance, les coordonnées des cinq points de repère et les coordonnées de la région englobant le visage dans l’image. Nous présentons dans l’image 3.3 une illustration des résultats fournis par le MTCNN. La région faciale détectée

est encadrée tout en précisant aussi les cinq points de repère (Figure 4.5). Les coordonnées des ces cinq points ainsi que les coordonnées de la région du visage et le score de confiance (Figure 3.3c).



FIGURE 3.3: Résultats fournis par le détecteur MTCNN [7]

Finalement, le MTCNN nous permet de ne garder que les images contenant clairement des régions faciales. En d'autres termes, si le MTCNN ne réussit pas à détecter les cinq points de repère sur une image, il considère que cette image ne contient pas un vrai visage. Cette condition nous permet de rejeter les images où les visages possèdent une large variation de pose ou une image de visage de faible qualité. Pour les images en vue latérale, si le détecteur ne capte pas les cinq points de repère alors le visage n'est pas reconnu. Même si nous avons des visages frontales, le déséquilibre de la luminosité ou le bruit peut cacher certains détails du visage et ces images seront rejetées. Nous voulons justifier notre choix du détecteur facial MTCNN. Pour cela, nous comparons le détecteur facial par rapports deux autres détecteurs fréquemment utilisées dans la littérature : Le détecteur Dlib [59] et Viola-Jones [60]. Nous avons trouvé que le détecteur MTCNN arrive à détecter un nombre de visages plus élevé que Dlib et Viola-jones pour une image contenant plusieurs visages même avec des faibles résolutions.

3.1.2 Classification des images

La classification faciale désigne le processus de regroupement des visages des personnes présentes sur un ensemble de photos ou de vidéos. L'objectif est de regrouper ces images non étiquetées en fonction de leur représentation en petits

sous-ensembles. Une classe pour chaque identité [61].

Deux problèmes majeurs ont été rencontrés dans les applications de classification de visages : la représentation des visages en des vecteurs caractéristiques et la mesure de la similarité à utiliser pour classifier les visages et le choix d'une mesure de similarité. Récemment, la description des visages grâce des réseaux profonds a connu un essor considérable. Nous pouvons notamment citer parmi les plus célèbres *FaceNet* [21].

FaceNet est basé sur un réseau convolutif profond pour extraire un vecteur de caractéristiques robuste de l'image faciale. Ces caractéristiques sont considérées comme étant robustes face aux différentes variations que peut subir le visage tels que la pose, conditions d'éclairage, résolution, émotions, ...etc.

Pour regrouper les visages par identité, nous suivons la démarche suivante : nous créons, pour chaque visage détecté, un vecteur de caractéristique de dimension égale à 128 en utilisant l'étape d'extraction de caractéristiques de FaceNet. Ensuite, en fonction de ces vecteurs, nous regroupons les visages en utilisant l'algorithme de classification par ordre de rang (*Rank-Order clustering algorithm* [62]). Cet algorithme est une forme de regroupement hiérarchique, qui utilise une mesure de distance basée sur le plus proche voisin. L'algorithme définit une distance appelée (Rank Order Distance) pour mesurer la similarité entre deux visages. Formellement, l'algorithme fonctionne comme suit :

1. Chaque visage est dans une classe à part.
2. Répéter Fusionner n'importe quelle paire de classes si leur distance Rank-Order est inférieure à un certain seuil.
3. Arrêtez si aucune classe ne peut être fusionnée ; sinon mise à jour des classes et distances entre classes, et repassez à 2.

La Figure 3.4 montre le principe de fonctionnement de l'algorithme Rank Order Clustering.

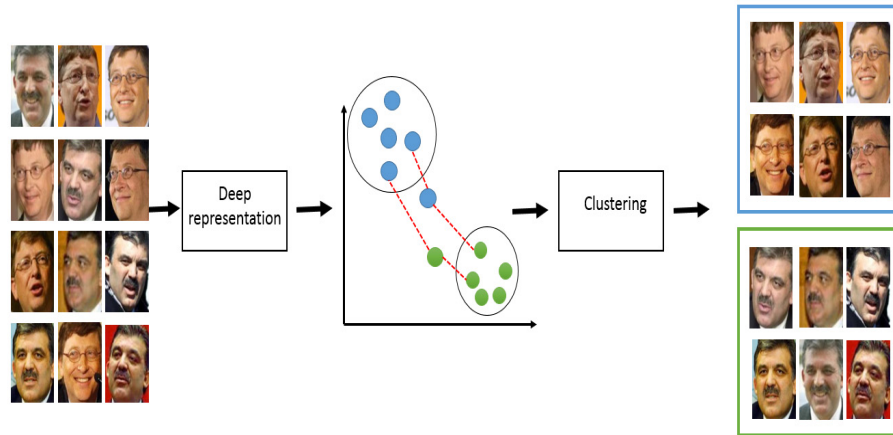


FIGURE 3.4: illustration de la classification par ordre de rang

3.1.3 Sélection des images candidates

Dans [63, 64], les auteurs utilisent toutes les images de la vidéo pour calculer la qualité d'image et choisir par la suite les images clés. En effet, le traitement de toutes les images demande énormément de temps et de mémoire. De plus, cela peut provoquer une redondance étant donné que les images successives peuvent avoir presque le même contenu. Dans [42], les auteurs ont utilisé une règle de fenêtrage qui consiste à sélectionner une image pour chaque seconde (FPS : Frame Per Second). Les auteurs considèrent qu'en une seconde, il n'y a pas de changement notable dans le contenu des images successives.

Malgré sa simplicité, cette méthode ne correspond pas à une extraction d'images clés basée sur les images faciales. L'utilisation du temps comme un critère de sélection en choisissant une image par seconde ne garantit pas l'obtention des meilleures images faciales. Pour cette raison, nous devons choisir une méthode qui prend en compte l'état du visage pour l'image sélectionnée. Il faut que le visage dans cette image soit frontal, neutre, de bonne qualité, ..etc. Nous utilisons le score de confiance fourni par le MTCNN comme critère pour sélectionner les images candidates à l'utilisation comme images clés. Le score de confiance reflète la probabilité que l'objet détecté soit un visage. Les images de visages ayant le score de confiance maximum ne sont pas seulement les plus frontales. D'autres facteurs

sont pris en compte, tels que les expressions faciales, la luminosité, la résolution, etc. Dans la figure 3.5, nous montrons un exemple de séquence complète (figure 3.5a) et les cinq images sélectionnées (figure 3.5c).



(a) Séquence d'image



(b) Séquence d'image triée (à modifier)



(c) Ensemble des images candidate

FIGURE 3.5: Sélection des images candidates

Nous commençons par trier les images en fonction du score de confiance (Figure 3.5b). Pour construire l'ensemble d'images candidates, nous choisissons les cinq premières images de chaque visage. En d'autres termes, les cinq images ayant les meilleurs scores de confiance des visages. Le principal avantage de choisir cinq images candidates pour chaque identité, est d'avoir le même nombre pour chaque groupe d'images de visage à traiter, ce qui garantit un temps de traitement similaire pour toutes les séquences. De plus, en utilisant le score de confiance, le critère de sélection des images ne dépend pas de l'apparence de l'image dans la séquence (ordre chronologique), mais seulement de sa valeur de confiance, qui reflète la probabilité que l'objet détecté soit un visage.

3.1.4 Calcul du score de qualité du visage

L'évaluation de la qualité d'image faciale pour chaque image est effectuée en calculant quatre mesures : La pose (Pose Score PS), la clarté (Sharpness Score SS), la luminosité (Brighness Score BS) et la résolution (Résolution Score RS).

1. **Pose de la tête :** La mesure de pose est considérée comme la mesure la plus importante dans l'estimation de la qualité d'image faciale. Cette mesure est définie par la détection de l'orientation de la tête par rapport à une caméra. Cette mesure permet d'évaluer à quel point ce visage est frontal. En effet, une grande variation de pose cache la plupart des détails utiles dans l'image de visage. De plus, les images frontales sont les plus recommandées dans les systèmes d'analyse faciale.

D'après [65], deux méthodes d'estimation de la pose existent dans la littérature : Les méthodes locales et globales. Les méthodes locales utilisent les composantes du visage (les sourcils, les yeux et les lèvres) pour estimer la pose. Mais, sur les images à faible résolution, la détection de ces composantes est assez difficile. La méthode globale utilise l'ensemble de l'image du visage pour estimer la pose. L'utilisation des méthodes globales permet d'éviter le problème des images à basse résolution. De plus, nous n'avons pas besoin de détecter ces petits éléments dans un tel visage, il suffit de détecter la région du visage.

La pose est définie comme la différence entre le centre de masse et le centre de région dans l'image faciale. Le centre de la région pour chaque visage est calculé à l'aide de l'équation 3.1 :

$$x_r = \frac{x_2 - x_1}{2}, y_r = \frac{y_2 - y_1}{2} \quad (3.1)$$

Où (x_1, y_1) sont les coordonnées du point supérieur gauche et (x_2, y_2) les coordonnées du point inférieur droit dans la région du visage.

En suivant le même principe de [57], nous définissons le centre de masse comme un point central dans le visage. Nous allons déterminer ce point sans

avoir besoin de calculer ses coordonnées. Dans ce travail, nous considérons que le centre de masse est le nez, défini par (x_m, y_m) . Si le visage est frontal, les coordonnées du nez sont proches du centre de la région. Avec la variation de la pose de la tête, la distance entre ces deux points augmente également (voir figure 3.6).



(a) Image du visage en position frontale

(b) Image du visage ayant subi une rotation

FIGURE 3.6: Distance ente le centre de région (*) et le centre de masse (+)

Finalement, nous calculons la distance entre le centre de masse et le centre de la région en utilisant l'équation 3.2.

$$PS = \frac{1}{1 + \sqrt{(x_r - x_m)^2 + (y_r - y_m)^2}} \quad (3.2)$$

2. **Clarté (Sharpness)** Nous appliquons un opérateur Gaussien à l'image du visage. Ensuite, la valeur de clarté est définie comme la différence absolue entre l'image d'origine et l'image opérée, divisée par la taille de l'image du visage.

$$SS = \frac{\sum_{i=1}^{i=W} \sum_{j=1}^{j=H} |I - G(I)|}{W * H} \quad (3.3)$$

avec I est l'image d'origine, $G(I)$ est l'image filtrée. W , H représentent les dimension de l'image de visage.

3. **Luminosité** Le score de luminosité de l'image du visage, décrit par l'équation 3.4 est égal à la somme de tous les pixels de l'image divisée par la résolution de l'image.

$$BS = \frac{\sum_{i=1}^{i=W} \sum_{j=1}^{j=H} I_{i,j}}{W * H} \quad (3.4)$$

4. **Résolution C** est la multiplication de la hauteur H et la largeur W de l'image du visage :

$$RS = H * W \quad (3.5)$$

Après le calcul des quatre mesures pour chaque image, les valeurs retenues seront combinées en une seule valeur. Tout d'abord, chaque mesure est normalisée en fonction de la valeur maximale obtenue pour une collection des images de la même identité. Ensuite, nous combinons les scores obtenues pour avoir un score de qualité global (équation 3.6).

$$Q = \sum_{i=1}^{i=N} \frac{S_i}{S_{max}} \quad (3.6)$$


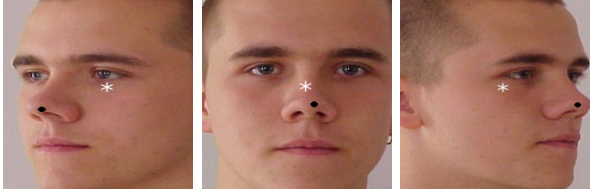




Où S est le score obtenu pour la i^{th} mesure calculée, et S_{max} est la valeur maximale obtenue de cette mesure dans une collection d'images donnée. N définit le nombre de mesures utilisées pour estimer la qualité d'image. Finalement, l'image du visage ayant le score le plus élevé sera sélectionnée comme image clé.

3.1.5 Évaluation de la méthode KS-FQA

Nous rappelons que la pose du visage est vue comme la différence entre le centre de masse et le centre de région. Dans notre travail, les coordonnées du nez sont utilisées comme le centre de masse puisque c'est le point central dans un visage. De cette manière, le centre de masse sera proche du centre de région si le visage est frontal. Avec la variation de la pose, la distance entre ces deux points augmente également.

Donc, nous pouvons estimer la variation de la pose dans les différentes rotations possibles du visage. Le tableau 3.1 illustre deux exemples de rotations du visage et les deux points détectés.

TABLE 3.1: Estimation de la pose par rapport à la rotation du visage

Rotation	Illustration	Exemples
Yaw		
Pitch		
Roll		

Dans la première ligne, et pour une image de visage frontale, les deux points sont très proches. Pour les deux autres images (les vues latérales), la distance entre ces deux points augmente autant que la rotation du visage augmente. Dans le deuxième exemple, et même avec les visages frontales, notre méthode est capable de détecter une grande variation de pose dans les deux sens (haut ou bas).

3.2 Extraction des images clés à base d'un apprentissage profond et la qualité d'image faciale (FQM-CNN) [4]

Nous avons conçu une deuxième méthode nommée FQM-CNN. Cette méthode a pour objectif d'extraire des images clés en se basant sur la qualité des images faciales tout en utilisant les descripteurs de visage et les réseaux de neurones convolutionnels (CNN).

