Neutralisation des expressions faciales pour améliorer la reconnaissance du visage

Baptiste Chu

N° d'ordre : 2015-05

Thèse de l'Université de Lyon

Délivrée par
l'Ecole Centrale de Lyon

présentée en vu d'obtenir le titre de

DOCTEUR
spécialité « Informatique »

préparée au
Laboratoire d'InfoRmatique en Image et Systèmes d'information

Neutralisation des expressions faciales pour améliorer la reconnaissance du visage

Ecole doctorale Informatique et Mathématiques

Thèse soutenue le 02 Mars 2015 par

M. Baptiste Chu

devant le jury composé de :

Rapporteurs :	Mme	Alice Caplier	Professeur Grenoble INP
	M.	Sébastien Marcel	Directeur du laboratoire biométrie IDIAP
Examinateurs :	M.	Kévin Bailly	Docteur Université Pierre et Marie Curie
	M.	Mohamed Daoudi	Professeur Télécom Lille1
	M.	Stéphane Gentric	Docteur Safran Morpho
Directeur de thèse :	M.	Liming Chen	Professeur Ecole Centrale de Lyon
Encadrant :	M.	Sami Romdhani	Docteur Safran Morpho

"Je ne cherche pas à connaître les réponses,
je cherche à comprendre les questions."

Confucius (-551/-479)

Remerciements

Je tiens à remercier, à travers cette page, toutes les personnes qui ont contribué, de près ou de loin, à la réalisation et à l'aboutissement de cette thèse.

Les travaux présentés dans cette thèse, ont fait l'objet d'une convention CIFRE entre la société Morpho-Groupe Safran et l'Ecole Centrale de Lyon (Laboratoire d'InfoRmatique en Image et Systèmes d'information).

En premier lieu, j'aimerai remercier mes deux directeurs de thèse, Liming Chen et Sami Romdhani, pour l'attention et le soutien qu'ils ont porté à mon travail durant ces trois années. Tout au long de ce travail, ils ont su m'apporter un soutient constant, une disponibilité, une écoute. Ils m'ont permis, à travers leurs connaissances, leurs critiques, les nombreux conseils, de mener à bien ce travail.

J'adresse également mes remerciements à Stéphane Gentric, responsable de la recherche Visage, ainsi qu'aux directeurs de l'unité de recherche et de technologie de Morpho, François Rieul et Géraldine Genest, pour avoir accepté de soutenir ces travaux.

J'associe à ces remerciements tous mes collègues de Morpho, et plus particulièrement Anouar, Anthony, Catherine, Christelle, Julien, Olivier, Pierre, Sarah, Thomas et Vincent pour leur aide précieuse durant ces trois années, leur disponibilité et leur bonne humeur.

Je tiens également à adresser mes remerciements aux membres et aux doctorants du laboratoire pour leur accueil et leur sympathie. Je souhaite remercier particulièrement Claire, Jean-Noël et Yoann pour les bons moments partagés ensemble ainsi qu'Isabelle et Colette pour leur aide et disponibilité.

Mes remerciements s'adressent également aux professeurs Alice Caplier et Sébastien Marcel pour avoir accepté de rapporter ma thèse et pour l'intérêt qu'ils ont porté à ces travaux. Je tiens à remercier également Kévin Bailly et Mohamed Daoudi d'avoir accepté de participer à mon jury de thèse.

Je termine par un immense merci à ma famille et mes amis qui ont toujours été présents au cours de ces trois ans, pour m'encourager et me soutenir.

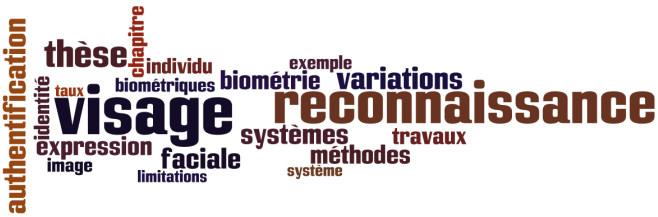
Table des matières

Liste des figures

Introduction

1

Sommaire



Nous allons, à travers cette introduction, exposer le contexte de cette thèse.

Nous commencerons par présenter de manière générale les différentes méthodes d'authentification utilisées dans notre vie quotidienne.

Parmi ces méthodes, la biométrie permet une authentification fiable et sûre. Nous en présenterons les principaux avantages et inconvénients. Cadre de notre thèse, la reconnaissance de visages est l'une des modalités biométriques couramment utilisée. Nous en ferons une présentation plus détaillée dans le but de préciser les enjeux et les motivations de cette thèse. L'exposé des limitations actuelles de cette technologie nous permettra de justifier notre approche ainsi que les motivations ayant conduit aux travaux présentés dans cette thèse.

