Reconnaissance automatique d'expressions faciales en présence d'occultations partielles du visage

Delphine Poux

▶ **To cite this version:**

Delphine Poux. Reconnaissance automatique d'expressions faciales en présence d'occultations partielles du visage. Vision par ordinateur et reconnaissance de formes [cs.CV]. Université de Lille, 2021. Français. NNT : . tel-03613718v1

HAL Id: tel-03613718

https://hal.science/tel-03613718v1

Submitted on 23 Feb 2022 (v1), last revised 18 Mar 2022 (v2)

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.

THÈSE de DOCTORAT

Délivrée par :

Université de Lille

École Doctorale Sciences pour l'Ingénieur

Unité de recherche Centre de Recherche en Informatique, Signal et Automatique de Lille, UMR 9189 - CRIStAL - F-59000 Lille, France

Thèse préparée et soutenue publiquement par :

Delphine Poux

le 01/09/2021 pour obtenir le grade de Docteur en Informatique

Reconnaissance automatique des expressions faciales en présence d'occultations partielles du visage

Devant le jury composé de :

Laetitia Jourdan	Prof., Université de Lille	**Présidente du jury**
Yannick Benezeth	MCF - HDR, Université de Bourgogne	**Rapporteur**
Frédéric Jurie	Prof., Université de Caen et Expert Scientifique à SAFRAN	**Rapporteur**
Saida Bouakaz	Prof., Université Claude Bernard Lyon 1	**Examinatrice**
Bruno Bachimont	Prof., Université de technologie de Compiègne - UTC	**Examinateur**
Hichem Sahbi	CR - HDR., Université Sorbonne	**Examinateur**
Chaabane Djeraba	Prof., Université de Lille	**Directeur de thèse**
Ioan Marius Bilasco	MCF - HDR, Université de Lille	**Co-encadrant de thèse (invité)**
Nacim Ihaddadene	MCF, Junia Lille	**Co-encadrant de thèse**
Matthieu Cord	Prof., Université Sorbonne	**Examinateur (invité)**

Remerciements

Je remercie tout d'abord mon directeur de thèse Chaabane Djeraba qui m'a guidée tout au long de ma thèse. Je le remercie également pour ses conseils pour préparer la suite de ma carrière. Je remercie vivement mes encadrants Ioan Marius Bilasco et Nacim Ihaddadene pour nos discussions, leurs conseils et leur soutien durant ces années.

Mes remerciements vont également à Yannick Benezeth et Frédéric Jurie d'avoir accepté d'être rapporteurs de ces travaux. Je remercie Laetitia Jourdan de m'avoir fait l'honneur de présider le jury de soutenance. Mes remerciements vont aussi à Saida Bouakaz, Bruno Bachimont, Matthieu Cord et Hichem Sahbi d'avoir accepté d'être les examinateurs de ce travail. Je remercie l'ensemble des membres du jury pour leurs retours, leurs conseils et leurs questions qui me permettent d'avoir une vue nouvelle sur mes travaux.

Un grand merci à toute l'équipe FOX qui m'a accompagnée, soutenue et aidée pendant ces années. Merci à José Mennesson, Pierre Tirilly et Jean Martinet pour leur bienveillance et nos discussions pendant lesquelles ils m'ont apporté de nombreux conseils. Je remercie vivement les post-doctorants Amel Aissaoui, Benjamin Allaert et Romain Belmonte ainsi que les doctorants de l'équipe FOX pour leur soutien, leur aide et pour nos très nombreuses discussions qui m'ont permis d'évoluer au long de ces années.

Je voudrai également remercier les membres de Junia-Isen et tout particulièrement le département Mathématiques et Informatique pour leur accompagnement au long de cette thèse.

Merci à ceux qui m'ont soutenue pendant ces années de thèse et en particulier aux membres du laboratoire CRIStAL et du groupe Image, aux enseignants et collègues du FIL et de l'IUT A de Lille ainsi qu'aux collègues de l'IRCICA.

Un grand merci à ma famille et particulièrement à mon mari Thomas qui ont tou-

jours été présents pour moi, m'ont fait confiance, m'ont toujours soutenue et poussée à me dépasser. Enfin, je pense beaucoup à mon papa qui a toujours cru en moi et aurait, je pense, été fier du travail accompli.

Abstract

Automatic facial expression recognition is useful to create applications in various domains such as health, road safety or marketing where feedback on user state is relevant. Despite very good results in controlled settings (frontal face, no occlusion, good illumination), facial expression recognition is still today challenging under unconstrained environment. Among the different challenges, occlusions are particularly difficult to handle as they add noisy elements on the images and hide parts of the information. Several solutions have been proposed to address this issue. These solutions can be roughly categorized in two : those which focus on visible regions of the face and those which try to reconstruct the hidden part. State-of-the-art solutions are mainly based on texture or, sometimes, geometry and few are based on movement. However, movement seems to be particularly adapted under occlusions thanks to different motion properties such as close range propagation and local coherency. In this manuscript, we show the interest of using movement to overcome the issue of occlusions for the task of facial expression recognition.

Résumé

La reconnaissance automatique des expressions faciales peut s'avérer très utile pour diverses applications dans des domaines variés tels que la santé, la sécurité routière ou encore le marketing. Bien que des algorithmes permettent aujourd'hui une très bonne reconnaissance des expressions faciales dans un environnement contraint (pose frontale, pas d'occultation, bonne luminosité), la reconnaissance automatique des expressions faciales reste encore complexe dans un cadre naturel où l'on se retrouve confronté à certains défis. Parmi ces défis, les occultations rendent la tâche particulièrement difficile car elles ajoutent du bruit aux images et masquent une partie de l'information. Pour répondre à cette problématique, différentes solutions ont été proposées. Ces solutions peuvent être regroupées sous deux catégories : les solutions qui se concentrent sur les régions visibles du visage et celles qui reconstruisent les zones cachées. Les solutions de l'état de l'art sont principalement basées sur des éléments de texture, certaines solutions s'intéressent à la géométrie du visage mais très peu s'intéressent au mouvement. Or, le mouvement semble particulièrement adapté grâce à ses différentes propriétés : sa propagation et sa cohérence locale. Les travaux proposés dans ce manuscrit démontrent l'intérêt du mouvement pour reconnaître les expressions faciales en présence d'occultations partielles du visage.

Table des matières

Avant-propos

400% : c'est l'amélioration possible du taux de conversion lorsque l'expérience utilisateur est améliorée sur un site web, c'est-à-dire lorsque son état émotionnel reste positif tout au long de son expérience [1]. Le risque d'accident de voiture est 9,8 fois plus élevé lorsque le conducteur est dans un état émotionnel négatif comme la colère ou la tristesse [2]. Les émotions ont un impact important sur notre comportement et ces quelques chiffres en témoignent. La reconnaissance automatique des émotions pourrait, alors, s'avérer particulièrement utile pour différentes applications dans des domaines divers (on peut citer, entre autres, la détection d'un état émotionnel négatif dans un cadre de sécurité routière ou encore l'étude de la satisfaction des clients à des fins marketing). Afin de mettre en place ces applications, il s'avère alors nécessaire que les machines puissent également reconnaître automatiquement les états émotionnels. Pour cela, il est nécessaire de s'appuyer sur des signaux permettant de les reconnaître. Les émotions sont, en effet, communiquées aux autres par différentes manifestations dont font partie nos expressions faciales. Ces expressions faciales représentent alors un signal visible particulièrement intéressant pour étudier et reconnaître les émotions. Elles sont d'ailleurs une modalité de reconnaissance souvent utilisée dans la littérature pour reconnaître automatiquement l'état émotionnel des personnes même si d'autres modalités sont aujourd'hui étudiées telles que la posture et la gestuelle du corps [3] ou des paramètres physiologiques [4]. Cependant, la reconnaissance automatique des expressions faciales reste encore aujourd'hui confrontée à un certain nombre de défis tels que : les variations de la pose de la tête ou encore les occultations partielles du visage.

Chapitre 1

Contexte et verrous scientifiques

Contents

Afin de mesurer l'impact sociétal et positionner les enjeux de mes travaux de thèse, nous commençons ce manuscrit par un aperçu global du contexte et des problématiques liées à la reconnaissance automatique des expressions faciales. Dans cette introduction, nous évoquons, dans un premier temps, l'intérêt de rendre les machines capables de reconnaître nos émotions, les applications concrètes qui peuvent en bénéficier et nous abordons le lien entre les émotions et les expressions faciales. Nous discutons, dans une seconde section, des avancées de la littérature pour reconnaître automatiquement les expressions faciales avant d'en exposer les limites et les verrous

scientifiques. Nous proposons, enfin, d'énoncer notre positionnement avant d'ouvrir sur nos contributions.

1.1 Enjeux de la reconnaissance automatique des expressions faciales

Notre état émotionnel a un impact important sur notre vie quotidienne, sur notre comportement et sur notre santé. Une personne triste ou en colère aura un risque plus élevé de devenir dépressive, malade [5] ou encore d'avoir un accident de voiture lorsqu'elle prend le volant [6]. À l'inverse, une personne ayant un état d'esprit positif, parce qu'elle est en train de vivre une expérience utilisateur agréable sur un site web par exemple, sera plus encline à acheter les produits proposés [7]. Sherman et al. [8] ont montré que l'état émotionnel est déterminant dans le comportement du consommateur et a une grande influence sur l'argent dépensé dans un magasin.

Ces diverses observations laissent penser que pouvoir automatiquement détecter différents états émotionnels pourrait s'avérer particulièrement utile dans de nombreux domaines :

Sécurité routière : reconnaître automatiquement de la colère ou de la tristesse pourrait, par exemple, aider à éviter des accidents de la route en réagissant avant qu'ils ne se produisent [1].

Santé : des applications d'aide au diagnostic permettraient de détecter certains signaux de maladies comme la dépression [9].

Robotique : les assistants personnels envahissent aujourd'hui notre quotidien mais sont encore insensibles à nos émotions, créant alors des frustrations chez les utilisateurs. L'analyse automatique de l'état émotionnel pourrait permettre d'adapter la

1. `http://go.affectiva.com/in-cabin-sensing`

réponse des assistants aux ressentis des utilisateurs [2]. Il est, en effet, important que la communication numérique s'approche d'un modèle d'interaction naturelle [10] où l'empathie et les émotions jouent un rôle crucial.

Marketing : la création d'une expérience utilisateur appropriée pour maintenir un bon état d'esprit des clients est cruciale pour avoir une entreprise florissante. La reconnaissance automatique des états émotionnels pourrait alors permettre d'améliorer les services proposés en fonction des retours positifs ou négatifs des clients. Dans le cadre de services à la personne comme des services de transport ou de soins esthétiques par exemple, ce type d'application pourrait permettre d'améliorer le confort des clients en adaptant automatiquement l'environnement direct de la personne (la musique, la luminosité, ...) [3].

Jeux vidéos : l'expérience d'un joueur pourrait être améliorée en adaptant automatiquement la difficulté du jeu en fonction du retour du joueur [11].

E-learning : lors de leçons à distance, la caméra face à l'apprenant pourrait permettre de détecter ses incompréhensions ou ses frustrations pour adapter automatiquement la méthode d'apprentissage [12].

Nous notons donc l'intérêt de pouvoir automatiquement reconnaître les émotions. Afin de déterminer comment mettre en place de telles applications, il est important de comprendre ce qu'est une émotion et comment la reconnaître.

1.2 État émotionnel et expressions faciales

Les chercheurs ont des points de vue différents sur la définition d'une émotion [13] et sont également en désaccord sur les fonctions premières d'une émotion et ses

2. `https://www.webempath.com/`
3. `http://go.affectiva.com/affdex-for-market-research`

activations (voir Annexe A.1). Cependant, les chercheurs se mettent plutôt d'accord sur le fait que les émotions impliquent des réactions physiologiques et comportementaux (ton de la voix, comportements du corps et expressions faciales) [14]. Par ces différentes manifestations, les émotions sont communiquées aux autres et font partie intégrante de nos interactions sociales (voir Annexe A.2). Albert Mehrabian [15] montre que la communication d'un état d'esprit d'une personne à l'autre passe majoritairement par de la communication non-verbale comme la voix ou les expressions faciales. Ainsi, l'auteur estime à 7% l'impact des mots, 38% l'impact de la voix et 55% l'impact des expressions faciales sur les ressentis de l'interlocuteur.

Les expressions faciales représentent donc un élément important de communication. Différents chercheurs, dont notamment Charles Darwin et Paul Ekman ont mis en lumière le principe d'universalité des expressions faciales (voir Annexe A.1) et ont défini les six expressions faciales universelles que sont : la joie, la peur, la surprise, le dégoût, la colère et la tristesse illustrées sur la Figure 1.1.

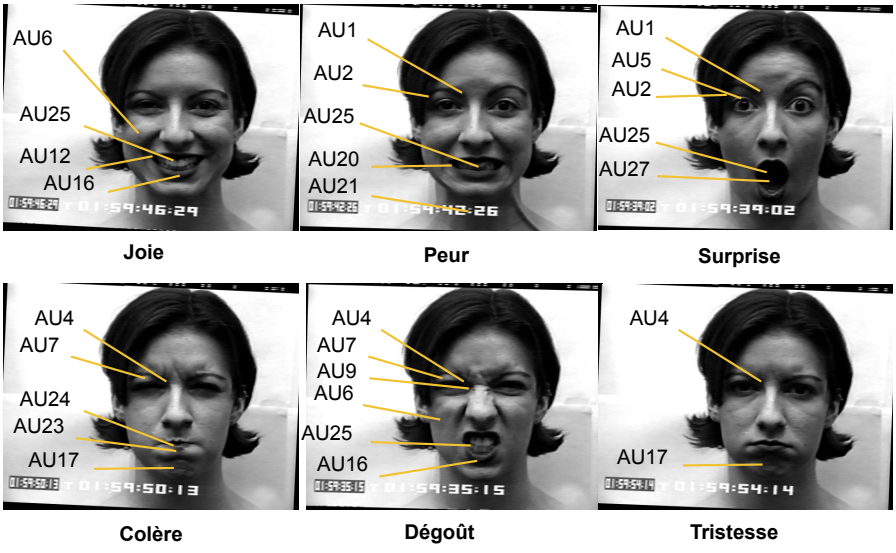


FIGURE 1.1 – Six expressions faciales basiques décrites par leurs catégories et par les unités d'actions qui apparaissent sur le visage.

Pour décrire plus finement les signaux envoyés par les différents muscles du visage en fonction de chaque catégorie, Ekman et Friesen ont proposé le système FACS (Facial Action Coding System) où chaque expression est décrite par une composition de mouvements faciaux appelés unités d'action (AU) [16]. Ainsi, les expressions peuvent être décrites soit par leur catégorie (joie, peur, surprise, colère, dégoût ou tristesse) soit par la combinaison de plusieurs unités d'actions comme illustré sur la Figure 1.1.

Une autre approche pour décrire les émotions consiste à les caractériser selon plusieurs dimensions [17, 18]. En particulier, les dimensions de valence (plaisir-déplaisir) et d'activation (actif-passif), sont les dimensions les plus utilisées. Dans ce cas, les états émotionnels ne sont pas considérés comme discrets mais continus et une émotion particulière est représentée par un point dans cet espace. Cette représentation a l'avantage de définir de nombreuses nuances au niveau des émotions. En contrepartie, la représentation dimensionnelle, avec ses nombreuses possibilités, rajoute une difficulté supplémentaire, notamment pour l'annotation des bases de données.

Bien que la représentation dimensionnelle permette une palette large et précise d'émotions, cette représentation est peu utilisée pour la reconnaissance automatique des expressions faciales en raison de la difficulté d'annotation des bases de données et, par conséquent, du faible nombre de bases de données disponibles. Parmi les bases de données recensées par Mollahosseini et al. [19] en 2017 lors de la publication de la base de données AffectNet, sur 14 bases de données classiques de la littérature (AffectNet compris), 10 contiennent des annotations discrètes (dont 8 selon les expressions faciales universelles et 5 avec des unités d'action) contre 5 qui proposent une annotation contenant des dimensions continues.

Les 6 expressions faciales universelles sont donc très utilisées pour la reconnaissance automatique des expressions faciales et nous nous concentrerons, dans ce manuscrit, uniquement sur cette représentation. Nous proposons, dans la section suivante, de donner un aperçu de l'évolution des méthodes proposées pour la reconnaissance automatique des expressions faciales.

1.3 Reconnaissance automatique des expressions faciales

La reconnaissance automatique des expressions faciales, dont les premiers travaux ont été publiés en 1978 [20], est un sujet riche en études, illustrant ainsi son importance sociétale. Bien que le sujet ait été assez négligé jusque dans les années 90 en raison d'un manque de techniques pour la détection des visages [21], de nombreux chercheurs se sont intéressés à ce sujet depuis. La base de données CK [22] publiée en 2000 et étendue en 2010 par CK+ [23] est l'une des premières bases de données importante qui a permis d'étudier la reconnaissance automatique des expressions faciales dans un environnement entièrement contrôlé [21] (un bon éclairage, une pose statique et frontale du visage, pas d'occultation, ...). De très bons résultats ont été obtenus sur de telles données [24, 25, 26]. Pour obtenir ces résultats, deux grandes approches sont utilisées dans la littérature pour reconnaître automatiquement les expressions faciales : les approches basées sur des descripteurs classiques et des approches basées sur des descripteurs appris par des architectures de réseaux de neurones. L'apprentissage par réseaux de neurones présente l'avantage d'apprendre les caractéristiques les plus appropriées pour la classification en apprenant conjointement les caractéristiques et le classifieur. Cependant, ces méthodes nécessitent de grandes bases de données afin de mener à bien ce double apprentissage. Les descripteurs classiques sont, quant à eux,

proposés pour extraire des caractéristiques voulues invariantes à différents éléments tels que les variations de luminosité ou de rotation.

Différentes solutions qui s'appuient sur ces méthodes sont d'ailleurs d'ores et déjà commercialisées par différentes structures comme Affectiva [4], le neurodatalab [5] ou encore Microsoft Azure [6].

Malgré leur commercialisation, ces solutions restent aussi très sensibles à différents éléments, dont notamment la variation de la pose ou les occultations du visage, lorsqu'elles sont utilisées en environnement naturel comme le montre une étude proposée par le neurodatalab [7]. Dans cette étude, les taux de reconnaissance obtenus par plusieurs solutions sur différentes bases de données sont comparés. Parmi ces bases de données, on retrouve AFEW [27] qui est constituée d'extraits de films. Cette étude montre que, sur AFEW, le taux de reconnaissance maximal obtenu par ces différents outils est de 45%, ce qui montre l'intérêt de poursuivre les études sur le sujet.

1.4 Verrous scientifiques

Les techniques récentes sont capables de reconnaître les expressions faciales de manière très précise dans un environnement totalement contrôlé. Néanmoins, dans un environnement naturel, certaines conditions affectent encore considérablement les résultats. Les principaux défis sont : les variations d'éclairage, les variations de l'intensité de l'expression, les variations de la pose de la tête et les occultations.

4. https://www.affectiva.com
5. https://api.neurodatalab.dev/
6. https://azure.microsoft.com/fr-fr/blog/tag/emotion-api/
7. https://medium.com/@neurodatalab/comparing-emotion-recognition-tech-microsoft-neurodata-lab-amazon-affectiva-cb1b00fd5f1e

Parmi ces différents défis, les occultations représentent un défi particulièrement important et difficile [28]. Les occultations peuvent être causées par des accessoires ou par des parties du corps de la personne comme l'illustre la Figure 1.2. Mais, une occultation peut également être engendrée par une mauvaise luminosité ou une variation de pose. Une luminosité trop faible peut ainsi cacher une partie du visage et une pose du visage avec un angle important peut rendre inaccessibles certaines zones du visage. Il semble donc particulièrement important d'essayer de répondre au problème des occultations en priorité car la résolution des défis tels que la pose ou l'illumination peut donc passer, en partie, par la résolution du problème des occultations.

FIGURE 1.2 – Exemple d'occultations rencontrées en conditions réelles.

1.5 Positionnement

Les solutions proposées dans la littérature pour faire face aux occultations peuvent être regroupées en deux catégories : les solutions qui reconstruisent les parties cachées du visage et celles qui concentrent l'attention sur les régions visibles. Toutes ces solutions sont essentiellement basées sur la texture ou la géométrie du visage.

Dans cette thèse, nous avons exploré l'utilisation du mouvement facial afin de proposer des solutions nouvelles pour répondre au défi des occultations. Il a été montré que la reconnaissance des expressions faciales par des humains avec des données dynamiques est plus facile que sur des images statiques [29]. Dans ce manuscrit, nous décri-

vons des propriétés qui montrent que l'étude du mouvement facial semble également complètement adaptée pour gérer les occultations. Parmi ces propriétés on dénombre : la propriété de propagation du mouvement et la similarité des motifs d'activation des expressions entre plusieurs personnes. Nous décrivons en parallèle les solutions proposées pour exploiter chacune de ces propriétés.

FIGURE 1.3 – A et B sont des images extraites d'une séquence vidéo de l'ensemble de données CK+ respectivement aux temps t et $t+1$. C illustre le flux optique calculé avec la méthode Deepflow [30] entre A et B.

La première propriété (illustrée par la Figure 1.3) est la propagation du mouvement, due à l'élasticité de la peau. Cette propriété permet de reconnaître les expressions faciales malgré les occultations partielles car, même si la partie la plus importante du visage est occultée, le mouvement se propage aux régions voisines. Comme le montre la Figure 1.3, bien qu'il soit difficile de déterminer visuellement la différence entre les images A et B, le mouvement calculé sous forme de flux optique, permet de conserver de précieuses informations malgré l'occultation.

La deuxième propriété est la similarité des motifs de mouvements entre différentes personnes. Comme l'illustre la Figure 1.4, plusieurs personnes faisant la même expression faciale ne sont pas similaires si on s'intéresse à la texture car la texture dépend fortement de l'identité des personnes. Au contraire, l'activation des muscles du visage d'une personne à l'autre reste similaire et, dans le domaine du mouvement, plusieurs personnes faisant la même expression font des mouvements assez semblables. On peut

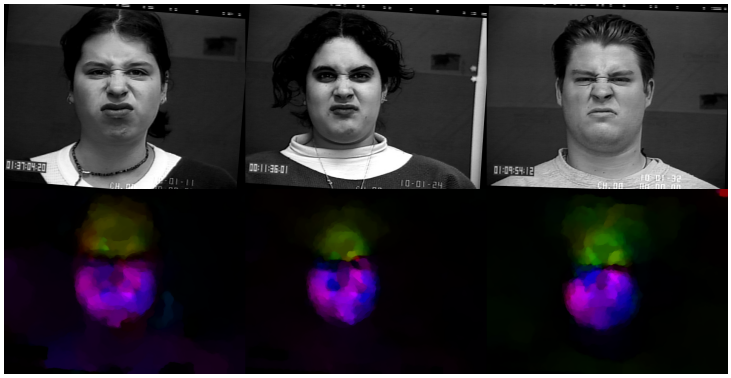


FIGURE 1.4 – La première ligne montre plusieurs personnes exprimant l'expression de dégoût extraite de la base de données CK+. La deuxième ligne représente les flux optiques correspondants calculés avec la méthode DeepFlow [30] entre l'image neutre et l'image apex, c'est-à-dire la plus grande intensité de l'expression, de ces différentes personnes.

noter ainsi une similarité forte malgré la différence d'identité. Cette similarité peut être particulièrement utile pour reconstruire directement le mouvement en se basant sur des motifs de mouvements similaires.

1.6 Plan du manuscrit

Dans la suite de ce manuscrit, nous exposons, dans une première partie, différents aspects de l'état de l'art. Le Chapitre 2 résume les méthodes de la littérature utilisées pour la reconnaissance automatique des expressions faciales dans un environnement contrôlé et les bases de données disponibles pour cette tâche.

Nous exposons, ensuite, dans le Chapitre 3, des travaux représentatifs de l'état de l'art pour répondre à la problématique des occultations et proposons un aperçu des protocoles expérimentaux utilisés afin de déterminer le plus approprié. Dans cette première partie, on remarque que les méthodes de reconnaissance des expressions faciales

exploitant la dimension temporelle semblent améliorer les résultats par rapport aux méthodes statiques. Nous voyons ensuite que les propositions de la littérature pour répondre aux occultations sont regroupées sous deux grandes catégories que sont : la reconstruction des zones cachées et les méthodes qui se concentrent sur les zones visibles. Nous notons surtout que, dans les deux cas, très peu de solutions se basent sur des éléments de mouvements.

La deuxième partie de ce manuscrit expose les contributions proposées pour répondre à la problématique des occultations en exploitant les propriétés liées au mouvement. Afin d'exploiter la propriété de propagation du mouvement, nous introduisons, dans le Chapitre 5, notre méthode de calcul des modèles faciaux intelligents en sélectionnant les régions les plus importantes pour reconnaître les expressions faciales malgré différentes occultations.

Nous proposons ensuite, au Chapitre 6, d'exploiter la propriété de similarité de mouvement entre différentes personnes en proposant une reconstruction des parties occultées directement dans le domaine du flux optique.

Pour évaluer nos propositions, nous proposons des protocoles expérimentaux clairs et reproductibles reprenant les occultations les plus représentatives de la littérature. Nous mettons à disposition le code [8] de découpage des données et d'occultation utilisé pour notre méthode de reconstruction afin de faciliter davantage la reproduction de ce protocole.

Enfin, une synthèse et les perspectives ouvertes par ce travail sont proposées dans le dernier chapitre.

8. `https://gitlab.univ-lille.fr/fox/occ_gen`

Première partie

Évolution des méthodes de reconnaissance d'expressions faciales et défis

Dans la première partie de ce manuscrit, nous donnons un aperçu des méthodes proposées dans la littérature pour reconnaître les expressions faciales.

Le Chapitre 2 récapitule l'ensemble du processus de reconnaissance des expressions faciales en environnement contrôlé (pose frontale et fixe, pas d'occultation, une bonne luminosité et un arrière-plan neutre). Nous proposons, tout d'abord, une vue d'ensemble du processus complet classique utilisé pour la reconnaissance des expressions faciales. Nous passons en revue ensuite les étapes de prétraitement avant d'exposer les descripteurs et les méthodes de classification de l'état de l'art.

Nous nous intéressons, d'un côté, aux descripteurs classiques (LBP, points caractéristiques, etc.) puis, d'un autre côté, aux approches basées sur un apprentissage par réseaux de neurones. Nous distinguons les approches statiques des approches dynamiques, que ce soit pour les descripteurs classiques comme pour les architectures d'apprentissage par réseaux de neurones.

Nous terminons ce chapitre par un aperçu des différentes bases de données disponibles ainsi que quelques résultats obtenus dans la littérature sur ces bases de données. Ces évaluations nous permettent de mettre en avant l'intérêt du mouvement pour reconnaître les expressions faciales ainsi que la difficulté, encore aujourd'hui, à travailler sur des données naturelles.

Après avoir défini l'ensemble des étapes proposées dans l'état de l'art dans un cadre contrôlé, nous nous concentrons, au Chapitre 3, sur la problématique des occultations. Nous proposons, dans un premier temps de discuter de l'impact des occultations sur la reconnaissance des expressions faciales.

Nous exposons, ensuite, les deux catégories de solutions proposées pour répondre à la problématique des occultations : les solutions qui se concentrent uniquement sur les parties visibles du visage et celles qui reconstruisent les zones cachées. Nous détaillons, enfin, les différents protocoles utilisés dans l'état de l'art pour évaluer ces méthodes.

Chapitre 2

Reconnaissance automatique d'expressions faciales

Contents

La reconnaissance automatique des expressions faciales dans un environnement totalement contrôlé est étudiée depuis de nombreuses années et certaines solutions permettent d'obtenir des résultats très satisfaisants. Dans ce chapitre, nous présentons, tout d'abord, l'ensemble du processus pour reconnaître automatiquement les expressions faciales. Nous énumérons, ensuite, les différentes étapes de ce processus, des étapes de prétraitement à la classification en passant par l'extraction des caractéristiques. Nous terminons ce chapitre par une présentation des bases de données existantes dans la littérature et des résultats obtenus dans l'état de l'art sur ces bases.

2.1 Processus de reconnaissance

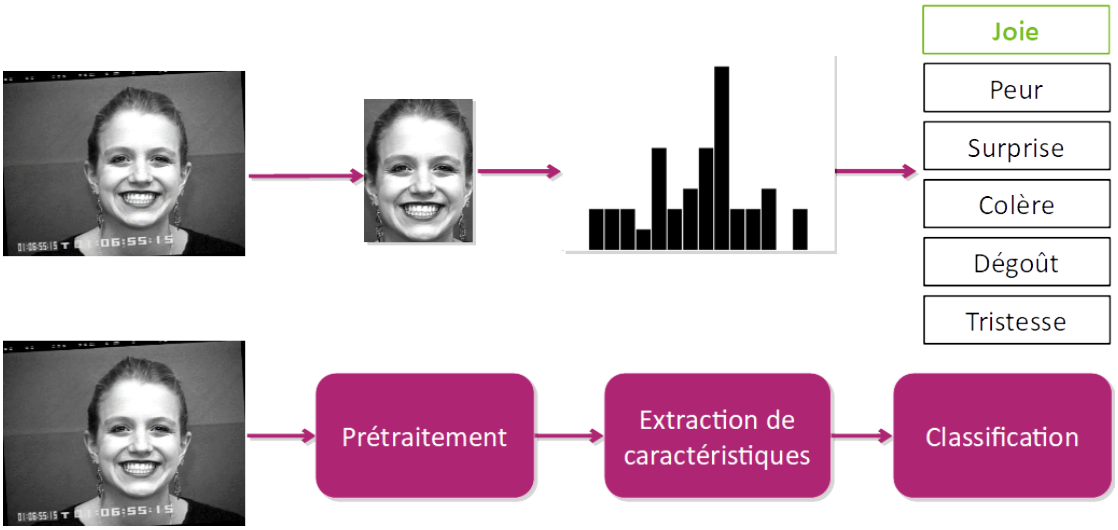


FIGURE 2.1 – Processus classique de reconnaissance des expressions faciales.