L'idée de base consiste à apprendre un CNN pour estimer la qualité d'une seule image faciale en entrée. En d'autres termes, l'estimation de la qualité se fait sans avoir besoin de calculer les scores pour toute une séquence ou ensemble d'image comme dans la méthode KS-FQA.

En effet, notre système comporte deux étapes principales. La première étape consiste à former la base de données d'apprentissage. La deuxième étape consiste à apprendre un nouveau CNN et à prédire le score de qualité des images faciale. L'apprentissage se fait grâce à la base d'images et les scores (FQS) générées dans la première étape. Nous détaillerons dans les paragraphes suivantes chaque étape. L'organigramme de cette méthode est présenté dans la figure 3.7.

3.2.1 Étape1 : préparation des données d'apprentissage

3.2.1.1 Sélection de l'image de référence

Dans la première étape, nous nous intéressons à la génération des étiquettes pour les images de l'ensemble d'apprentissage. Nous commençons par diviser l'ensemble des images faciales en deux sous ensembles : un ensemble des images de références (*Template*) et le reste des images (*Probe*) [66]. Le premier sous ensemble comprend une seule image par identité. En fait, ces images représentent les images les plus neutres pour chaque identité que nous l'utiliserons par la suite pour calculer le score de la qualité des images de l'autre ensemble (*probe set*). Néanmoins, les images de ce sous ensemble ne peuvent pas être choisi manuellement.

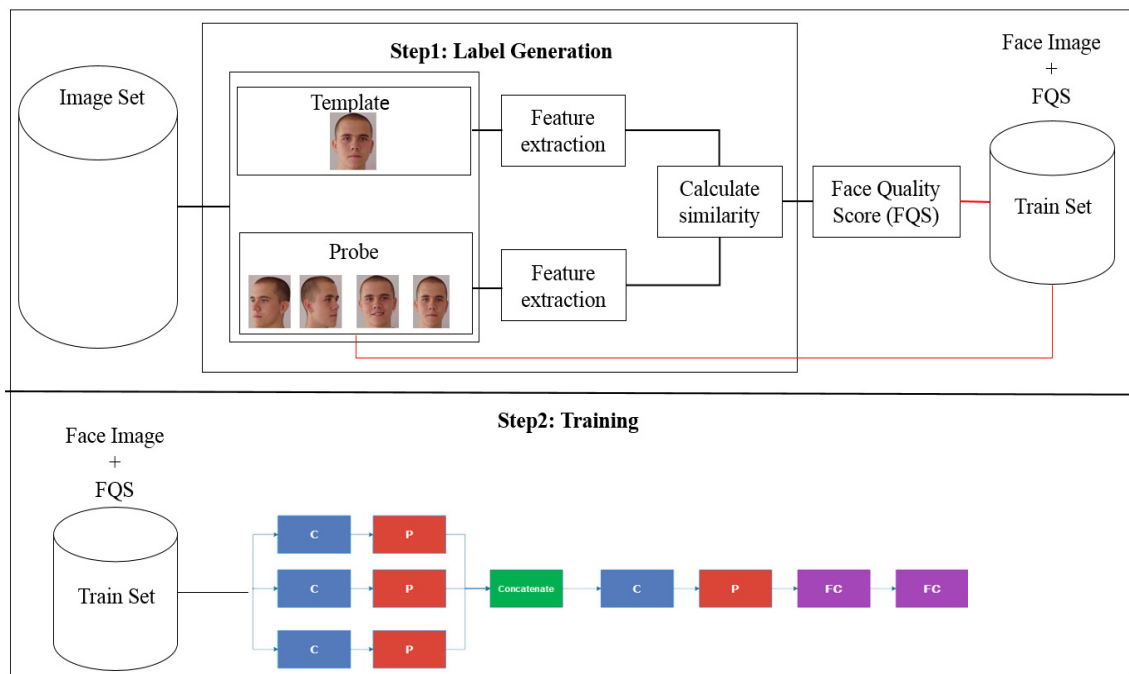


FIGURE 3.7: Schéma de la méthode FQM-CNN [4]

Grâce à l'utilisation du détecteur MTCNN (voir section 3.1.1), nous choisissons l'image référence de manière objective en fonction du score de confiance. Ce score reflète la probabilité que l'objet détecté soit un visage. C'est-à dire, l'image du visage ayant le score de confiance le plus élevé sera choisie comme référence.

La figure 3.8 présente des ensembles d'images et les références choisies.

A partir de ces exemples, nous pouvons remarquer que l'image ayant le score de confiance maximal n'est pas l'image la plus frontale. D'autres facteurs sont pris en considération à savoir la neutralité, la distribution uniforme de luminosité, la clarté, ...etc. Le standard choisit dans le premier exemple est une image frontale avec la meilleure résolution. Ceci nous aide à visualiser les détails du visage. L'image choisie dans le deuxième exemple est l'image la plus frontale et neutre.

3.2.1.2 Génération des étiquettes et description globale de la région du visage

Inspiré par les travaux de [67] et [63], au lieu d'utiliser des métriques et des systèmes de pondération, nous allons utiliser des descripteurs globaux dans le but



FIGURE 3.8: Exemple des images de références sélectionnées

de calculer la qualité de l'image. Nous utilisons trois descripteurs de visage pour l'extraction des caractéristiques faciales : LBP, Gabor, HoG. Ensuite, nous calculons la similarité entre les descripteurs de l'image de référence et chaque descripteur des images de l'autre ensemble *Probe*. Le score final obtenu représente la qualité de l'image faciale. Notre contribution au niveau de cette partie d'extraction des caractéristiques est la combinaison des trois descripteurs HOG, LBP et Gabor. Ces trois descripteurs sont complémentaires dans la description du visage. Le HoG a prouvé sa robustesse face aux expressions faciales et la variation de la pose. D'autre part, le LBP est caractérisé par sa robustesse face aux changements monotones de luminosité. Ceci, malgré qu'il est sensible aux bruits et aux grandes variations des niveaux de gris. Le filtre de Gabor est connu pour sa robustesse face aux petites translations et rotations, les changements de luminosité. Surtout, Gabor est efficace pour éliminer le bruit dans les images du visages où nous le trouvons fréquemment dans les vidéos.

3.2.1.3 Calcul du score de qualité

L'image 3.9 présente l'organigramme du processus d'extraction des descripteurs et la génération de score de qualité.

Le calcul de la qualité consiste à comparer l'image en entrée avec une image de référence. La différence entre ces deux images donne le score de qualité de l'image.

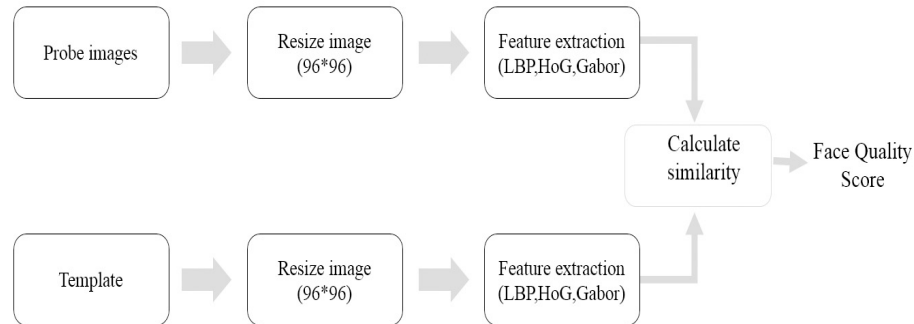


FIGURE 3.9: Module de génération des étiquettes

Nous commençons par re-dimensionner les images de niveaux de gris en imagerie de taille 96 pixel * 96 pixel. Puis, nous procédons à l'extraction des descripteurs des images. Ensuite, nous calculons la similarité entre les deux descripteurs obtenus en utilisant la mesure de similarité distance du cosinus [68] présentée dans l'équation 3.7.

$$sim(I1, I2) = \frac{\sum_{i=1}^N I1_i * I2_i}{\sqrt{\sum_{i=1}^N I1_i^2} * \sqrt{\sum_{i=1}^N I2_i^2}} \quad (3.7)$$

. Le score final est obtenu en faisant la somme des trois mesures de similarité obtenues avec chacun des descripteurs cités plus haut.

3.2.2 Étape2 : Apprentissage

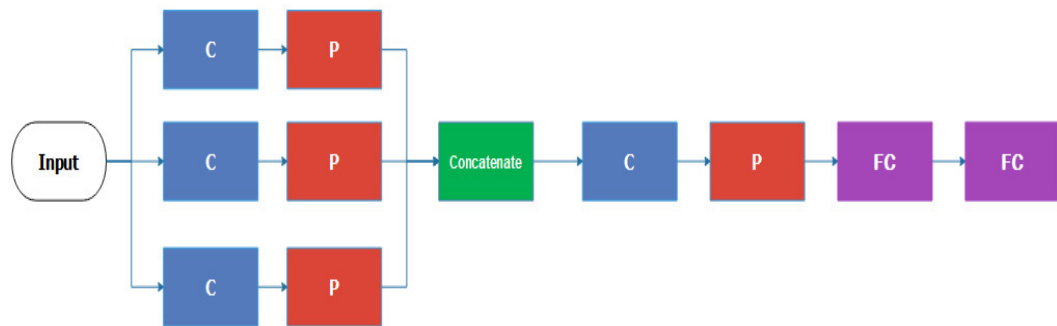
L'idée de notre architecture CNN est basée sur l'utilisation du modèle d'inception. Le principe de ce modèle est de concaténer différentes sorties provenant de différentes couches de convolution ayant des filtres de tailles différentes. L'intérêt d'utiliser ce modèle, est d'améliorer la précision en ajoutant plus de couches en largeur bien qu'en profondeur.

Une architecture CNN profonde comportant un modèle d'inception possède un nombre élevé des paramètres. Ce qui peut causer un sur-apprentissage. Particulièrement, si la base ne comporte pas assez de données étiquetées. De plus, ce nombre

élevé de paramètres augmente aussi la complexité de calcul [69].

Dans notre travail, nous allons tirer profit des fonctionnalités offertes par les CNN à savoir les fonctions d'activation, les couches de sous-échantillonnage et les couches de convolution avec des filtres de taille 1×1 . Ces fonctionnalités nous aident à éviter les problèmes mentionnés au-dessus. Notre modèle d'inception est différent de ceux proposés dans la littérature, bien que les modèles classiques utilisent plusieurs couches en largeur : généralement trois ou quatre, nous allons utiliser trois couches de convolution en largeur. Chaque couche est suivie d'une fonction d'activation non linéaire PReLU [70] et une couche de pooling (max-pooling).

L'architecture globale de notre réseau est décrite dans la figure 3.10. Nous avons d'abord appliqué un modèle d'inception pour pouvoir extraire plus d'informations à partir de l'image initiale. Notre modèle est composé seulement de deux blocs : le premier est le module d'inception. Le deuxième est un bloc de convolution. Ce qui nous donne un nombre raisonnable de paramètres. Dans ce travail, nous allons aussi intégrer les fonctions d'activation non linéaires et les couches de sous-échantillonnage dans le réseau. Le sous échantillonnage nous aide à diminuer la taille des données. Les fonctions d'activation contrôlent le nombre de paramètres dans le réseau. L'architecture proposée est illustrée dans la figure 3.10.



C: Convolution layer

P:Max-pooling layer

FC : Fully connected layer

FIGURE 3.10: Architecture du CNN proposé [4]

Dans ce modèle, nous utilisons la rectifieur non linéaire PReLU. Nous appliquons

le modèle d'inception avec trois couches de convolution ayant trois paramètres différents : La première couche de convolution avec 16 filtres de taille 5×5 . La deuxième couche a également 16 filtres mais de taille 3×3 . La dernière couche possède seulement 16 filtres avec la taille 1×1 . Chaque couche de convolution est suivie par une fonction d'activation PReLU et une couche de sous-échantillonnage (max-pooling) de taille égale à 2. Ce module est suivi par un autre bloc comprenant une couche de convolution de 128 filtres de taille 5×5 , suivie par une couche PReLU et une couche de pooling maximale de taille 2×2 . L'architecture proposée se termine par deux couches entièrement connectées. Nous utilisons une fonction sigmoïde dans le dernier nœud. Cette fonction est utilisée pour générer une sortie numérique du score prédit, dans une plage de 0,0 à 1,0 [71]. L'optimisation est réalisée à l'aide de la méthode de Adam [72]. Nous utilisons des sous-ensembles (*mini-batches*) de 128 échantillons. Le modèle est régularisé en utilisant Dropout appliqué avant les deux couches entièrement connectées avec un taux de 0,5. Le taux d'apprentissage a été initialement fixé à 0,001. La phase d'apprentissage dure 500 itérations. Dans le but de concevoir une base de données ayant un nombre d'images intéressant et un contenu varié nous allons utiliser cinq bases de données différentes. Cette variété de contenu est très utile pour l'apprentissage de notre CNN. De plus, notre base d'apprentissage décrit la majorité des difficultés rencontrées dans des environnements encombrés réels.

Ces images sont collectés à partir de cinq bases de données différentes : FRI CVL [73], Face Recognition Data [74], AT&T [75], Yale Face Database B [76] et Yale Extended Face Database B [76].

3.2.3 Évaluation de l'architecture CNN proposée

Pour tester l'architecture que nous avons proposé, nous allons effectuer deux tests différents :

Nous commencerons par évaluer l'utilité des fonctions d'activation au sein du modèle d'inception. Par la suite, le modèle ayant la meilleur performance en terme

de précision sera comparé aux architectures des méthodes de la littérature.