Pour conclure ce chapitre, le plan de cette thèse sera présenté.

1.1 Enjeux et motivations

Le visage est la caractéristique biométrique couramment utilisée par les hommes pour se reconnaitre. Certaines maladies peuvent entrainer la perte de cette capacité. Selon Kennerknecht *et al.* [41], la prosopagnosie (incapacité à identifier une personne à partir de son visage) toucherait environ 2.5% de la population. Observée pour la première fois en 1947 sur des soldats, cette maladie peut être acquise (présente dès la naissance) ou être la conséquence d'une lésion cérébral (suite à un accident vasculaire cérébrale par exemple).

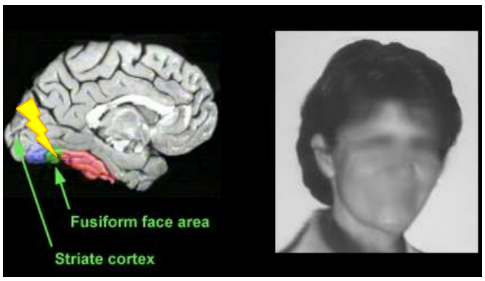


Figure 1.1 – *Prosopagnosie provoquée par une lésion de la FFA*

Les personnes atteintes d'inacuité visuelle souffrent également de cette incapacité à reconnaitre les personnes par leur visage. Actuellement, un certain nombre de recherches ont lieu sur ces problématiques. Nous pouvons par exemple citer l'utilisation de la reconnaissance faciale à travers des *Google Glass* pour améliorer le quotidien des personnes dans l'incapacité à reconnaitre un visage.

Plus généralement, la problématique d'identification des personnes occupe aujourd'hui une place prépondérante au sein de notre société. Assurer l'identité d'une personne est nécessaire dans de nombreux domaines : Contrôle d'accès à des zones réglementées, passage aux frontières, opérations bancaires, ...

Afin de garantir l'identité d'un individu en limitant les risques d'usurpation, de nombreux systèmes d'authentification sont mis en place dans la vie quotidienne. Traditionnellement, ces systèmes sont classés en trois catégories :

En premier lieu, les systèmes basés sur une connaissance a priori de l'individu (ce que l'on sait de lui). Ils sont regroupés sous le terme de Knowledge Based Authentification (KBA). Ce type de méthode est couramment utilisée pour prouver notre identité avant d'accéder à un service (Messagerie internet, téléphone portable, ...). L'authentification est permise grâce à un mot de passe, un code PIN (Personal Identification Number) ou une question secrète (par exemple : Quel est le nom de jeune fille de votre mère ?). Ces informations, enregistrées par le fournisseur de service lors du premier contact avec l'utilisateur, permettent ensuite de s'assurer que l'individu est bien celui que l'on pense être.

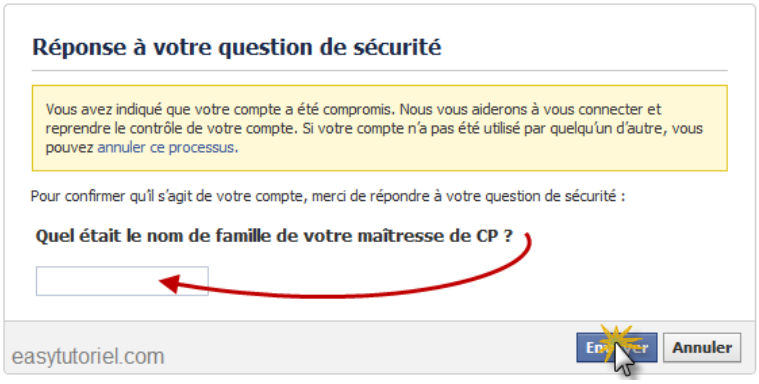


FIGURE 1.2 – *Exemple d'authentification par "question secrète"*

Ces méthodes présentent l'avantage d'être bien acceptées (car couramment utilisées). Cependant, elles présentent un certain nombre de limitations parmi lesquels un risque de vol du mot de passe (Par hameçonnage par exemple) ou bien un oubli de ce même mot de passe par l'utilisateur.

La deuxième catégorie regroupe les systèmes basés sur la possession d'un objet. L'accès à une zone réglementée en utilisant une clef ou un badge en sont des exemples. Le risque de duplication de l'objet permettant l'authentification constitue l'une des limitations principales de ces méthodes.