Un processus classiquement utilisé dans la littérature, illustré sur la Figure 2.1, comporte les étapes suivantes :

Prétraitement : une étape de prétraitement est d'abord nécessaire pour normaliser les images et ainsi réduire au maximum le bruit pour se concentrer sur l'information essentielle.

Extraction de caractéristiques : une extraction de caractéristiques est ensuite utilisée pour réduire les dimensions des données et concentrer l'analyse sur les informations liées aux expressions.

Classification : grâce à l'extraction de caractéristiques qui permettent une réduction de la dimension, des méthodes classiques de classification telles que le SVM, le KNN ou un réseau de neurones comme le MLP peuvent ensuite être appliquées. Parmi ces classifieurs, le SVM (Séparateur à Vastes Marges ou Support Vector Machine) [31] est le classifieur le plus souvent utilisé pour la reconnaissance automatique des expressions faciales car il a l'avantage de construire des modèles complexes avec assez peu de données tout en permettant une bonne généralisation [32].

L'extraction de caractéristiques et la classification peuvent se faire selon deux méthodologies : les méthodes basées sur des descripteurs classiques ou les méthodes basées sur un apprentissage de caractéristiques. D'un côté, les méthodes à base de descripteurs permettent d'extraire les caractéristiques génériques à partir d'observations et de calculs décrits explicitement. De l'autre, les caractéristiques et le classifieur sont appris, souvent, par des architectures à base de réseaux de neurones. En fonction des données d'entrée (images fixes ou séquences vidéo), différentes approches ont été proposées, qu'elles soient classiques ou basées sur un apprentissage par réseau de neurones.

Dans les sections suivantes, nous passons en revue les différentes étapes du processus de reconnaissance des expressions faciales. Nous décrivons, dans une première section les différentes étapes classiques de prétraitement. Nous nous concentrons ensuite sur les méthodes proposées dans la littérature pour l'extraction de caractéristiques et la classification. Pour ce faire, nous proposons, d'une part, de détailler les descripteurs classiques et, d'autre part, les méthodes basées sur un apprentissage par réseaux de neurones. Pour ces deux approches, nous distinguons l'analyse statique et l'analyse dynamique du visage.

2.2 Prétraitement

Un processus de reconnaissance des expressions faciales nécessite quelques étapes de prétraitement afin de supprimer les informations inutiles et de minimiser le bruit présent dans les données.

Les étapes de prétraitement les plus utilisées sont illustrées sur la Figure 2.2. Une première étape consiste à détecter le visage afin de concentrer l'attention uniquement sur la région contenant le visage et ignorer l'arrière-plan. À partir du visage, des points caractéristiques (landmarks) peuvent être détectés afin d'être directement utilisés pour la classification ou comme guide pour normaliser le visage. Enfin, si nécessaire, une étape de normalisation peut être utilisée pour réduire les variations causées par l'environnement ou les conditions d'acquisition telles que les variations d'éclairage ou de pose de la tête. Pour cette dernière étape, on peut utiliser les points caractéristiques pour aligner le visage ou encore une normalisation photométrique (égalisation d'histogramme par exemple) pour diminuer les effets néfastes d'une mauvaise luminosité. Nous décrivons, dans les sous-sections suivantes, ces différentes étapes de prétraitement.

(a) Image initiale (b) Détection du visage (c) Points caractéristiques

FIGURE 2.2 – Étapes de prétraitement habituellement utilisées lors d'un processus de reconnaissance des expressions faciales.

2.2.1 Détection du visage

La détection des visages est une première étape cruciale pour diverses applications telles que la reconnaissance de personnes, le suivi des visages, la reconnaissance de l'âge ou de l'expression faciale.

De nombreux détecteurs de visages ont été proposés dans la littérature. Ces détecteurs peuvent être classés en quatre catégories [33] :

- **Méthodes de classifieurs en cascade** inspirées par la méthode proposée par Viola et Jones [34, 35],
- **Méthodes par parties** principalement basées sur les modèles de pièces déformables (DPM) initialement proposés en reconnaissance de formes. Les DPM voient le visage comme une collection de différents éléments ainsi que des connexions entre ces éléments [36, 37],
- **Méthodes de caractéristiques par canal** qui calculent différents canaux à partir de l'image initiale tels que la magnitude du gradient et les histogrammes de gradients orientés. Ces différents canaux sont ensuite concaténés et utilisés lors de l'apprentissage du classifieur [38],
- **Méthodes basées réseaux de neurones** déjà exploitées dans les années 90 [39], elles sont de plus en plus nombreuses et se basent aujourd'hui principalement sur des réseaux de neurones convolutionnels qui permettent d'apprendre automatiquement les caractéristiques les plus appropriées pour la détection de visage [40].

2.2.2 Détection de points caractéristiques

Les points caractéristiques du visage, illustrés sur la Figure 2.3, décrivent la structure du visage comme un graphe dont les noeuds délimitent les contours et les éléments qui le constituent comme les yeux, le nez ou la bouche.

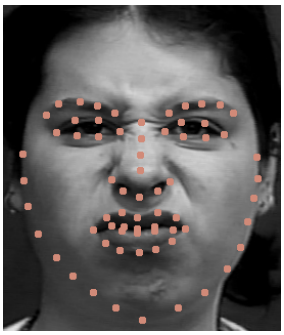


FIGURE 2.3 – 68 points caractéristiques qui délimitent les différents éléments qui constituent le visage.

Dans le cadre d'un processus de reconnaissance des expressions faciales, les points caractéristiques peuvent être utiles pour les étapes suivantes :

- extraction de caractéristiques : prises telles quelles ou en y appliquant différents calculs pour décrire plus précisément la géométrie du visage. Ces points permettent, en effet, de caractériser les déformations du visage fortement liées aux expressions faciales.
- découpage du visage : la reconnaissance des expressions faciales peut se faire en étudiant le visage par zone. Les points caractéristiques permettent alors de délimiter ces zones.
- normalisation/alignement du visage : les points caractéristiques permettent de définir la localisation de différentes zones du visage et servent alors de guides pour minimiser les transformations (d'échelle, de rotation, de position) afin d'obtenir une forme de référence. Un alignement simple et très courant consiste à aligner les yeux et à mettre tous les visages à la même échelle de taille.

Pour détecter les points caractéristiques, différentes solutions ont été proposées dans l'état de l'art. Ces solutions peuvent être regroupées sous deux grandes catégories [41] : les approches génératives et les approches discriminatives.

Approches génératives : ces méthodes reposent sur le calcul d'une modélisation de la distribution des probabilités jointes des modèles d'apparence et de forme. La méthode Active Appearance Model (AAM) proposée par Cootes et al. [42] fait partie des travaux fondateurs pour la détection de points caractéristiques et a été très amplement étudiée. Cette méthode repose sur l'apprentissage d'un modèle de forme, d'un modèle d'apparence et d'un modèle de mouvement. Le modèle de forme, également appelé PDM (Point Distribution Model) est construit à partir de données de formes d'apprentissage normalisées et dont la dimension a été réduite en passant par une analyse en composantes principales (ACP). L'algorithme repose ensuite sur l'apprentissage d'un modèle de mouvement qui apprend la déformation de la texture par rapport à un modèle de forme de référence à laquelle sont ajoutés des paramètres de transformation. Pour l'apprentissage de ce modèle, une fonction de coût représentant la reconstruction de la texture déformée par le modèle de mouvement pour coller à la forme est minimisée.

Approches discriminatives : contrairement aux approches génératives qui apprennent un espace de probabilités jointes entre le modèle de forme et d'apparence, les approches discriminatives apprennent directement la fonction de transformation entre l'apparence et la forme [43]. Pour apprendre ce modèle, une approche largement utilisée dans la littérature est la régression en cascade. Pour mettre en place cette approche, on calcule la forme initiale qui correspond à la forme moyenne des données d'apprentissage. À partir de cette forme initiale, différentes fonctions de régression sont appliquées pour coller à la forme du visage traité. Chaque fonction de régression est apprise in-

dépendamment et permet, au fur et à mesure, d'affiner la forme par rapport au visage. Les premières régressions permettent ainsi de d'adapter la forme du visage en fonction de la pose du visage et les fonctions de régression suivantes permettent d'affiner les points pour suivre les contours du visage. Plus récemment, les réseaux de neurones ont permis de conjointement apprendre les caractéristiques et la régression [44, 45].

Aujourd'hui, les approches discriminatives sont les plus utilisées en raison de leur robustesse, notamment par leur capacité de généralisation, mais aussi pour leur rapidité de calcul.

2.3 Descripteurs

À partir des images de visages prétraitées, différents descripteurs ont été proposés. Ces descripteurs permettent d'extraire des caractéristiques de ces images permettant de réduire la dimension tout en conservant les informations pertinentes liées aux expressions faciales. Dans cette section, nous donnons un bref aperçu des descripteurs proposés dans la littérature pour extraire explicitement des caractéristiques à partir de données statiques ou dynamiques. En fin de section, nous nous intéressons également aux architectures neuronales capables d'apprendre des caractéristiques pertinentes pour la tâche de reconnaissance d'expressions dans un contexte statique ou dynamique.

2.3.1 Descripteurs statiques

Les caractéristiques statiques sont extraites d'images et peuvent être regroupées en deux catégories : celles qui décrivent la géométrie du visage et celles qui décrivent la texture.

Géométrie : les descripteurs de géométrie sont principalement basés sur les points caractéristiques qui correspondent aux contours de zones importantes du visage telles que le nez, les yeux, la bouche et le contour du visage comme illustré sur la Figure 2.3. Les coordonnées de ces points peuvent servir telles quelles de caractéristiques ou différentes mesures peuvent être extraites à partir de ces points pour caractériser davantage la géométrie du visage. Les descripteurs géométriques ont l'avantage d'être beaucoup moins sensibles à l'identité des personnes et permettent ainsi de se concentrer sur l'expression faciale. Néanmoins, ces descripteurs sont généralement basés sur 68 points caractéristiques, ce qui est réducteur pour distinguer correctement les différentes expressions. Les déformations du visage liées aux expressions faciales sont, en effet, d'intensités variables et difficilement mesurables en ne conservant que quelques points de contours. De plus, ces descripteurs sont complètement dépendants de la détection des points caractéristiques qui ne sont pas toujours précis, surtout dans certaines conditions naturelles.

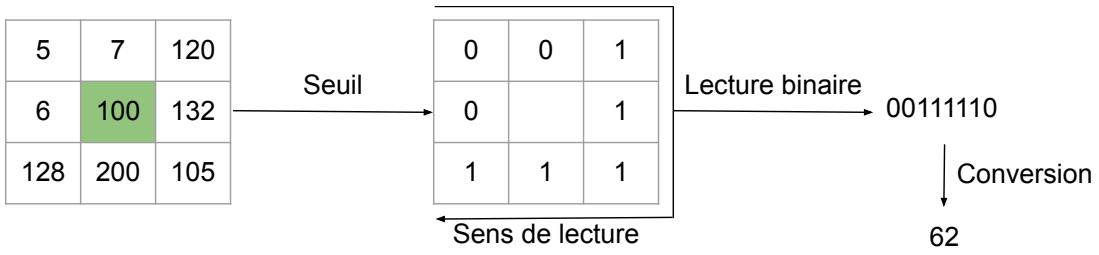


FIGURE 2.4 – Opérateur LBP appliqué sur chaque pixel de l'image afin de calculer l'histogramme correspondant.

Texture : le descripteur d'apparence le plus populaire dans le domaine de l'analyse faciale, qui a été introduit par Ojala et al. [46] est le LBP (Local Binary Pattern). Pour chaque pixel de l'image, un voisinage est considéré, et l'intensité du pixel central est comparée aux intensités des voisins comme l'illustre la Figure 2.4. Un histogramme est

ensuite construit à partir des valeurs obtenues. Un tel descripteur met en évidence les informations de forme telles que les coins ou les contours de l'image.

Pour affiner les caractéristiques calculées en fonction des différentes zones, le visage est souvent divisé en régions. Les descripteurs sont alors calculés indépendamment dans chaque région et sont ensuite concaténés pour définir le descripteur final comme illustré sur la Figure 2.5.

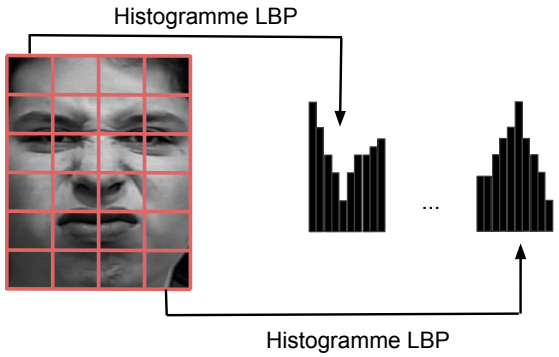


FIGURE 2.5 – Découpage en régions pour le calcul de caractéristiques LBP.

D'autres descripteurs de texture ont également été proposés comme le LDP (Local Directional Pattern) [47] qui utilise un masque de Kirsch sur chaque pixel pour déterminer les informations de contours ou encore le descripteur HOG (Histogram of Oriented Gradient) [48] qui utilise les gradients de l'image pour récupérer les orientations et les magnitudes des contours présents dans l'image.

2.3.2 Descripteurs dynamiques

Nous savons que l'exploitation de séquences d'images facilite la reconnaissance des expressions faciales, notamment, grâce à l'étude de Bassili [29]. Afin d'exploiter l'information temporelle, la plupart des descripteurs encodant des caractéristiques au sein d'images 2D ont été étendus.

Géométrie : les descripteurs géométriques ont ainsi évolués pour prendre en considération le suivi des points caractéristiques [49, 50].

Texture : les descripteurs de texture ont été étendus pour ajouter une dimension dans le domaine temporel. Pour ajouter la dimension temporelle, différents axes sont considérés : l'axe XY qui correspond à l'axe en 2 dimensions d'une image, l'axe XT qui correspond à l'axe X à travers le temps et qui encode l'évolution sur l'axe X des pixels au cours du temps et l'axe YT qui considère l'axe Y à travers le temps et qui encode l'évolution des pixels selon l'axe Y au cours du temps. Une extension très connue est le LBP-TOP proposée en 2007 par Zhao et al. [51].

Approches dense : en plus de ces deux catégories, des caractéristiques spécifiques au mouvement sont utilisées, notamment basés sur des calculs de flux optiques denses, qui caractérisent les mouvements entre deux images successives. Le flux optique quantifie le mouvement et se montre particulièrement adapté pour prendre en considération les changements apparus dans le temps sur un visage. L'idée du flux optique, illustrée sur la Figure 2.6 est d'identifier les déplacements des différents pixels entre deux images dans le temps.

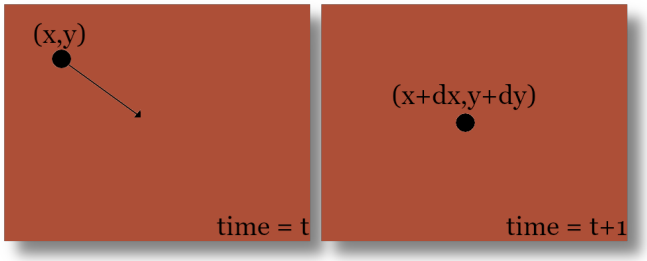


FIGURE 2.6 – Principe du flux optique. Ici les deux images montrent un déplacement du point noir de (dx, dy).

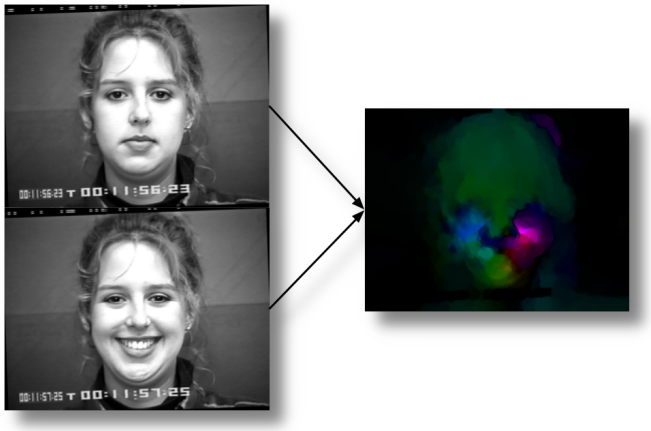


FIGURE 2.7 – Illustration d'un flux optique avec la méthode Deepflow [30] entre la première et la dernière image d'une séquence de CK+ [23] lorsque la personne fait une expression de joie.

La Figure 2.7 présente un flux optique calculé entre une image neutre et une image d'apex (plus grande intensité de l'expression faciale). Le flux ainsi calculé rend compte des mouvements entre les deux images. Différentes approches sont proposées pour calculer le flux optique. Les approches dominantes sont les approches variationnelles, introduites par Horn et Schunck en 1981 [52]. Ces approches consistent à résoudre un problème d'optimisation, généralement basé sur l'équation : $I(x+w, t+1) - I(x, t) = 0$ où $I(x, t)$ est l'intensité d'un pixel x à un temps t et w est le déplacement du pixel au temps $t + 1$. Parmi les approches proposées, la plus populaire est celle proposée par Farnebäck [53] qui cible le déplacement de voisinages locaux. Pour couvrir les larges déplacements, Farnebäck utilisent une approche pyramidale en étudiant les images à différentes échelles. D'autres approches de la littérature se basent plutôt sur une association de motifs (pattern matching)[54, 30] qui cherche à associer une fenêtre de pixels au temps t à une fenêtre de pixels au temps $t + 1$.

Ces différentes approches évoluent, aujourd'hui, vers des solutions basées sur des réseaux de neurones. DeepFlow, proposé par Weinzaepfel et al. [30], se base sur des

algorithmes d'association de motifs. FlowNet, proposé par Fischer et al. [55], utilise une architecture d'auto-encodeur U-Net qui prend en entrée les deux images et calcule en sortie le flux optique.

Le flux optique peut être utilisé tel quel comme descripteur ou servir d'entrée pour des descripteurs de mouvement plus complexes comme les descripteurs HOOF [56] ou LMP [57]. HOOF (Histogram of Oriented Optical Flow), proposé par Chaudhry et al., calcule les histogrammes des flux optiques calculés entre chaque image regroupés par tranches d'angles appelés bin et pondérés par leur magnitude. Le LMP (Local Motion Pattern) proposé par Allaert et al. en 2019 exploite la cohérence locale de mouvement au sein de chaque région du visage et la propagation naturelle du mouvement liée aux propriétés d'élasticité de la peau. Le principe du LMP est représenté dans la Figure 2.8 où CMR représente l'épicentre du mouvement et NMR les régions voisines où la propagation du mouvement implique une cohérence de mouvement. La concaténation des directions principales de chaque région représente alors le vecteur de caractéristiques.

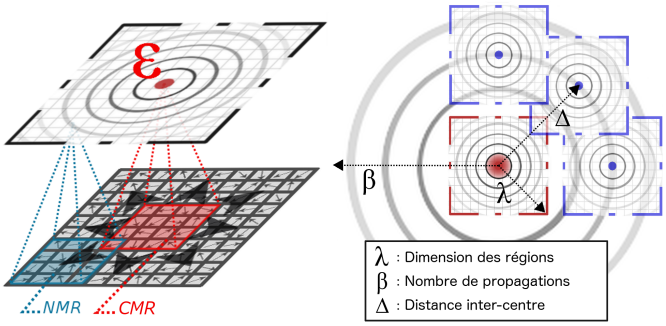


FIGURE 2.8 – Description du LMP. Image extraite de l'article de Allaert et al. [57]

Un comparatif des différentes méthodes de flux ainsi que des différents descripteurs basés sur du flux optique a été proposé en 2019 par Allaert et al. [58]. Dans cette étude, les auteurs mettent en évidence le fait que les méthodes les plus récentes n'offrent pas forcément les meilleurs résultats pour la reconnaissance automatique des expressions faciales. Les auteurs expliquent cela par la présence de filtres qui lissent le flux.

Les opérations de lissage sont adaptées lorsque l'on étudie des scènes en mouvement (MPI Sintel Dataset [1]), mais sur un visage, le lissage risque de supprimer des informations utiles. Les expressions faciales impliquent, en effet, généralement, la présence de discontinuités induites par l'activation musculaire (apparition d'une fossette par exemple).

2.3.3 Caractéristiques apprises statiques

Les réseaux de neurones profonds se sont montrés particulièrement efficaces dans la reconnaissance automatique et ont trouvé naturellement leur place dans le cadre de la reconnaissance automatique des expressions faciales [59].

Dans le cas des images ainsi que dans le cas des expressions faciales [60, 61, 62, 63], l'architecture la plus commune est l'architecture CNN (Convolutional Neural Network) constituée de couches de convolution et de pooling pour apprendre dans un premier temps les caractéristiques des images. Ces couches sont suivies de couches entièrement connectées pour la classification, comme illustré sur la Figure 2.9.

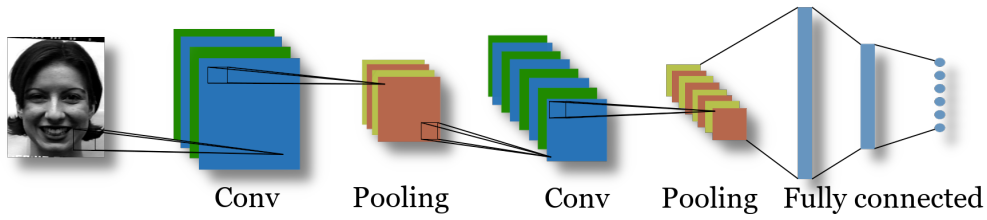


FIGURE 2.9 – Exemple d'une architecture CNN.

Les couches de convolution ont un rôle d'extraction de caractéristiques en apprenant des filtres de convolutions par lesquels passent les données de la couche précé-

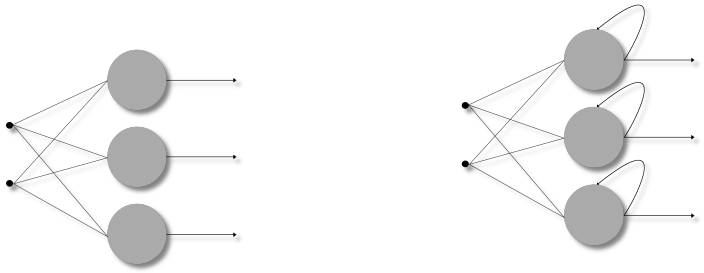
1. http://sintel.is.tue.mpg.de/

dente. La couche de convolution permet d'apprendre la corrélation entre les pixels voisins tout en réduisant considérablement le nombre de paramètres grâce au partage des poids de la carte de caractéristiques calculée [59]. Les couches de pooling permettent quant à elles de réduire la dimension de la donnée en gardant les informations importantes des différentes zones de l'image, souvent en ne gardant que la valeur maximale ou la valeur moyenne de chaque fenêtre. Le pooling permet également une invariance spatiale [64]. La sortie de la couche précédant les couches entièrement connectées représente alors les caractéristiques extraites de la donnée.

2.3.4 Caractéristiques apprises spatio-temporelles

Afin d'ajouter une dimension temporelle au réseau de neurones, différentes solutions ont été proposées. Une première solution consiste à s'intéresser à une succession de caractéristiques issues d'images successives d'une séquence vidéo (RNN). Une deuxième catégorie de solutions propose une description conjointe spatio-temporelle en mesurant les correlations des pixels dans le temps et l'espace (3D-CNN). Une dernière possibilité est de combiner ces deux méthodologies en étudiant en parallèle les données statiques et l'évolution spatio-temporelle (Ensemble network).

RNN : pour prendre en considération la dimension temporelle, des architectures de réseaux de neurones récurrents ont été proposés en ajoutant une mémoire à l'architecture [65] : soit en récupérant entièrement la sortie de l'activation à $t-1$ pour un RNN traditionnel, soit en ajoutant des fonctions permettant de conserver ou d'oublier certaines parties de l'information afin d'éviter la disparition du gradient. La Figure 2.10 illustre la différence d'apprentissage entre un algorithme classique dit "feedforward" et un réseau de neurones récurrent. Dans le cas d'un réseau "feedforward" l'apprentissage se fait toujours d'une couche à la suivante sans aucun retour en arrière alors qu'un réseau récurrent va permettre de cumuler les informations. Dans le cadre d'une



(a) Réseau "feedforward" (b) Réseau récurrent

FIGURE 2.10 – Différences de propagation entre un réseau classique "feedforward" et un réseau récurrent permettant de garder en mémoire des informations sur les données précédentes.

séquence vidéo, le cumul se fera sur les différentes images de la séquence. Un réseau de neurones récurrent permet alors de caractériser l'évolution temporelle des caractéristiques statiques.

3D-CNN : une autre solution proposée dans l'état de l'art est d'utiliser une architecture 3D-CNN, qui étend l'architecture CNN en remplaçant les convolutions spatiales par des convolutions spatio-temporelles qui partagent les mêmes poids au cours du temps [66]. Cette dernière solution a été particulièrement utilisée pour la reconnaissance automatique des expressions faciales dans un cadre dynamique [67, 68, 69]. Les convolutions 3D permettent alors de caractériser simultanément l'évolution spatiale et temporelle en caractérisant les changements locaux au sein des zones de convolution.

Ensemble network : une autre possibilité proposée dans l'état de l'art est d'apprendre à partir de données statiques et temporelles. Pour ce faire, une première possibilité est d'entraîner deux réseaux simultanément :un réseau qui apprend dans le domaine spatial et un réseau entraîné dans le domaine temporel. Des connexions résiduelles entre les deux réseaux peuvent également être ajoutées pour améliorer le lien entre les réseaux durant l'apprentissage. En 2017, par exemple, Zhang et al. [70] proposent

une architecture temporelle qui analyse la trajectoire des points caractéristiques ainsi qu'un réseau qui analyse l'image en apex statique. Les sorties de ces deux réseaux sont fusionnées pour obtenir une prédiction finale qui prend alors en considération les informations temporelles grâce à la trajectoire des points caractéristiques et les informations spatiales avec l'analyse de l'image. Une autre solution est de donner en entrée d'un même réseaux des données statiques et des données temporelles tels que le flux optique par exemple. En 2019, Sun et al. [71] proposent un apprentissage sur trois canaux : l'image de l'expression en apex, le composant en X du flux optique calculé entre l'image neutre et l'image apex et le composant en Y de ce même flux optique. Un mécanisme de fusion est ensuite appliqué pour permettre d'obtenir des caractéristiques qui permettent de décrire les informations spatiales grâce à l'image et les informations temporelles du flux optique.

Parmi les différentes approches proposées pour encoder l'information temporelle, les approches qui prennent en entrée des informations de mouvements semblent assez intéressantes car elles obtiennent des résultats de reconnaissance satisfaisants en nécessitant relativement peu de données. En effet, ces approches calculent au préalable une partie de l'information temporelle en extrayant les points caractéristiques ou le flux optique [59].

2.4 Bases de données et évaluations

Les méthodes de reconnaissance automatique des expressions faciales ainsi que les bases de données disponibles pour évaluer ces méthodes ont fortement évolué ces dernières années. L'explosion du nombre de données disponibles, notamment grâce à l'essor d'Internet, a permis une évolution des bases de données disponibles, avec des données de plus en plus naturelles et en nombre de plus en plus conséquent. Alors que les premières bases de données disponibles ont été principalement acquises dans des

conditions de laboratoire, ce qui limite fortement le nombre de sujets et les conditions d'acquisition, les bases de données les plus récentes sont maintenant souvent créées en collectant des images sur Internet. Cela permet d'obtenir des bases de données beaucoup plus conséquentes et qui reflètent des conditions d'acquisition beaucoup plus naturelles.

Pour donner un aperçu de l'évolution des bases de données, certaines notions importantes sont résumées dans le Tableau 2.1. Dans ce tableau, nous pouvons observer que les bases de données récentes et bien connues ne sont plus acquises dans des conditions de laboratoire et que, d'une part, la taille des bases de données est beaucoup plus importante, et, d'autre part, les défis que posent ces données sont désormais multiples. Nous pouvons également remarquer, à partir de ce tableau, que les bases de données les plus récentes sont plus souvent constituées de données statiques. Cette évolution peut s'expliquer par le fait qu'il est plus compliqué de récupérer des données vidéos annotées sur Internet que des images. Ces bases de données ont donc l'avantage d'être plus conséquentes mais ne permettent pas une analyse dynamique qui se montre, pourtant, plus efficace.

Le Tableau 2.2 présente les résultats obtenus par différentes méthodes de l'état de l'art sur ces bases de données, extraits du recueil proposé par Li et al. [59]. Comme le montre ce tableau, alors que les expressions faciales dans les bases de données contrôlées sont maintenant presque parfaitement reconnues, les bases de données plus naturelles comme AFEW sont encore aujourd'hui confrontés à des difficultés et les résultats sont encore inférieurs à 60%. Parmi les défis rencontrés, les variations de la pose de la tête et les occultations sont considérées comme les principaux défis [59, 28].