Dans le premier test, nous ferons recours aux trois fonctions d'activation les plus connues et les plus utilisées : ReLU [77], PReLU [70] et ELU[78].

1. **ReLU** : est l'une des fonctions d'activation les plus connus. Cette fonction est définie comme suit :

$$a_{i,j,k} = \max(0, z_{i,j,k}) \quad (3.8)$$

Où $z_{i,j,k}$ est l'entrée de la fonction d'activation de coordonnées i, j sur le k^{me} canal. ReLU est une fonction linéaire par morceaux, qui remplace les valeurs négatives par zéro et conserve la partie positive. L'utilisation d'un simple opérateur (max) de ReLU permet une complexité de calcul inférieure à d'autres fonctions d'activation comme sigmoïde, tanh ou certaines autres plus récentes [79].

2. **PReLU** : La fonction PReLU est une autre alternative de ReLU. Le point commun entre PReLU et LReLU (Leakly ReLU) est dans l'utilisation d'un paramètre λ . Par contre, PReLU fait apprendre ce paramètre de manière adaptative et LReLU utilise un paramètre statique. Cette fonction est utilisée afin d'améliorer la précision. PReLU est défini mathématiquement comme suit :

$$a_{i,j,k} = \max(0, z_{i,j,k}) + \lambda_k \min(0, z_{i,j,k}) \quad (3.9)$$

Où λ_k est le paramètre de k^{me} canal.

Malgré que PReLU introduit un nombre supplémentaire de paramètres, ceci n'engendre pas un sur-apprentissage. De plus, le coût de calcul supplémentaire est quasiment négligeable [79].

3. **ELU** : Permet un apprentissage rapide pour les réseaux profonds et offre des taux de précision élevés pour des tâches de classification grâce à l'utilisation de la fonction exponentielle. La fonction Elu garde la partie négative contrairement aux autres fonctions déjà présentées. Cette partie négative est bénéfique

pour un apprentissage rapide [79]. La fonction ELU est définie comme suit :

$$a_{i,j,k} = \max(0, z_{i,j,k}) + \min(\lambda(e^w - 1), 0) \quad (3.10)$$

Avec $w = z_{i,j,k}$ qui représente l'entrée de la fonction d'activation.

Cette évaluation a été réalisée en utilisant les trois bases de données standards : MNIST [80], Cifar 10 [81] et Cifar 100 [81].

L'évaluation des performances de notre CNN est faite en deux étapes : tester d'abord la fonction d'inception. ensuite, nous allons déterminer le modèle ayant la meilleure performance en terme de précision par rapport aux différentes modèles utilisées dans la littérature pour l'extraction des images clés à base d'un apprentissage profond. Les résultats en terme de précision et la description de chaque architecture sont présentés dans le tableau 3.2.

TABLE 3.2: Taux de précision obtenus en utilisant différentes architectures CNN

Description de l'architecture	Précision
Architecture CNN basique. Composée de 4 couches de convolution suivie des couches de sous-échantillonnage, une couche entièrement connectée avec une régression linéaire.[82].	11 %
bloc d'inception composé de 4 couche de convolution dont deux couches avec des filtres de taille 1×1 , les deux autres couches (3×3) , (5×5) sont aussi suivis des couche de pooling. Ce bloc est suivis de deux autres bloc de convolution et deux couche entièrement liées et un dropout a l'intermédiaire. Finalement une sortie sigmoïde. [83].	87.14 %
Architecture VGG-16 : Fréquemment utilisée pour l'estimation de la qualité des images particulièrement pour les méthodes basées sur un apprentissage profond [84].	92,40 %
Architecture FQM-CNN proposée [4].	99,01 %

À partir de du tableau 3.2, nous pouvons remarquer une différence énorme des pourcentages de précision. L'architecture utilisée dans le travail de [82] est relativement simple, ce qui signifie le faible taux obtenu. L'utilisation du modèle d'inception comme [83] avec différents paramètres (nombre et taille des filtres), l'utilisation du dropout et la fonction sigmoïde pour la sortie, peut améliorer le résultat final. Le dernier résultat concerne la structure VGG-16 [84] qui utilise une architecture profonde avec 11 couches de convolution suivies des fonctions d'activations. Cette architecture prend longtemps à converger et à fournir une valeur de précision plus intéressante.

3.2.4 Extraction des images clés à base de la qualité d'image faciale

Dans cette section, nous allons montrer l'efficacité des méthodes proposées pour l'extraction des images clés (FQM-CNN et KS-FQA). Notre évaluation se base sur une combinaison de critères subjectifs (la qualité) et objectifs (la quantité). En outre, l'évaluation subjective vise à juger si les images clés extraites représente convenablement les identités en question. L'évaluation objective porte sur la mesure des performances des systèmes proposés en termes de précision de reconnaissance, précision/rappel. Deux bases de données sont utilisées dans le processus d'évaluation des méthodes d'extraction des images clés : FRI CVL Dataset [73] et YouTube Face (YTF).

Pour l'évaluation subjective, l'objectif est de trier les images de la séquence en fonction du score de qualité. Nous comparons les classement fournies par KS-FQA, FQM-CNN et le module de génération des étiquettes de la méthode FQM-CNN par rapport à une vérité terrain de la base fournie par [85] et d'autres méthodes de la littératures. Deux exemples seront présentés dans les Tableau 3.3 et 3.4. Nous désignons par la valeur 0 l'image choisie comme référence dans cet ensemble.

TABLE 3.3: Classement des images de la base CVL











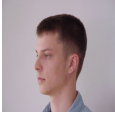



Frames							
Vérité-terrain [85]	4	2	3	1	5	5	4
Nasrollahi et.al [63]	5	1	2	1	6	5	3
Chen et al.,2015[86]	4	2	3	1	-	-	-
KS-FQA [3]	4	3	1	2	Rejetée	Rejetée	Rejetée
générateur des étiquettes	3	1	0	2	Rejetée	Rejetée	Rejetée
FQM-CNN[4]	4	2	3	1	Rejetée	Rejetée	Rejetée

TABLE 3.4: Rang du résultat pour la séquence CVL

Frames							
Vérité terrain [85]	6	4	1	5	6	3	2
Nasrollahi et al [65]	7	4	1	5	6	3	2
Qi et al. [67]	6	4	5	2	7	3	1
Nasrollahi et al [63]	7	5	3	4	6	2	1
KS-FQA [3]	Rejected	5	1	4	Rejected	2	3

Nous pouvons remarquer que le rang augmente, alors que la qualité de l'image diminue. KS-FQA ne traite pas les expressions faciales. C'est pourquoi nous trouvons que le classement obtenu ne prend pas ce facteur en considération. En utilisant le générateur des étiquettes, le classement est influencé par la qualité de la référence choisie. Nous remarquons aussi que les visages qui ont subi une rotation plus importante ont les derniers rangs. Le classement fourni par FQM-CNN correspond au même classement de la vérité terrain. Sauf que celle-ci considère toutes les images, même ceux avec une variation importante de la pose.

Dans nos méthodes, et avec l'utilisation du détecteur MTCNN, ces images sont rejetées, ce qui réduit le nombre d'images à traiter.

Nous allons maintenant faire une évaluation objective des images clés obtenues pour une tâche de vérification faciale. Pour ce faire, nous allons suivre ce démarche : nous commençons par définir une images clé pour chaque identité. Ensuite, nous utilisons ces images clés comme entrée pour l'algorithme de reconnaissance faciale.

TABLE 3.5: Comparaison par rapport aux méthodes classiques

Méthode	Précision
Fourney et al. [87]	69,82%
Wong et al. [88]	79,92 %
Nikitin et al. [89]	74.46%
Anantharajah et al. [90]	89.7 %
Qi et al. [67]	92,6 %
KS-FQA [3]	95.4 %

TABLE 3.6: Comparaison par rapport aux méthodes profondes

Méthode	# Image	# Réseau	Précision
Deep face [91]	4M	3	91.4%
Deep Id+ [92]	0.2M	25	93.2%
FaceNet [21]	200 M	1	95.1%
VGG-face [93]	2.6M	1	97.3%
Center loss [94]	0.7M	1	94.9%
DFCL [95]	4.7 M	1	96.06%
FQM-CNN [4]	0.22 M	1	95.2 %

D'après les résultats du tableau 3.5, l'avantage de la méthode KS-FQA réside dans les formules utilisées pour calculer les métriques du visage. Les autres méthodes combinent les métriques en une seule valeur. Généralement la mesure de symétrie ne détecte pas tous les types de rotations possibles ("Yaw" voir tableau 3.1). De plus, l'utilisation des poids influe les résultats tout en accordant plus de priorité aux métriques qu'aux autres.

Les résultats du tableau 3.6 montrent que FQM-CNN permet d'obtenir des taux de précision plus élevés que certains modèles profonds de la littérature. Par contre, notre approche donne des résultats inférieures à des méthodes telles que [93] et [95], qui sont caractérisées par une grande capacité d'apprentissage due à l'utilisation d'un grand ensemble de données et d'un réseau très profond. Augmenter la taille de l'ensemble d'apprentissage et la durée d'apprentissage peut nous donner

des résultats plus performants. D'autre part, FQM-CNN présente un avantage majeur par rapport à KS-FQA, c'est que nous pouvons traiter toute image de façon indépendante sans avoir besoin d'avoir toute la séquence.

Conclusion

L'objectif principal des méthodes proposées dans ce deuxième axe de recherche, est l'extraction des images clés de visages d'une vidéo où le visage est dans une position neutre, dans de bonnes conditions d'illumination et sans émotions. Nous avons tout d'abord appliqué la Méthode MTCNN pour la localisation de la région du visage dans une image. Ensuite, nous avons appliqué différents descripteurs basés sur la qualité du visage et avons proposé une architecture CNN afin d'extraire les meilleures images clés pour faciliter la reconnaissance de personnes dans une base de vidéos dans une seconde étape. Nous pouvons soumettre au moteur de recherche le visage d'une personne et le moteur de recherche retournera toutes les vidéos où cette personne apparaît.

Chapitre 4

Construction de résumés vidéos par les descripteurs 2D et 3D du visage : application à la reconnaissance faciale

Introduction

Afin d'extraire les images clés de visages les plus représentatives de la vidéo dans des environnements encombrés, et dans l'objectif d'améliorer les performances de la reconnaissance faciale, nous nous concentrons sur la reconnaissance des visages en se basant sur l'extraction des caractéristiques faciales. Une nouvelle méthode a été proposée. Cette méthode Deep 3D-LBP propose d'utiliser la bio modalité 2D/3D du visage et extraire un descripteur plus robuste face aux variations de pose, d'expression du visage et de luminosité dans l'image. Ce travail a fait l'objet des publications suivantes : [5] et [22]. Ce travail correspond à l'encadrement en thèse de Rahma Abed.

4.1 Description des images faciales par la modélisation 3D du visage et la description de la texture (Deep 3D-LBP) [5]

Nous proposons dans cette partie un descripteur de visage afin de mieux représenter un visage et le reconnaître par la suite. L'idée de base consiste à fusionner les informations issues des descripteurs de forme et de texture en un seul descripteur facial qui sera utilisé pour entraîner un réseau de neurones pour une reconnaissance plus efficace. Pour ce faire, nous proposons d'utiliser à la fois la forme 3D du visage pour atténuer la variation de la pose et les émotions, avec le descripteur LBP vu sa robustesse face aux variations d'illumination. En d'autres termes, nous construisons un système de reconnaissance des visages robuste aux variations de pose, illumination et d'expression faciale. Pour cela, nous utilisons un modèle 3D de représentation et d'alignement facial. Nous calculons un descripteur LBP construit sur le maillage 3D pour obtenir une représentation du visage qui combine les informations de forme et de texture. Enfin, nous utilisons un CNN pour avoir un système de reconnaissance facial efficace.

La méthode proposée (appelée Deep 3D-LBP) est composée de trois étapes :

- Tout d'abord, nous effectuons la détection des visages et la localisation des points de repère qui caractérisent le visage.
- Ensuite, un modèle générique de visage 3D est utilisé pour faire correspondre les images 2D. Ceci est accompli en modélisant la différence dans la carte de texture des images d'entrée et de référence alignées en 3D. Nous utilisons le mesh-LBP [96] comme extracteur de caractéristiques de visage.
- Enfin, les descripteurs obtenus sont utilisés pour entraîner une architecture CNN basée sur la fusion des informations de pose, de texture et de forme et reconnaître les visages.

L'organigramme de la méthode proposée Deep 3D-LBP est illustré dans la figure 4.1.