Pour faire face à ce risque, des méthodes appelées Token Based Authentification (TBA) ont été développées. Ces méthodes sont basées sur un système d'identification par jeton unique. Ce jeton, utilisé ensuite de manière similaire à un mot de passe, ne peut être utilisé qu'une seule fois (l'unicité du jeton est garanti par un algorithme cryptographique). De tels dispositifs peuvent être un objet physique ou dématérialisé sous forme de logiciel. Ils peuvent, par exemple, être utilisés par des banques pour valider un paiement par carte bancaire en ligne :

FIGURE 1.3 – *Exemple de dispositif d'authentification de type Token Based Authentification*

Le troisième type d'authentification est basé sur ce que l'on est. Ce sont les caractéristiques biométriques d'un individu (empreintes digitales, visage, iris, ...) qui lui permettent d'être authentifié. Celles-ci ont la particularité d'être universelles (tout le monde les possède), uniques (deux personnes ne peuvent partager une même caractéristique biométrique), permanentes (elles ne varient pas ou peu dans le temps). Ces méthodes permettent de garantir un niveau de sécurité plus élevé. En effet, les données biométriques sont beaucoup plus difficiles à falsifier en comparaison

aux moyens présentés précédemment. De plus, les risques liés à l'oubli de mot de passe ou à la perte de l'objet permettant l'authentification sont évités.

FIGURE 1.4 – *Exemple de dispositifs d'authentification biométriques*

L'authentification par biométrie, couramment appelée biométrie, permet donc de garantir un niveau de sécurité élevé tout en facilitant la tâche de l'utilisateur. Elle est définie ainsi par la CNIL [20] :

"La biométrie regroupe l'ensemble des techniques informatiques permettant d'identifier un individu à partir de ses caractéristiques physiques, biologiques voire comportementales. Les données biométriques ont la particularité d'être uniques et permanentes. Elles permettent de ce fait le "traçage" des individus et leur identification certaine."

Cette définition met en avant les deux spécificités principales de la biométrie : l'unicité et la fiabilité. La fait que deux personnes ne peuvent partager les mêmes caractéristiques permet de garantir un taux de fausses acceptions faible (risque de considérer à tort deux jeux de données biométriques comme issus de la même personne). Ce faible taux permet donc permet de garantir un système d'authentification fiable.

Différentes modalités peuvent être utilisées dans le but de garantir l'identité d'un individu. Chacune d'elles possède des avantages lui permettant d'être utilisée de façon optimale dans des scénarios spécifiques :

- Biométrie à grande fiabilité : Iris
- Biométrie à traces : Empreintes digitales
- Biométrie non coopérative : Visage

L'étude des empreintes digitales est la biométrie la plus répandue. Sa grande fiabilité lui a permis d'être largement déployée. Aujourd'hui, cette technologie s'est ouverte au grand public par exemple pour permettre le déblocage d'un ordinateur ou d'un téléphone portable.

FIGURE 1.5 – *Déverrouillage Exemple de dispositifs d'authentification biométriques*

Bien que largement utilisée de nos jours, cette technologie est contrainte par une limitation importante : Un contact physique est nécessaire entre le capteur d'acquisition et le doigt de l'individu. D'autres biométries permettent de s'affranchir de cette contrainte. En effet, la reconnaissance du visage, de l'iris [24] ou bien encore du réseau veineux de la paume de la main [48] peut être effectuée sans contact.

Dans cette thèse, nous nous intéressons à la reconnaissance de visage. La grande flexibilité de cette technologie en termes de sources d'acquisition constitue un réel avantage. La comparaison de deux visages peut par exemple être effectuée à partir d'un flux vidéo provenant d'un système de vidéo-protection ou de la photo contenue dans un passeport biométrique.

Le niveau de maturité atteint par cette technologie lui permet d'être mise en application dans de nombreuses applications (Contrôle d'accès, passage aux frontières, vidéo-surveillance, . . .).

Les premiers travaux sur la reconnaissance faciale ont été effectués au cours des années 1960. Dès 1966, Bledsoe repère les principales limitations de la reconnaissance faciale :

"Le problème de reconnaissance faciale est rendu difficile par la grande variabilité dans la pose du visage, l'intensité de l'éclairage et l'expression du visage."

En 1973, Kanade [39], considéré comme l'un des pères de la reconnaissance faciale, effectue l'une des premières tentatives de reconnaissance de visages lors de sa thèse de doctorat. Depuis ces premiers travaux, la reconnaissance faciale a été l'objet de nombreuses recherches et restent toujours un sujet important dans la communauté de la vision par ordinateur.