Base de données	Année	Statique/Dynamique	Environnement d'acquisition	Vue de profil	Variations de luminosité	Occultations	No. de sujets
JAFFE [72]	1998	statique	lab	non	non	non	10
CK [22]	2000	dynamique	lab	non	non	non	97
MMI [73]	2005	les deux	lab	oui	non	non	19
BU-3DFE [74]	2006	statique	lab	non	non	non	100
Oulu-CASIA [75]	2008	dynamique	lab	non	oui	non	50
BU-4DFE [76]	2008	dynamique	lab	non	non	non	101
CK+ [23]	2010	dynamique	lab	non	non	non	123
MUG [77]	2010	dynamique	lab	non	non	non	86
AFEW [27]	2011	dynamique	films	oui	oui	oui	220
FER [78]	2013	statique	Internet	oui	oui	oui	35,887
RAF-DB [79]	2017	statique	Internet	oui	oui	oui	29,672
AffectNet [19]	2017	statique	Internet	oui	oui	oui	$\sim$ 1,800,000

TABLE 2.1 – Évolution des bases de données proposées dans la littérature pour la reconnaissance des expressions faciales.

Bases de données	Architectures statiques	Architectures dynamiques
JAFFE	95.8%	/
CK+	98.9%	**99.6%**
MMI	78.5%	**91.5%**
AFEW	/	58.8%

TABLE 2.2 – Résultats de reconnaissance des expressions faciales obtenus avec des approches d'apprentissage profonds de l'état de l'art à partir d'architectures statiques ou dynamiques [59].

Par ailleurs, ce tableau semble confirmer que les informations temporelles aident à la reconnaissances des expressions faciales : les approches dynamiques donnent ici les meilleurs résultats.

2.5 Conclusion

Nous avons étudié dans ce chapitre l'ensemble des étapes du processus classique de reconnaissance des expressions faciales. Après un aperçu du processus complet de reconnaissance des expressions faciales dans la section 2.1, nous avons étudié chaque étape de ce processus.

La première étape est l'étape de prétraitement exposée à la section 2.2 qui se compose généralement d'une étape de détection du visage puis il peut s'avérer utile de détecter les points caractéristiques du visage et, si nécessaire, une dernière étape de normalisation est appliquée pour rendre les données uniformes et minimiser le bruit apporté par différentes variations liées à la captation.

À partir de ces images prétraitées, deux catégories de méthodes peuvent être employées. D'une part, des caractéristiques classiques (LBP, points caractéristiques, etc.)

peuvent être extraites afin d'entraîner un classifieur. D'autre part, des architectures de réseaux de neurones peuvent apprendre conjointement les caractéristiques les plus pertinentes et le classifieur. Nous avons vu que l'information temporelle semble particulièrement adaptée pour mieux reconnaître les expressions faciales. Enfin, nous avons vu que les méthodes de l'état de l'art restent encore sensibles à différents défis en environnement naturel.

Chapitre 3

Occultations partielles du visage : impacts et solutions

Contents

Dans un environnement naturel, des accessoires ou la personne elle-même peuvent souvent cacher une partie du visage. Un foulard, un masque de chirurgie ou la main de la personne peuvent masquer la partie inférieure du visage. Des lunettes de soleil, des cheveux ou un chapeau peuvent occulter la partie supérieure du visage. De plus, ces occultations peuvent être aggravées par un mauvais éclairage, une mauvaise résolution de l'image ou par des éléments ajoutés artificiellement comme un flou cinétique ou des éléments d'anonymisation par exemple. Tous ces éléments rendent la reconnaissance difficile en raison d'une perte d'informations et du bruit ajouté par ces occultations.

Afin de répondre à cette problématique, différentes solutions ont été proposées dans la littérature. Ces solutions peuvent être classées en deux catégories :

1. les méthodes qui se concentrent sur les régions visibles du visage. Ces méthodes s'intéressent aux informations disponibles et tendent à ignorer les zones cachées du visage.
2. les méthodes basées sur une reconstruction de l'information perdue par les occultations. Ces méthodes permettent de retrouver une donnée qui contient toutes les informations d'un visage complet.

Dans la première catégorie, les premières solutions proposées [80, 81] divisent le visage en différentes régions et s'intéressent principalement aux régions visibles. Plus récemment, ces méthodes ont évolué vers des approches neuronales grâce aux mécanismes d'attention qui apprennent à se focaliser sur les parties visibles du visage. Ces mécanismes utilisent des poids selon l'importance des régions dans le processus de reconnaissance [82, 83]. Dans le cas des occultations, les poids se retrouvent naturellement plus élevés sur les zones du visage non occultées.

Pour la deuxième catégorie, les premières approches [84, 85] étaient basées sur le suivi des points caractéristiques. Dans le cas d'une occultation, les points caractéristiques qui ne sont pas détectés, car la zone n'est pas visible, sont déduits avec des

approches statistiques. Avec l'arrivée des descripteurs d'apparence, les solutions proposées ont plutôt eu tendance à reconstruire directement l'apparence en employant des algorithmes basés sur un RPCA (Robust Principal Component Analysis) [86] ou à partir d'algorithmes génératifs [87]. Ces derniers ont connu une grande évolution ces dernières années avec l'essor des architectures d'apprentissage profond [88].

Les sections suivantes décrivent, dans un premier temps, l'impact des occultations partielles dans le cadre d'un processus de reconnaissance des expressions faciales. Dans un second temps, nous détaillons davantage les deux catégories de solutions proposées dans la littérature pour répondre à la problématique des occultations. Nous proposons, ensuite, un aperçu de la méthodologie utilisée pour évaluer ces solutions dans la littérature afin de mettre en évidence la diversité des protocoles proposés et la difficulté de comparer ces solutions.

3.1 Impact des occultations

À notre connaissance, l'étude la plus complète de l'impact des occultations sur la reconnaissance des expressions faciales a été réalisée par Kotsia et al. en 2008 [89] sur la base de données CK. Dans cette étude, les auteurs étudient l'importance des régions du visage pour reconnaître les expressions faciales. Cette étude permet d'avoir un aperçu des régions du visage dont l'occultation impacte le plus la reconnaissance des expressions. Ils ont ensuite étudié différentes occultations importantes : de la partie gauche du visage, de la partie droite du visage, des yeux et de la bouche.

Ces travaux montrent que les occultations gauche et droite ont relativement peu d'impact. Cette observation semble indiquer une certaine symétrie au niveau des expressions faciales. Pourtant, différentes études ont montré l'asymétrie des expressions faciales [90, 91]. Les auteurs mettent également en avant le fait que l'importance des régions du visage dépend fortement de l'expression faciale considérée. Enfin, ils notent

que l'occultation des yeux et de la bouche ont un impact important sur les taux de reconnaissance des expressions faciales.

En 2012, Azmi et Yegane [92] étudient l'impact des occultations sur la base de données JAFFE et notent également une différence notable de l'impact des occultations en fonctions des expressions faciales. Ils notent également une différence de l'impact des occultations en fonction des caractéristiques extraites. L'occultation du haut du visage, et en particulier des yeux, a un impact plus important que les occultations du bas du visage en utilisant Gabor alors qu'ils notent l'inverse en utilisant des descripteurs tels que LBP ou LGBP.

On peut noter de ces différents travaux la difficulté à déduire l'importance des régions du visage et l'impact que peuvent avoir les occultations sur ces régions. Ces travaux montrent ainsi qu'en fonction des descripteurs utilisés, l'impact peut être différent. On observe, néanmoins, que les yeux et la bouche sont des zones importantes et que les occultations de ces zones sont particulièrement néfastes pour la reconnaissance des expressions faciales. Ces travaux montrent également que les occultations engendrent des impacts différents en fonction des expressions faciales. Une analyse par expression semble alors adéquate pour s'adapter aux différentes occultations du visage.

3.2 Exploitations des régions visibles du visage

Une première catégorie de solutions concentre l'attention sur les régions visibles du visage et ignore partiellement, voire totalement, les zones occultées.

Certaines solutions ont été proposées sur la base d'un découpage en régions du visage à partir d'une représentation éparse d'images de visages. Ces méthodes sont inspirées des classifieurs SRC (Sparse Representation Classifier)[93] qui consistent à

créer des dictionnaires à partir d'images de visages. Chaque dictionnaire est constitué d'images d'une même classe (dans notre cas des images représentant une même expression faciale). Lors de la phase de test, la donnée d'entrée est représentée en calculant une combinaison linéaire des données d'apprentissage d'un même dictionnaire. Ce calcul est réalisé afin d'obtenir une combinaison linéaire la plus similaire possible à la donnée initiale. Le dictionnaire utilisé pour représenter la donnée de test permet alors automatiquement la classification à partir du label du dictionnaire.

Cotter et al. [94, 95] proposent en 2010 de s'appuyer sur ce type de classifieur en découpant le visage en différentes régions et en constituant des dictionnaires pour chacune de ces régions. La classification finale se fait ensuite en combinant les résultats de classification par région. Lors du calcul de la combinaison linéaire, la différence entre l'image originale et la combinaison linéaire (constituée d'images non occultées) est plus importante dans le cas d'une occultation. Afin d'exploiter cette propriété, Cotter et al. proposent une pondération des différentes régions du visage lors de la fusion des classifieurs.

Cette observation a également été utilisée par Huang et al. en 2012 [96] afin de permettre une détection de l'occultation. Ce type de solutions basées sur des dictionnaires d'images nécessite cependant des données d'apprentissage suffisamment proches de données de test pour permettre de calculer une combinaison linéaire pertinente.

D'autres approches, qui divisent également le visage en plusieurs régions, utilisent un mécanisme de fusion. Certaines solutions proposent une fusion de décision pour chaque région du visage. Liu et al. [81] proposent un découpage du visage en parts égales et une classification indépendante de chaque région du visage. Une fusion de décision est ensuite appliquée en sélectionnant la classe la plus représentative parmi les classifications des différentes régions.

Pour pondérer l'influence des différentes régions du visage, des solutions, comme celle proposée par Dapogny et al. [97], proposent une méthode basée sur un calcul de poids des régions. Dans leurs travaux, les auteurs ont utilisé un auto-encodeur entraîné à donner un poids de confiance aux différentes régions. Cette confiance détermine automatiquement si la région permet d'obtenir des informations pertinentes et, implicitement, si la région est occultée ou non.

Plus récemment, ces techniques tendent à être automatisées par des mécanismes d'attention dont le but est de se concentrer sur les régions les plus pertinentes du visage [82, 83]. La Figure 3.1 illustre les cartes d'attention obtenues par ces mécanismes. Sur cette figure, les touches de couleur montrent les zones d'attention, les zones les plus rouges dénotent les zones où l'attention est la plus forte. Les poids d'importance permettent ensuite de pondérer les pixels de l'image en fonction de l'importance que ces mécanismes lui portent. Dans le contexte des occultations, ces mécanismes sont utilisés pour se concentrer sur les régions visibles du visage.

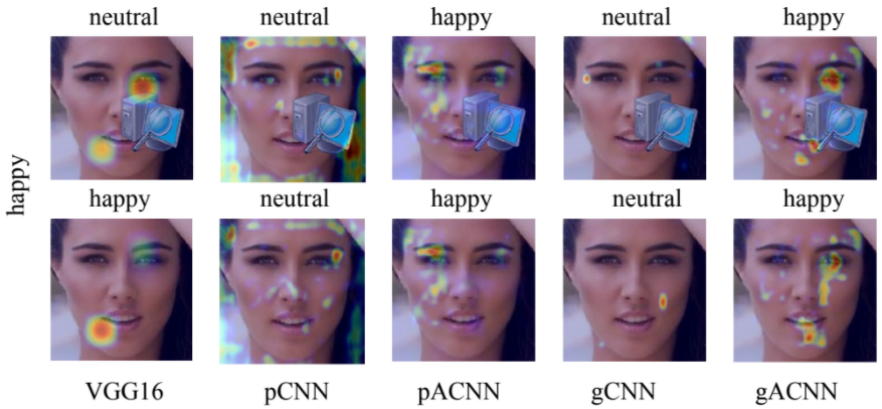


FIGURE 3.1 – Cartes d'attention obtenues avec différentes méthodes sur les images originales (deuxième ligne) et en rajoutant une occultation (première ligne). Le label de cette image est indiqué à gauche (happy) et les résultats de classification de chaque méthode sont indiqués au dessus de chaque image. Cette figure est extraite de l'article de Li et al. [82]

3.3 Reconstruction des zones cachées du visage

Une seconde catégorie de solutions s'appuie sur une reconstruction des zones cachées du visage. Différentes méthodes ont été proposées dans ce sens. Ces méthodes peuvent être basées sur une reconstruction des caractéristiques ou sur une reconstruction de la texture.

3.3.1 Reconstruction de caractéristiques

Des méthodes ont été proposées au début des années 2000, basées sur le suivi des points caractéristiques du visage. Bourel et al. [84, 85] ont proposé de retrouver les points caractéristiques perdus en utilisant le tracker Kanade-Lucas. Plus tard, des solutions statiques ont été proposées, basées sur une reconstruction des points caractéristiques. Towner et Slater [98] ont proposé, dans ce but, une méthode basée sur l'analyse en composantes principales (ACP) tandis que Zhang et al. [99] ont proposé une combinaison du point itératif le plus proche (ICP) et des Fuzzy-C-Means (FCM). Ce type de méthode est fortement dépendant de la détection des points caractéristiques. Cependant, la détection des points caractéristiques est encore aujourd'hui complexe en présence d'occultations [33].

3.3.2 Reconstruction de texture

Plutôt que reconstruire les caractéristiques, d'autres méthodes s'appuient sur la reconstruction de la texture. En reconstruisant directement la texture, ces méthodes permettent d'éviter les difficultés liées à l'extraction des caractéristiques en présence d'occultations. Ces méthodes de reconstruction sont souvent basées sur une analyse robuste en composantes principales (RPCA) [100] qui a été proposée comme une amélioration de l'ACP [101] et qui s'avère plus robuste aux occultations. Dans ce cas, une

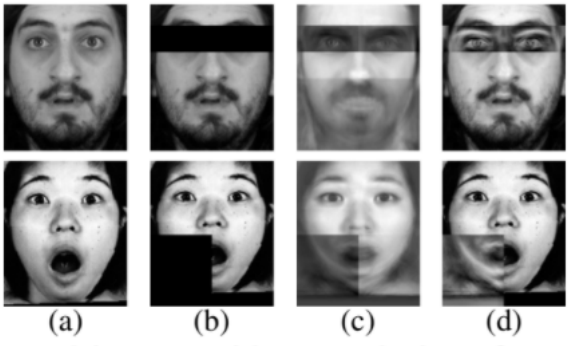


FIGURE 3.2 – Processus de reconstruction par RPCA extrait de l'article proposé par Cornejo et al. [86]. (a) illustre les images originales de MUG [77] et JAFFE [72] sans occultation. (b) représente les images occultées. (c) illustre la reconstruction RPCA de l'image occultée et (d) présente l'image originale reconstruite en remplaçant la zone occultée par la reconstruction.

reconstruction basée sur la RPCA est calculée et la partie occultée de l'image originale est remplacée par la partie reconstruite [86]. Cette méthode a cependant tendance à ajouter des artefacts sur l'image originale comme le montre la Figure 3.2 extraite de l'article proposé par Cornejo et al. [86]. Ces artefacts peuvent alors déformer l'expression faciale et ainsi rendre la reconnaissance plus complexe.

Plus récemment, les solutions se tournent vers une reconstruction basée sur une architecture de réseaux de neurones. Parmi ces typologies d'architectures, nous retrouvons les auto-encodeurs (AE) et les réseaux antagonistes génératifs (GAN).

L'auto-encodeur dont l'architecture est schématisée sur la Figure 3.3, est constituée d'une couche d'entrée, d'une couche cachée et d'une couche de sortie. L' auto-encodeur crée une compression en sortie de l'encodeur et reconstruit la donnée initiale en sortie du décodeur à partir de sa représentation compressée (avec un nombre de neurones généralement similaire en entrée et en sortie).

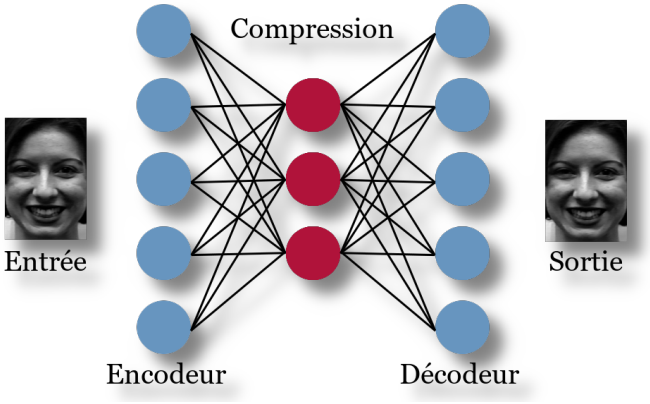


FIGURE 3.3 – Illustration d'une architecture classique d'auto-encodeur.

La structure de l'auto-encodeur est composée d'un nombre de neurones en entrée qui est fonction de la dimension des données. Cette couche d'entrée est suivie d'une couche cachée composée d'un nombre réduit de neurones. La sortie de cette couche peut être vue comme une compression de la donnée initiale, car cette représentation contient les éléments nécessaires à sa reconstruction. Enfin, la sortie du décodeur est composée du même nombre de neurones que le nombre de neurones en entrée de l'architecture afin de proposer une reconstruction de la donnée. Afin d'entraîner l'auto-encodeur, une erreur quadratique moyenne (EQM ou MSE en anglais) est souvent appliquée pour quantifier l'erreur de reconstruction de la sortie par rapport à la donnée d'entrée. Un auto-encodeur sans couche de non-linéarité se comporte comme une ACP (Analyse en Composantes Principales) mais la possibilité d'ajouter de la non-linéarité à l'auto-encodeur rend cet outil plus puissant [102]. Différentes variantes de l'auto-encodeur ont été proposées. Parmi ces différentes variantes, on trouve, notamment, l'auto-encodeur débruitant qui reconstruit une donnée sans bruit à partir d'une donnée bruitée.

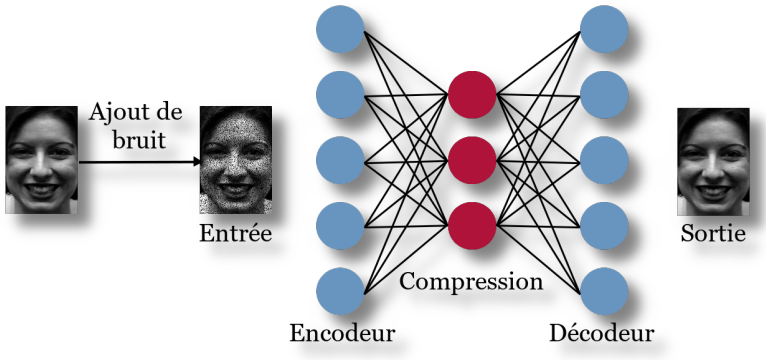


FIGURE 3.4 – L'auto-encodeur débruitant consiste à ajouter du bruit sur la donnée initiale avant d'entraîner l'auto-encodeur à reconstruire la donnée. Cette architecture permet d'entraîner un auto-encodeur plus robuste aux bruits et légères variations.

Un auto-encodeur débruitant, illustré sur la Figure 3.4, est entraîné afin d'être plus robuste au bruit et donc, éviter que l'auto-encodeur apprenne la fonction identité. Dans ce cas, un bruit aléatoire est ajouté aux données d'entrée et l'auto-encodeur est entraîné à reconstruire des données sans bruit. Pour ce faire, prenons x les données d'entrée, $\hat{x}$ la version bruitée de x. $\hat{x}$ est donné en entrée à l'auto-encodeur et l'erreur quadratique moyenne est calculée en comparant la reconstruction avec x et non $\hat{x}$. Dans ce cas, le calcul de l'erreur ne prend pas en compte l'entrée donnée à l'auto-encodeur mais sa version non bruitée.

Dans le cadre des occultations, on peut considérer l'occultation comme un bruit. L'auto-encodeur débruitant est alors entraîné à reconstruire l'image sans occultation. Gondara et al. [103], par exemple, utilisent une architecture d'auto-encodeur débruitant pour supprimer le bruit (artefacts de captation, etc.) présent sur des images médicales . Ces architectures d'auto-encodeurs peuvent également être empilées pour affiner la reconstruction des zones occultées, comme le font Zhang et al. [104] qui proposent une reconstruction des régions occultées du visage dans un processus de reconnaissance de personnes.

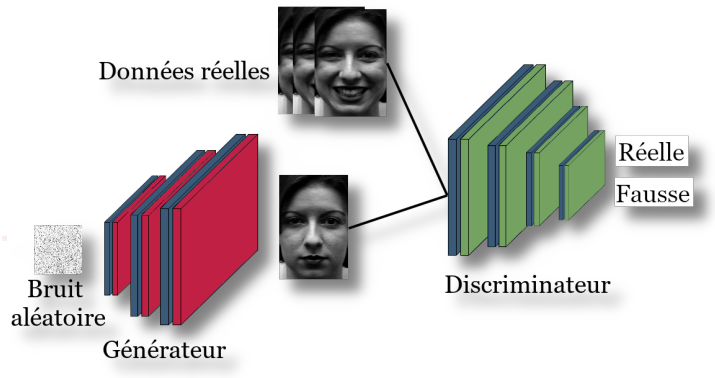


FIGURE 3.5 – Illustration d'une architecture type GAN.

Les réseaux antagonistes génératifs représentent une deuxième catégorie d'architectures de reconstruction de plus en plus utilisées dans l'état de l'art [87, 105, 106]. Les réseaux antagonistes génératifs, introduits en 2014 par Goodfellow et al. [107], sont principalement utilisés comme algorithmes génératifs mais peuvent également être utilisées pour une reconstruction de données bruitées.

Un réseau antagoniste génératif est basé sur une relation entre deux réseaux nommés générateur et discriminateur comme l'illustre la Figure 3.5. Le générateur est entraîné à créer de nouvelles données tandis que le discriminateur essaie d'évaluer le réalisme des données générées par le générateur. La sortie du discriminateur est directement utilisée par le générateur comme fonction de coût.

L'apprentissage parallèle de ces deux réseaux entraîne alors une complexité beaucoup plus importante que pour l'auto-encodeur seul car il faut trouver un équilibre d'apprentissage, ce qui implique une paramétrisation plus complexe que pour l'apprentissage d'une architecture d'auto-encodeur seule.

On peut noter qu'une architecture d'auto-encodeur est, en général, utilisée comme générateur. Li et al. [108] proposent, en effet, une reconstruction de visages occultés qui utilise un auto-encodeur en tant que générateur et deux discriminateurs permettant de vérifier la cohérence locale et la cohérence globale de l'image reconstruite.

Les réseaux antagonistes génératifs sont souvent utilisés pour l'édition [109] ou la complétion d'images [106, 105, 108, 110]. Ces nouvelles architectures sont, cependant, encore assez peu utilisées dans le cadre d'une reconstruction des zones occultées du visage pour la reconnaissance des expressions faciales. Dans un cadre de reconnaissance des expressions faciales, à notre connaissance, Ranzato et al. [111] ont été les premiers, en 2011, à proposer une architecture profonde pour reconstruire un visage occulté. Plus récemment, Lu et al. [87] ont proposé une architecture GAN basée sur deux discriminateurs : un discriminateur classique qui tente de distinguer les données réelles des données générées et un qui évalue la capacité à reconnaître les expressions faciales à partir des images générées. Cette solution prend donc en considération la capacité à bien reconstruire les informations liées à l'expression faciale.

3.4 Protocoles d'évaluation existants proposés dans la littérature

Dans cette section, nous comparons les protocoles expérimentaux proposés dans la littérature pour évaluer les méthodes décrites dans les paragraphes précédents. Pour ce faire, nous proposons de comparer différents points du protocole expérimental que sont : les bases de données utilisées, les occultations générées et le protocole utilisé pour découper la base de données en une base d'apprentissage et une base de test.

3.4.1 Bases de données

Nous avons référencé dans le Tableau 3.1 les bases de données utilisées pour évaluer les méthodes présentées dans les sections précédentes. Comme le montre ce tableau, on peut dénombrer 10 bases de données différentes utilisées de manière hétérogène. Ainsi, il est alors assez complexe de comparer ces méthodes entre elles. Cependant, on

	Article	Base de données
Régions visibles	Cotter et al. [94, 95]	JAFFE
	Huang et al. [96]	CK+
	Liu et al. [81]	JAFFE
	Zhang et al. [80]	CK, JAFFE
	Dapogny et al. [97]	CK+, BU-4D, SFEW
	Li et al. [82]	RAF-DB, AffectNet, CK+, MMI, Oulu-Casia, SFEW
	Wang et al. [83]	FERPlus, AffectNet, RAF-DB, SFEW
Reconstruction	Bourel et al. [84, 85]	CK
	Towner and Slater [98]	CK
	Cornejo et al. [86]	CK+, JAFFE, MUG
	Ranzato et al. [111]	CK, TFD
	Lu et al. [87]	AffectNet, RAF-DB

TABLE 3.1 – Bases de données utilisées dans la littérature pour évaluer les méthodes proposées pour répondre à la problématique des occultations en se concentrant sur les régions visibles du visage ou en passant par une reconstruction.

remarque que quelques bases de données comme CK, CK+ ou JAFFE sont les bases de données les plus fréquemment utilisées.

Le Tableau 3.2 propose un comparatif en termes de nature de données et conditions de capture des bases les plus utilisées. Ce tableau reprend les bases de données qui apparaissent au moins dans 3 articles cités dans le tableau précédent. CK+ [23] étant une extension de la base de données CK [22], nous les considérons ensemble. CK et CK+ sont les bases de données les plus fréquemment utilisées dans la littérature avec 8 occurrences sur les papiers cités dans le tableau précédent. Ce sont également les seules bases de données qui contiennent des séquences vidéo, ce qui permet d'exploiter directement le mouvement induit par l'expression faciale. On remarque enfin que, parmi

ces bases de données, on retrouve des bases de données contrôlées et non contrôlées. On peut cependant noter que les bases de données contrôlées sont plus utilisées que les bases de données non contrôlées. Ce point s'explique facilement par le fait qu'une base de données contrôlée permet de se concentrer entièrement sur la problématique des occultations sans avoir de bruit introduit par une mauvaise luminosité ou des variations de pose. De plus, ces bases de données étant entièrement contrôlées, les occultations seront obligatoirement simulées, ce qui permet de comparer facilement les résultats sans occultation avec les résultats obtenus en appliquant une solution particulière pour répondre à la problématique des occultations. La différence entre ces résultats est alors entièrement dûe aux occultations.

Base de données	**# / Travaux**	**Vidéo**	**Contrôlé**
CK-CK+	8 / [96, 97, 82, 84, 85, 98, 80, 86, 111]	**oui**	**oui**
JAFFE	4 / [94, 95, 81, 80, 86]	non	**oui**
AffectNet	3 / [82, 83, 87]	non	non
SFEW	3 / [97, 82, 83]	non	non
RAF-DB	3 / [82, 83, 87]	non	non

TABLE 3.2 – Comparatif des bases de données parmi les plus fréquentes présentées dans le Tableau 3.1.

Au regard de ces observations et des possibilités laissées par la base CK+ de travailler à la fois dans le domaine spatial et temporel, nous approfondissons, dans la suite de cette section, la manière dont cette base de données a été utilisée dans les méthodes s'intéressant à la reconnaissance d'expressions en présence d'occultations.

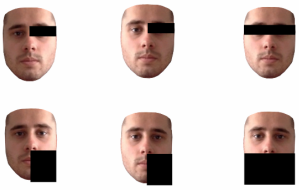
Article	Occultations	Simulation
Dapogny et al. [97]		Superposition d'un rectangle de bruit poivre et sel
Li et al. [82]		Superposition d'images de 4000 objets collectés à partir d'Internet basé sur 50 recherches de mots-clés
Cornejo et al. [86]		Superposition de boites noires

FIGURE 3.6 – Simulations des occultations utilisées pour évaluer différentes solutions de la littérature.

3.4.2 Stratégies d'occultation

En reprenant les différents papiers exposés dans les tableaux précédents et en se concentrant sur ceux qui s'évaluent sur la base de données CK+, nous proposons de comparer la façon dont les auteurs simulent les occultations étudiées lors de leurs évaluations.

Les simulations d'occultations sont présentées sur la Figure 3.6. Comme le montre cette figure, bien que ces différentes solutions soient évaluées sur la même base de données, les occultations proposées sont simulées de façon très différente. Ces différences reposent sur : la façon de simuler l'occultation (boîtes noires, objets, rectangles poivre et sel) mais aussi sur l'emplacement de ces occultations. Notons que, bien que l'occultation des yeux et de la bouche soient des occultations fréquemment utilisées, on constate l'absence d'un cadre unifié permettant de juger de manière impartiale des capacités de reconnaissance en présence d'occultations.

3.4.3 Apprentissage et tests

Nous avons vu que les bases de données utilisées pour évaluer les méthodes sont variées, que sur une même base les occultations générées et étudiées diffèrent également. Nous proposons, enfin, de comparer les découpages des bases en apprentissage et test utilisés dans la littérature en nous concentrant de nouveau sur les méthodes évaluées sur la base de données CK+. On peut enfin constater que sur ces trois méthodes proposées, trois protocoles expérimentaux différents sont également proposés.