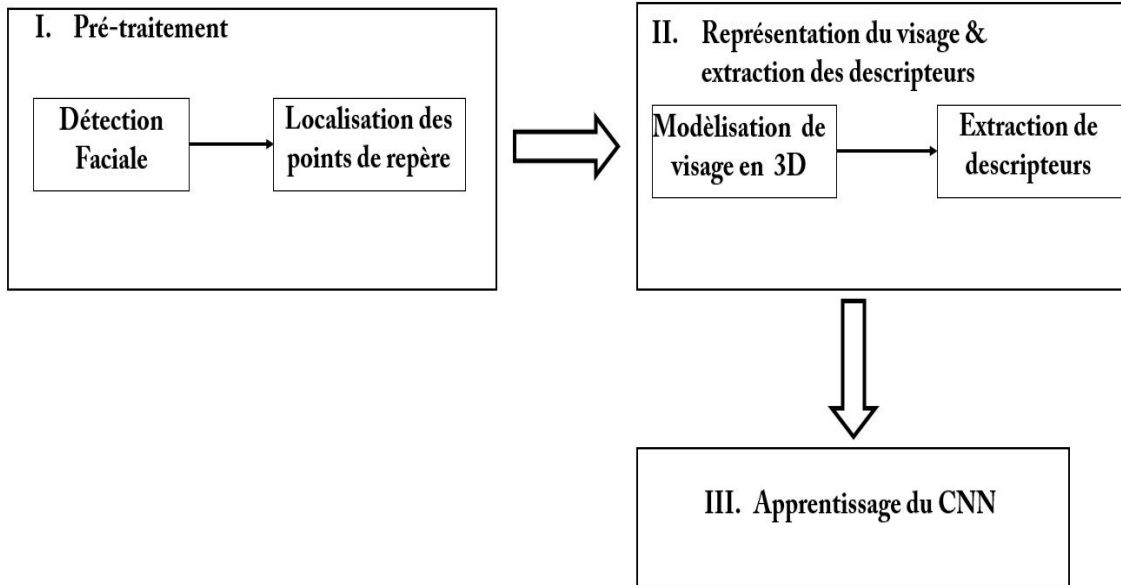


FIGURE 4.1: Schéma de la méthode Deep 3D-LBP

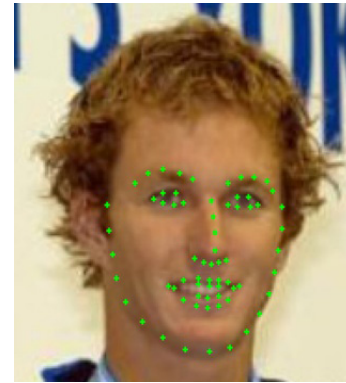
4.2 Pré-traitement : Détection faciale et localisation des points de repère

Nous utilisons le détecteur de visage Dlib [8] pour détecter et localiser les visages des images. Dlib est utilisé car en plus de la localisation et la détection des régions faciales, il localise 68 points de repère dans la région faciale qui permettent de construire le modèle 3D du visage.

Pour la localisation des points de repère, la bibliothèque dlib met en œuvre l'algorithme de Khazemi et Sullivan afin de détecter précisément un ensemble de 68 points de repère pour les visages en utilisant des ensembles d'arbres de régression [97]. Pour conclure, le processus de détection des points de repère se déroule en deux étapes comme montré dans la figure 4.2 : la première étape vise à détecter et localiser le visage (figure 4.2a) tandis que la deuxième détecte les 68 points de repères. Ces 68 points désignent les principales zones du visage à savoir la bouche, les yeux, les sourcils, le nez et la mâchoire.



(a) détection du visage



(b) détection des points de repère du visage

FIGURE 4.2: Exemple d'utilisation du détecteur Dlib [8]

Dlib estime l'emplacement de 68 coordonnées (x, y) qui correspondent aux points de repères, comme le montre la figure 4.2. À partir de cette représentation, nous pouvons identifier les sept régions du visage : Points de la mâchoire [1-17], points du sourcil droit [18-22], points du sourcil gauche [23-27], points du nez [28-36], points de l'œil droit [37-41], points de l'œil gauche [43-48], points de la bouche [49-68]. Dlib a prouvé sa robustesse de détection faciale dans des milieux encombrés. La figure 4.3 illustre un exemple de détection de plusieurs visages dans des conditions non contrôlées. Dlib détecte moins de visages dans des milieux encombrés que le MTCNN,. Mais, il donne une meilleure représentation grâce aux 68 points caractéristiques qu'il détecte dans la région du visage.

4.3 Représentation faciale en 3D et extraction des caractéristiques

Dans cette étape, un modèle 3D de visage est utilisé pour faire la correspondre avec les images 2D. Ensuite, nous utilisons le mesh-LBP [96] comme extracteur de caractéristiques de visage.

Dans ce travail, le modèle 3D utilisé est le modèle de visage de Surrey [98]. La première composante est l'ajustement de la pose (de la caméra). Étant donné un ensemble de points de repère en 2D et leurs correspondances connues dans le mo-

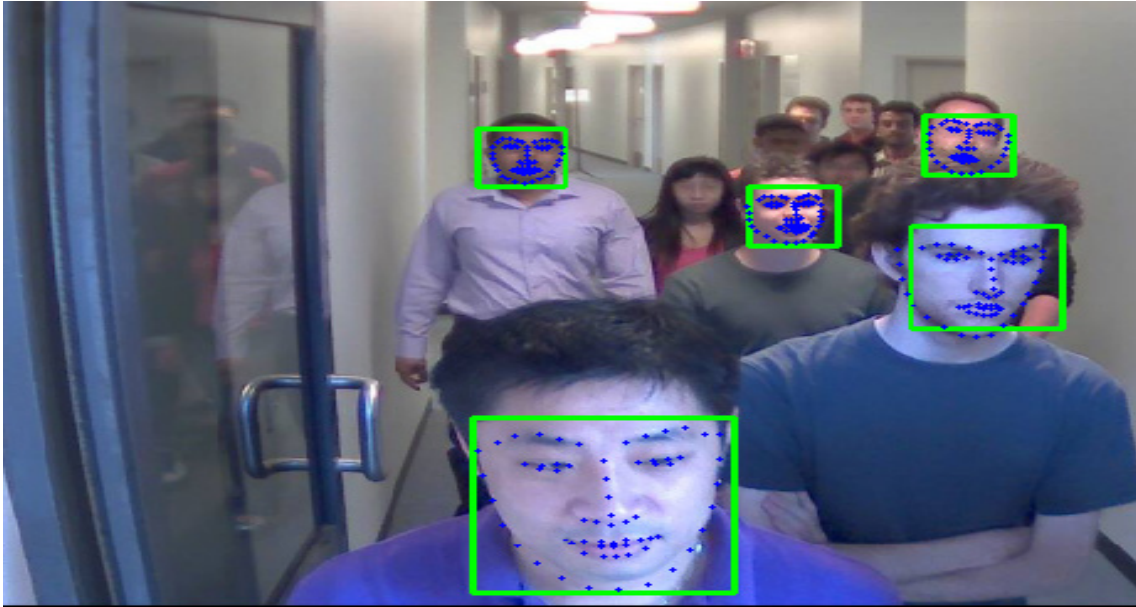


FIGURE 4.3: Exemple de détection de visages dans des milieux encombrés avec Dlib [8]

dèle morphable 3D, le but est d'estimer la meilleure pose du visage. Le deuxième composant consiste à reconstruire la forme 3D en utilisant la matrice estimée de la caméra. Il met en œuvre un ajustement simple de la forme aux points de repère, similaire à l'algorithme d'Aldrian et Smith [99]. Les étapes d'estimation de la pose et d'ajustement de la forme peuvent être itérées si nous le souhaitons afin de renouveler les estimations. Le modèle 3D du visage est extrait et stocké dans un *isomap*, comme montré dans la figure 4.4.

La figure 4.5 montre un exemple d'ajustement des points de repère pour l'image

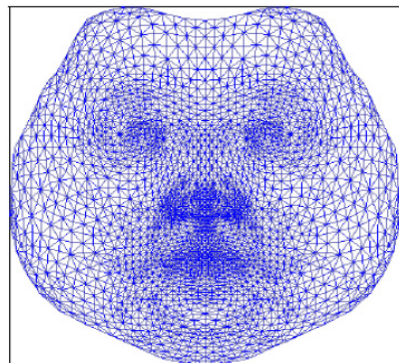


FIGURE 4.4: Représentation de la texture sous la forme d'un isomap

d'entrée (Figure 4.5a), la forme résultante et l'ajustement du modèle de caméra

sont illustrés à la figure 4.5c. Sur la figure 4.5b, les régions d’auto-occlusion sont représentées par des taches blanches.



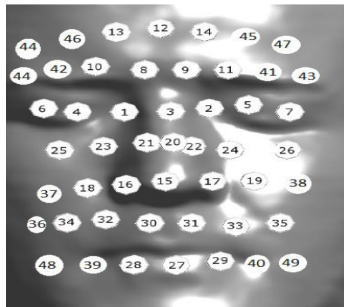
FIGURE 4.5: Un exemple de résultat de l’ajustement de point de repère [9].

Maintenant, nous allons présenter le descripteur mesh-LBP qui est utilisé pour fusionner les caractéristiques géométriques et d’apparence extraites des modèles de visage 3D. Dans la représentation standard du visage basée sur les LBP [100], une image de visage 2D est divisée en une grille des blocs rectangulaires, puis les histogrammes des descripteurs LBP sont extraits de chaque bloc et concaténés pour former une description globale du visage. Pour étendre ce schéma au modèle 3D du visage, nous devons d’abord partitionner la surface du visage en une grille de régions (l’équivalent des blocs dans 2D-LBP), calculer les histogrammes correspondants, puis les regrouper en une seule structure.

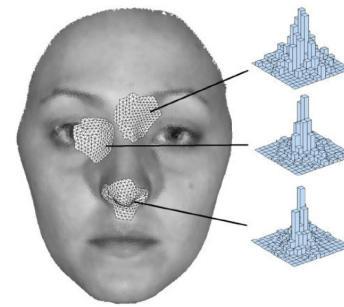
Le calcul du descripteur LBP sur une région 3D s’effectue selon les étapes suivantes :

- Tout d’abord, le plan formé par le bout du nez et les deux points de repère du coin interne des yeux est initialement calculé. Nous avons utilisé ces trois points de repère car ils sont les plus précis et détectables sur le visage. Ils sont également assez robustes aux expressions faciales [8]. À partir de ces points de repère, nous dérivons, par un simple calcul géométrique, un ensemble ordonné et régulièrement espacé de points sur ce plan.

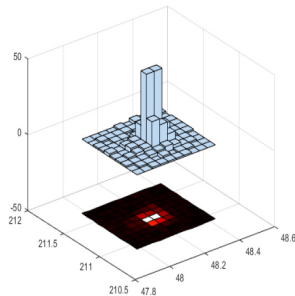
- Ensuite, le plan est légèrement incliné, d'une quantité constante, pour le rendre plus aligné avec l'orientation du visage, puis nous projetons cet ensemble de points sur la surface du visage, le long de la direction normale du plan. Le résultat de cette procédure est une grille ordonnée de points, qui définit un atlas pour les régions du visage qui diviseront la surface du visage. La grille contient 49 points formant 7×7 constellation comme le montre la figure 4.6a.
- Une fois la grille de points définie, nous extrayons un voisinage de facettes autour de chaque point de la grille. Chaque voisinage peut être défini par l'ensemble des facettes confinées dans une sphère, centré sur un point de la grille (Figure 4.6b).
- Calculer le descripteur mesh-LBP correspondant à chaque région (figure 4.6c).
- Regrouper les descripteurs calculés en une seule structure (image) (figure 4.6d).



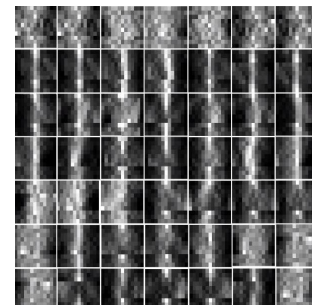
(a) Grille résultat de 49 points



(b) Extraction des Facets voisines



(c) Calcul du descripteur mesh-LBP multi-résolutions



(d) Final face descriptor image

FIGURE 4.6: Construction de l'image du descripteur.

4.4 Phase d'apprentissage

L'originalité que nous proposons dans cette partie est que l'entrée du CNN n'est pas l'image du visage mais plutôt l'image de descripteur du visage. Nous entraînons notre réseau CNN afin de classifier l'image du descripteur de visage créé à l'aide de mesh-LBP en entrée. L'architecture du CNN proposée est composée de deux couches de convolution (C), deux couches entièrement connectées (F), une couche de max-pooling (M) et une couche localement connectée (*Locally Connector Layer*) (L).

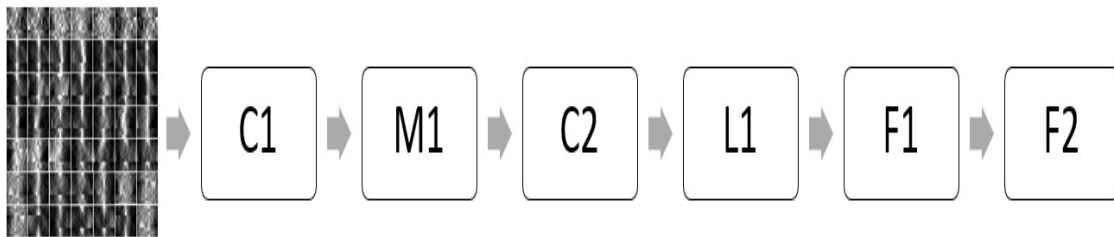


FIGURE 4.7: Architecture du réseau de neurones convolutionnel proposé.

La principale différence entre la couche de convolution et la couche localement connectée est que le filtre dans la couche de convolution est le même pour tous les neurones de sortie. En d'autres termes, nous utilisons un seul filtre pour calculer tous les neurones. Alors que, dans la couche localement connectée, chaque neurone possède son propre filtre. Ce type de couche permet au réseau d'apprendre différents types de caractéristiques pour différentes régions de l'image d'entrée. En fait, plusieurs chercheurs ont tiré profit de cette propriété, notamment pour les tâches de vérification des visages. Par exemple, les zones situées entre les yeux et les sourcils présentent un aspect très différent et ont une capacité de discrimination beaucoup plus élevée que les zones situées entre le nez et la bouche [101].