Dapogny et al. [97] utilisent l'erreur out-of-bag qui est un protocole spécifique aux algorithmes de random forests et n'est donc pas forcément applicable sur toutes les méthodes.

Cornejo et al. [86] réalisent un découpage avec 80% de données pour l'apprentissage et 20% pour le test. 50% des données d'apprentissage sont occultées ainsi que 50% des données de test. Ce protocole expérimental dépend fortement du découpage originel et des données sur lesquelles sont appliquées les occultations. Sans connaître les données contenues dans ces différents découpages, ce protocole est difficilement comparable à d'autres méthodes.

Enfin, Li et al. [82] utilisent une validation croisée en 10 plis. Même si nous n'avons pas accès au découpage effectué, la validation croisée en 10 plis permet cependant de moyenner les résultats sur l'ensemble des données et réduit ainsi l'impact du découpage initial.

3.4.4 Récapitulatif et résultats d'évaluation des méthodes

Le Tableau 3.3 reprend les évaluations des méthodes proposées dans la littérature et donne un aperçu des résultats obtenus en présence d'occultations les régions des yeux et de la bouche. On note avec ces résultats que l'occultation des yeux semble avoir

plus d'impact sur JAFFE alors que l'occultation de la bouche et du bas du visage a plus d'impact sur CK et CK+. Li et al. [82] obtiennent des meilleurs résultats, cependant, nous notons la différence de protocole assez importante avec l'ajout d'objets sur le visage de taille différente ainsi que la difficulté d'autant plus importante à reproduire et comparer les résultats avec cette méthode.

	Article	**Occultation**	**Yeux**	**Bouche**	**Haut**	**Bas**
JAFFE	(1) Cotter et al. [94, 95]	Boites noires	/	93.4%	/	/
	(1) Liu et al. [81]	Boites noires	87.23%	89.47%	/	/
	(1) Zhang et al. [80]	Boites blanches	80.3%	78.4%	/	/
CK	(2) Bourel et al. [84, 85]	Points supprimés	/	/	$\simeq 80\%^*$	$\simeq 55\%^*$
	(2) Towner et Slater [98]	Points supprimés	/	/	82%	70%
	(2) Ranzato et al. [111]	Boites grises	$\simeq 88\%^*$	$\simeq 87\%^*$	$\simeq 77\%^*$	$\simeq 72\%^*$
	(1) Zhang et al. [80]	Boites blanches	95.1%	90.8%	/	/
CK+	(1) Huang et al. [96]	Boites noires	93%	79.08%	/	73.54%
	(1) Dapogny et al. [97]	Bruit poivre et sel	$\simeq 76\%^*$	67.1%	/	/
	(1) Li et al. [82]	Objets	96.57%	92.93%	/	/

TABLE 3.3 – Résultats de la littérature obtenus par zone d'occultation. Dans la première colonne, (1) représente une méthode qui se concentre sur les régions visibles du visage et (2) représente une méthode qui passe par une reconstruction. $(^*)$ Ces résultats ont été lus sur les graphiques disponibles.

Le Tableau 3.4 récapitule, enfin, les résultats obtenus sur des bases de données naturelles. Les bases de données, bien que naturelles, ne sont pas spécifiques aux occultations et il est donc nécessaire également de déterminer comment étudier les occultations sur ces données. On remarque également une disparité dans les protocoles utilisés : des objets rajoutés, du bruit poivre et sel, des boites noires ou encore en récupérant spécifiquement les images occultées parmi toutes les images de la base de données. Les taux de reconnaissance obtenus sont alors difficilement comparables. Ces résultats se montrent encore largement perfectibles, notamment pour AffectNet et SFEW.

	Article	**Occultation**	**Reconnaissance**
AffectNet	(1) Li et al. [82]	Objets ajoutés	54.84%
	(1) Wang et al. [83]	Sélection des images occultées	58.5%
	(2) Lu et al. [87]	Boites noires	51.21%
RAF-DB	(1) Li et al. [82]	Objets ajoutés	80.54%
	(1) Wang et al. [83]	Sélection des images occultées	86.9%
	(2) Lu et al. [87]	Boites noires	78.35%
SFEW	(1) Li et al. [82]	Objets ajoutés	/
	(1) Wang et al. [83]	Sélection des images occultées	56.4%
	(1) Dapogny et al. [97]	Bruit poivre et sel	37.1%

TABLE 3.4 – Résultats de la littérature obtenus sur des données naturelles. Dans la première colonne, (1) représente une méthode qui se concentre sur les régions visibles du visage et (2) représente une méthode qui passe par une reconstruction.

3.5 Conclusion

Dans ce chapitre, nous avons donné un aperçu de l'impact des occultations du visage sur la reconnaissance des expressions faciales, des méthodes proposées pour y ré-

pondre ainsi que des protocoles utilisés dans la littérature pour évaluer ces méthodes. Dans une première partie, nous avons étudié l'impact des occultations sur la reconnaissance des expressions faciales. Les travaux de la littérature montrent l'importance des régions de la bouche et des yeux avec des impacts différents selon les bases de données étudiées. En plus de ces observations, l'occultation des yeux et de la bouche sont des occultations qui semblent assez fréquentes. On peut notamment citer les occultations liées aux lunettes, à une anonymisation ou à une frange de cheveux, qui apparaissent dans les régions des yeux. Pour les occultations de la bouche, on peut citer les masques, un foulard ou encore la main de la personne (les personnes peuvent avoir tendance à cacher leurs dents quand ils rient, à poser la tête sur la main s'ils s'ennuient ou encore à poser la main sur la bouche lorsqu'elles réfléchissent). Cette première partie permet alors de définir les zones primordiales à étudier dans le cadre des occultations partielles du visage.

Nous avons également vu que l'impact dépendait de chaque expression faciale et du descripteur utilisé. Il semble alors approprié de faire une étude par expression faciale et, en fonction du descripteur, de déterminer à nouveau les régions du visage les plus pertinentes.

Dans une seconde partie, nous avons étudié les différentes solutions proposées dans la littérature pour répondre à la problématique des occultations. Ces solutions sont principalement basées sur des informations de texture ou de géométrie et peuvent être regroupées sous deux principales catégories : les solutions qui se concentrent sur les régions visibles du visage et celles qui reconstruisent les informations cachées. Ces deux catégories de solutions ont connu des avancées importantes ces dernières années et semblent toutes deux prometteuses. Les solutions qui se concentrent sur les régions visibles du visage exploitent l'information disponible et n'ajoutent ainsi aucun biais aux données initiales alors que les méthodes de reconstruction permettent de revenir sur une donnée qui contient des informations sur l'ensemble du visage, qui est un cadre

parfaitement maîtrisé pour obtenir d'excellents taux de reconnaissance.

Pour exploiter les régions visibles du visage, les méthodes se basent généralement sur un découpage du visage afin de s'appuyer, en priorité, sur les régions non occultées.

Pour la reconstruction des zones cachées, les méthodes récentes sont basées sur deux typologies de réseaux de neurones : les auto-encodeurs et les réseaux antagonistes génératifs. Les réseaux antagonistes génératifs s'avèrent plus coûteux et complexes à mettre en place. Une architecture d'auto-encodeur s'avère alors un bon compromis entre complexité, qualité de reconstruction et nombre de données nécessaires.

Dans la dernière partie de ce chapitre, nous avons étudié les protocoles expérimentaux utilisés dans la littérature pour évaluer les méthodes proposées pour tenter de répondre à la problématique des occultations partielles du visage. Nous avons pu mettre en avant l'absence de protocole expérimental unifié pour équitablement comparer les différentes solutions. Cette section nous a, cependant, permis de noter certains points communs de la littérature permettant de mettre en place un protocole le plus représentatif possible. Nous pouvons noter, dans cette dernière section, que CK+ est une base de données qui a été la plus fréquemment utilisée et se montre particulièrement adaptée. Nous avons également pu noter que les occultations proposées dans la littérature, bien que différentes, sont principalement situées sur les régions des yeux ou de la bouche. Nous avons également vu que le découpage en données d'apprentissage et de test n'est pas non plus uniformisé. Nous avons, toutefois, noté qu'un découpage en 10 plis permettait une comparaison plus équitable car, à tour de rôle, l'ensemble des données servent à la fois pour l'entraînement et pour le test, tout en limitant les effets de bord des tirages aléatoires.

En ce qui concerne les données provenant de bases de données non contrôlées, il reste encore de larges améliorations possibles. Des marges de progression sont également attendues dans un cadre contrôlé, les taux de reconnaissance sont encore d'en-

viron 20% plus bas lorsque la bouche est occultée. Il nous semble alors nécessaire de poursuivre les travaux sur des bases de données contrôlées, qui permettent d'étudier les données de façon dynamiques grâce aux bases de données disponibles mais, surtout, qui permettent de s'abstraire, pour le moment, d'autres difficultés engendrées par des défis supplémentaires tels quel les variations de pose notamment.

Chapitre 4

Synthèse

Dans cette première partie du manuscrit, nous avons étudié les solutions existantes de l'état de l'art pour reconnaître automatiquement les expressions faciales. Nous avons vu, au Chapitre 2, les différentes étapes utilisées lors d'un processus de reconnaissance automatique des expressions faciales.

Nous avons identifié deux catégories de méthodes utilisées pour reconnaître les expressions faciales : les méthodes basées sur une extraction de caractéristiques classiques et les méthodes basées sur des caractéristiques apprises (à partir de réseaux de neurones).

Nous avons noté que, que ce soit avec une extraction de caractéristiques classiques ou par l'utilisation de réseaux de neurones, les informations de mouvement semblent s'avérer particulièrement importantes pour reconnaître de façon plus efficace les expressions faciales.

Compte tenu de ces observations, nous avons étudié, au Chapitre 3, l'impact et les méthodes proposées dans la littérature pour répondre au défi des occultations. Nous avons pu voir, dans une première section, que la bouche et les yeux, en plus d'être fréquemment occultés, sont des zones qui contiennent des informations importantes per-

mettant de reconnaître les expressions faciales. Les solutions proposées doivent donc en priorité répondre à ce type d'occultations. Pour traiter ces occultations, deux catégories de méthodes ont été proposées dans la littérature : les solutions qui se concentrent sur les régions visibles du visage et celles qui reconstruisent les zones cachées. Ces deux catégories de méthodes ont chacune leurs avantages et leurs inconvénients et ont évolué toutes deux avec, notamment, les avancées récentes des architectures de réseaux de neurones. Les méthodes qui exploitent les régions visibles du visage se sont récemment tournées vers les mécanismes d'attention qui permettent une automatisation de la sélection des zones à étudier alors que les méthodes de reconstruction se tournent davantage vers des architectures de reconstruction comme des réseaux antagonistes adverses. Différentes typologies d'architectures peuvent être utilisées dans un cadre de reconstruction dont notamment les auto-encodeurs et les réseaux antagonistes génératifs.

Nous notons que, malgré l'importance du mouvement pour reconnaître les expressions faciales, le mouvement est très peu utilisé dans le cadre des solutions proposées pour répondre au défi des occultations. Dans une dernière section, nous avons également étudié les protocoles expérimentaux utilisés dans le cadre des méthodes de reconnaissance des expressions faciales en présence d'occultations. Dans cette section, nous avons remarqué l'hétérogénéité et la difficulté de reproduction des protocoles expérimentaux, ce qui ne permet pas une comparaison équitable des méthodes.

Deuxième partie

Exploitation des propriétés de mouvement pour la reconnaissance des expressions faciales en présence d'occultations partielles du visage

Dans le cadre des occultations, une étude dynamique basée sur l'analyse de mouvement semble particulièrement adaptée. Le mouvement dispose, en effet, de différentes propriétés pertinentes pour répondre au défi des occultations dont notamment : la propagation du mouvement liée à l'élasticité de la peau et la similarité inter-personnes des mouvements liés aux expressions faciales.

Propagation du mouvement : comme l'illustre la Figure 4.1, malgré une occultation de la bouche lors d'une expression de joie, où l'épicentre du mouvement, situé aux coins des lèvres, est donc masqué, les informations de mouvement restent très informatives dans les régions aux alentours. Cette propriété de propagation de mouvement s'avère particulièrement utile dans le cadre d'une solution pour répondre au défi des occultations en s'appuyant sur les régions visibles du visage. Ces régions conservent, en effet, beaucoup d'informations malgré les occultations, même lorsqu'elles concernent l'épicentre du mouvement.

FIGURE 4.1 – Illustration avec les données de CK+ des propriétés du mouvement, sur la base de l'occultation d'une partie du visage (bouche et une partie du nez)

Similarité de mouvement : comme le montre la Figure 4.1, les informations de texture sont très différentes entre plusieurs personnes qui expriment la même expression, car la texture est fortement liée à l'identité de la personne. À l'inverse, dans le domaine du mouvement, les informations liées à l'identité de la personne semblent avoir un impact limité et le mouvement induit par une même expression semble relativement similaire d'une personne à une autre. Cette propriété peut s'avérer particulièrement utile dans le cadre d'une méthode de reconstruction de données corrompues par une occultation. Dans notre cas, une zone du visage occultée pourra ainsi être reconstruite en nous basant sur les mouvements effectués par d'autres personnes pour une même expression.

Dans cette seconde partie du manuscrit, nous proposons deux méthodes qui prennent avantage des propriétés évoquées ci-dessus :

a) une méthode qui se concentre sur les régions visibles du visage en se basant sur des caractéristiques de mouvement - Afin de se concentrer sur les régions visibles du visage tout en conservant le plus d'informations possible concernant l'expression faciale, nous proposons de calculer des modèles faciaux composés d'un sous-ensemble de régions qui se concentrent sur les régions les plus informatives. Pour exploiter la propriété de propagation du mouvement, la solution proposée repose entièrement sur un descripteur de mouvement. L'importance des régions du visage est calculée pour chaque expression faciale. Des modèles faciaux composés des régions les plus pertinentes parmi les régions visibles sont calculés pour chaque type d'occultation. Une fusion des décisions de classifieurs par expression permet de déterminer l'expression en présence d'une occultation donnée. Nous aboutissons ainsi à un classifieur par occultation. Cette première méthode a montré son efficacité en présence d'occultations très sévères.

b) une méthode qui reconstruit les zones cachées par l'occultation en effectuant une reconstruction directement dans le domaine du flux optique. - Cette deuxième solution permet de minimiser les informations liées à l'identité de la personne en reconstruisant des informations de mouvement, plutôt liées à l'expression faciale. Pour ce faire, nous avons utilisé une architecture d'auto-encodeur. Comme nous l'avons vu, l'auto-encodeur permet une reconstruction de qualité tout en ayant une complexité modérée (notamment en comparaison à une architecture de type GAN). Cette deuxième solution s'est montrée également efficace en permettant d'obtenir des résultats compétitifs à ceux de l'état de l'art en ayant l'avantage de construire un modèle unique pour toutes les expressions faciales ainsi qu'un seul classifieur quelle que soit l'occultation considérée.

Ces deux solutions sont expliquées et détaillées dans les chapitres suivants :

Le Chapitre 5 reprend en détail les différentes étapes qui nous ont permis de construire les modèles faciaux qui privilégient les régions visibles. Pour évaluer cette première méthode, nous avons proposé différentes occultations les plus sévères possibles afin de construire des modèles faciaux permettant de répondre à un très large panel d'occultations.

Le Chapitre 6 explique la démarche complète proposée pour reconstruire les données corrompues par des occultations en reconstruisant les flux optiques associés. Cette solution se base sur un auto-encodeur débruitant qui prend en entrée des flux optiques calculés à partir d'images corrompues par des occultations simulées. La fonction de coût s'appuie sur les flux optiques calculés à partir des mêmes images non occultées.

Pour chacune de ces deux méthodes, nous avons proposé un protocole expérimental clair incluant une étape de simulation d'occultations. Le code de simulation proposée au Chapitre 6 est publiquement mis à disposition [1]

1. `https://gitlab.univ-lille.fr/fox/occ_gen`

Chapitre 5

Modèles faciaux intelligents pour la reconnaissance d'expressions faciales en présence d'occultations partielles du visage

Contents

5.1 Introduction

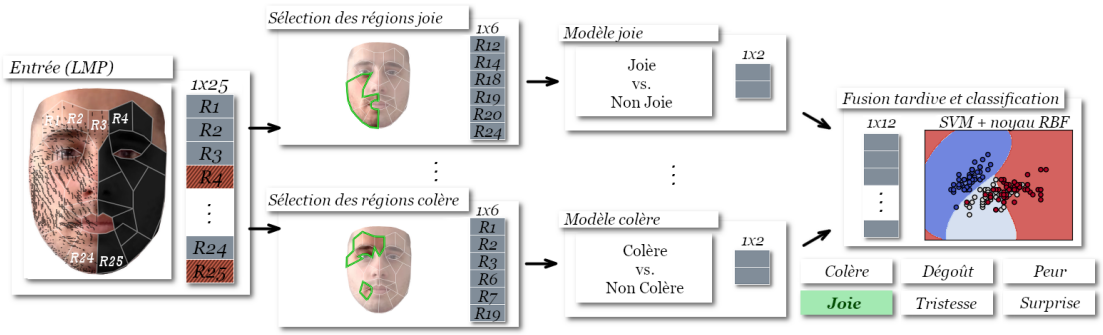


FIGURE 5.1 – Processus complet proposé pour reconnaître les expressions faciales malgré des occultations partielles du visage en se concentrant sur les régions visibles du visage.

Nous proposons d'exploiter la propriété de propagation du mouvement pour proposer une solution de reconnaissance des expressions faciales en présence d'occultations partielles du visage en exploitant les mouvements perçus dans les régions visibles. La Figure 5.1 illustre le processus complet proposé. Pour exploiter la propagation du mouvement, nous extrayons des flux optiques à partir de la séquence vidéo initiale. Nous avons vu au Chapitre 3 que chaque occultation a un impact différent en fonction de l'expression faciale sous-jacente. Nous proposons alors une étude par expression faciale et par occultation. Nous construisons, ainsi, des modèles faciaux optimaux qui prennent en considération les régions du visage qui contribuent le plus à la reconnaissance de chaque expression faciale. Les modèles faciaux sont d'abord calculés en prenant en

considérations toutes les régions du visage. Ces premiers modèles permettent alors d'avoir un aperçu de l'importance des régions du visage pour la reconnaissance de chaque expressions faciale. Ces modèles permettent également de connaître le nombre minimal de régions nécessaire pour reconnaître correctement les expressions faciales. Dans un second temps, différentes occultations sont prises en compte afin de se concentrer uniquement sur les régions visibles.

Ces modèles faciaux sont calculés grâce à un système de pondération de chaque région. Ces poids sont calculés en fonction de l'importance de chaque région pour la reconnaissance de chaque expression faciale. Les régions visibles dont les poids sont les plus importants sont ainsi sélectionnées pour former le modèle facial utilisé. La seconde étape consiste à entraîner des classifieurs binaires à reconnaître chaque expression faciale en considérant uniquement les régions du visage sélectionnées à l'étape précédente. Ces classifieurs permettent de définir si la donnée d'entrée peut être classée ou non selon l'expression faciale considérée. Enfin, un mécanisme de fusion est appliqué pour définir l'expression faciale finale et ainsi obtenir un modèle unifié pour chaque occultation qui, pour une donnée d'entrée, détermine l'expression.

Dans la suite de ce chapitre, nous expliquons, dans la section 5.2, la méthodologie utilisée pour la construction des modèles faciaux pour chaque occultation et chaque expression faciale ainsi que le mécanisme de fusion utilisé. L'évaluation, exposée à la section 5.3, est proposée en deux temps. Dans un premier temps, nous évaluons les modèles faciaux en utilisant des classifieurs binaires par expression faciale. Dans un deuxième temps, les résultats présentés utilisent la méthode complète avec le mécanisme de fusion et permettent alors une comparaison avec les méthodes de la littérature. Nous finissons ce chapitre par un bilan et une discussion sur la méthodologie proposée.

5.2 Importance des régions du visage pour la reconnaissance des expressions faciales

La méthode proposée vise à se concentrer sur un nombre restreint de régions du visage en sélectionnant les régions les plus informatives parmi les régions visibles. Il est donc nécessaire d'établir le meilleur compromis entre le nombre de régions du visage requis pour reconnaître les expressions faciales et la performance obtenue. Pour trouver ce compromis, nous proposons, en première partie de cette section, de détailler le découpage du visage en régions, ainsi que la manière dont les poids sont calculés. Dans une deuxième partie, nous optimisons le nombre de régions nécessaires pour reconnaître de façon optimale les expressions faciales.

Pour calculer l'importance de chaque région du visage, des configurations, c'est-à-dire des sous-ensembles de régions du visage, sont étudiées. Nous étudions les régions en utilisant des configurations afin de prendre en considération la propagation du mouvement, c'est-à-dire en prenant l'ensemble du mouvement plutôt que se concentrer sur son épicentre. De plus, pour reconnaître les expressions faciales et éviter toute ambiguïté, il peut s'avérer nécessaire d'avoir les informations de différentes régions simultanément. Pour chaque configuration, un classifieur est alors entraîné à reconnaître chaque expression faciale en ne conservant que les régions figurant dans la configuration étudiée. À partir des résultats obtenus sur ces différentes configurations, un poids est attribué à chaque région du visage en fonction des performances des différentes configurations. Ces différentes étapes sont détaillées dans la sous-section 5.2.1.

Les poids obtenus sur les différentes régions du visage permettent de calculer les modèles faciaux optimaux contenant un nombre minimal de régions permettant de reconnaître les expressions faciales. La sous-section 5.2.2 explique les étapes utilisées pour optimiser ces modèles faciaux grâce aux poids calculés pour chaque région.

La sous-section 5.2.3 expose, enfin, notre démarche pour fusionner les décisions de chaque classifieur binaire et obtenir une classification unifiée par occultation.

5.2.1 Calcul de poids

Le calcul de poids de chaque région du visage se fait en trois étapes :

1. constituer des sous-ensembles de régions, appelées configurations, contenant un nombre de régions inférieur à celui du modèle facial initial. Pour cette première étape, nous nous inspirons du modèle facial en 25 régions proposé par Allaert et al. [112] qui permet un découpage en régions en fonction des différents muscles du visage.
2. évaluer chaque configuration C_j en étudiant les performances de reconnaissance de chaque configuration en utilisant uniquement les informations de mouvement comprises dans les régions R_k incluses dans C_j.
3. le taux de reconnaissance obtenu pour chaque configuration C_j contribue au calcul de poids de chaque région R_k qui le composent.

Les prochaines sous-sections détaillent ces différentes étapes.

Calcul des configurations

La première étape, permettant de calculer les poids des régions du visage, explore des configurations représentant des sous-ensembles de régions. Le modèle facial initial utilisé est illustré sur la Figure 5.2, découpe le visage en 25 régions.

La génération de toutes les combinaisons possibles parmi les 25 régions du visage aurait été trop gourmand en temps et en puissance de calcul. Afin de tirer partie de la propriété de propagation du mouvement, nous avons alors décidé de ne considérer que les configurations qui contiennent des régions connectées entre elles. Ainsi,

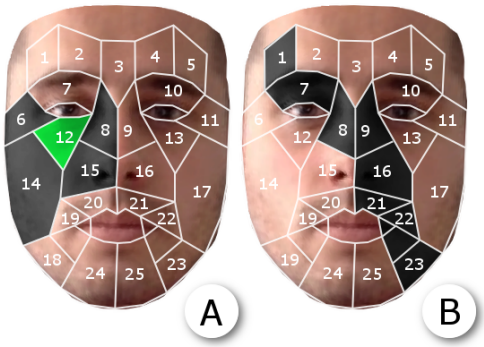


FIGURE 5.2 – Configurations du visage à partir de régions voisines.

comme l'illustre la Figure 5.2-A, à partir de la région 12, la configuration de taille 1 est $\{R_{12}\}$ et les configurations de taille 2 sont : $\{R_{12}, R_6\}$, $\{R_{12}, R_8\}$, $\{R_{12}, R_{14}\}$ et $\{R_{12}, R_{15}\}$ car les régions R_6, R_8, R_{14} et R_{15} sont directement connectées à la région R_{12}. Les configurations de tailles supérieures sont construites de la même façon : en suivant la connectivité entre les régions. Nous avons décidé d'étudier toutes les configurations possibles suivant cette méthodologie jusqu'aux configurations de taille 8. Huit régions permettent ainsi d'obtenir des configurations permettant de parcourir les différentes régions des yeux, les régions du nez, les régions de la bouche mais également de parcourir ces différentes régions grâce à des configurations verticales ou diagonales comme l'illustre la Figure 5.2-B. En allant jusqu'à 8 régions, nous permettons alors d'obtenir des configurations qui contiennent les différents éléments du visage ce qui permet d'établir des corrélations entre ces différents éléments. Nous obtenons un total de 21294 configurations différentes.

Évaluation des configurations

Pour chaque configuration, les informations de mouvement sont extraites des régions sélectionnées. À partir de ces informations de mouvement, un classifieur est entraîné à reconnaître chaque expression faciale pour chaque configuration. Les taux

de reconnaissance obtenus sont ensuite directement utilisés pour calculer les poids de chaque région du visage.

Pondération des régions du visage en absence d'occultations

Pour chaque région, le poids est initialisé à zéro. Un score est ensuite calculé pour chaque configuration C_j en fonction du taux de reconnaissance obtenu, mais aussi en fonction du taux de reconnaissance moyen de toutes les configurations contenant le même nombre de régions que C_j. Le score est, en effet, normalisé par l'écart-type des configurations de même nombre de régions que C_j. Enfin, pour accorder une importance plus grande aux configurations permettant d'obtenir de très bons résultats avec très peu de régions, le score final est normalisé par l'exponentiel du nombre de régions de C_j. Le score est ainsi obtenu par la formule suivante :

$$\omega(C_j, emo) = exp((a(C_j, emo) - \mu_i)/\sigma_i)/exp(i) \tag{5.1}$$

où $j \in [1, 21294]$, $i = |C_j| \in [1, 8]$, $a(C_j, emo)$ est le taux de reconnaissance obtenu par la configuration C_j évalué sur l'expression faciale emo. μ_i et σ_i représentent, respectivement, la moyenne et l'écart-type des taux de reconnaissance obtenus par les configurations contenant i régions. Enfin, $exp(i)$ est l'exponentielle de i et permet de modérer le score en fonction de la taille de la configuration.

Le score ainsi calculé est cumulé aux poids de chaque région du visage R_k comprise dans la configuration C_j. Les poids de chaque région du visage sont, finalement, normalisés en divisant les poids obtenus par le nombre de configurations qui contenaient chaque région. Ces poids finaux représentent ainsi notre score d'importance de chaque région du visage pour reconnaître chaque expression faciale.

La Figure 5.3 présente les poids calculés sous forme de carte de chaleur sur la base de données CK+ [23]. Pour caractériser le mouvement nous utilisons le descripteur LMP [57]. Nous choisissons ce descripteur car, d'une part, il est basé sur du flux optique dense, qui permet de conserver les mouvements subtils liés à la propagation et, d'autre part, ce descripteur a été pensé et créé spécialement pour reconnaître les expressions faciales grâce à un filtrage du flux optique en fonction de la cohérence de mouvement au sein des régions du visage.

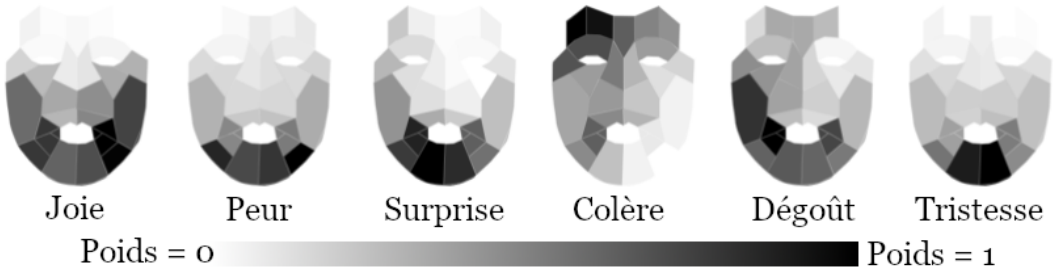


FIGURE 5.3 – Cartes de chaleur représentant le poids d'importance de chaque région du visage pour reconnaître chaque expression faciale. Ces cartes de chaleur sont calculés sur la base de données CK+ [23]

Pour mesurer les taux de reconnaissance nous utilisons un classifieur SVM et un protocole de validation croisée en 10 plis. Nous pouvons observer sur cette figure que les régions les plus importantes du visage sont, en général, situées autour de la bouche, excepté pour la colère où les régions importantes se situent plutôt au niveau des sourcils. Nous pouvons également noter la légère asymétrie des cartes de chaleur, ce qui tend à montrer que les expressions faciales ne sont pas symétriques. Cette asymétrie semble complètement cohérente car elle a déjà été démontrée dans différents travaux [90, 91].

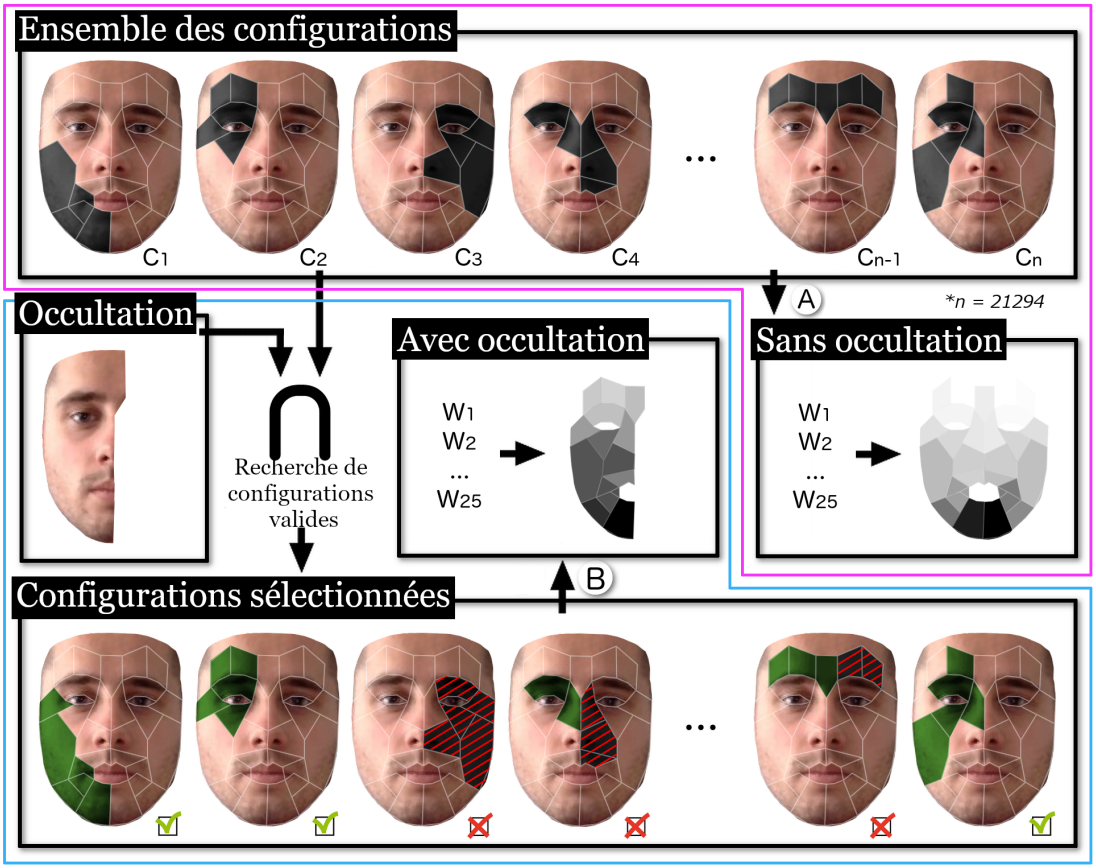


FIGURE 5.4 – Processus de filtrage des configurations pour le calcul de poids des régions en présence ou en absence d'occultation. Les cartes de chaleur représentent les poids obtenus pour l'expression faciale de tristesse sans occultation et avec une occultation de la partie droite du visage.

Pondération des régions du visage en présence d'occultations

Pour prendre en considération différentes occultations et calculer des poids qui prennent en considération le fait que certaines régions ne sont pas visibles, un processus similaire est mis en place, illustré dans la Figure 5.4, où la carte de chaleur représente la carte obtenue pour l'expression de tristesse dans l'encadré bleu.

Plutôt que de prendre en considération toutes les configurations calculées, comme nous l'avons fait sans occultation (dans l'encadré rose), lorsque nous considérons une occultation particulière, un filtrage des configurations est mis en place pour ne pas prendre en compte les configurations qui contiennent au moins une région occultée comme illustré dans la partie basse de la figure. Ainsi, lors du calcul des poids, les régions du visage occultées gardent alors un poids nul.

5.2.2 Optimisation des modèles faciaux par expression et par occultation

À partir des poids calculés, les régions du visage sont triées selon leur ordre d'importance : de la région la plus importante pour reconnaître une expression faciale à la moins importante. Ce classement des régions par expression permet alors de construire différents modèles entraînés pour reconnaître les expressions faciales avec un nombre restreint de régions.

Chaque modèle est ainsi entraîné avec les n régions les plus importantes avec $n \in [1, 25]$. Les scores de reconnaissance obtenus par ces modèles permettent alors de mettre en lumière le nombre minimal de régions permettant de reconnaître chaque expression faciale de façon optimale, c'est-à-dire avec un taux de reconnaissance similaire au score obtenu avec toutes les régions du visage.

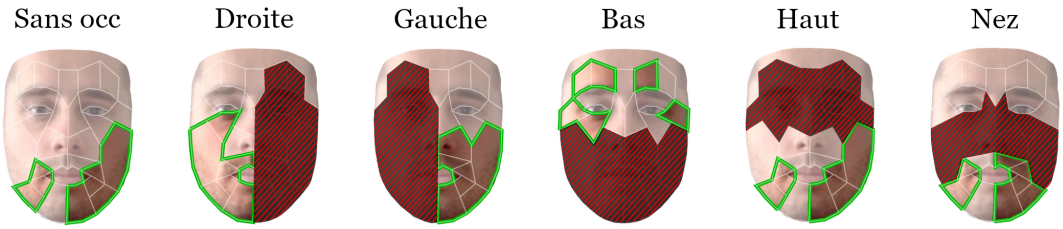


FIGURE 5.5 – Modèles faciaux avec les 6 meilleures régions calculés pour l'expression faciale de joie en fonction de différentes occultations.

Dans la Figure 5.5, on illustre les modèles faciaux calculés sur la base de données CK+ en gardant les 6 meilleures régions. Nous utilisons 6 régions, car, suite aux expérimentations détaillées dans la section 5.3.2 nous avons observé qu'en absence d'occultation, avec 6 régions, nous obtenons des résultats comparables à ceux obtenus en utilisant l'ensemble des régions du visage. Lorsqu'une occultation est considérée, les poids sont recalculés en tenant compte de cette occultation particulière.

Cette démarche a été appliquée sur les six expressions faciales afin de construire un modèle binaire pour chaque occultation et chaque expression. Ces modèles permettent, pour une occultation et une expression faciale données, de déterminer si la donnée peut être classifiée selon l'expression faciale étudiée ou non.

5.2.3 Optimisation des modèles faciaux par occultation

Le processus de construction de modèles faciaux décrit dans la section précédente permet de construire un modèle de reconnaissance par expression faciale et par occultation. Ces modèles donnent lieu à des classifieurs binaires pour chaque expression faciale permettant de déterminer si la donnée d'entrée correspond à l'expression faciale étudiée ou non. L'importance des régions du visage étant spécifique aux expressions faciales, il semblait, en effet, pertinent de faire dans un premier temps une étude par expression. Pour permettre de reconnaître les différentes expressions faciales à partir

de la donnée initiale, nous proposons un mécanisme de fusion des classifieurs binaires qui permet alors d'obtenir en sortie l'expression faciale.

Le processus d'apprentissage du modèle de fusion se décompose en deux principales étapes. Dans un premier temps, les six classifieurs binaires sont entraînés à reconnaître chaque expression faciale en prenant en considération les x régions sélectionnées à partir du mécanisme d'optimisation des régions du visage.

Ces différents modèles fournissent, en sortie, la probabilité que la donnée d'entrée soit de la classe étudiée ou non. À partir de ces probabilités, un nouveau modèle est entraîné en prenant, en entrée, la concaténation des probabilités de chaque classifieur binaire et permet d'obtenir, en sortie, l'expression faciale. Pour ce faire, les classifieurs binaires sont gelés et le deuxième apprentissage ne se fait que sur le classifieur de fusion.

Ce mécanisme de fusion permet d'obtenir un classifieur unifié par occultation qui permet de déterminer l'expression faciale.

5.3 Évaluation

Nous évaluons, dans cette section, les différentes étapes de notre méthode. Pour ce faire, nous proposons une évaluation en deux étapes. Dans un premier temps, nous évaluons les classifieurs binaires construits par expression faciale et par occultation. La deuxième partie de l'évaluation permet une évaluation par occultation après avoir appliqué le mécanisme de fusion.

5.3.1 Données

Nous présentons, dans un premier temps, la base de données utilisée ainsi que les occultations étudiées.

Base de données

Nous avons choisi d'évaluer la méthode sur la base de données CK+ qui est la base de données la plus fréquemment utilisée dans la littérature pour évaluer les méthodes de reconnaissance des expressions faciales en présence d'occultations partielles du visage. De plus, c'est une base de données qui contient des séquences vidéo, ce qui la rend totalement adaptée pour exploiter le mouvement facial. La base de données CK+ est une base de données complètement contrôlée qui contient 374 séquences vidéo annotées. Chaque vidéo commence avec un visage neutre et se termine sur le visage en phase d'apex.

Étant donné que CK+ est une base de données entièrement contrôlée, il n'existe pas d'occultation dans cette base. Nous présentons, dans le paragraphe suivant, les occultations sélectionnées à partir des occultations qui ont le plus d'impact sur la reconnaissance des expressions faciales et qui sont les plus représentatives des occultations proposées dans la littérature.

Simulation des occultations partielles du visage

Comme nous l'avons vu au Chapitre 3, les occultations simulées dans la littérature sont diverses et il est difficile d'établir une base commune permettant de comparer et de se comparer à différentes solutions de l'état-de-l'art. Néanmoins, les régions des yeux ou de la bouche sont des régions fréquemment étudiées dans la littérature. Ces régions sont, en effet, plus couramment occultées en situation réelle mais elles représentent également des zones importantes pour la reconnaissance des expressions faciales.

Pour évaluer notre méthode, nous considérons les occultations les plus étudiées dans la littérature : occultations des yeux et de la bouche illustrées sur la Figure 5.6. Afin de calculer des modèles faciaux robustes à un large panel d'occultations, nous étudions les occultations les plus étendues possibles. Nous proposons, en plus des occultations classiques de la littérature, d'étudier les occultations des parties gauche et droite du visage ainsi que des régions du nez afin de connaître l'impact de tous les éléments du visage.

(a) Pas d'occultation

(b) Droite

(c) Gauche

(d) Bas

(e) Haut

(f) Nez

FIGURE 5.6 – Illustration d'un visage sans occultation et des différentes occultations considérées pour l'évaluation.

Dans les sous-sections suivantes, nous construisons et évaluons les modèles faciaux en nous basant sur les données et les occultations décrites ci-dessus. Dans une première sous-section, nous construisons les modèles faciaux optimaux pour chaque expression faciale et pour chaque occultation. Cette première étude permet d'évaluer les taux de reconnaissance obtenus avec ces modèles faciaux en utilisant une approche par expression. Dans une seconde sous-section 5.3.3, nous évaluons le mécanisme de fusion des modèles des différentes expressions faciales. Nous calculons les taux de reconnaissance obtenus avec un modèle de fusion pour chaque occultation.

5.3.2 Évaluation par expression et par occultation

Nous construisons, dans cette section, les modèles faciaux par expression. Dans cette première partie de l'évaluation, nous commençons par étudier le nombre minimal de régions nécessaires pour reconnaître les différentes expressions faciales en absence d'occultation. Ce nombre minimal de régions sert ensuite de repère pour sélectionner les régions du visage en présence d'occultations.

Répartition des données

Afin d'étudier chaque expression faciale individuellement, nous devons, dans un premier temps, générer des sous-ensembles de données de CK+ par expression. Pour cela, nous sélectionnons, pour chaque sous-ensemble de données, toutes les données de l'expression faciale étudiée pour définir une première classe et, pour la seconde classe, nous sélectionnons aléatoirement une combinaison stratifiée des données des cinq autres expressions faciales. Par exemple, le sous-ensemble de données pour l'expression faciale de joie contient deux classes différentes : la joie et la non-joie. Toutes les données initiales de CK+ annotées "joie" sont sélectionnées et la classe "non-joie" est composée d'un sous-ensemble stratifié des cinq autres expressions.

La répartition des données pour chaque sous-ensemble est détaillée dans le Tableau 5.1.

	Joie	Peur	Surprise	Colère	Dégoût	Tristesse	Total
Sous-ensemble joie	95	19	19	19	19	19	190
Sous-ensemble peur	10	50	10	10	10	10	100
Sous-ensemble surprise	16	16	80	16	16	16	160
Sous-ensemble colère	7	7	7	35	7	7	70
Sous-ensemble dégoût	8	8	8	8	40	8	80
Sous-ensemble tristesse	13	13	13	13	13	65	130

TABLE 5.1 – Répartition des données par expression faciale pour chaque sous-ensemble généré.

Calcul du nombre minimal de régions

Cette première étude permet de définir le nombre minimal de régions permettant de reconnaître les différentes expressions faciales de façon optimale. Pour cela, nous avons généré, pour chaque expression faciale, les 21294 configurations décrites dans les sections précédentes. Après avoir extrait les caractéristiques de mouvement liées à ces configurations, nous les avons évaluées en entraînant des classifieurs SVM avec noyau RBF. Le taux de reconnaissance retenu est calculé avec une validation croisée en 10 plis. Les 25 régions sont triées par ordre d'importance en fonction des poids. À partir de ce tri, différents modèles sont entraînés en sélectionnant un nombre variable de régions du visage : entre une seule région du visage (la plus importante pour chaque expression faciale) jusqu'aux 25 régions.

La Figure 5.7 présente les résultats obtenus en entraînant différents classifieurs avec des nombres variables de régions du visage en fonction du tri par importance pour chaque expression faciale. Les cartes de chaleur correspondantes sont illustrées sur la partie basse de la figure. Ces résultats montrent qu'avec uniquement 6 régions du visage pour chaque expression faciale, les taux de reconnaissance sont quasiment

optimaux. Nous pouvons même constater qu'avec une seule région, les expressions de joie, de surprise et de dégoût sont déjà reconnues de façon quasiment optimale.

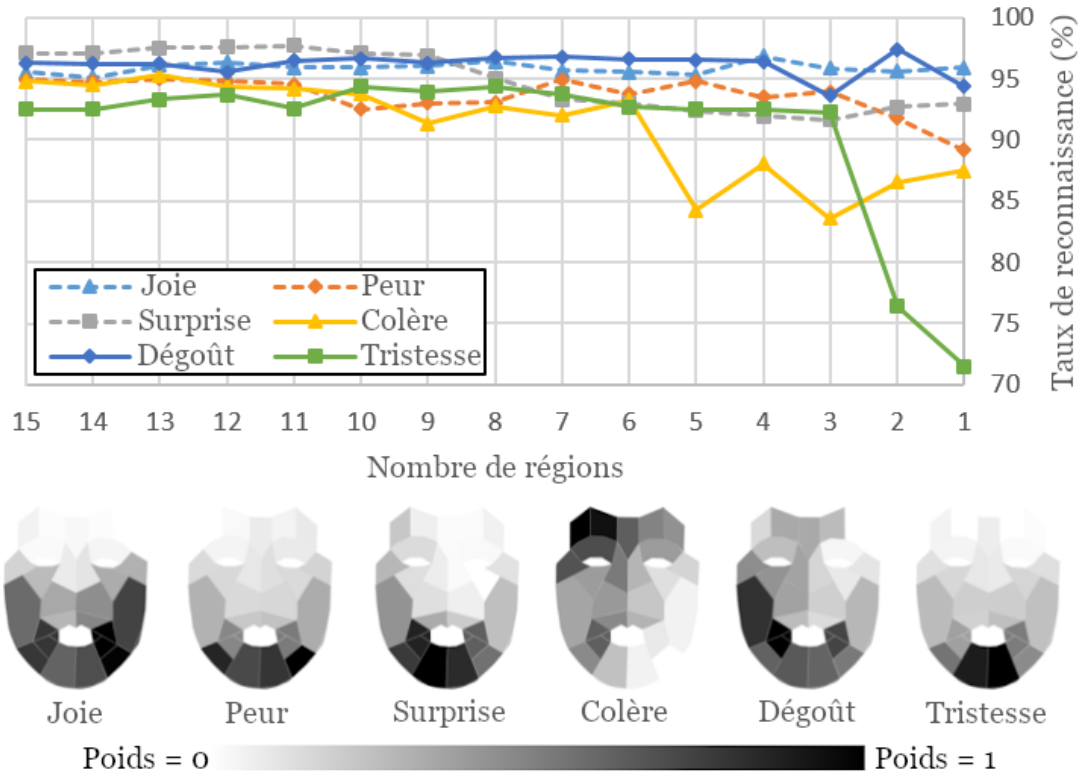


FIGURE 5.7 – Évolution des taux de reconnaissance par expression faciale en fonction du nombre de régions du visage.

La tristesse nécessite au moins 3 régions pour être reconnue, ce qui laisse penser qu'il est nécessaire d'avoir le mouvement du menton ainsi que du coin des lèvres pour reconnaître la tristesse. La colère est l'expression faciale qui semble la plus complexe car elle est la seule à nécessiter réellement 6 régions du visage pour obtenir des taux proches des taux optimaux. En observant la carte de chaleur calculée sur la colère, on note alors qu'en plus des zones du haut du visage, il est nécessaire de voir également le mouvement du coin de la bouche. Cela peut s'expliquer car les mouvements du haut du visage liés à la colère peuvent être confondus avec ceux de la tristesse qui

se caractérisent également par un froncement de sourcils.

Ces observations rappellent les observations faites par Kotsia et al. [89] qui, en étudiant l'impact des occultations sur la reconnaissance des expressions faciales, notaient également une plus grande complexité à reconnaître la colère et soulignaient que la surprise était l'expression faciale la moins confondue avec les autres.

Évaluation de l'approche par expression faciale en présence d'occultations

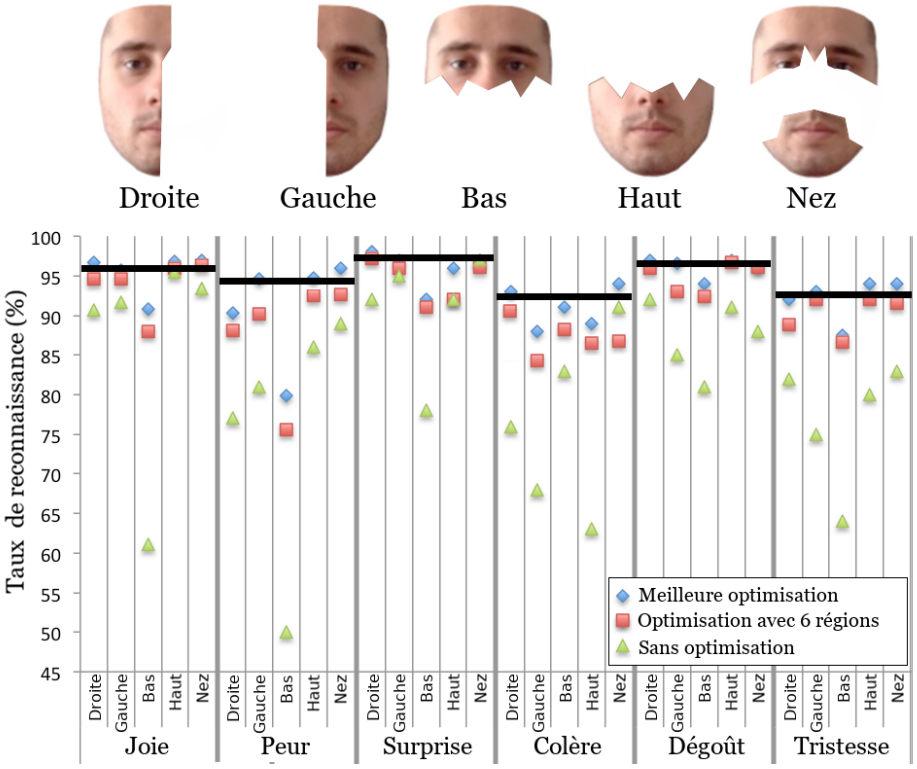


FIGURE 5.8 – Taux de reconnaissance obtenus pour chaque expression faciale en présence de différentes occultations sévères du visage.

Afin de calculer les modèles faciaux adaptés pour les différentes occultations partielles du visage considérées, de nouveaux poids sont calculés en ne conservant que les configurations qui ne contiennent que des régions visibles. Ces nouveaux poids

permettent un nouveau tri des régions du visage non impactées par l'occultation et de nouveaux modèles sont entraînés à reconnaître les expressions faciales en se basant sur un sous-ensemble de régions du visage déterminé par les poids d'importance.

Les résultats obtenus pour chaque occultation étudiée et pour chaque expression faciale sont présentés sur la Figure 5.8.

Sur cette figure :

— les *traits noirs*, pour chaque expression faciale, représentent les résultats obtenus en absence d'occultation : les modèles sont appris sur les 25 régions du visage et les tests sont effectués sur les 25 régions non occultées ;

— les *triangles verts* concernent les résultats obtenus sur un modèle utilisant 25 régions du visage en testant sur des données occultées. Ces résultats permettent de fixer une base de comparaison pour évaluer les améliorations apportées par notre approche ;

— les *losanges bleus* représentent les meilleurs résultats obtenus par notre méthode, avec un nombre de régions variable, dépendant de chaque expérimentation ;

— les *carrés rouges* représentent les résultats obtenus avec notre approche en fixant le nombre de régions du visage à 6, 6 étant le nombre minimal de régions permettant de reconnaître de manière efficace les différentes expressions faciales comme nous l'avons noté dans la sous-section précédente.

L'approche proposée démontre son efficacité en améliorant nettement les résultats en présence d'occultations. On peut également noter que les résultats obtenus avec notre approche se révèlent très proches de ceux obtenus en absence d'occultations.

Concernant l'importance des différentes zones du visage, on remarque l'importance du bas du visage, mise à part pour la colère, pour pouvoir reconnaître les différentes expressions. Les taux de reconnaissance sont, en effet, très bas avec une occultation du bas du visage sans utiliser de modèle facial adapté.On constate, par ailleurs, des améliorations très nettes en utilisant notre méthode. La peur reste, cependant, plus

complexe à reconnaître uniquement en considérant les zones du haut du visage.

Concernant l'importance du haut du visage, on remarque également que les résultats obtenus avec un modèle facial adapté pour l'expression faciale de colère, pour laquelle l'occultation du haut du visage a un impact important, se rapprochent très clairement des résultats optimaux et notre approche permet de très largement améliorer les résultats obtenus en présence d'occultation du haut du visage.

Concernant les occultations gauche et droite du visage, on relève, sans utiliser notre approche, une baisse parfois significative. Les occultations gauche et droite n'ont pas le même impact sur les taux de reconnaissance. Ces observations semblent confirmer l'asymétrie des expressions faciales déjà montrées dans différents travaux [90, 91] et qui s'observe également sur les cartes de chaleur (voir Figure 5.7).

Enfin, les résultats obtenus sans utiliser de modèle facial adapté permettent de souligner une baisse des résultats de reconnaissance en présence d'occultation du nez et des joues, notamment pour les expressions faciales de tristesse et de dégoût. Ces résultats laissent penser que, bien que l'épicentre des mouvements se situe plutôt vers les yeux ou la bouche, les mouvements ont tendance à se propager et la perte de ces zones implique effectivement une perte d'information.

Nous avons présenté jusqu'ici les résultats de la première étape de notre approche en sélectionnant des régions optimales pour chaque expression faciale en présence ou non d'occultations partielles du visage. Cette première étape a permis de fixer le nombre minimal de régions nécessaires pour reconnaître chacune des expressions faciales et a permis de montrer que la sélection des régions permet d'améliorer nettement les résultats de reconnaissance par rapport à une méthodologie sans sélection de région. Ces paramètres étant fixés et validés, nous évaluons, dans la section suivante, le processus complet comprenant l'étape de fusion.

5.3.3 Évaluation par occultation

Dans cette section, nous présentons l'évaluation de notre approche complète en ajoutant l'étape de fusion. Nous évaluons ici l'efficacité de notre approche pour reconnaître les expressions faciales en construisant un seul modèle par occultation, puis nous comparons les résultats obtenus avec ceux de la littérature.

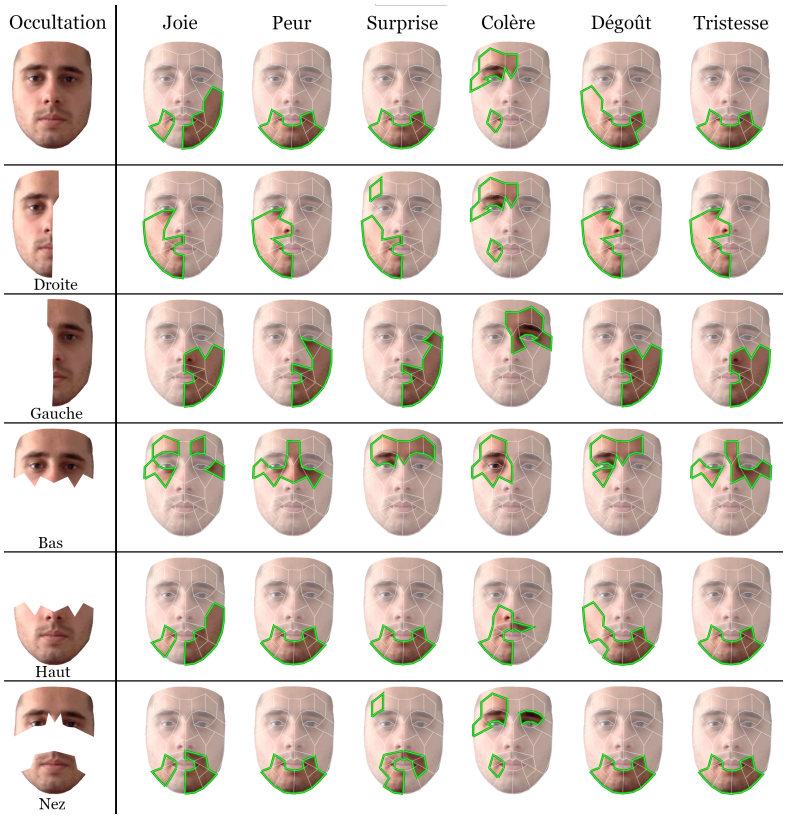


FIGURE 5.9 – Sélection des régions du visage par expression en fonction des occultations étudiées. Le nombre de régions conservées est fixé à 6.

En partant des conclusions des évaluations précédentes, pour chaque expression faciale et pour chaque occultation, nous sélectionnons les 6 meilleures régions (illustrées sur la Figure 5.9). Utiliser 6 régions permet de reconnaître individuellement les expressions faciales de façon efficace.

Répartition des données

Pour évaluer le processus complet nous avons découpé la base de données en deux sous-ensembles, un sous-ensemble permettant d'entraîner les classifieurs binaires et l'autre sous-ensemble permettant d'entraîner le classifieur de fusion. Pour obtenir un compromis entre la taille des deux sous-ensembles, nous avons découpé la base de données en prenant 40% de la base pour entraîner les classifieurs par expression et les 60% de données restantes pour entraîner le classifieur de fusion. Les sous-ensembles par expression sont construits selon la même méthodologie que celle exposée dans la section 5.3.2. Comme pour la reconnaissance par expression, les taux de reconnaissance pour l'étape de fusion sont calculés avec une validation croisée en 10-plis. La répartition des données pour cette évaluation est présentée dans le Tableau 5.2.

	Par expression (40%)	Fusion (60%)	Total
Joie	38	57	95
Peur	21	31	53
Surprise	33	49	83
Colère	14	22	37
Dégoût	16	24	41
Tristesse	26	39	65

TABLE 5.2 – Nombre de séquences vidéo utilisées pour l'évaluation des étapes de la méthode proposée.

Les classifieurs par expression sont alors entraînés sur les 40% de CK+ et sont fixés. Pour ce premier apprentissage, un processus d'optimisation de type *gridsearch* est employé afin de choisir les paramètres du SVM en utilisant une validation croisée en 10 plis. Les modèles par expression utilisés dans la deuxième étape sont appris en utilisant, pour chaque expression, les paramètres retenus par le *gridsearch* et en apprenant sur le

sous-ensemble complet par expression. Ces modèles sont ensuite fixés pour l'apprentissage du classifieur de fusion (SVM avec noyau RBF). Les résultats de reconnaissance suite à la fusion sont calculés grâce à une validation croisée en 10 plis sur les 60% de données restantes.

Afin de minimiser le biais lié au découpage des données 40/60, les résultats présentés sont calculés en constituant 10 répartitions différentes de la base de données CK+ en 10 plis. Les résultats retenus sont alors calculés en moyennant les résultats obtenus sur ces 10 répartitions.

Analyse des résultats

Le Tableau 5.3 récapitule les résultats obtenus avec le processus complet de notre approche sans et avec différentes occultations du visage.

Sans occultation					
91.3%	73.4%	88.8%	89.0%	89.3%	90.7%

TABLE 5.3 – Taux de reconnaissance avec notre approche évaluée sur la base de données CK+ avec et sans occultations.

Ces résultats montrent la robustesse de notre approche face à des occultations importantes du visage. Nous remarquons que, comme le montrait les résultats précédents de la section 5.3.2 (voir Figure 5.8), le bas du visage est une zone critique pour reconnaître les expressions faciales. Les taux de reconnaissance sont, en effet, à leur plus bas niveau lorsque cette zone est occultée.