L'utilisation de couches locales ne complique pas l'extraction des caractéristiques, mais elle a un impact sur le nombre de paramètres soumis à l'apprentissage [91]. Cela signifie que le nombre de paramètres sera multiplié par le nombre de neurones de sortie, ce qui pourrait augmenter le nombre de paramètres dans notre réseau.

Cependant, dans une architecture CNN plus petite comme la nôtre, nous pourrions éviter ces problèmes. La taille de l'image du descripteur de visage est de 91×91 pixels. Ces images sont introduites dans notre CNN. La première couche convolutive (C1) possède 32 filtres de taille 11×11 . Ensuite, une couche de max-pooling (M1) avec un stride de 2. Cette dernière est suivie par une autre couche convolutive (C2) avec 16 filtres de taille 9×9 . La couche suivante (L1) est une couche localement connectée composée de 16 filtres. Enfin, deux couches entièrement connectées : F5 et F6. Ces couches sont capables de capturer les corrélations entre des caractéristiques de visage distantes. La sortie de la première couche entièrement connectée (F1) du réseau est utilisée comme notre vecteur de caractéristiques qui représentent le visage. La sortie de la dernière couche entièrement connectée (F2) introduit une fonction softmax à K classe. Si nous désignons par o_k la k^{me} sortie du réseau pour une entrée donnée, la probabilité assignée à la k -ième classe est la sortie de la fonction k-softmax :

$$P_k = \exp o_k / \sum_h^K o_h \quad (4.1)$$

Il est important de mentionner l'utilisation de la fonction d'activation ReLU [32] après la convolution, les couches connectées localement et entièrement connectées (sauf la dernière L6). De plus, nous utilisons la fonction de perte d'entropie croisée afin de maximiser la probabilité de la classe correcte (identification du visage). Nous entraînons notre architecture avec environ 500,000 images provenant de la base CASIA-WebFace [102], qui contient 494,414 images de 10,575 personnes collectées sur Internet. Puisque l'entrée du réseau CNN est une image de descripteur du visage, nous entraînons le réseau pour 10 époques sur l'ensemble des données.

4.5 Évaluation de la méthode Deep 3D-LBP

Dans cette évaluation, nous utilisons quatre jeux de données : Multi-PIE [10], Labeled faces in the wild LFW [11] et YouTube Face Dataset [12]. Notre évaluation comporte trois axes. Premièrement, nous évaluons l'efficacité de notre méthode pour

reconnaître des visages en variant la pose et l'illumination en utilisant le jeu de données de visages Multi-PIE. Ensuite, nous évaluons la robustesse de notre méthode face aux variations d'expressions faciales en mesurant le taux de reconnaissance sous 6 variations d'expressions faciales où on fait l'analyse 3D du visage. Nous testons notre méthode dans un environnement encombré par rapport principalement à la variation de la pose.

Nous allons utiliser la base Multi-PIE pour la reconnaissance des visages en présence de variations de la pose. Aussi, en combinant les variations de la pose et de l'illumination. La base de données de visages Multi-PIE contient plus de 750 000 images de 337 personnes enregistrées en quatre sessions sur une période de cinq mois. Les visages ont été acquis sous 15 angles de vue et 19 conditions d'éclairage différentes tout en affichant une gamme définie d'expressions faciales. En plus, des images frontales à haute résolution ont également été acquises pour chaque visage. Nous présentons dans la figure 4.10 un exemple de visage de la base Multi-PIE [10].



FIGURE 4.8: Exemple de de visage de la base Multi-PIE acquis dans des milieux encombrés et des conditions non contrôlées [10]

4.5.1 Reconnaissance des visages en présence de la variation à la pose

Nous pouvons classer les méthodes de la littérature en deux catégories : les méthodes 2D et 3D. Les méthodes 2D traitent la pose et l'illumination en utilisant les pixels de l'image ou les caractéristiques de l'image pour la reconnaissance. Alors que les méthodes 3D modélisent les variations de pose en se basant sur une approche d'analyse par synthèse [103]. La reconnaissance faciale 3D a le potentiel d'atteindre une meilleure précision que son homologue 2D en mesurant la géométrie des caractéristiques rigides sur le visage. Elle est plus efficace que reconnaissance faciale 2D surtout lors du changement d'éclairage, les différentes expressions faciales, le maquillage et l'orientation de la tête (pose) [104]. En d'autres termes, ces méthodes visent à faire correspondre un modèle de visage 3D à une image 2D annotée en entrée. Nous notons que l'annotation consiste essentiellement en la détection des points de repère du visage. Les résultats globaux sont résumés dans le tableau 4.1.

Globalement, les résultats montrent que la plupart des méthodes 3D peuvent obtenir

TABLE 4.1: Taux de reconnaissance (%) sous différentes poses avec la base de données Multi-PIE [10]

Méthodes & dates		-45°	-30°	-15°	+15°	+30°	+45°
2D	DAE [100],2009	69.0	81.2	91.0	91.9	86.5	74.3
	GMA [105],2012	75.0	74.5	82.7	92.6	87.5	65.2
	MRFs [106],2013	86.3	89.7	91.7	91	89	85.7
	SPAE [107],2014	84.9	92.6	96.3	95.7	94.3	84.4
	RFG [108],2014	86.4	91.2	96.0	96.1	90.90	85.4
	SF-VF +LBP [109],2020	91.43	93.88	91.14	90.91	92	87.14
3D	asthana [110],2011	74.1	91.0	95.7	95.7	89.5	74.8
	MDF [111],2012	78.7	94.0	99.0	98.7	92.2	81.8
	PAF [112],2013	84	99	99.33	99.67	99.67	98.33
	HPEN+PCA [113], 2015	88.5	95.4	97.2	98	95.7	89
	U-3DMM +(PCA) [114], 2016	91.2	95.7	96.8	96.9	95.3.	90.9
	U-3DMM +(HDF) [114], 2016	96.5	98.4	99.2	98.2	98.9	97.9
	ESO+LPQ [115],2017	91.7	95.3	96	96.7	95.3	90.3
Deep 3D-LBP	97.4	99.5	99.5	99.7	99.0	96.7	

des taux de précision plus élevés que les méthodes 2D, en particulier en présence d'une grande variation de pose ($\pm 45^\circ$). Néanmoins, SPAE [107], qui se caractérise par une capacité de modélisation non linéaire robuste, fonctionne plus efficacement que les autres méthodes 2D et fournit de meilleurs résultats que certaines méthodes 3D telles que Asthana [110] et MDF [111]. Mais cela reste limité, et nous pouvons noter que d'autres méthodes 3D (HPEN [113], et ESO [115]) dépassent ces résultats avec une grande marge de différence. HPEN [113], qui génère une image de visage neutre et frontale basée sur le 3DMM et la transformation 3D préservant l'identité, et qui utilise l'ACP pour la classification, n'obtient de bons résultats qu'en cas de faible variation de pose. En outre, le modèle U-3DMM peut modéliser à la fois la pose et la forme du visage plutôt que la pose uniquement [110, 111]. Nous présentons le résultat de l'U-3DMM en utilisant des caractéristiques de Gabor à haute dimension (HDF) [116] et des coefficients de l'ACP. L'U-3DMM (HDF) fonctionne beaucoup mieux que l'U-3DMM (PCA) en raison de la capacité de la caractéristique HDF à capturer les informations faciales globales et locales.

Cependant, la méthode que nous proposons est plus performante que les méthodes 2D et 3D. Cela est dû à la capacité de notre méthode à modéliser et à apprendre à la fois la variation de la pose et la variation de l'illumination. En résumé, notre méthode est plus fiable que les autres méthodes car elle traite les limitations de ces dernières. Nous utilisons des données 3D provenant de 3DMM pour résoudre les problèmes de variation importante de la pose, comme [114, 100, 111], et la caractéristique LBP pour être plus robuste contre l'illumination et l'expression. De plus, l'utilisation de CNN pour l'apprentissage nous permet d'obtenir des taux de précision élevés, plutôt que d'utiliser d'autres méthodes de classification [114, 115]. En outre, la frontalisation du visage [113] améliore le résultat, mais elle semble limitée, en particulier dans les degrés supérieurs, ce qui pourrait être causé par le classificateur utilisé (PCA).

4.5.2 Reconnaissance des visages invariante en fonction de la pose et de l'illumination

Les résultats présentés dans le tableau 4.2 comparent notre méthode Deep 3D-LBP à d'autres méthodes en variant simultanément l'illumination et la pose.

Les résultats du tableau 4.2 montrent que l'utilisation des méthodes du sous-espace

TABLE 4.2: Taux de reconnaissance Recognition (%) sous différentes poses et illumination avec la base de données Multi-PIE [10]

Méthodes & dates		-45°	-30°	-15°	+15°	+30°	+45°
Subspace learning	Li [117],2011	63.5	69.3	79.7	75.6	71.6	54.6
Deep learning	DNN-RL [101],2013	67.1	74.6	86.1	83.3	75.3	61.8
	MVP [118],2014	84.9	92.6	96.3	95.7	94.3	84.4
	DNN-CPF [119],2015	73	81.7	98.4	89.5	80.4	70.3
	LNFF-LRA [120],2017	77.2	87.7	94.9	94.8	88.1	76.4
	HPN [121], 2017	71.3	78.8	82.2	86.2	77.8	74.3
3D	3DMM	74.1	91	95.7	95.7	89.5	74.8
	U-3DMM,2016 [114]	73.1	86.9	93.3	91.3	81.2	69.7
	ESO-3DMM,2016 [115]	80.8	88.9	96.7	97.6	93.3	81.1
	GM-3DMM [122],2018	84.3	89.4	97.4	99	96.8	92
	Deep 3D-LBP	97.4	99.5	99.5	99.7	99.0	96.7

[117] offre les pires résultats. En outre, nous avons pu constater que notre méthode surpasse à la fois l'apprentissage profond et les méthodes basées sur la 3D. De plus, il est évident que l'utilisation d'un 3DMM est bien adaptée pour résoudre les problèmes causés par les variations extrêmes de pose et d'illumination. Par exemple, nous pouvons remarquer que la valeur maximale obtenue par les méthodes basées sur la DL est de 96,3% et 95,7% avec une variation de pose de -15° et $+15^\circ$ respectivement. Néanmoins, pour les méthodes basées sur la 3D, GM-3DMM [122] obtient 97,4 % et 99 % avec des variations de pose de $+15^\circ$ et $+15^\circ$ respectivement. En utilisant la Deep 3D-LBP, nous sommes en mesure d'obtenir des résultats beaucoup plus intéressants, et ceci est plus notable dans les grandes variations d'installation (97,4 %, 96,7 % obtenus en -45° et $+45^\circ$ respectivement par notre méthode contre 84,9 et 92 comme valeur maximale obtenue pour toutes les méthodes de l'état de l'art). Nous pouvons également remarquer la différence

en prenant l'exemple de U-3DMM [114] dans le tableau 4.1 et le tableau 4.2. Dans lequel, U-3DMM offre des résultats plus intéressants en variant la pose (96.5% dans -45° plutôt que pour la variation de l'illumination et de la pose (73,1% en -45°). Par contre, notre méthode reste fiable et donne de meilleurs résultats dans les deux cas.

4.5.3 Reconnaissance de visages avec variation des expressions faciales

Pour inclure des cas plus difficiles, nous avons testé notre méthode sur l'ensemble de données Bosphorus, qui présente une énorme variation dans les expressions faciales. Le tableau 3 fournit une comparaison complète entre notre méthode et les méthodes récemment publiées. Le tableau 4.3 fournit une comparaison complète entre notre méthode et les méthodes publiées.

TABLE 4.3: Taux de reconnaissance (%) en présence des expressions faciales avec la base de données Multi-PIE [10]

Méthodes & dates	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise
Li et al, 2011 [123]	100	88.7	76.8	92.9	95.3	95.5	98.6
Berretti et al,2013 [124]	97.9	85.9	81.2	90	92.5	93.95	91.5
Li et al,[125],2015	100	97.18	86.96	98.57	98.11	100	98.59
Azazi et al, [126],2015	81.25	82.5	90	86.25	97.5	67.5	83.75
Lei et al,[127],2016	98.96	94.12	88.24	98.55	98.08	96.08	96.92
Deng et al,[128],2017	100	95.8	92.8	97.7	95.3	98.5	98.6
Hariri et al ,2017, [129]	87.5	86.25	85.25	81	93	79.75	90.5
Abbad et al,[130],2018	100	95.77	88.41	81.41	88.68	96.97	92.96
Zhang et al,[131],2019	100	81.69	79.71	88.57	96.23	90.91	95.77
Deng et al, [132],2020	100	97.2	94.2	97.1	96.2	98.5	98.6
Liang et al ,2020, [133]	100	94.37	85.51	97.14	69.23	98.59	98.52
Atik et al,2021, [134]	98.68	78.87	81.16	80	97.17	95.45	95.77
Deep 3D-LBP	100	97.18	96.75	100	97.63	98.88	100

Nous concluons que notre solution est efficace par rapport aux méthodes de l'état de l'art. Plusieurs méthodes atteignent un taux de 100 % lorsque le visage est neutre. Mais, cette précision diminue lorsque les expressions changent. Par exemple, Zhang et al [131] obtiennent de meilleurs résultats en observant les expressions

émotionnelles Neutre et Heureux (100% et 96.23% respectivement). Cependant, il y a une baisse significative lorsqu'il s'agit d'une émotion de COLÈRE (81,6 %), de DÉCROI (79,71 %) et de PEUR (88,5 %). Cependant, notre méthode conserve des résultats supérieurs à ceux de la même catégorie.