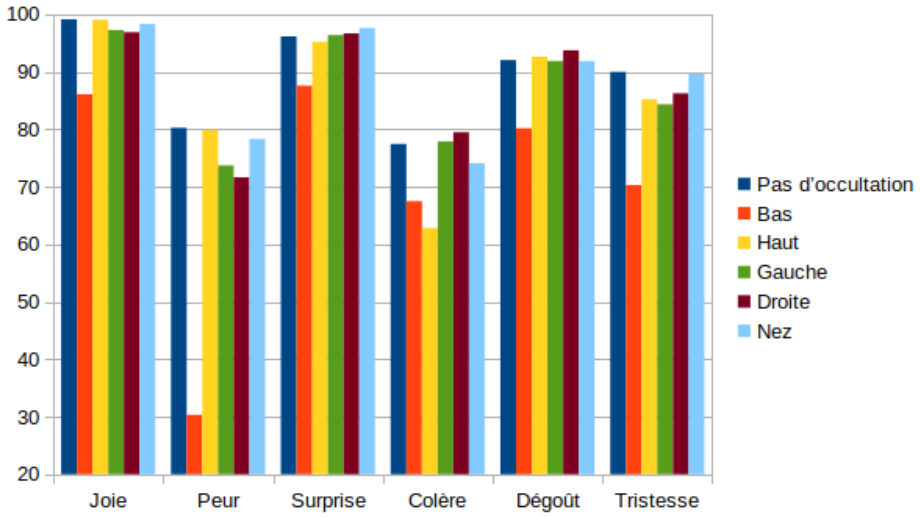


FIGURE 5.10 – Résultats de notre approche évaluée sur la base de données CK+ par expression avec et sans occultation.

Les résultats obtenus par expression sont présentés sur la Figure 5.10. Comme pour les résultats précédents, on voit que l'occultation du bas du visage est l'occultation la plus complexe à maîtriser, notamment pour l'expression faciale de peur. On souligne également que l'expression de colère est plus affectée par l'occultation du haut du visage.

Les occultations du haut et du bas du visage ayant un impact plus fort, nous illustrons, pour rendre compte des effets de ces occultations sur les taux de reconnaissance, les matrices de confusion sur la Figure 5.11. La matrice de confusion de l'occultation du bas du visage montre que la peur est confondue avec plusieurs expressions faciales, ce qui peut laisser penser qu'il y a assez peu de mouvement caractérisant spécifiquement la peur dans le haut du visage. On note également que la peur est surtout confondue avec la tristesse. Cette confusion pourrait s'expliquer par le fait que ces deux expressions s'expriment sur le haut du visage par un froncement de sourcils dont la dynamique est commune aux deux expressions. Concernant l'occultation du haut du visage,

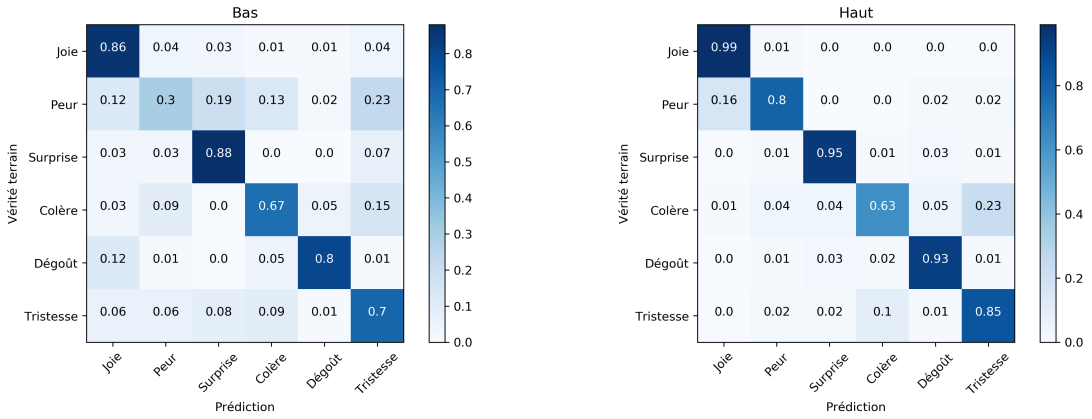


FIGURE 5.11 – Matrices de confusion obtenues en présence d'occultation de la moitié basse et de la moitié haute du visage.

on remarque surtout que la colère est principalement confondue avec la tristesse. Ces deux expressions partagent un mouvement du coin des lèvres vers le bas et nécessitent d'autres informations sur le haut du visage afin d'être départagés.

Comparaison à l'état de l'art

Dans cette partie, nous comparons les scores obtenus aux scores de différentes approches de l'état de l'art par rapport à un certain nombre d'occultations communément rencontrées dans la littérature. Dans le cadre de notre approche, nous proposons six modèles permettant d'être robuste à un grand spectre d'occultations. Dans la littérature, les résultats sont fréquemment proposés dans le cadre d'occultation de la bouche ou des yeux. Afin de s'approcher davantage des occultations simulées dans la littérature, nous ajoutons l'occultation de la bouche en conservant les joues visibles.

Nous comparons alors les modèles correspondant à ces occultations avec les résultats de l'état de l'art. Ces résultats sont présentés sur la Figure 5.12. Les résultats obtenus par l'approche proposée sont comparés aux résultats obtenus par Kotsia et al. [89], Ranzato et al. [111], Huang et al. [96] et Dapogny et al. [97].

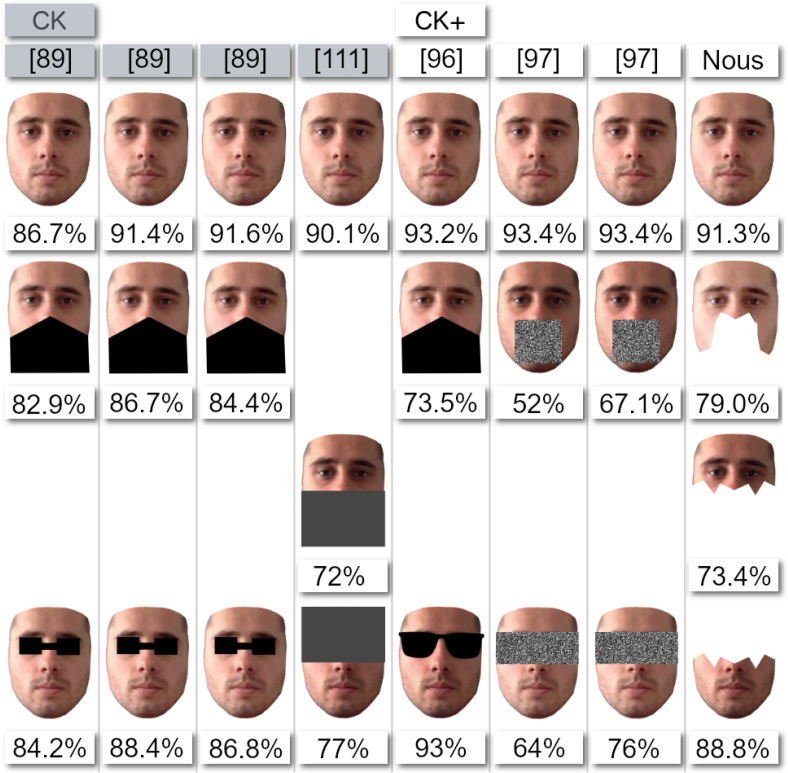


FIGURE 5.12 – Comparaison de nos résultats avec les résultats obtenus par d'autres approche de la littérature.

Kotsia et al. étudient l'impact des occultations sur la reconnaissance des expressions faciales en utilisant différents outils : l'algorithme DNMF (première colonne), un descripteur basé sur du suivi de points caractéristiques (deuxième colonne) et des filtres de Gabor (troisième colonne).

Ranzato et al. (colonne 4) proposent un algorithme génératif pour reconstruire les zones occultées du visage. Huang et al. (colonne 5) proposent une détection de l'occultation grâce à une représentation éparse, une analyse basée sur un découpage en différentes régions du visage et une fusion de caractéristiques.

Enfin, Dapogny et al. [97] ont proposé des descripteurs locaux basés sur une forêt d'arbres décisionnels calculés sur différentes régions du visage avant d'appliquer une fusion. Dans ces travaux, les auteurs proposent également un auto-encodeur entraîné à définir un poids de confiance sur les différentes régions qui permet de représenter un poids d'occultation de la région. Dans ces travaux, ils proposent alors de présenter les résultats en pondérant avec les poids calculés ainsi que sans pondération. Les résultats présentés sur la Figure 5.12 montrent, dans la sixième colonne les résultats sans pondération obtenus par Dapogny et al. et, sur la septième colonne, les résultats avec pondération.

On note les résultats très compétitifs de notre approche par rapport aux résultats des méthodes de l'état de l'art malgré des occultations plus importantes. La différence importante des résultats obtenus sur l'occultation du bas du visage en incluant et en excluant les joues montrent qu'une grande partie de l'information liée aux expressions faciales se retrouve dans ces zones alors que l'épicentre du mouvement principal est plutôt situé au niveau de la bouche. Cette différence tend à montrer également l'intérêt de tenir compte de la propagation du mouvement. Bien que la propagation résulte dans des mouvements de moindre intensités, ces derniers conservent des informations importantes.

Application de l'approche sur des occultations réelles

Les résultats de notre approche sur CK+ montrent l'intérêt de la méthode en permettant d'obtenir des résultats compétitifs par rapport à l'état de l'art. Afin d'avoir un aperçu qualitatif sur la capacité de notre approche à travailler sur des données réelles, nous appliquons notre méthode sur des données captées par nos soins. Nous avons capturé des séquences vidéo avec des occultations statiques générées par des lunettes de soleil ou une main sur la bouche. Pour l'analyse, nous récupérons les modèles entraînées sur les données de CK+ comme indiqué dans les sections précédentes et nous les appliquons sur ces nouvelles données. La Figure 5.13 montre les résultats obtenus. Sur cette figure sont représentées les visualisations HSV des flux optiques des zones sélectionnées par notre approche sur les différentes données ainsi que les probabilités de sortie des classifieurs binaires de chaque expression faciale. En dernière colonne de cette figure, nous indiquons la prédiction de notre classifieur.

5.4 Conclusion

Dans ce chapitre, nous avons exploité la propriété de propagation de mouvement en proposant une méthode qui se concentre sur les régions visibles du visage. Pour cela, nous avons proposé un processus pour déterminer les poids d'importance des différentes régions du visage. Étant donné que les zones de plus haute importance dépendent fortement de chaque expression faciale, nous avons étudié chaque expression faciale indépendamment afin de construire des modèles faciaux optimaux par expression. Ces modèles faciaux ont été optimisés de façon à sélectionner un nombre minimal de régions du visage. En sélectionnant un nombre de régions les plus petit possible, nous construisons ainsi des modèles robustes à un grand spectre d'occultations qui peuvent concerner toutes les régions non utilisées dans le modèle. Ces modèles faciaux

FIGURE 5.13 – Évaluation qualitative des résultats obtenus avec notre approche sur des données en présence d'occultations réelles.

ont été calculés sans et avec occultations partielles du visage et permettent d'entraîner des classifieurs binaires pour reconnaître individuellement les différentes expressions faciales. Afin d'obtenir une classification finale unifiée, l'approche proposée a été complétée par une fusion tardive.

Nous notons des résultats très compétitifs en comparaison à ceux de l'état de l'art en dépit des occultations du visage très sévères concernant les zones importantes du visage. Notre approche permet, en effet, d'obtenir des taux de reconnaissance de 88.8% en présence d'une occultation du haut du visage. Notre approche atteint 79.0% en présence d'une occultation de la bouche et 73.4% avec une occultation de l'ensemble des régions du bas du visage. Enfin, les résultats qualitatifs proposés sur des occultations réelles permettent de démontrer la généralisation de notre approche.

Chapitre 6

Reconstruction des flux optiques

Contents

Au chapitre précédent nous avons exploré la propriété de propagation du mouvement en proposant une solution qui se concentre sur les régions visibles du visage. Cette première solution permet de réduire considérablement les effets des occultations sur les taux de reconnaissance. Cependant, elle nécessite d'entraîner un classifieur spécifique adapté aux différentes occultations. Nous proposons, dans ce chapitre, une solution basée sur une reconstruction afin de ramener les visages occultés dans des conditions d'analyse idéale en compensant les effets des occultations. Cette reconstruction permet de ne conserver alors qu'un classifieur unique. Dans ce chapitre, nous proposons d'exploiter la propriété de similarité de mouvement en proposant de reconstruire les informations bruitées dues à une occultation directement dans le domaine du mouvement.

6.1 Introduction

Pour reconstruire les données cachées en exploitant la propriété de similarité, nous proposons une nouvelle approche qui reconstruit le flux optique calculé à partir de séquences vidéo de visages occultés. L'approche proposée, illustrée sur la Figure 6.1, se base sur une architecture d'auto-encodeur débruitant pour reconstruire les flux optiques calculés sur des données occultées. Un classifieur entraîné à reconnaître les expressions faciales à partir de flux optiques de données non occultées prend ensuite en entrée les flux optiques reconstruits pour l'étape de classification.

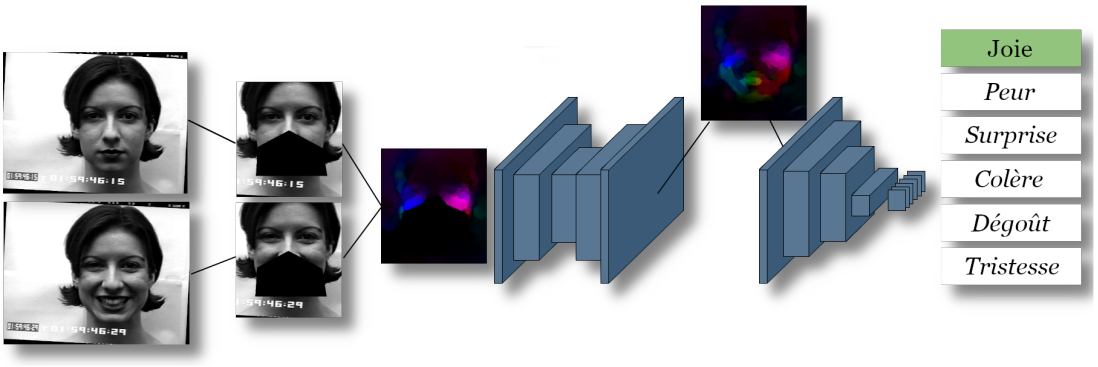


FIGURE 6.1 – Processus complet de la méthode de reconstruction de flux optique proposée.

Dans la suite de ce chapitre, nous détaillons les différentes étapes de notre approche. La section 6.2 décrit, dans un premier temps, les étapes de préparation des données nécessaires pour apprendre le modèle de reconstruction. La section 6.3 détaille ensuite notre méthode de reconstruction en exposant l'architecture de l'auto-encodeur utilisée ainsi que la fonction de coût associée pour l'apprentissage. La section 6.4 présente le protocole d'évaluation ainsi que différentes expérimentations menées pour optimiser les méta-paramètres de notre approche. Notre approche permet d'obtenir des résultats très compétitifs par rapport à ceux de l'état de l'art, notamment dans le cadre d'une occultation de la bouche qui est une occultation qui a un impact important sur les données de CK+.

6.2 Préparation des données

L'apprentissage d'un auto-encodeur débruitant nécessite d'avoir la vérité terrain du flux optique calculé sur des données non occultées ainsi que le flux optique calculé sur ces mêmes images occultées. Cette section décrit la méthodologie utilisée pour l'occultation des données ainsi que le calcul du flux optique normalisé afin d'obtenir

une dimension du flux optique adaptée à la première couche de l'architecture de l'auto-encodeur.

6.2.1 Génération des occultations

Afin de proposer des occultations des régions les plus importantes pour la reconnaissance des expressions faciales, nous proposons différentes occultations des yeux et de la bouche comme l'illustre la Figure 6.2. Les yeux et la bouche sont des régions souvent occultées pour évaluer des méthodes proposées pour répondre à la problématique des occultations. Ces occultations sont simulées en ajoutant des boites noires statiques sur toutes les images d'une séquences vidéo.

FIGURE 6.2 – Occultations sélectionnées pour évaluer notre approche, appliquées sur la base de données CK+.

6.2.2 Calcul du flux optique

Etant données les bases de données disponibles dans la littérature, nous ne disposons actuellement pas de grands volumes de données pour évaluer notre approche. Par conséquent, il s'avère alors nécessaire de limiter le nombre de paramètres à apprendre en passant par des architectures faiblement profondes. En complément, nous travaillons également sur des flux optiques de taille réduite en entrée du système. Afin de préserver autant que possible la qualité du calcul de flux, nous calculons le flux optique sur les images en haute résolution avant d'y appliquer un processus de réduction

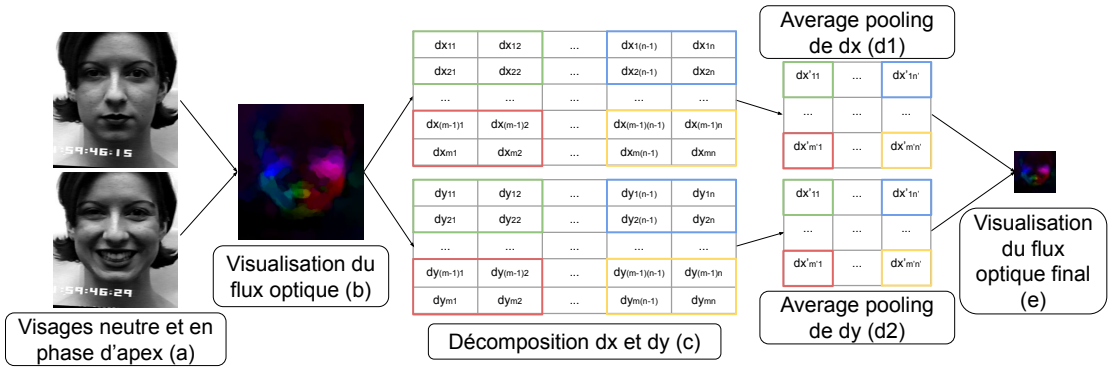


FIGURE 6.3 – Processus proposé pour calculer et réduire la taille des flux optiques calculés pour les utiliser directement dans le cadre de notre méthodologie.

permettant d'obtenir une taille standard qui permet de satisfaire les critères d'entrées normalisées de l'auto-encodeur.

Pour réduire le flux optique, nous proposons une réduction inspirée des méthodes de pooling. Le calcul du flux optique est effectué en trois étapes illustrées sur la Figure 6.3. Les images initiales en début de ce processus sont rognées pour ne conserver que la zone de l'image contenant le visage. Le rognage de l'image est calculé à partir de la position des yeux et de la distance inter-oculaire. Le flux optique est ensuite calculé sur ces images rognées avec la méthode de Farnebäck [53]. Enfin, afin de normaliser les tailles de flux optique et pour réduire les coûts de calculs et le nombre de paramètres dans l'architecture neuronale, le flux optique est réduit sur les dimensions x et y. Pour cette réduction, des nouvelles valeurs sont calculées à partir de fenêtres glissantes. Les nouvelles valeurs sont alors calculées grâce à la formule suivante :

$$resize(OF, i, j) = (\mu(OF[\lfloor dx * i \rfloor, ..., \lfloor dx * (i+1) + \lceil dx \rceil - 1 \rfloor])/dx, \\ \mu(OF[\lfloor dy * j \rfloor, ..., \lfloor dy * (j+1) + \lceil dy \rceil - 1 \rfloor])/dy)$$

avec dx et dy les coefficients de réduction entre la taille initiale et la taille finale et μ le calcul de la moyenne.

Afin de tenir compte du nombre limité de données, nous choisissons une taille finale de flux offrant un bon compromis entre performance de reconnaissance et efficacité de calcul. Ces expérimentations sont présentées dans la sous-section 6.4.2.

6.3 Reconstruction du flux optique

Notre approche se base sur une architecture d'auto-encodeur débruitant pour reconstruire le flux optique calculé sur des données occultées. Les flux optiques calculés par la méthodologie décrite dans la section précédente sont alors directement utilisés en entrée de l'architecture. Dans cette section, nous décrivons, dans un premier temps, l'architecture de l'auto-encodeur utilisée pour la reconstruction. Dans un second temps, nous discutons des différentes fonctions de coût qui nous semblent adaptées pour entraîner l'auto-encodeur.

6.3.1 Architecture

L'architecture de l'auto-encodeur utilisée dans le cadre de notre approche est inspirée par les architectures Hourglass et U-Net [113, 114], qui proposent des architectures d'auto-encodeurs symétriques avec des skip connexions. L'architecture U-Net a, par ailleurs, été spécifiquement proposée pour apprendre sur relativement peu de données. L'architecture proposée est composée de couches successives de convolutions avec des noyaux de taille 3x3 qui est la taille minimale possible d'un noyau de convolution permettant de caractériser les changements spatiaux. Étant donnée la taille réduite des entrées, nous nous limitons à une taille de 3x3 afin de caractériser les changements locaux. Ces convolutions sont suivies d'une couche de ReLu et de max-pooling pour l'encodeur et de couches successives de convolutions et de up-sampling pour le décodeur. Les dimensions x et y du flux optique représentent les déplacements en x et en

y de chaque point et peuvent alors prendre des valeurs négatives selon la direction de ces points. La dernière convolution n'est donc pas suivie d'une couche de ReLu afin de permettre la reconstruction d'un flux optique qui peut contenir des éléments négatifs.

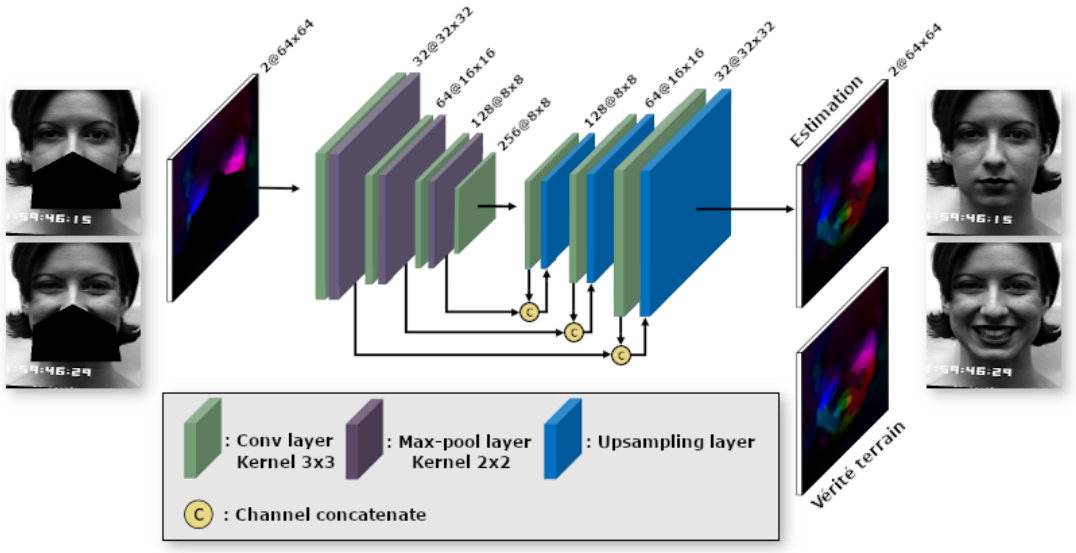


FIGURE 6.4 – Illustration de l'auto-encodeur de reconstruction qui prend en entrée des flux optiques bruités par la présence d'occultations sur les images d'origine et qui calcule une reconstruction de ces flux optiques. La fonction de coût est calculée en comparant la reconstruction obtenue avec la vérité terrain, obtenue en calculant les flux optiques sur les données en absence d'occultation.

Comme l'illustre la Figure 6.4, l'auto-encodeur prend, en entrée, les flux optiques calculés entre deux images occultées et calcule, en sortie, la reconstruction.

Contrairement à l'Hourglass qui additionne les sorties des couches, l'architecture U-Net propose plutôt de les concaténer, ce qui permet de réutiliser des caractéristiques de plus haut niveau. Nous proposons plutôt ce type de connexion qui semble plus adapté pour reconstruire plus finement le flux optique en permettant de réutiliser les caractéristiques provenant des régions non occultées notamment.

La vérité terrain utilisée pour entraîner l'auto-encodeur correspond au calcul du flux optique entre les images de la donnée initiale sans occultation. L'auto-encodeur apprend ainsi à reconstruire un flux optique qui s'approche du flux optique calculé sur des images non occultées.

6.3.2 Fonction de coût

La fonction de coût permet de comparer la reconstruction par l'auto-encodeur avec la vérité terrain. Différentes fonctions de coût ont été considérées afin de trouver la fonction la plus appropriée pour préserver les informations liées aux expressions faciales.

Nous considérons dans les formules suivantes deux ensembles de flux optiques U et V (prédiction/vérité terrain) contenant n flux optiques. Chaque flux optique est composé des déplacements en x et en y.

MSE : l'erreur quadratique moyenne est une fonction de coût classique. Elle calcule le carré de la distance euclidienne entre la reconstruction et la vérité terrain en chaque point.

$$MSE(U,V) = \frac{1}{2n}\sum_{i=1}^{n}(u_{ix} - v_{ix})^2 + (u_{iy} - v_{iy})^2 \tag{6.1}$$

Wing : initialement proposée pour localiser les points caractéristiques [115], la fonction de coût *wing* pénalise davantage les erreurs moyennes et petites en utilisant une erreur basée sur un logarithme dans le cas d'erreurs inférieures à un certain seuil. Grâce à cette adaptation, les données pour lesquelles l'erreur est assez faible ont encore un poids assez important dans l'apprentissage, ce qui permet d'affiner la prédiction. Nous proposons d'étudier cette fonction de coût car elle pourrait permettre une reconstruction plus détaillée du flux optique.

$$loss(U,V) = \frac{1}{2n}\sum_{i=1}^{n} wing(u_i - v_i) \tag{6.2}$$

$$wing(u,v) = \begin{cases} \omega ln(1 + |u-v|/\epsilon), & \text{si } |u-v| < \omega \\ |u-v| - C, & \text{sinon} \end{cases} \tag{6.3}$$

où ω définit la partie non linéaire, ϵ la courbure de la fonction et C une constante de lissage entre les parties linéaires et non linéaires.

Endpoint : nous étudions également l'erreur endpoint proposée pour évaluer les calculs de flux optiques [116] qui analyse l'erreur de calcul de flux optique en chaque point dans les directions x et y en calculant les distances euclidiennes.

$$endpoint(U,V) = \frac{1}{2n}\sum_{i=1}^{n} \sqrt{(u_{ix} - v_{ix})^2 + (u_{iy} - v_{iy})^2} \tag{6.4}$$

Ces différentes métriques présentent des atouts pour reconstruire le flux optique et nous proposons de mesurer l'intérêt de chacune au sein de notre méthode. Ces fonctions de coût sont évaluées dans la sous-section 6.4.3 de la section d'évaluation 6.4.

6.4 Évaluation

Nous évaluons notre approche en nous concentrant sur la capacité à reconnaître les expressions faciales sur les données reconstruites et non pas sur la précision de la reconstruction.

Nous définissons tout d'abord le protocole expérimental en sélectionnant la base de données et le classifieur utilisé pour mesurer l'impact de notre reconstruction. Nous proposons, ensuite, une optimisation des différents méta-paramètres ainsi qu'une évaluation en plusieurs étapes.

Dans un premier temps, nous proposons une analyse de la dimension de sortie de la réduction du flux optique afin d'optimiser la reconnaissance des expressions faciales en tout minimisant la complexité du réseau. Pour ce faire, nous étudions les résultats de reconnaissance des expressions faciales avec l'architecture CNN proposée en prenant en entrée différentes tailles de flux optiques. La dimension du flux optique optimal ainsi déterminé est fixé pour la suite de l'évaluation.

Dans un second temps, nous étudions les différents méta-paramètres liés à la reconstruction du flux optique afin d'optimiser notre processus. Dans cette seconde partie, nous évaluons d'abord l'impact de la fonction de coût utilisée afin d'utiliser la fonction la plus adaptée à notre tâche. Nous étudions, ensuite, l'impact des skip connexions pour optimiser l'architecture proposée. Nous étudions, ensuite, différentes stratégies pour calculer les flux optiques à partir des séquences vidéo utilisées pour l'apprentissage de l'auto-encodeur. Ces stratégies permettent d'étudier l'impact des données d'apprentissage sur la reconstruction lorsque l'on ajoute différentes intensités de mouvements et que l'on augmente le nombre de données d'apprentissage.

Dans un troisième temps, nous évaluons la capacité de généralisation de notre approche à d'autres classifieurs. Nous utilisons alors la paramétrisation optimale calculée à partir du CNN et nous évaluons les taux de reconnaissance des expressions faciales à partir des flux reconstruits en utilisant un SVM.

En dernier, à partir du processus optimisé, nous comparons, enfin, les résultats obtenus avec ceux de l'état de l'art.

Dans les sections qui suivent, nous exposons dans un premier temps le protocole expérimental utilisé pour évaluer notre méthode puis nous étudions les différents méta-paramètres. Enfin, nous comparons les résultats de notre approche avec ceux de l'état de l'art.

6.4.1 Protocole expérimental

La première étape pour mettre en place un protocole expérimental clair est de définir la base de données utilisée. Comme nous l'avons vu au Chapitre 3, la base de données la plus fréquemment utilisée dans la littérature pour évaluer les méthodes de reconnaissance en présence d'occultations partielles du visage est la base de données CK+. CK+ est une base de données complètement adaptée, d'une part, pour étudier les occultations car c'est une base de données entièrement contrôlée et, d'autre part, complètement adaptée à notre proposition car c'est une base de donnée dynamique qui nous permet alors facilement de calculer des flux optiques. Choisir une base de données entièrement contrôlée permet, en effet, de se concentrer sur la problématique des occultations car c'est une base de données qui ne comporte pas d'autre défi particulier. De plus, l'absence d'occultation réelle sur la base de données et le fait de simuler les occultations permet de plus facilement quantifier l'impact des occultations sur le processus de reconnaissance. Enfin, cela permet également de quantifier le gain obtenu par une méthode proposée par rapport aux résultats obtenus en présence d'occultation mais aussi de quantifier la perte de performance par rapport aux résultats sans occultation.

Protocole expérimental pour la reconnaissance automatique des expressions faciales

Pour reconnaître automatiquement les expressions faciales à partir de flux optique, une solution a été proposée par Allaert et al. [58]. Nous proposons alors de reprendre l'architecture proposée par Allaert et al. car c'est une architecture qui a déjà montré son efficacité dans un cadre de reconnaissance des expressions faciales en utilisant un apprentissage basé sur du flux optique.

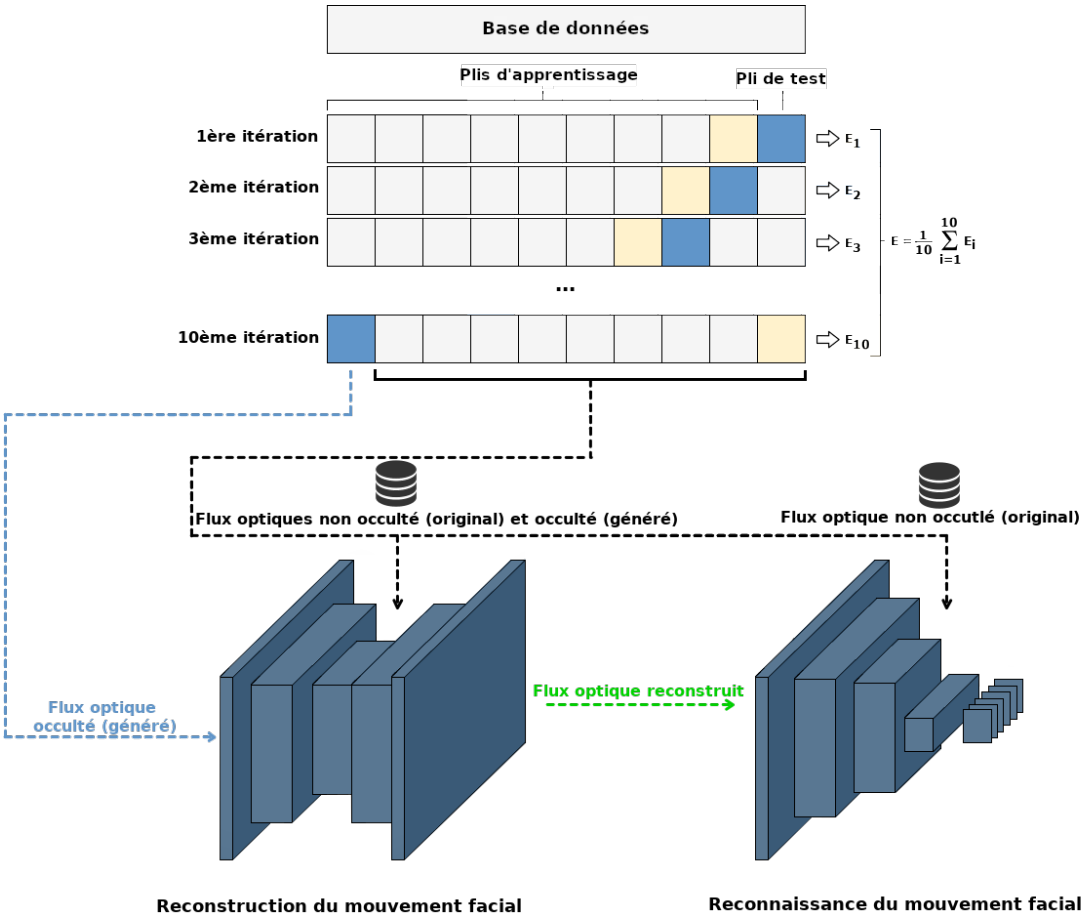


FIGURE 6.5 – Évaluation de l'approche avec un protocole expérimental en validation croisée en 10 plis. Les résultats sont obtenus en testant successivement sur chaque pli, en validant sur un deuxième pli et en apprenant sur les 8 plis restants.

Afin d'entraîner ce CNN, chaque pli est évalué successivement en entraînant le CNN sur 8 plis, en validant sur le 9ème pli et en testant sur le 10ème comme l'illustre la Figure 6.5. Les taux de reconnaissance des expressions faciales présentés représentent alors la moyenne des taux de reconnaissance sur les 10 plis de test successifs. Selon l'expérimentation, le pli de test pourra être constitué :

- des flux optiques originaux pour calculer les taux de reconnaissance sans occultation
- des flux optiques corrompus par l'occultation sur les images originales pour calculer les taux de reconnaissance en présence d'occultations sans utiliser la méthode
- des flux optiques reconstruits grâce à notre approche pour évaluer notre méthode et situer cette méthode par rapport aux résultats obtenus sans occultation et sans reconstruction.

Protocole expérimental pour l'étape de reconstruction

Le protocole expérimental utilisé pour une expérimentation lors de reconstruction est le même que celui utilisé pour la reconnaissance. Nous utilisons alors, pour une même expérimentation, les mêmes plis d'apprentissage, de validation et de test pour garantir une totale impartialité. Ainsi, on s'assure que les données de test utilisés pour la reconstruction n'ont jamais été rencontrées lors de l'apprentissage du classifieur.

De la même façon que pour la reconnaissance, chaque pli est évalué successivement en ayant préalablement calculé les flux optiques sur les données occultées.

6.4.2 Paramétrisation du processus de reconnaissance

Nous proposons, dans le cadre de l'évaluation de notre approche, d'étudier chaque paramètre successivement afin de les fixer au fur et à mesure de notre évaluation. Afin

d'optimiser dans un premier temps l'étape de classification, nous étudions la dimension du flux optique dans le but de trouver un compromis qui permette d'obtenir des scores de reconnaissance satisfaisants tout en minimisant la complexité. Une taille de flux optique plus grande augmente, en effet, le nombre de paramètres de notre réseau, ce qui augmente la complexité de son apprentissage.

En plus de fixer la taille des flux optiques lors de leur normalisation, cette première évaluation permet également de calculer un score de reconnaissance sans occultation qui servira de base de comparaison. Nous calculerons également les résultats de reconnaissance en présence de différentes occultations, ce qui permettra de comparer notre approche à ces deux scores.

Dans cette première partie, nous étudions différentes tailles de flux optiques lors de la normalisation du flux. Afin d'étudier un large panel de tailles, nous étudions les dimensions finales suivantes : 24x24, 48x48, 64x64, 96x96 et 128x128. Le protocole expérimental permettant de récupérer les résultats est le protocole détaillé dans les sections précédentes. Pour minimiser le biais lié à l'initialisation aléatoire du réseau, nous évaluons chaque taille sur 100 graines aléatoires différentes.

Les résultats obtenus sur ces 100 graines pour les différentes tailles sont présentés sur la Figure 6.6. Les résultats présentés sont calculés en effectuant la moyenne et la médiane des scores obtenus sur les 100 graines. Comme le montre la figure de gauche, les tailles optimales qui permettent d'obtenir les meilleurs scores sont les tailles 64x64 et 96x96. Ce graphique montre également qu'une taille supérieure à 96x96 ne semble pas permettre d'obtenir de meilleurs résultats, ce qui peut s'expliquer par une complexité plus forte avec une taille plus grande mais qui permet également de constater que l'information pertinente est déjà présente dans une taille inférieure. Les tailles 64x64 et 96x96 offrant des scores similaires, nous proposons de fixer la taille à 64x64 qui permet une complexité un peu plus faible.

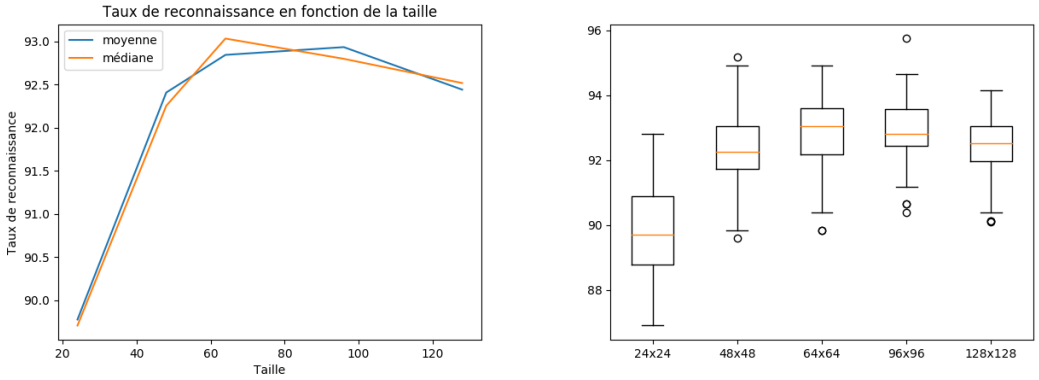


FIGURE 6.6 – Résultats obtenus en fonction des tailles des flux optiques utilisés en entrée du CNN de reconnaissance. Le graphique de gauche présente les résultats médians et moyens obtenus en moyennant sur 100 graines pour les tailles : 24x24, 48x48, 64x64, 96x96 et 128x128. Le graphique de droite présente les boites à moustaches correspondants. Ces graphiques montrent qu'une taille de 64x64 semble un compromis pertinent avec cette architecture en permettant un taux de reconnaissance optimal.

Nous fixons la dimension du flux optique réduit à 64x64 et nous calculons les résultats obtenus sans et avec les occultations étudiées afin d'avoir des premiers éléments comparatifs de notre approche. Pour calculer ces scores, nous fixons la graine afin d'obtenir un taux de reconnaissance moyen pour une taille de 64x64. Les scores sont obtenus selon le protocole expérimental détaillé dans les sections précédentes. Ces scores sont présentés sur le Tableau 6.1 et permettent un premier point de repère.

			
92.8%	73.8%	71.1%	46.8%

TABLE 6.1 – Base de comparaison dont les scores sont obtenus sans utiliser notre approche. Ces résultats sont obtenus avec l'architecture CNN proposée en entraînant sur des visages non occultés et en testant d'une part, sans occultation et, d'autre part, en présence de différentes occultations partielles du visage.

6.4.3 Paramétrisation de l'auto-encodeur de reconstruction

Dans la sous-section précédente, la taille du flux optique redimensionné ainsi que la graine utilisé pour l'architecture CNN de reconnaissance des expressions faciales ont été fixées. Cette première évaluation a permis d'obtenir un point de repère pour les expérimentations suivantes. Dans cette sous-section l'architecture de reconstruction est évaluée afin d'optimiser la fonction de coût utilisée, l'architecture utilisée en fonction des connexions skip ainsi que la stratégie de calcul des flux à partir des séquences vidéo.

Évaluation de la méthode proposée en fonction de la fonction de coût utilisée

Différentes fonctions de coût ont été proposées, chacune présentant des intérêts différents. Dans cette partie, les différentes fonctions de coût sont évaluées afin de déterminer la plus adéquate. Pour rappel, les fonctions de coût proposées sont : l'erreur quadratique moyenne (*MSE*) qui correspond à une fonction de coût classique en apprentissage automatique, l'erreur *wing* qui permet de pénaliser davantage des petites et moyennes erreurs et pourrait ainsi permettre une reconstruction plus fine et l'erreur *endpoint* qui est une fonction de coût classique pour comparer des flux optiques.

Pour évaluer ces différentes fonctions de coût, le protocole expérimental détaillé dans les sections précédentes est appliqué avec la base de données CK+. Pour cette première évaluation, les flux optiques utilisés sont des flux optiques calculés entre la première (neutre) et la dernière image (apex) de chaque séquence de CK+.

			
MSE	86.2%	74.0%	67.3%
Wing	86.1%	79.5%	69.0%
EndPoint	**87.1%**	**80.1%**	**70.2%**

TABLE 6.2 – Résultats obtenus avec notre approche en fonction de la fonction de coût utilisée pour la rétropropagation de l'architecture auto-encodeur de reconstruction.

				Gain moyen
MSE	+12.4%	+2.9%	+20.5%	+11.9%
Wing	+12.3%	+8.4%	+22.2%	+14.3%
EndPoint	**+13.3%**	**+9.0%**	**+23.4%**	**+15.2%**

TABLE 6.3 – Gains obtenus avec notre approche en fonction de la fonction de coût utilisée. Les gains sont obtenus en calculant la différence entre les résultats obtenus avec notre approche (présentés sur le Tableau 6.2) avec les résultats obtenus sans utiliser de reconstruction (présentés sur le Tableau 6.1.

Le Tableau 6.2 présente les résultats obtenus après reconstruction des différentes occultations étudiées en utilisant les différentes fonctions de coût. Le Tableau 6.3 présente le gain obtenu en comparant les résultats du Tableau 6.2 avec les résultats obtenus sans utiliser de reconstruction. Ces premiers résultats permettent, tout d'abord, de

montrer l'efficacité de la méthode proposée avec un gain significatif par rapport aux résultats sans reconstruction. En comparant les résultats obtenus avec les différentes fonctions de coût, on peut observer que la fonction de coût *endpoint* est clairement la plus adaptée à notre tâche avec des résultats de reconnaissance supérieurs pour les différentes occultations étudiées. En observant le gain moyen des différentes fonctions de coût, on peut noter que les fonctions *wing* et *endpoint* s'avèrent être plus pertinentes que la *MSE* pour reconstruire une données qui permet de reconnaître les expressions faciales. Cette observation peut s'expliquer par le fait que la *wing* semble, comme attendu, reconstruire de façon plus fine les flux optiques. La fonction *endpoint*, quant à elle, est spécifique au flux optique ce qui peut expliquer la cohérence évidente entre cette fonction de coût et notre approche.

À la vue de ces résultats, la fonction de coût *endpoint* semble complètement adaptée et nous fixons cette fonction de coût pour la suite des expérimentations.

Évaluation de la méthode proposée en fonction des skip connexions ajoutées à l'architecture de l'auto-encodeur de reconstruction

Afin d'obtenir des reconstructions plus fines, nous nous inspirons encore de l'architecture Hourglass et ajoutons des connexions entre les couches de l'encodeur et celles du décodeur. Ces connexions permettent alors de récupérer, au niveau du décodeur, des caractéristiques de plus haut niveau présentes au niveau de l'encodeur. Les différentes connexions étudiées sont illustrées sur la Figure 6.7.

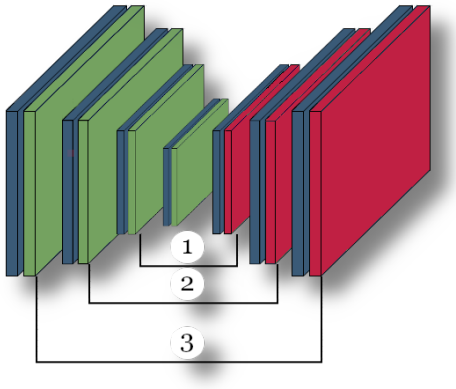


FIGURE 6.7 – Illustration des différentes connexions étudiées.

			
/	87.1%	80.1%	70.2%
1	88.4%	81.1%	71.1%
2	89.3%	81.8%	71.9%
3	89.4%	81.0%	70.8%
1+2	88.9%	82.6%	71.4%
1+3	89.3%	82.5%	71.5%
2+3	89.5%	82.6%	72.1%
1+2+3	**89.6%**	**83.1%**	**72.5%**

TABLE 6.4 – Taux de reconnaissance obtenus en fonction de différentes connexions ajoutées sur l'architecture de reconstruction.

Le Tableau 6.4 résume l'ensemble des résultats obtenus en ajoutant différentes connexions sur l'auto-encodeur de reconstruction. Les résultats de ce tableau montrent que l'ajout de ces connexions permettent de récupérer de l'information utile pour la reconnaissance des expressions faciales. Ces résultats permettent également de noter que la meilleure architecture est celle qui contient des connexions résiduelles entre toutes

les couches cachées du réseau. Pour la suite des expérimentations, l'architecture est alors fixée en ajoutant des connexions résiduelles sur les trois couches cachées.

Évaluation de la méthode proposée en fonction des stratégies de calcul de flux optiques

Dans les sections précédentes, les flux optiques utilisés pour l'apprentissage de l'auto-encodeur étaient uniquement les flux optiques calculés entre la première (image neutre) et la dernière image (image d'apex) de chaque séquence vidéo.

Afin d'augmenter le nombre de données mais aussi d'élargir le panel des intensités d'expressions, différentes stratégies sont proposées. Quatre stratégies de calcul des flux sont proposées, résumées sur la Figure 6.8 :

- apex : la stratégie apex calcule uniquement le flux optique entre la première et la dernière image de la séquence
- trois dernières images : la stratégie des trois dernières images calcule les flux optiques entre la première image et les trois dernières images de la séquence
- flux successifs : calcule les flux optiques entre les différentes images successives
- mid-séquence : est la stratégie qui permet d'obtenir le plus grand nombre de flux optiques tout en conservant un mouvement minimal. Les flux optiques sont calculés entre toutes les paires d'images $prvs$ et nxt tels que $prvs < nxt$ et $nxt >= n/2$ avec n le nombre d'images de la séquence. Les expressions faciales étant activées en général à partir du milieu de la séquence sur la base de données CK+, cette stratégie permet d'éviter de calculer des flux optiques entre deux images où il n'y a pas ou très peu de mouvement. La stragégie permet également d'obtenir un large panel d'intensités d'expressions.

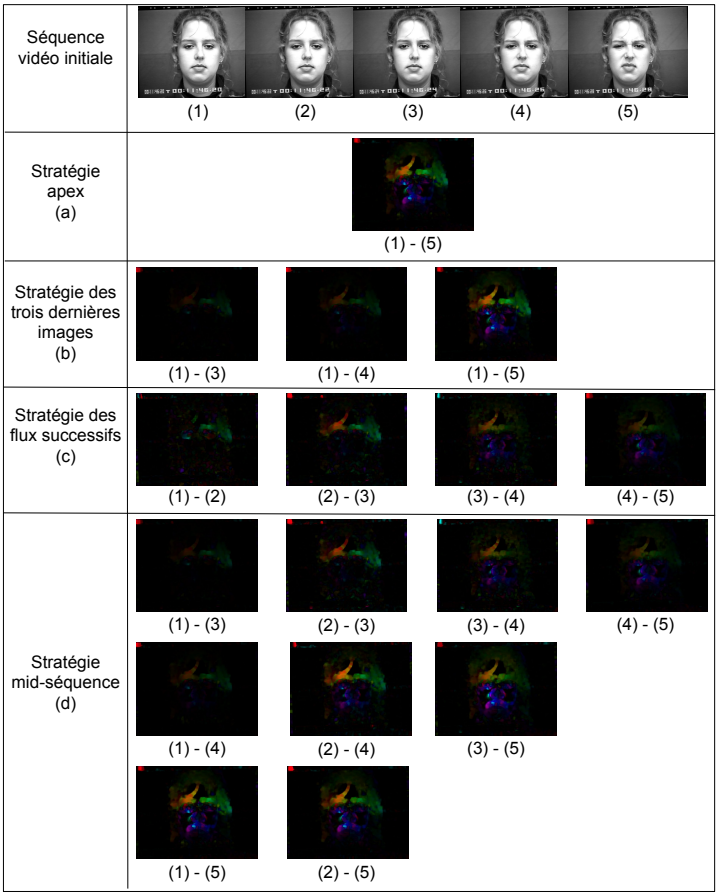


FIGURE 6.8 – Représentation des différentes stratégies de calcul des flux optiques pour l'apprentissage de l'auto-encodeur de reconstruction. Les stratégies reposent sur une sélection des images de chaque séquence pour le calcul des flux optiques.

Ces différentes stratégies sont utilisées pour augmenter le nombre de données lors de l'apprentissage de l'auto-encodeur. Les flux optiques utilisés pour la validation et le test restent uniquement les flux optiques calculés entre la première et la dernière image, ce qui correspond au calcul de données utilisé pour le CNN de reconnaissance.

Les résultats obtenus en utilisant ces différentes stratégies lors de l'apprentissage de l'auto-encodeur sont résumés dans le Tableau 6.5.

			
Apex	89.6%	83.1%	72.5%
Trois dernière images	90.2%	83.6%	74.4%
Flux successifs	85.5%	77.6%	66.8%
Mid-séquence	**91.1%**	**85.9%**	**75.2%**

TABLE 6.5 – Impact de la stratégie de calcul des flux optiques d'apprentissage sur la reconnaissance des expressions faciales en utilisant notre approche.

Les résultats montrent qu'en augmentant le nombre de flux optiques tout en conservant des intensités fortes comme le fait la stratégie des trois dernières images permet d'améliorer légèrement les résultats. La stratégie des flux successifs, qui ne considère que des flux optiques d'images successives, se montre beaucoup moins efficace, ce qui s'explique facilement par le fait que cette stratégie ne permet pas de généraliser car l'apprentissage ne se fait que sur des flux optiques d'assez faible intensité. Ces résultats montrent donc que, malgré le nombre beaucoup plus important de données d'apprentissage, la variation des intensités des flux optique s'avère particulièrement importante pour permettre une bonne généralisation. Enfin, la stratégie mid-séquence, qui permet d'obtenir un large panel d'intensités de mouvements tout en générant un nombre de données d'apprentissage le plus conséquent est la stratégie qui s'avère la plus efficace et semble alors permettre une meilleure généralisation.

6.4.4 Évaluation de la capacité de généralisation

Nous proposons maintenant d'évaluer la généralisation de notre approche et le choix de ces méta-paramètres en utilisant un autre classifieur.

Pour ce faire, pour l'étape de reconnaissance, nous utilisons un classifieur SVM avec noyau RBF entraîné avec le même protocole que précédemment en intégrant, pour chaque itération, le pli de validation aux plis d'apprentissage. Les flux optiques sont vectorisés et utilisés en entrée du SVM.

	SVM				CNN			
								
Sans reco.	90.9%	74.9%	45.5%	41.7%	92.8%	73.8%	71.1%	46.8%
Apex	/	86.4%	78.1%	73.8%	/	89.6%	83.1%	72.5%
Gain	/	+11.5%	+32.6%	+32.1%	/	+15.8%	+12%	+25.7%
Mid-séquence	/	89.0%	82.1%	74.3%	/	91.1%	85.9%	75.2%
Gain	/	+14.1%	+36.7%	+32.6%	/	+17.3%	+14.8%	+28.4%

TABLE 6.6 – Résultats obtenus avec un classifieur SVM en comparaison avec les résultats obtenus avec le CNN.

Le Tableau 6.6 présente les résultats obtenus avec ce nouveau classifieur. La première ligne de ce tableau présente les résultats obtenus sans utiliser de reconstruction, c'est-à-dire en entraînant sur des données non occultées et en testant sur les données occultées. La première colonne indique les résultats obtenus en absence d'occultation. Les lignes suivantes présentent les résultats obtenus en testant sur des données reconstruites en utilisant notre approche de reconstruction entraînée selon les stratégies apex et mid-séquence.

Nous notons que le CNN permet d'extraire des caractéristiques plus robustes aux occultations que le flux brut en entrée du SVM, notamment pour l'occultation de la

bouche. Les occultations ont, en effet, un impact plus important lors de cette expérimentation. Pourtant, nous remarquons que la reconstruction permet de revenir sur des résultats comparables à ceux obtenus avec le CNN. Le gain obtenu grâce à la reconstruction, noté en dessous de chaque stratégie est alors d'autant plus important. Ces résultats montrent la généralisation de la reconstruction à d'autres classifieurs que le CNN.

6.4.5 Comparatif avec l'état de l'art

Le Tableau 6.7 propose une comparaison entre les résultats obtenus par notre approche avec les paramètres optimisés dans les sections précédentes et les résultats d'autres approches de l'état de l'art évalués sur CK+.

					
Huang et al. [96]	93.2%	**93% / -0.2%**	73.5% / -19.7%	-19.9%	-
Dapogny et al. [97]	**93.4%**	76% / -17.4%	67.1% / - 26.3%	-43.7%	-
Notre méthode	92.8%	91.1% / -1.7%	**85.9% / -6.9%**	**-8.6%**	**75.2% / -17.6%**

TABLE 6.7 – Comparatif des résultats de la méthode proposée avec des résultats de l'état de l'art obtenus sur la base de données CK+ pour les occultations des yeux, de la bouche et du bas du visage. Ce tableau illustre l'impact de différentes occultations sur les performances des méthodes et permet d'étudier la perte entre les résultats sans occultation et les résultats avec les différentes approches. La perte cumulée des occultations des yeux et de la bouche est également présentée à la quatrième colonne de ce tableau pour comparer plus facilement les différentes approches.

Nous mettons en avant les meilleurs résultats dans ce tableau en indiquant, pour chaque occultation, le meilleur résultat en gras et nous soulignons le deuxième. Étant

données les légères différences sur les résultats sans occultations, nous ajoutons dans le tableau la perte engendrée par les différentes occultations, c'est-à-dire, la différence entre les résultats sans occultation et les résultats obtenus par les différentes approches. Pour comparer plus facilement ces différentes méthodes, nous ajoutons également le cumul des pertes liées aux occultations de la bouche et des yeux. En regardant ces cumuls, la méthode proposée est clairement compétitive avec celles de l'état de l'art. Nous soulignons également que, bien que Huang et al. et Dapogny et al. ne proposent pas d'occultation aussi importante que celle proposée dans notre évaluation avec notre occultation du bas du visage, nous obtenons des résultats plus élevés pour une occultation de la bouche.

Les différences de résultats entre ces différentes solutions pourraient s'expliquer, entre autres, par le choix des caractéristiques exploitées dans chacune de ces méthodes. Contrairement à notre approche, celles de Huang et al., Dapogny et al. se basent sur une analyse statique des images et n'utilisent pas d'information temporelle, ce qui pourrait expliquer la perte un peu plus importante au niveau des résultats obtenus. Huang et al. proposent une analyse basée sur des descripteurs temporels de forme et de texture. La dimension temporelle de ces descripteurs semble permettre de récupérer des informations supplémentaires malgré les occultations. Cependant, Huang et al. ne se basent pas sur des descripteurs de mouvement denses, ce qui ne permet pas de détecter des déformations subtiles du visage.

Or, lors d'une occultation de l'épicentre des déformations du visage, il peut ne rester que des déformations subtiles. Notre approche se base sur un calcul de flux optique qui permet alors de conserver ces informations subtiles malgré les occultations, ce qui peut, en partie, expliquer les performances de notre approche. De plus, notre approche est la seule, parmi ces trois approches, à se baser sur une reconstruction. La reconstruction pourrait alors s'avérer également plus efficace qu'une analyse basée sur les régions visibles. Notons, enfin, que nous avons sélectionné les approches compara-

tives en fonction des protocoles expérimentaux les plus proches de ceux proposés avec notre méthode.

Comme nous l'avions souligné, des légères différences au niveau des protocoles expérimentaux restent présentes entre ces trois approches. Le protocole expérimental utilisé étant rendu disponible, les prochaines approches pourront se comparer à notre méthode de façon complètement équitable.

6.5 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche pour répondre à la problématique des occultations dans un cadre de reconnaissance automatique des expressions faciales. Cette approche s'appuie sur une reconstruction des zones cachées du visage. Pour cela, nous avons proposé de reconstruire directement dans le domaine du flux optique afin d'exploiter la propriété de similarité inter-personnes du mouvement. Cette reconstruction s'appuie sur une architecture d'auto-encodeur débruitant qui est une architecture complètement adaptée dans un cadre de reconstruction de données bruitées. Dans ce chapitre, nous avons exposé le processus complet, du prétraitement des données jusqu'à la classification et avons, ainsi, pu comparer les résultats avec ceux d'autres approches de l'état de l'art.

L'approche proposée s'est montrée particulièrement efficace en permettant d'obtenir des scores généralement supérieurs à ceux de l'état de l'art (91.1% avec une occultation des yeux, 85.9% en présence d'une occultation de la bouche et d'une partie du nez et 75.2% lors d'une occultation du bas du visage). Dans le cadre de cette méthode, nous avons également proposé un protocole expérimental clair et disponible afin de faciliter la comparaison des futurs travaux.

Chapitre 7

Synthèse

7.1 Résumé des contributions

La reconnaissance automatique des expressions faciales reste, encore aujourd'hui, confrontée à différents défis en environnement naturel. La présence d'occultations partielles du visage est considérée comme l'un des défis majeurs pour reconnaître les expressions faciales. Différentes solutions ont été proposées dans la littérature pour répondre à cette problématique. Ces solutions peuvent être regroupées sous deux grandes catégories : les solutions qui se concentrent sur les régions visibles du visage et celles qui reconstruisent les zones cachées. Ces solutions sont principalement basées sur des descripteurs de texture ou de géométrie. Pourtant, le mouvement semble adapté pour répondre à la problématique des occultations grâce à certaines propriétés. Les propriétés du mouvement que nous mettons en avant dans ce manuscrit sont : la propriété de propagation du mouvement, qui permet de conserver de l'information liée à l'expression faciale malgré les occultations et la similarité des mouvements liés aux expressions faciales inter-personnes, qui permet de s'abstraire de l'identité de la personne d'une part et, d'autre part, permet de plus facilement s'appuyer sur une donnée simi-

laire pour pouvoir reconstruire les zones cachées d'une donnée occultée. Durant ma thèse, nous avons exploité ces propriétés pour proposer des solutions dans les deux catégories.