4.5.4 Reconnaissance faciale sur l'ensemble de données LFW

La base LFW [11] contient des visages qui ont été extraits dans des conditions non contrôlées à partir de vidéo-surveillance. Les conditions d'éclairage ne sont pas maîtrisées et peuvent varier d'un visage à un autre. La résolution du visage change avec la distance par rapport à la caméra où ce visage a été pris. L'arrière plan varie également car il peut influencer le processus de reconnaissance faciale. Dans les images de visages de la base LFW la pose change ainsi que les expressions faciales. LFW peut être considéré aujourd'hui comme le plus important jeu de données de visage dans des conditions non contrôlées. Il contient plus de 13 000 images de visages collectés sur le Web. Il s'agit de visages souvent dans une position frontale et avec une bonne luminosité, mais l'occlusion par objets et l'expression des visages ne sont pas contrôlés. Nous montrons dans la figure 4.9 un exemple du même visage qui a été acquis dans des conditions non contrôlées. Nous avons testé la méthode

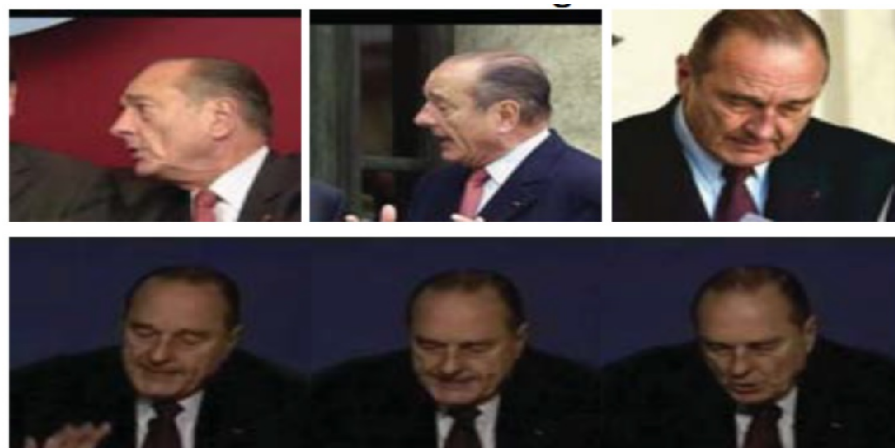


FIGURE 4.9: Exemple de de visage de la base LFW acquis dans des milieux encombrés et des conditions non contrôlées [11]

Deep 3D-LBP [5] sur cette base. Le résultat de reconnaissance de visages sur le jeu

de données LFW est montré dans le tableau 4.4.

TABLE 4.4: Taux de reconnaissance (%) sur le jeu de données LFW [11]

Méthodes & dates	Précision
Light-CNN [135],2018	98.13 %
M2FPA [136],2019	99.41 %
FR-CNN [137],2020	99.2 %
Fan-Face [138],2020	99.56 %
Deep 3D-LBP [5], 2021	99.59 %

Toutes les méthodes de l'état de l'art citées dans le tableau 4.4 utilisent le même organigramme que la méthode Deep 3D-LBP.

Ils sont composés d'une étape de pré-traitement de détection du visage. Ensuite, vient une étape de représentation et d'extraction des caractéristiques robustes et fiables face aux différents changements que peut subir un visage. A la fin la reconnaissance faciale se fait par un apprentissage CNN.

4.5.5 Reconnaissance faciale sur l'ensemble de données YTF

Nous validons ensuite notre méthode d'extraction des images clés Deep 3D-LBP sur le jeu de données YouTube Faces Dataset (YTF) [12]. Il s'agit d'une base de données d'images issues de vidéos contenant des visages conçue pour étudier le problème de la reconnaissance faciale en conditions réelles (qualités de résolution, de luminosité non contrôlées). La qualité des images vidéo YouTube est généralement moins bonne que celle de l'ensemble LFW. Ceci est du principalement au mouvement de la caméra et à la distance entre les individus et la caméra.

Le jeu de données contient 621 126 images obtenues à partir de 3 425 vidéos de 1 595 personnes différentes. Toutes les vidéos ont été téléchargées à partir de YouTube. En moyenne, 2.15 vidéos sont disponibles pour chaque personnalité présente dans la base. Nous présentons dans la figure quelques exemples de visages issus de la base YTF. Les résultats sont résumés dans le Tableau 4.5.



FIGURE 4.10: Exemple de visage de la base YTF acquis dans des milieux encombrés et des conditions non contrôlées [12]

TABLE 4.5: Taux de reconnaissance sur le jeu de données YTF [13]

Méthodes & Dates	#image	# réseau	Précision
DeepID2+,2015 [139]	0.3 M	25	93.2 %
Center loss ,2016[140]	0.7 M	1	94.9 %
NormFace,2017 [141]	1.5 M	1	94.72 %
DFCL,2017 [142]	4.7M	1	96.06 %
Deep 3D-LBP [5],2021	0.5	1	94.97 %

Notre méthode Deep 3D-LBP, a un taux de précision supérieur aux autres méthodes similaires de la littérature. Par contre, Deep 3D-LBP nous donne un taux de précision inférieur à celui de [142] qui utilise des ensembles de données d'apprentissage très grand de l'ordre de 4 millions d'images de visage avec un réseau beaucoup qui contient un nombre de couches beaucoup plus élevé que le notre. Malgré ceci, nous

avons un taux de précision qui s'approche malgré la grande différence dans la taille des échantillons d'apprentissage plus petits pour l'apprentissage. Nous devons songer dans le futur de tester notre méthode avec une base d'apprentissage très grande.

En outre, les tests précédents ont prouvé l'efficacité de notre méthode Deep 3D-LBP pour la tâche de classification et de reconnaissance faciale. Une amélioration est toujours possible pour répondre aux défaillances dues au mouvement de la caméra et la grande distance entre les individus et la caméra. Néanmoins, nous restons meilleurs que les méthodes similaires dans la littérature.

Conclusion

Nous avons proposé dans ce quatrième chapitre qui fait partie des contributions présentées autour du deuxième axe de recherche, une nouvelle méthode d'extraction des images clés basée sur la description 2D/3D du visage. Ceci dans le principal objectif de décrire au mieux le visage qui peut subir différentes transformations dues à la pose, les émotions et le changement d'illumination. Nous avons appliqué l'opérateur de détection de visage Dlib qui permet de détecter 68 points caractéristiques sur le visage. Ces points seront utilisés pour construire un modèle 3D du visage pour le remettre dans une position neutre. Une fois fait, nous calculons un nouveau descripteur proposé Mesh LBP et nous appliquons enfin un apprentissage par CNN pour la reconnaissance et la classification du visage. La méthode Deep 3D-LBP proposée a donnée des résultats très encourageants comparé aux méthodes similaires dans la littérature.

Travaux en cours et perspectives

Depuis l'obtention de mon doctorat en informatique en mars 2011, mes travaux de recherche se sont focalisés sur les résumés de vidéo afin de permettre la recherche d'objets génériques ou la reconnaissance faciale dans des grandes bases de vidéos. L'objectif étant d'introduire en requête une image d'un objet ou un visage et de récupérer les vidéos contenant cet objet ou les vidéos où le visage apparaît.

En ce qui concerne le premier axe qui consiste à construire des résumés de vidéos sous forme d'images clés pour faciliter la recherche d'objets génériques, mes travaux actuels et futurs visent à extraire les objets saillants à partir d'une vidéo afin de construire un résumé visuel de toute la base de vidéos. Ce résumé visuel va contenir les objets les plus présents dans la base et sera un point de départ pour la recherche par la suite. Le système visuel humain reconnaît particulièrement des régions d'intérêt dans des scènes complexes appelées régions d'intérêts ou régions saillantes. La détection d'objets saillants dans des vidéos (SOD) est un domaine de recherche qui a pour objectif de détecter l'objet saillant qui attire le plus l'attention visuelle. Le résultat de la détection d'objets saillants dans une vidéo est une carte de saillance dans laquelle chaque pixel est étiqueté par une valeur réelle prise dans l'intervalle $[0,1]$ pour indiquer sa probabilité d'appartenir à un objet saillant. Plus la valeur est élevée, plus la saillance est élevée.

Dans la littérature, les méthodes traitant de la détection des objets saillants peuvent être classées en deux grandes approches : les approches basées image et les ap-

proches basées vidéo. Les approches basées image effectuent une segmentation de l'image. D'un autre côté, le système visuel humain est très sensible aux mouvements. Les objets en mouvement dans une vidéo ont le plus de probabilité d'être des objets saillants. Les approches basées vidéo détectent l'objet saillant en utilisant des indices, à la fois, du domaine spatial que du domaine temporel. Enfin, une carte de saillance spatio-temporelle globale est obtenue en combinant la carte de saillance spatiale et la carte de saillance temporelle. Le type d'applications que nous pouvons viser dans le futur est par exemple celui de la conduite autonome et la reconnaissance automatique des panneaux de signalisation. Ceci contribue à l'amélioration de la sécurité de la conduite. Un autre domaine application est celui de la vidéo surveillance. Les objets tels que les humains, les voitures attirent généralement un grand intérêt et doivent être soigneusement observé. Pour capter l'évolution de ces objets spécifiques, le calcul de la saillance fournit un indice important pour localiser les objets cibles. Les méthodes de calcul de saillance basées vidéo insistent uniquement sur l'étiquetage de chaque pixel de l'image vidéo en indiquant "sailant" ou "non sailant". Enfin, les objets vidéo saillants sont défini à l'échelle de toute la vidéo entière comme étant les objets qui apparaissent le plus tout au long de la vidéo. Les approches de détection d'objets saillants basées vidéo peuvent aussi être classées en deux grandes méthodes : méthodes les traditionnelles et méthodes utilisant l'apprentissage profond. Les méthodes traditionnelles de détection d'objet saillants sont basées sur la segmentation des images de la vidéo en deux classes : avant plan (objet saillant) et l'arrière-plan. Avec ce genre de méthodes nous pouvons rater des régions saillantes lorsque l'objet saillant touche les bords de l'image ou lorsque la région saillante est composée de deux objets distants. Les méthodes d'extraction d'objets saillants basées apprentissage profond sont récemment devenues un axe de recherche très actif. Ces techniques ont considérablement amélioré la détection des objets saillants dans les vidéos. Les méthodes basées sur l'apprentissage profond peuvent atteindre des performances élevées, mais elles sont largement dépendantes des jeux de données d'apprentissage. Actuellement, une approche est en cours de

proposition et de validation à travers des résultats expérimentaux. Un article dans une revue internationale est aussi en cours de préparation.

Par ailleurs, il nous semble important de poursuivre les efforts déployés dans le second axe de recherche. Dans nos travaux futurs, nous nous focaliserons sur la technique de reconstruction des images faciales à partir d'une image donnée. L'objectif de la reconstruction est de fournir une image faciale frontale et neutre à partir d'une image originale qui n'est ni dans une position frontale ni neutre en émotions. Actuellement, nous sommes entrain d'étudier la possibilité de contribution dans cette étape qui d'après nos premières conclusions réside essentiellement dans deux volets : premièrement, la reconstruction peut améliorer la qualité des images clés extraites des vidéos dans le cas où on n'arrive pas à trouver des images frontales pour certaines identités. De cette façon, on peut très nettement améliorer le processus de reconnaissance faciale. Deuxièmement, les images frontales qui seront générées par notre méthode seront exploitées pour représenter l'identité des individus qui apparaissant dans une vidéo avec leurs images frontales neutres en émotions. Ceci sera très utile dans la phase de recherche. Ces ensemble d'images frontales extraits à partir de chaque vidéo peuvent servir aussi pour d'autre applications à savoir le suivie des mouvements des personnes, création des Croquis du visage (Face Sketch). Pour ce faire, nous proposons d'utiliser Les Réseaux Antagonistes Génératifs (Generative Adversarial Network GAN). En effet, ce réseau permet de générer une image faciale, dans une position frontale et neutre. Dans ce travail, nous utiliserons une variante de ces modèles générateurs appelés les Réseaux Antagonistes Génératifs Conditionnelles (Conditional Generative Adversarial Network), ou CGAN en abrégé. Les CGAN est un type de GAN qui implique la génération conditionnelle d'images. C'est à dire, fournir des informations supplémentaires à l'entrée du modèle pour lui guider à générer les images demandées. Cette condition, ou information peut être une étiquette de classe, une information supplémentaire à savoir une mesure ou pose, descripteur de forme, ou des coefficients, ou une image. Ces informations permettent la génération ciblée d'images d'un type donné. Ce modèle est constitué

de deux parties. La première partie du réseau CGAN nommé Générateur qui vise à produire une image faciale de la même identité à partir d'une image non frontale. Tandis que la partie discriminateur va prédire si l'image générée est une image réelle ou image reconstruite. Le générateur continue à créer des nouvelles images et d'affiner son processus jusqu'à ce que le discriminateur ne puisse plus faire la différence entre les images générées et les images réelles. Pour apprendre ce réseau CGAN à construire des images frontales et neutres, on donne en entrée trois images faciales pour la même identité (image frontale, image où le visage est tourné à droite, image où le visage est tourné à gauche). Ces images seront fournies au modèle générateur sous la forme de deux couples (image frontale, image du visage tourné à droite), (image frontale, image du visage à gauche). L'image frontale est considérée comme l'information supplémentaire qui va servir à guider le générateur CGAN à générer à partir de l'image du visage tournée, une image frontale qui devrait être similaire à l'image donnée en entrée. Ce générateur fonctionne à base des auto-encodeurs. Ce dernier est un algorithme d'apprentissage non supervisé à base de réseaux de neurones artificiels. Les auto-encodeurs sont utilisés principalement pour la compression de données, pour l'extraction de caractéristiques, filtrer des images, ou encore pour générer des nouvelles images. L'architecture d'un auto-encodeur est constituée de deux parties. En premier lieu, nous trouvons un encodeur. L'encodeur permet de condenser l'information disponible initialement (image, texte, audio, etc.) en extrayant des descripteurs pour décrire l'information initiale. Le vecteur obtenu par l'encodeur est de taille beaucoup plus petite que l'information initiale. La deuxième composante est appelée décodeur. Il se charge de reconstruire l'information de départ, à partir du vecteur compressé. Nous sommes entrain de tester expérimentalement cette approche proposée. Les premières expériences montrent des résultats très encourageants. Nous avons effectivement réussi à générer des images de visage frontales, neutres en émotions à partir d'images non frontales et non neutres en émotions. Nous projetons aussi dans des travaux futurs, d'étudier les vidéos du domaine du e-learning. En effet, nous voulons estimer le pourcentage

où un élève a bien suivi sa séance en ligne. Nous estimons qu'il ne suit pas la séance lorsque son visage n'est pas dans une position frontale, lorsque son visage a certaines émotions comme les yeux fermés ou lorsque son visage est occulté par un objet externe. Une première étape consistera tout d'abord à former un échantillon d'apprentissage qui sont malheureusement très rares dans ce domaine de recherche qui reste encore ouvert à beaucoup de perspectives.

Nous vivons actuellement dans un monde entouré d'objets connectés qui utilisent de modèles d'apprentissage automatique. Au cours de nos journées, nous utilisons ces modèles de machine learning plus que nous ne le pensons. Les tâches quotidiennes telles que parcourir les réseaux sociaux, prendre une photo, consulter des vidéos, dépendent toutes des modèles d'apprentissage automatique. Nous savons tous que pour l'exécution de ces modèles, nous avons besoin de systèmes informatiques suffisamment puissants pour les gérer. Ainsi, la plupart de ces modèles fonctionnent sur d'énormes centres de données avec des clusters de CPU et de GPU (même des TPU dans certains cas).

TinyML est un domaine d'étude de l'apprentissage automatique et des systèmes embarqués qui explore les types de modèles que nous pouvons exécuter sur de petits appareils connectés à faible puissance comme les micro-contrôleurs. Il permet l'exécution des architectures profondes directement sur des objets connectés et non par requête à un serveur sur le cloud. L'idée est de permettre aux appareils TinyML de fonctionner et de prendre des décisions en mode autonome et en temps réel. La conception des architectures profondes entre grandement en jeu. Il faut concevoir des architectures efficaces et qui ne soient pas très coûteuses en temps de calcul. Dans ce cas figure, l'appareil TinyML pourra s'appuyer sur l'apprentissage par renforcement afin de prendre les bonnes décisions. Ainsi, le temps de réponse sera rapide. Nos données ne seront stockés nulle part et on consommera moins de bande passante internet.

Bibliographie

- [1] Bahroun S. Gharbi H. and Zagrouba E. A novel key frame extraction approach for video summarization. In *International Conference on Computer Vision Theory and Applications*, volume 4, pages 146–153. SCITEPRESS, 2016. doi:DOI:10.5220/0005725701460153.
- [2] Zagrouba E. Gharbi H., Bahroun S. Key-frame extraction for video summarization using local description and repeatability graph clustering. In *Multimedia Tools and Applications*, volume 13, pages 507–515. Springer, 2019. doi:https://doi.org/10.1007/s11760-018-1376-8.
- [3] Zagrouba E. Bahroun S., Abed R. Ks-fqa : Keyframe selection based on face quality assessment for efficient face recognition in video. In *IET Image Processing*, page https://doi.org/10.1049/ipr2.12008, 2020.
- [4] Zagrouba E. Abed R., Bahroun S. Keyframe extraction based on face quality measurement and convolutional neural network for efficient face recognition in videos. In *Multimedia Tools and Applications*, volume 80, pages 23157–23179. Springer, 2021. doi:https://doi.org/10.1007/s11042-020-09385-5.
- [5] Zagrouba E. Bahroun S., Abed R. Deep 3d-lbp : Cnn-based fusion of shape modelling and texture descriptors for accurate face recognition. In *The Visual*

- Computer*, pages 1–16. Springer, 2021. doi:<https://doi.org/10.1007/s00371-021-02324-x>.
- [6] Bahroun S. Gharbi H. and Zagrouba E. Robust interest points matching based on local description and spatial constraints. In *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, volume 9443, 2015. doi:<https://doi.org/10.1117/12.2179923>.
 - [7] K.Zhang, Z.Zhang, Z.Li, Y.Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. In *IEEE Signal Processing Letters*, volume 23, pages 1499 – 1503, 2016.
 - [8] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5) :75–174, 2010.
 - [9] Manuel Grand-Brochier. *Descripteurs 2D et 2D+ t de points d'intérêt pour des appariements robustes*. PhD thesis, 2011.
 - [10] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5) :807–813, 2010.
 - [11] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild : A database for studying face recognition in unconstrained environments. 2008.
 - [12] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.
 - [13] Michalis Vrigkas, Christophoros Nikou , Ioannis A. Kakadiaris. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
 - [14] Amit Bora and Shanu Sharma. A review on video summarization approaches : Recent advances and directions. In *2018 International Conference on Ad-*

- vances in Computing, Communication Control and Networking (ICACCCN)*, pages 601–606. IEEE, 2018.
- [15] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110, 2004.
- [16] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf : Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.
- [17] Messaoudi M. Gharbi H., Bahroun S. and Zagrouba E. Key frames extraction using graph modularity clustering for efficient video summarization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1502–1506. IEEE, 2017. doi:10.1109/ICASSP.2017.7952407.
- [18] Bahroun S. Messaoudi M. and Zagrouba E. Video summarization based on local features. In *25th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision in Cooperation with EUROGRAPHICS Association*, pages 2464–4617, 2017. doi:https://dSPACE5.zcu.cz/handle/11025/29606.
- [19] Bahroun S. Gharbi H., Massaoudi M. and Zagrouba E. Key frames extraction based on local features for efficient video summarization. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 275–285. Springer, 2016.
- [20] Gharbi H. Bahroun S. and Zagrouba E. Local query on satellite images based on interest points. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 4508–4511, 2014. doi:doi:10.1109/IGARSS.2014.6947494.
- [21] F.Schroff,D.Kalenichenko, J.Philbin. FaceNet : A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

- [22] Rahma Abed, Sahbi Bahroun, and Ezzeddine Zagrouba. Toward a robust shape and texture face descriptor for efficient face recognition in the wild. In *International Conference on Computer Analysis of Images and Patterns*, pages 319–328. Springer, 2021. doi:https://doi.org/10.1007/978-3-030-89131-2_29.
- [23] Rahma Abed, Sahbi Bahroun, and Ezzeddine Zagrouba. Face retrieval in videos using face quality assessment and convolution neural networks. In *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 399–405. IEEE, 2020. doi:[10.1109/ICCP51029.2020.9266253](https://doi.org/10.1109/ICCP51029.2020.9266253).
- [24] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, pages 404–420. Springer, 2000.
- [25] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. Local binary patterns for still images. In *Computer vision using local binary patterns*, pages 13–47. Springer, 2011.
- [26] Luo Juan and Oubong Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, 3(4) :143–152, 2009.
- [27] PM Panchal, SR Panchal, and SK Shah. A comparison of sift and surf. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2) :323–327, 2013.
- [28] Ebrahim Karami, Siva Prasad, and Mohamed Shehata. Image matching using sift, surf, brief and orb : performance comparison for distorted images. *arXiv preprint arXiv :1710.02726*, 2017.
- [29] Jinke Li and Geng Wang. An improved sift matching algorithm based on geometric similarity. In *2015 IEEE 5th International Conference on Electronics Information and Emergency Communication*, pages 16–19. IEEE, 2015.

- [30] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59, 1996.
- [31] Bahroun Mohamed Sahbi. *Construction de thesaurus pour la recherche interactive dans une base d'images satellitaires*. PhD thesis, INRIA Saclay, 2011.
- [32] Ee Sin Ng and Nick G Kingsbury. Matching of interest point groups with pairwise spatial constraints. In *2010 IEEE International Conference on Image Processing*, pages 2693–2696. IEEE, 2010.
- [33] Thomas O Binford and Tod S Levitt. Quasi-invariants : Theory and exploitation. In *Proc. DARPA Image Understanding Workshop*, pages 819–829, 1993.
- [34] Ali Ismail Awad and Mahmoud Hassaballah. Image feature detectors and descriptors. *Studies in Computational Intelligence. Springer International Publishing, Cham*, 2016.
- [35] Gabor J. Székely et Maria L. Rizzo. Hierarchical clustering via joint between-within distances : Extending ward's minimum variance method. In *Journal of Classification*, volume vol. 22, no 2, pages p. 151–183, 2005. doi : DOI10.1007/s00357-005-0012-9.
- [36] Jean-Paul Benzécri. Pratique de l'analyse des données. In *Edition Dunod*, volume T1 (analyse des correspondances. Exposé élémentaire), 1984.
- [37] Chinh Dang, Abdolreza Moghadam, and Hayder Radha. Rpca-kfe : key frame extraction for consumer video based robust principal component analysis. *arXiv preprint arXiv :1405.1678*, 2014.
- [38] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5) :75–174, 2010.

- [39] Stéphan Cléménçon, Hector De Arazoza, Fabrice Rossi, and Viet Chi Tran. Hierarchical clustering for graph visualization. *arXiv preprint arXiv :1210.5693*, 2012.
- [40] J-J Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graphs. *Statistics and computing*, 18(2) :173–183, 2008.
- [41] Hugo Zanghi, Christophe Ambroise, and Vincent Miele. Fast online graph clustering via erdős–rényi mixture. *Pattern recognition*, 41(12) :3592–3599, 2008.
- [42] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2) :026113, 2004.
- [43] Satu Elisa Schaeffer. Graph clustering. *Computer science review*, 1(1) :27–64, 2007.
- [44] Gaurav Agarwal and David Kempe. Modularity-maximizing graph communities via mathematical programming. *The European Physical Journal B*, 66(3) :409–418, 2008.
- [45] Fabrice Rossi and Nathalie Villa-Vialaneix. Représentation d’un grand réseau à partir d’une classification hiérarchique de ses sommets. 2011.
- [46] Padmavathi Mundur, Yong Rao, and Yelena Yesha. Keyframe-based video summarization using delaunay clustering. *International Journal on Digital Libraries*, 6(2) :219–232, 2006.
- [47] Marco Furini, Filippo Geraci, Manuela Montangero, and Marco Pellegrini. Stimo : Still and moving video storyboard for the web scenario. *Multimedia Tools and Applications*, 46(1) :47, 2010.
- [48] Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. Vsumm : A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1) :56–68, 2011.

- [49] Marcos Vinicius Mussel Cirne and Helio Pedrini. Viscom : A robust video summarization approach using color co-occurrence matrices. *Multimedia Tools and Applications*, 77(1) :857–875, 2018.
- [50] Ruxandra Tapu and Titus Zaharia. A complete framework for temporal video segmentation. In *2011 IEEE International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, pages 156–160. IEEE, 2011.
- [51] Cai Bo, Zhang Lu, and Zhou Dong-ru. A study of video scenes clustering based on shot key frames. *Wuhan University Journal of Natural Sciences*, 10(6) :966–970, 2005.
- [52] Omaima NA Al-Allaf. review of face detection systems based artificial neural networks algorithms. *International Journal of Multimedia Its Applications (Ijma)*, 6(1), 2014.
- [53] Yuzhen Niu, Yini Zhong, Wenzhong Guo, Yiqing Shi, and Peikun Chen. 2d and 3d image quality assessment : A survey of metrics and challenges. *IEEE Access*, 7 :782–801, 2018.
- [54] Pavel Korshunov and Wei Tsang Ooi. Video quality for face detection recognition and tracking. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 7(3) :1–21, 2011.
- [55] Lingling Ke Niu Yuzhen and Wenzhong Guo. Evaluation of visual saliency analysis algorithms in noisy images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 27(6) :915–927, 2016.
- [56] Emmanouil Kafetzakis, Christos Xilouris, Michail Alexandros Kourtis, Marcos Nieto, Iveel Jargalsaikhan, and Suzanne Little. The impact of video transcoding parameters on event detection for surveillance systems. *IEEE International Symposium on Multimedia*, 7(3) :333–338, 2013.