Nous avons proposé une première solution qui se concentre sur les régions visibles du visage en calculant des modèles faciaux optimisés. Cette première solution s'appuie sur un descripteur de mouvement afin de tirer partie de la propriété de propagation du mouvement, qui permet de conserver de l'information sur les régions visibles malgré l'occultation.

Nous avons proposé une seconde solution basée sur une reconstruction du flux optique corrompu calculé à partir d'images de visages occultées. Cette seconde méthode se base sur une architecture d'auto-encodeur débruitant qui permet de reconstruire un flux optique non bruité par une occultation.

7.2 Discussion

Occultation/Méthode	Exploitation des régions visibles [117]	Reconstruction [118]
/	91.3%	92.8%
Yeux	88.8%	91.1%
Bouche	79%	85.9%
Bas du visage	73.4%	75.2%

Table 7.1 – Comparaison des résultats obtenus par les deux approches proposées dans ce manuscrit en fonction des différentes occultations étudiées.

Le Tableau 7.1 propose une comparaison des résultats obtenus entre ces deux approches sur des occultations similaires. En comparant ces deux méthodes, on peut noter que les résultats obtenus par reconstruction semblent légèrement plus prometteurs

sur des occultations similaires, notamment pour l'occultation de la bouche qui est une région clé pour reconnaître différentes expressions faciales.

7.3 Perspectives

Nous proposons, dans cette section, de discuter les perspectives ouvertes par ce travail de thèse.

7.3.1 Mécanismes d'attention

Pour notre première approche, nous avons proposé une méthodologie qui se base sur des étapes manuelles afin d'en tirer toutes les conclusions et d'avoir suffisamment de recul sur les différents résultats de chaque étape. Cette première approche a permis de mettre en avant différentes informations sur l'importance des régions du visage et la façon dont les régions les plus importantes peuvent se répartir en présence d'occultation. Ces étapes peuvent aujourd'hui être davantage automatisées grâce aux mécanismes d'attention, de plus en plus utilisés pour répondre à la problématique des occultations [82, 83]. Des mécanismes d'attention peuvent, en effet, être utilisés pour calculer automatiquement des poids d'attention sur chaque élément de la donnée d'entrée. Ces méthodes sont donc totalement appropriées dans le cadre des occultations afin de se concentrer sur les régions visibles. Nous souhaitons, pour de futurs travaux lier l'exploitation de la propagation du mouvement avec l'efficacité de ces nouvelles méthodes. Les solutions actuellement proposées dans la littérature sont basées sur la texture des visages. Une perspective à nos travaux serait d'automatiser la sélection des régions du visage sur lesquelles s'appuyer grâce aux mécanismes d'attention, en conservant, en entrée de l'apprentissage des informations de mouvement.

7.3.2 Reconstruction

Notre seconde proposition repose sur un auto-encodeur entraîné à reconstruire les flux optiques corrompus par des occultations partielles du visage.

Une première évolution de notre approche serait un apprentissage avec différentes occultations afin d'apprendre à l'auto-encodeur à généraliser pour reconstruire les flux optiques bruités quelque soit l'occultation rencontrée.

Par la suite, afin de créer une reconstruction plus fine, nous souhaitons poursuivre ces travaux en ajoutant différentes architectures d'auto-encodeurs empilées. Cette étude permettra également d'étudier l'intérêt de complexifier le réseau afin de déterminer s'il est nécessaire d'affiner la reconstruction ou s'il est plus intéressant d'améliorer davantage l'architecture initiale.

Pour la suite de ces travaux, inspirés par les travaux de Lu et al. [87], nous aimerions inclure une fonction de coût liée directement aux expressions faciales. Le but final de notre approche n'étant pas de reconstruire un flux optique parfait mais surtout de récupérer et reconstruire les informations liées aux expressions faciales. Parmi les solutions possibles pour mettre en place cette solution, une possibilité est de faire évoluer la solution vers une architecture de type GAN avec un réseau adverse entraîné à reconnaître les expressions faciales.

L'auto-encodeur étant, initialement, proposé pour calculer des caractéristiques de façon non supervisée, une autre approche que nous aimerions également explorer serait d'étudier directement la sortie de l'encodeur. La reconstruction est, en effet, superflue, si l'information liée à l'expression est présente en sortie d'encodeur. La sortie de l'encodeur pourrait ainsi être utilisée comme caractéristiques d'entrée d'un classifieur.

Inspirés par Pan et al. [119], l'apprentissage de l'encodeur peut également être guidé par un réseau parallèle qui apprend à extraire les caractéristiques de flux optiques issues d'images non occultées afin d'obtenir des caractéristiques qui s'approchent le plus des caractéristiques obtenues sur des données non occultées.

7.3.3 Occultations dynamiques

Les travaux et les approches proposées dans ce manuscrit se concentrent sur des occultations statiques. Les occultations étudiées sont, en effet, fixes entre la première et la dernière image de chaque séquence. Ainsi, les flux optiques calculés sur ces séquences vidéo contenaient des zones sans aucun mouvement. Dans un contexte plus naturel, les occultations peuvent être dynamiques : une main qui passe devant le visage par exemple. Pour aller plus loin dans nos travaux, nous aimerions étudier l'impact de ces occultations sur les flux optiques calculés. Les occultations dynamiques engendrent, en effet, un mouvement supplémentaire qui n'a pas de lien direct avec l'expression faciale. Plutôt qu'une zone sans mouvement facilement détectable, ces mouvements génèrent alors des flux optiques bruités qui peuvent perturber la reconnaissance des expressions faciales.

7.3.4 Variations de pose

Les solutions proposées dans ce manuscrit pour répondre à la problématique des occultations sont basées sur l'analyse du mouvement à partir du flux optique. Bien que le mouvement semble complètement adapté pour répondre à la problématique des occultations, il est également très sensible aux variations de pose. Une variation de pose va, en effet, générer un mouvement parasite lié au mouvement du visage et ainsi bruiter le mouvement plus faible de l'expression faciale. Afin d'étudier ce défi, Allaert et al. ont mis en place la base de données SNaP-2DFe [120] permettant d'avoir les

images simultanées de visage avec et sans variation de pose. Pour ce faire, les auteurs ont installé deux caméras : une caméra fixée sur un casque posé sur la tête du sujet et une caméra fixée face au sujet. Ainsi, la caméra du casque suit les mouvements de la tête et ne filme alors que les mouvements liés à l'expression faciale. En parallèle, la caméra posée en face du sujet filme tous les mouvements de la personne.

Nous aimerions donc exploiter la base de données permettant d'avoir les variations de pose ainsi que la vérité terrain pour reconstruire les informations de mouvement en ne gardant que les mouvements liés aux expressions faciales. Pour cela, nous souhaitons retravailler l'architecture d'auto-encodeur proposée dans ce manuscrit pour reconstruire les données non occultées en l'entraînant, cette fois-ci, sur des données contenant des variations de pose. La base de données proposée par Allaert et al. est alors totalement adaptée car elle permet un apprentissage supervisé en comparant les flux optiques reconstruits avec les flux optiques calculés sur les images de vérité terrain.

7.3.5 Reconnaissance en environnement naturel

À plus long terme, ces méthodes pourront être incluses dans un processus complet pour reconnaître les expressions faciales en environnement complètement naturel. Pour ce faire, il est alors également nécessaire d'obtenir une méthode robuste aux modifications de luminosité ainsi qu'aux changements d'arrière plan.

7.3.6 Ouverture à d'autres applications

Les méthodes proposées pourraient également être réutilisées dans le cadre d'autres applications. Les deux méthodes proposées peuvent être appliquées dans le cadre de la lecture sur les lèvres [121] en présence d'occultations partielles des lèvres avec un

micro par exemple. Plus généralement, la méthode de reconstruction du flux optique pourrait être incluse dans un processus de reconnaissance de mouvements (ex. mouvement d'actions [122], mouvement de foules [123]) en présence d'occultation partielle.

7.4 Publications

7.4.1 Journaux

Poux, D., Allaert, B., Mennesson, J., Ihaddadene, N., Bilasco, I. M., & Djeraba, C. (2020). Facial expressions analysis under occlusions based on specificities of facial motion propagation. Multimedia Tools and Applications, 1-23 (Impact factor 2.313 - 2019).

7.4.2 Conférences internationales

Poux, D., Allaert, B., Mennesson, J., Ihaddadene, N., Bilasco, L. M., & Dieraba, C. (2018, September). Mastering occlusions by using intelligent facial frameworks based on the propagation of movement. In 2018 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-6). IEEE.

7.4.3 Conférences nationales

Poux, D., Allaert, B., Ihaddadene, N., Bilasco, I. M., & Djeraba, C. (2018, November). Étude de l'apport de la reconstruction des régions occultées du visage pour la reconnaissance des expressions. In 2018 COmpression et REprésentation des Signaux Audiovisuels (CORESA).

7.4.4 Soumissions en cours

Poux, D., Allaert, B., Ihaddadene, N., Bilasco, I. M., Djeraba, C., & Bennamoun, M. (2020). Dynamic Facial Expression Recognition under Partial Occlusion with Optical Flow Reconstruction. arXiv preprint arXiv :2012.13217 (le papier a été accepté après la soutenance dans la revue IEEE TIP, DOI : 10.1109/TIP.2021.3129120).

Annexe A

Annexes

A.1 Émotions et universalité des expressions faciales

Définitions d'une émotion

En 1981, Kleinginna et Kleinginna [124] ont recensé 92 définitions différentes proposées dans la littérature. Ces définitions ont été triées en 11 catégories parmi lesquelles figurent les définitions dites affectives, cognitives ou encore physiologiques. Parmi ces définitions, on peut citer, entre autres, Charles Brenner en 1974 qui définit une émotion comme une sensation de plaisir, de déplaisir ou les deux, plus les idées, conscientes ou inconscientes associées à cette sensation [125]. Richard L. Bruce définit plutôt une émotion comme un état de réactions du corps face à des systèmes variés qui surviennent brutalement, altéré par leurs activités [126].

Plus récemment, Izard a tenté de déterminer les éléments à propos desquels les chercheurs tombaient d'accord concernant les émotions. Il a donc proposé un sondage à 35 scientifiques de quatre nationalités différentes ayant fait des travaux significatifs sur le sujet de l'émotion dans des disciplines variées [127]. Ce sondage contient 6

questions, parmi lesquelles : "Qu'est-ce qu'une émotion ?". En analysant les réponses à ce sondage, Izard et al. ont noté que, d'une part les scientifiques étaient encore assez réticents à répondre à cette question et, d'autre part, les définitions proposées ne permettaient toujours pas de mettre en lumière une définition unifiée.

Universalité des expressions faciales

Charles Darwin [128] a été le premier à s'intéresser à l'universalité des expressions faciales en comparant les expressions de personnes issues de différentes cultures. Il s'est également intéressé à l'aspect inné de ces expressions faciales en incluant, dans ses études, de très jeunes enfants ou des personnes nées non-voyantes. Charles Darwin a noté de fortes similarités entre les différentes cultures et en a donc conclu à l'universalité des expressions [129]. À partir de ces conclusions, Paul Ekman a poursuivi ces travaux [130, 131] en mettant en avant six expressions faciales universelles que sont : la joie, la peur, la surprise, la colère, le dégoût et la tristesse

A.2 Expressions faciales et communication affective

Catherine Belzung, Professeure de neuroscience à l'Université de Tour, voit les émotions comme un élément central du lien social car elles permettent notamment l'empathie [132]. La communication de nos émotions aux autres permet alors d'adapter nos interactions aux ressentis d'autrui. Le fait de savoir qu'une personne se retrouvera blessée par certains mots ou certaines de nos actions nous poussent à maîtriser nos actes et nos paroles. En 2000, Dimberg et al. [133] étudient les réactions faciales inconscientes de sujets face à des visages qui expriment certaines émotions. Lors de cette étude, des électrodes sont posées sur des muscles clés du visage : le grand zygomatique, qui permet le haussement des lèvres lors d'un sourire, et le corrugateur du sourcil, qui

permet le froncement de sourcils. Les sujets visionnent des images de visages neutres et, pendant des brefs instants de 30ms, un visage exprimant une expression faciale est affiché. Trois groupes sont aléatoirement répartis : un groupe qui ne verra que des visages neutres, un groupe qui voit apparaître brièvement des expressions de colère et un dernier qui voit apparaître des expressions de joie. En étudiant les signaux électriques mesurés par les électrodes, les auteurs ont pu remarquer une activation plus importante du grand zygomatique pour le groupe ayant vu des expressions de joie et une activation plus forte du corrugateur du sourcil chez les personnes ayant vu des expressions de colère. Cette étude tend à montrer qu'une émotion peut se transmettre de façon inconsciente en provoquant les mêmes expressions d'un individu à l'autre. Les émotions peuvent donc se communiquer d'une personne à l'autre par différents signaux et, la communication entre les individus comporte donc une dimension affective importante.

Bibliographie

[1] M. Gualtieri, "Best practices in user experience (ux) design," Design Compelling User Experiences to Wow your Customers, pp. 1–17, 2009.

[2] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," Proceedings of the National Academy of Sciences, vol. 113, no. 10, pp. 2636–2641, 2016.

[3] A. Crenn, A. Meyer, H. Konik, R. A. Khan, and S. Bouakaz, "Generic body expression recognition based on synthesis of realistic neutral motion," IEEE Access, vol. 8, pp. 207758–207767, 2020.

[4] R. M. Sabour, Y. Benezeth, F. Marzani, K. Nakamura, R. Gomez, and F. Yang, "Emotional state classification using pulse rate variability," in 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), pp. 86–90, IEEE, 2019.

[5] P. Salovey, A. J. Rothman, J. B. Detweiler, and W. T. Steward, "Emotional states and physical health.," American psychologist, vol. 55, no. 1, p. 110, 2000.

[6] G. Maycock, "Sleepiness and driving : the experience of uk car drivers," Journal of Sleep Research, vol. 5, no. 4, pp. 229–231, 1996.

[7] R. P. Bagozzi, M. Gopinath, and P. U. Nyer, "The role of emotions in marketing," Journal of the academy of marketing science, vol. 27, no. 2, pp. 184–206, 1999.

[8] E. Sherman, A. Mathur, and R. B. Smith, "Store environment and consumer purchase behavior : mediating role of consumer emotions," Psychology & Marketing, vol. 14, no. 4, pp. 361–378, 1997.

[9] B. Bediou, M. Saoud, C. Harmer, and P. Krolak-Salmon, "L'analyse des visages dans la dépression," L'Evolution psychiatrique, vol. 74, no. 1, pp. 79–91, 2009.

[10] B. Bachimont, "Separating bodies, synchronising minds : The role of digital technology in mediating distance," img journal, no. 3, pp. 54–69, 2020.

[11] B. Bontchev, "Adaptation in affective video games : a literature review. cybernetics information technols [internet]. 2016 sep [cited 2017 mar 22] ; 16 (3) : 3-34."

[12] K. Bahreini, R. Nadolski, and W. Westera, "Towards multimodal emotion recognition in e-learning environments," Interactive Learning Environments, vol. 24, no. 3, pp. 590–605, 2016.

[13] R. Adolphs, L. Mlodinow, and L. F. Barrett, "What is an emotion ?," Current Biology, vol. 29, no. 20, pp. R1060–R1064, 2019.

[14] I. B. Mauss and M. D. Robinson, "Measures of emotion : A review," Cognition and emotion, vol. 23, no. 2, pp. 209–237, 2009.

[15] A. Mehrabian, "Communication without words," Communication theory, vol. 6, pp. 193–200, 2008.

[16] R. Ekman, What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

[17] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response : dimensional covariation in the evaluation of pictorial stimuli.," Journal of psychophysiology, 1989.

[18] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," Journal of research in Personality, vol. 11, no. 3, pp. 273–294, 1977.

[19] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet : A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2017.

[20] N. S. M. Suwa and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," IJCPR, p. 408–410, 1978.

[21] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition : History, trends, and affect-related applications," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 8, pp. 1548–1568, 2016.

[22] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp. 46–53, IEEE, 2000.

[23] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression," in 2010 ieee computer society conference on computer vision and pattern recognition-workshops, pp. 94–101, IEEE, 2010.

[24] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre, "Continuous au intensity estimation using localized, sparse facial feature space," in 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pp. 1–7, IEEE, 2013.

[25] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in European conference on computer vision, pp. 425–442, Springer, 2016.

[26] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," The Visual Computer, vol. 34, no. 12, pp. 1691–1699, 2018.

[27] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," 2011.

[28] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion : A survey," ACM Computing Surveys (CSUR), vol. 51, no. 2, pp. 1–49, 2018.

[29] J. N. Bassili, "Emotion recognition : The role of facial movement and the relative importance of upper and lower areas of the face.," Journal of personality and social psychology, vol. 37, no. 11, p. 2049, 1979.

[30] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow : Large displacement optical flow with deep matching," in Proceedings of the IEEE international conference on computer vision, pp. 1385–1392, 2013.

[31] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18–28, 1998.

[32] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition : A survey," Symmetry, vol. 11, no. 10, p. 1189, 2019.

[33] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face : A face detection benchmark," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5525–5533, 2016.

[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1, pp. I–I, IEEE, 2001.

[35] J. Li and Y. Zhang, "Learning surf cascade for fast and accurate object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3468–3475, 2013.

[36] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in 2012 IEEE conference on computer vision and pattern recognition, pp. 2879–2886, IEEE, 2012.

[37] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," Image and Vision Computing, vol. 32, no. 10, pp. 790–799, 2014.

[38] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in IEEE international joint conference on biometrics, pp. 1–8, IEEE, 2014.

[39] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 1, pp. 23–38, 1998.

[40] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5325–5334, 2015.

[41] X. Jin and X. Tan, "Face alignment in-the-wild : A survey," Computer Vision and Image Understanding, vol. 162, pp. 1–22, 2017.

[42] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 6, pp. 681–685, 2001.

[43] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1078–1085, IEEE, 2010.

[44] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem ?(and a dataset of 230,000 3d facial landmarks)," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1021–1030, 2017.

[45] A. Dapogny, K. Bailly, and M. Cord, "Decafa : Deep convolutional cascade for face alignment in the wild," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6893–6901, 2019.

[46] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern recognition, vol. 29, no. 1, pp. 51–59, 1996.

[47] T. Jabid, M. H. Kabir, and O. Chae, "Facial expression recognition using local directional pattern (ldp)," in 2010 IEEE International Conference on Image Processing, pp. 1605–1608, IEEE, 2010.

[48] J. Chen, Z. Chen, Z. Chi, H. Fu, et al., "Facial expression recognition based on facial components detection and hog features," in International workshops on electrical and computer engineering subfields, pp. 884–888, 2014.

[49] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in Proceedings of the 5th international conference on Multimodal interfaces, pp. 258–264, 2003.

[50] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," IEEE transactions on image processing, vol. 16, no. 1, pp. 172–187, 2006.

[51] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 915–928, 2007.

[52] B. K. Horn and B. G. Schunck, "Determining optical flow," Artificial intelligence, vol. 17, no. 1-3, pp. 185–203, 1981.

[53] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in Scandinavian conference on Image analysis, pp. 363–370, Springer, 2003.

[54] W. Li and E. Salari, "Successive elimination algorithm for motion estimation," IEEE transactions on image processing, vol. 4, no. 1, pp. 105–107, 1995.

[55] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet : Learning optical flow with convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 2758–2766, 2015.

[56] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1932–1939, IEEE, 2009.

[57] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," IEEE Transactions on Affective Computing, 2019.

[58] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Optical flow techniques for facial expression analysis : Performance evaluation and improvements," arXiv preprint arXiv :1904.11592, 2019.

[59] S. Li and W. Deng, "Deep facial expression recognition : A survey," IEEE Transactions on Affective Computing, 2020.

[60] B. Fasel, "Robust face analysis using convolutional neural networks," in Object recognition supported by user interaction for service robots, vol. 2, pp. 40–43, IEEE, 2002.

[61] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," Neural Networks, vol. 16, no. 5-6, pp. 555–559, 2003.

[62] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," Journal of Electronic Imaging, vol. 25, no. 6,

p. 061407, 2016.

[63] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, "Facial expression recognition with faster r-cnn," Procedia Computer Science, vol. 107, pp. 135–140, 2017.

[64] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, and M. Alazab, "Facial sentiment analysis using ai techniques : State-of-the-art, taxonomies, and challenges," IEEE Access, vol. 8, pp. 90495–90519, 2020.

[65] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[66] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 4489–4497, 2015.

[67] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3d convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1609–1618, 2017.

[68] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–450, 2016.

[69] J. Zhao, X. Mao, and J. Zhang, "Learning deep facial expression features from image and optical flow sequences using 3d cnn," The Visual Computer, vol. 34, no. 10, pp. 1461–1475, 2018.

[70] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4193–4203, 2017.

[71] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," Pattern Recognition Letters, vol. 119, pp. 49–61, 2019.

[72] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in Proceedings Third IEEE international conference on automatic face and gesture recognition, pp. 200–205, IEEE, 1998.

[73] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in 2005 IEEE international conference on multimedia and Expo, pp. 5–pp, IEEE, 2005.

[74] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in 7th international conference on automatic face and gesture recognition (FGR06), pp. 211–216, IEEE, 2006.

[75] M. Taini, G. Zhao, S. Z. Li, and M. Pietikainen, "Facial expression recognition from near-infrared video sequences," in 2008 19th International Conference on Pattern Recognition, pp. 1–4, IEEE, 2008.

[76] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–6, IEEE, 2008.

[77] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, pp. 1–4, IEEE, 2010.

[78] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., "Challenges in representation learning : A report on three machine learning contests," in International conference on neural information processing, pp. 117–124, Springer, 2013.

[79] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2584–2593, IEEE, 2017.

[80] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random gabor based templates for facial expression recognition in images with facial occlusion," Neurocomputing, vol. 145, pp. 451–464, 2014.

[81] S. Liu, Y. Zhang, and K. Liu, "Facial expression recognition under partial occlusion based on weber local descriptor histogram and decision fusion," in Proceedings of the 33rd Chinese Control Conference, pp. 4664–4668, IEEE, 2014.

[82] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439–2450, 2018.

[83] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," IEEE Transactions on Image Processing, vol. 29, pp. 4057–4069, 2020.

[84] F. Bourel, C. C. Chibelushi, and A. A. Low, "Recognition of facial expressions in the presence of occlusion.," in BMVC, pp. 1–10, 2001.

[85] F. Bourel, C. C. Chibelushi, and A. A. Low, "Robust facial expression recognition using a state-based model of spatially-localised facial dynamics," in Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 113–118, IEEE, 2002.

[86] J. Y. R. Cornejo and H. Pedrini, "Emotion recognition from occluded facial expressions using weber local descriptor," in 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–5, IEEE, 2018.

[87] Y. Lu, S. Wang, W. Zhao, and Y. Zhao, "Wgan-based robust occluded facial expression recognition," IEEE Access, vol. 7, pp. 93594–93610, 2019.

[88] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks : Algorithms, theory, and applications," arXiv preprint arXiv :2001.06937, 2020.

[89] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," Image and Vision Computing, vol. 26, no. 7, pp. 1052–1067, 2008.

[90] P. Ekman, "Asymmetry in facial expression," Science, vol. 209, no. 4458, pp. 833–834, 1980.

[91] W. G. Dopson, B. E. Beckwith, D. M. Tucker, and P. C. Bullard-Bates, "Asymmetry of facial expression in spontaneous emotion," Cortex, vol. 20, no. 2, pp. 243–251, 1984.

[92] R. Azmi and S. Yegane, "Facial expression recognition in the presence of occlusion using local gabor binary patterns," in 20th Iranian Conference on Electrical Engineering (ICEE2012), pp. 742–747, IEEE, 2012.

[93] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," Proceedings of the IEEE, vol. 98, no. 6, pp. 1031–1044, 2010.

[94] S. F. Cotter, "Weighted voting of sparse representation classifiers for facial expression recognition," in 2010 18th European Signal Processing Conference, pp. 1164–1168, IEEE, 2010.

[95] S. F. Cotter, "Recognition of occluded facial expressions using a fusion of localized sparse representation classifiers," in 2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), pp. 437–442, IEEE, 2011.

[96] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," Pattern Recognition Letters, vol. 33, no. 16, pp. 2181–2191, 2012.

[97] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," International Journal of Computer Vision, vol. 126, no. 2-4, pp. 255–271, 2018.

[98] H. Towner and M. Slater, "Reconstruction and recognition of occluded facial expressions using pca," in International Conference on Affective Computing and Intelligent Interaction, pp. 36–47, Springer, 2007.

[99] L. Zhang, K. Mistry, M. Jiang, S. C. Neoh, and M. A. Hossain, "Adaptive facial point detection and emotion recognition for a humanoid robot," Computer Vision and Image Understanding, vol. 140, pp. 93–114, 2015.

[100] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 1, pp. 362–369, IEEE, 2001.

[101] I. T. Jolliffe, "Principal components in regression analysis," in Principal component analysis, pp. 129–155, Springer, 1986.

[102] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

[103] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 241–246, IEEE, 2016.

[104] Y. Zhang, R. Liu, S. Zhang, and M. Zhu, "Occlusion-robust face recognition using iterative stacked denoising autoencoder," in International Conference on Neural Information Processing, pp. 352–359, Springer, 2013.

[105] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in 2017 IEEE International Conference on Image Processing (ICIP), pp. 1202–1206, IEEE, 2017.

[106] A. Dapogny, M. Cord, and P. Pérez, "The missing data encoder : Cross-channel image completion with hide-and-seek adversarial network," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10688–10695, 2020.

[107] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, pp. 2672–2680, 2014.

[108] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3911–3919, 2017.

[109] V. Vielzeuf, C. Kervadec, S. Pateux, and F. Jurie, "The many variations of emotion," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1–7, IEEE, 2019.

[110] Q. Wang, H. Fan, L. Zhu, and Y. Tang, "Deeply supervised face completion with multi-context generative adversarial network," IEEE Signal Processing Letters, vol. 26, no. 3, pp. 400–404, 2018.

[111] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in CVPR 2011, pp. 2857–2864, IEEE, 2011.

[112] B. Allaert, I. M. Bilasco, and C. Djeraba, "Advanced local motion patterns for macro and micro facial expression recognition," arXiv preprint arXiv :1805.01951, 2018.

[113] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision, pp. 483–499, Springer, 2016.

[114] O. Ronneberger, P. Fischer, and T. Brox, "U-net : Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, pp. 234–241, Springer, 2015.

[115] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245, 2018.

[116] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet : Learning optical flow with convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 2758–2766, 2015.

[117] D. Poux, B. Allaert, J. Mennesson, N. Ihaddadene, I. M. Bilasco, and C. Djeraba, "Facial expressions analysis under occlusions based on specificities of facial motion propagation," Multimedia Tools and Applications, pp. 1–23, 2020.

[118] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Dynamic facial expression recognition under partial occlusion with optical flow reconstruction," arXiv preprint arXiv :2012.13217, 2020.

[119] B. Pan, S. Wang, and B. Xia, "Occluded facial expression recognition enhanced through privileged information," in Proceedings of the 27th ACM International Conference on Multimedia, pp. 566–573, 2019.

[120] B. Allaert, J. Mennesson, I. M. Bilasco, and C. Djeraba, "Impact of the face registration techniques on facial expressions recognition," Signal Processing : Image Communication, vol. 61, pp. 44–53, 2018.

[121] J. Shiraishi and T. Saitoh, "Optical flow based lip reading using non rectangular roi and head motion reduction," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–6, IEEE, 2015.

[122] A. Mazari and H. Sahbi, "Coarse-to-fine aggregation for cross-granularity action recognition," in 2020 IEEE International Conference on Image Processing (ICIP),

pp. 1541–1545, IEEE, 2020.

[123] Y. Benabbas, N. Ihaddadene, and C. Djeraba, "Motion pattern extraction and event detection for automatic visual surveillance," EURASIP Journal on Image and Video Processing, vol. 2011, pp. 1–15, 2011.

[124] P. R. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," Motivation and emotion, vol. 5, no. 4, pp. 345–379, 1981.

[125] C. Brenner, "On the nature and development of affects : A unified theory," The Psychoanalytic Quarterly, vol. 43, no. 4, pp. 532–556, 1974.

[126] R. L. Bruce, Fundamentals of physiological psychology. Holt, Rinehart, and Winston, 1977.

[127] C. E. Izard, "The many meanings/aspects of emotion : Definitions, functions, activation, and regulation," Emotion Review, vol. 2, no. 4, pp. 363–370, 2010.

[128] C. Darwin and P. Prodger, The expression of the emotions in man and animals. Oxford University Press, USA, 1998.

[129] K. Wolf, "Measuring facial expression of emotion," Dialogues in clinical neuroscience, vol. 17, no. 4, p. 457, 2015.

[130] P. Ekman and D. Keltner, "Universal facial expressions of emotion," Segerstrale U, P. Molnar P, eds. Nonverbal communication : Where nature meets culture, pp. 27–46, 1997.

[131] P. Ekman, "Strong evidence for universals in facial expressions : a reply to russell's mistaken critique.," 1994.

[132] C. Belzung, Neurobiologie des émotions. Uppr, 2016.

[133] U. Dimberg, M. Thunberg, and K. Elmehed, "Unconscious facial reactions to emotional facial expressions," Psychological science, vol. 11, no. 1, pp. 86–89, 2000.