- [57] Yuming Fang, Yuan Yuan, Leida Li, Jinjian Wu, Weisi Lin, and Zhiqiang Li. Performance evaluation of visual tracking algorithms on video sequences with quality degradation. *IEEE Access*, 5 :2430–2441, 2013.
- [58] A.Kumar,A.Kaur,M.Kumar. Face detection techniques : a review. In *Artificial Intelligence Review*, pages 1–22, 2018.
- [59] D.King. Dlib-ml : A machine learning toolkit. In *Journal of Machine Learning Research*, volume 10, pages 1755–1758, 2009.
- [60] P.Viola, M.Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Society Conference on Computer Vision and Pattern Recognition CVPR* , volume 10, pages 511–518, 2001.
- [61] F. Wen C. Zhu and J. Sun. A rank-order distance based clustering algorithm for face tagging. In *CVPR*, volume doi : 10.1109/CVPR.2011.5995680, pages pp. 481–488, 2011.
- [62] Chunhui Zhu, Fang Wen, and Jian Sun. A rank-order distance based clustering algorithm for face tagging. In *CVPR 2011*, pages 481–488. IEEE, 2011.
- [63] K.Nasrollahi, Th B.Moeslund. Face quality assessment system in video sequences. In *Biometrics and Identity Management. Springer Berlin Heidelberg*, 2008.
- [64] Xuan Qi and Chen Liu. Gpu-accelerated key frame analysis for face detection in video. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 600–605. IEEE, 2015.
- [65] K.Nasrollahi, Th B.Moeslund. Summarization of surveillance video sequences using face quality assessment. In *International Journal of Image and Graphics*, volume 11, pages 207–233, 2011.
- [66] Agarwal Vishal. Deep face quality assessment. *arXiv preprint arXiv :1811.04346*, page 6, 2018.

- [67] X.Qi, Ch.Liu. GPU-accelerated key frame analysis for face detection in video. In *IEEE workshop on Delay Sensitive Video Computing in the Cloud DSVCC*, 2015.
- [68] V.Struc, J.Z.Gros, S.Dobrisek, N.Pavesic. Exploiting representation plurality for robust and efficient face recognition. In *International Electrotechnical and Computer Science Conference*, page 121–124, 2013.
- [69] Ch.Szegedy, W.Liu, Y.Jia, P.Sermanet, S.Reed, D.Anguelov, D.Erhan, V.Vanhoucke, A.Rabinovich. Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*, page 1–9, 2015.
- [70] K.He, X.Zhang, Sh.Ren, J.Sun. Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision*, page 1026–1034, 2015.
- [71] X.Qi, C.Liu, S.Schuckers. Boosting Face in Video Recognition via CNN based Key Frame Extraction. In *international Conference of Biometrics(ICB)*, 2018.
- [72] D.P.Kingma, J.Ba. Adam : A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [73] F.Solina, P.Peer, B.Batagelj, S.Juvan, J.Kovac. Colorbased face detection in the 15 seconds of fame art installation. In *International Conference on Computer Vision/Computer Graphics Collaboration for Model-based Imaging, Rendering, Image Analysis and Graphical special Effects*, 2003.
- [74] UK University ofEssex. Face recognition data. URL : <https://cswww.essex.ac.uk/mv/allfaces/index.html>.
- [75] ATT Laboratories Cambridge. Face database. URL : <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>.

- [76] D.J.Kriegman K.C.Lee, J.Ho. Acquiring linear subspaces for face recognition under variable lighting. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 27, pages 684–698, 2005.
- [77] V. Nair, G.E.Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning (ICML)*, page 807–814, 2010.
- [78] S.Hochreiter D.A.Clevert, Th.Unterthiner. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations ICLR*, 2016.
- [79] G.Jiuxiang, W.Zhenhua, K.Jason, M.Lianyang, S.Amir, S.Bing, L.Ting, W.Xingxing, W.Gang, C.Jianfei. Recent advances in convolutional neural networks. In *International journal of Pattern Recognition*, volume 77, pages 354–377, 2017.
- [80] Y.LeCun, L.Bottou, Y.Bengio, P.Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.
- [81] A.Krizhevsky. Learning multiple layers of features from tiny images. In *Technical report, University of Toronto*, 2009.
- [82] S.Vignesh, K.M.Priya, S.Channappayya. Face image quality assessment for face selection in surveillance video using convolutional neural networks. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 577–581, 2015.
- [83] C.Liu X.Qi and S.Schuckers. Boosting face in video recognition via cnn based key frame extraction. *international Conference of Biometrics(ICB)*, pages 132–139, 2018.
- [84] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Arxiv*, pages 1409–1556, 2014.

- [85] K.Nasrollahi. *From face detection to face super-resolution using face quality assessment*. PhD thesis, Aalborg : Faculty of Engineering and Science, Aalborg University, 2009.
- [86] J.Chen, Y.Deng,G.Bai and G.Su,. Face image quality assessment based on learning to rank. In *Signal Processing Letters, IEEE*, volume 22, pages 90–94, 2015.
- [87] Adam Fourney and Robert Laganieri. Constructing face image logs that are both complete and concise. In *Fourth Canadian Conference on Computer and Robot Vision (CRV'07)*, pages 488–494. IEEE, 2007.
- [88] Y.Wong, Sh.Chen, S.Mau, C.Sanderson and B.C. Lovell. Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition. In *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 81–88, 2011.
- [89] Mikhail Nikitin, Vadim Konushin, and Anton Konushin. Face quality assessment for face verification in video. In *'2014*, pages 111–114, 2014.
- [90] K.Anantharajah, S.Denman, D.Tjondronegoro, S.Sridharan, C.Fookes, X.Guo. Quality Based Frame Selection for Face Clustering in News Video. In *International Conference on Digital Image Computing : Techniques and Applications (DICTA)*, 2013.
- [91] Y.Taigman, M.Yang, M.A.Ranzato and L.Wolf. DeepFace : Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1701–1708, 2014.
- [92] Y.Taigman, M.Yang, M.A.Ranzato, L.Wolf. Deep Learning Face Representation from Predicting 10,000 Classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1891–1898, 2014.
- [93] O.M.Parkhi, A.Vedaldi and A.Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

- [94] Y.Wen, K.Zhang, Z.LiYu and Y.Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.
- [95] W.Deng, B.Chen, Y.Fang and J.Hu. Deep correlation feature learning for face verification in the wild. In *IEEE Signal Processing Letters*, volume 24, pages 1877–1881, 2017.
- [96] Naoufel Werghi, Claudio Tortorici, Stefano Berretti, and Alberto Del Bimbo. Boosting 3d lbp-based face recognition by fusing shape and texture descriptors on the mesh. *IEEE Transactions on Information Forensics and Security*, 11(5) :964–979, 2016.
- [97] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2014.
- [98] P. Huber, G. Hu, R. Tena, P. Mortazavian, W.P. Koppen, W. Christmas and M. Rätzsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 79–86, 2016.
- [99] O. Aldrian and W. A. P. Smith. Inverse rendering of faces with a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5), 2013.
- [100] Yoshua Bengio. Learning deep architectures for ai. 2009.
- [101] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. pages 113–120, 2013.
- [102] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv*, 2014.

- [103] Guosheng Hu, Fei Yan, Josef Kittler, William Christmas, Chi Ho Chan, Zhenhua Feng, and Patrik Huber. Efficient 3d morphable face model fitting. *Pattern Recognition*, 67 :366–379, 2017.
- [104] Q; Niu S.; et al. Okuwobi, I. P.; Chen. Three-dimensional (3d) facial recognition and prediction. In *Signal, Image and Video Processing.*, page 1151–1158, 2016.
- [105] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis : A discriminative latent space. pages 2160–2167, 2012.
- [106] Ho Huy Tho and Rama Chellappa. Pose-invariant face recognition using markov random fields. *IEEE transactions on image processing*, pages=1573–1584, year=2012.
- [107] Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. pages 1883–1890, 2014.
- [108] Mehran Kafai, Le An, and Bir Bhanu. Reference face graph for face recognition. *IEEE Transactions on information forensics and security*, 9(12) :2132–2143, 2014.
- [109] Chayanut Petpairote, Suthep Madarasmi, and Kosin Chamnongthai. 2d pose-invariant face recognition using single frontal-view face database. *Wireless Personal Communications*, pages 1–17, 2020.
- [110] Akshay Asthana, Tim K Marks, Michael J Jones, Kinh H Tieu, and MV Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. pages 937–944, 2011.
- [111] Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan. Morphable displacement field based image matching for face recognition across pose. pages 102–115, 2012.

- [112] Dong Yi, Zhen Lei, and Stan Z Li. Towards pose robust face recognition. pages 3539–3545, 2013.
- [113] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015.
- [114] Guosheng Hu, Fei Yan, Chi-Ho Chan, Weihong Deng, William Christmas, Josef Kittler, and Neil M Robertson. Face recognition using a unified 3d morphable model. In *European Conference on Computer Vision*, pages 73–89. Springer, 2016.
- [115] Guosheng Hu, Fei Yan, Josef Kittler, William Christmas, Chi Ho Chan, Zhenhua Feng, and Patrik Huber. Efficient 3d morphable face model fitting. *Pattern Recognition*, 67 :366–379, 2017.
- [116] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality : High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3025–3032, 2013.
- [117] Annan Li, Shiguang Shan, and Wen Gao. Coupled bias–variance tradeoff for cross-pose face recognition. *IEEE Transactions on Image Processing*, 21(1) :305–315, 2011.
- [118] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning multi-view representation for face recognition. *arXiv preprint arXiv :1406.6947*, 2014.
- [119] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015.

- [120] Weihong Deng, Jiani Hu, Zhongjun Wu, and Jun Guo. Lighting-aware face frontalization for unconstrained face recognition. *Pattern Recognition*, 68 :260–271, 2017.
- [121] Changxing Ding and Dacheng Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66 :144–152, 2017.
- [122] Paul Koppen, Zhen-Hua Feng, Josef Kittler, Muhammad Awais, William Christmas, Xiao-Jun Wu, and He-Feng Yin. Gaussian mixture 3d morphable face model. *Pattern Recognition*, 74 :617–628, 2018.
- [123] Huibin Li, Di Huang, Pierre Lemaire, Jean-Marie Morvan, and Liming Chen. Expression robust 3d face recognition via mesh-based histograms of multiple order surface differential quantities. pages 3053–3056, 2011.
- [124] Stefano Berretti, Naoufel Werghi, Alberto Del Bimbo, and Pietro Pala. Matching 3d face scans using interest points and local histogram descriptors. *Computers & Graphics*, 37(5) :509–525, 2013.
- [125] Huibin Li, Di Huang, Jean-Marie Morvan, Yunhong Wang, and Liming Chen. Towards 3d face recognition in the real : a registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision*, 113(2) :128–142, 2015.
- [126] Amal Azazi, Syaheerah Lebai Lutfi, Ibrahim Venkat, and Fernando Fernández-Martínez. Towards a robust affect recognition : Automatic facial expression recognition in 3d faces. In *Expert Systems with Applications*, volume 42, pages 3056–3066. Elsevier, 2015.
- [127] Yinjie Lei, Yulan Guo, Munawar Hayat, Mohammed Bennamoun, and Xinzhi Zhou. A two-phase weighted collaborative representation for 3d partial face recognition with single sample. *Pattern Recognition*, 52 :218–237, 2016.

- [128] Xing Deng, Feipeng Da, and Haijian Shao. Efficient 3d face recognition using local covariance descriptor and riemannian kernel sparse coding. *Computers & Electrical Engineering*, 62 :81–91, 2017.
- [129] Walid Hariri, Hedi Tabia, Nadir Farah, Abdallah Benouareth, and David Declercq. 3d facial expression recognition using kernel methods on riemannian manifold. In *Engineering Applications of Artificial Intelligence*, volume 42, pages 25–32. Elsevier, 2017.
- [130] Abdelghafour Abbad, Khalid Abbad, and Hamid Tairi. 3d face recognition : Multi-scale strategy based on geometric and local descriptors. *Computers & Electrical Engineering*, 70 :525–537, 2018.
- [131] Ziyu Zhang, Feipeng Da, and Yi Yu. Data-free point cloud network for 3d face recognition. *arXiv*, pages arXiv–1911, 2019.
- [132] Xing Deng, Fepeng Da, Haijian Shao, and Yingtao Jiang. A multi-scale three-dimensional face recognition approach with sparse representation-based classifier and fusion of local covariance descriptors. *Computers & Electrical Engineering*, 85 :106700, 2020.
- [133] Yan Liang, Jia-Cheng Liao, and Jiahui Pan. Mesh-based scale-invariant feature transform-like method for three-dimensional face recognition under expressions and missing data. In *Journal of Electronic Imaging*, volume 29, page 053008. International Society for Optics and Photonics, 2020.
- [134] Muhammed Enes Atik and Zaide Duran. Deep learning-based 3d face recognition using derived features from point cloud. In *Innovations in Smart Cities Applications Volume 4 : The Proceedings of the 5th International Conference on Smart City Applications*, pages 797–808. Springer International Publishing, 2021.
- [135] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11) :2884–2896, 2018.

- [136] Peipei Li, Xiang Wu, Yibo Hu, Ran He, and Zhenan Sun. M2fpa : a multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10043–10051, 2019.
- [137] Hana Ben Fredj, Safa Bouguezzi, and Chokri Souani. Face recognition in unconstrained environment with cnn. pages 1–20, 2020.
- [138] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Fan-face : a simple orthogonal improvement to deep face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12621–12628, 2020.
- [139] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.
- [140] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2) :137–154, 2004.
- [141] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface : L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.
- [142] Weihong Deng, Binghui Chen, Yuke Fang, and Jiani Hu. Deep correlation feature learning for face verification in the wild. *IEEE Signal Processing Letters*, 24(12) :1877–1881, 2017.