Au cours de ces dernières années, un certain nombre de publications ont été consacrées aux problématiques décrites par Bledsoe. Celles-ci sont intrinsèquement liées au caractère non coopératif de cette biométrie. En contraignant l'utilisateur à respecter un certain nombre de critères [36], les taux de performances de la reconnaissance atteignent un taux de vérification de l'ordre de 95% pour un taux de fausses acceptances de 0.1% [32].

Ce niveau de performance peut être obtenu dans des scénarios dits coopératifs, comme par exemple dans le cas d'une personne souhaitant être identifiée par le système afin d'avoir accès à une zone réglementée. Afin d'être plus facilement authentifié, l'utilisateur coopère de façon à se placer dans les conditions d'acquisition permettant un fonctionnement

optimal de la reconnaissance faciale [76] : Une prise de vue frontale uniformément éclairée avec un visage neutre, sans expression.

D'un autre coté, cette biométrie peut être utilisée dans un contexte non coopératif (par exemple dans le cas de la vidéo-protection). Dans ce scénario, des variations dans l'apparence du visage peuvent apparaitre. La reconnaissance du visage est alors beaucoup plus délicate. Ces variations peuvent avoir de multiples causes :

- Variations de la pose : le visage n'est pas face au dispositif d'acquisition
- Variations de l'illumination : le visage n'est pas uniformément éclairé
- Variations d'expression : le visage n'est pas sous une expression neutre

FIGURE 1.6 – *Variations dans l'apparence d'un visage (Base CMU MultiPIE [31])*

Dans cette thèse, nous axerons notre travail sur les problèmes liés aux variations d'expressions, l'objectif final étant de rendre les systèmes actuels de reconnaissance faciale robustes aux variations simultanées d'expression et de pose du visage.

1.2 PROBLÉMATIQUE

La reconnaissance de visages peut être effectuée à partir d'un visage capturé en deux ou trois dimensions. Récemment, suite aux progrès observés dans les capteurs en trois dimensions (laser ou par lumière structurée), de nombreux travaux de recherche ont eu lieu dans le domaine de la reconnaissance 3D de visage.

Clairement, l'utilisation de scans 3D permet de diminuer la sensibilité aux variations de pose et d'illumination [17]. Certaines approches ont également permis de la rendre robuste aux variations d'expressions du visage [26]. Grâce à ces nombreux travaux, le taux de reconnaissance de visage en trois dimensions est proche de 100% [26].

Ces performances ont permis le déploiement de systèmes de contrôle d'accès sécurisés basés sur la forme 3D du visage. Ces systèmes très fiables autorisent une authentification sans contact. Cela garantit une plus grande

fluidité (plus besoin de s'arrêter au niveau du capteur) ainsi qu'une hygiène maximale (système d'accès à un bloc opératoire par exemple).

Cependant, la nécessité d'avoir enregistré au préalable la forme 3D de l'individu constitue une limitation importante. En effet, dans un certain nombre de scénarios, seule une image en deux dimensions du visage est disponible. Pour illustrer ce propos, nous pouvons citer deux exemples d'applications où la reconnaissance 3D du visage est problématique :

- Certains aéroports, confrontés à une augmentation constante du nombre de voyageurs, souhaitent fluidifier le trafic des passagers. Pour cela, des systèmes sont mis en place pour accélérer certaines étapes critiques telles que le contrôle aux frontières. Au cours de cette étape, l'identité de chaque voyageur est vérifiée par la police des frontières. Celle-ci garantit son identité en comparant son visage avec la photo présente sur le passeport. Un système automatisé de vérification d'identité peut alors être utilisé pour l'accélérer.

FIGURE 1.7 – *Portiques automatiques de passage aux frontières MorphoWay*

Le système MorphoWay (Figure 1.7), développé par Safran Morpho, est composé d'un sas que le passager traverse. Une image du visage du voyageur est capturée au cours de la traversée du sas. Cette image est comparée avec la photo contenue dans le passeport biométrique. Si la similarité entre ces deux images est suffisamment élevée, l'identité de l'individu est confirmée. Celui-ci est alors autorisé à franchir le sas et ainsi passer la frontière.

- De nos jours, de plus en plus de systèmes de vidéo-protection sont déployés tant dans des lieux publics (Gares, aéroports, ...) ou privés (Entreprises, banques, ...). Traditionnellement, les images acquises par une caméra de vidéo-surveillance sont enregistrées puis affichées sur un écran de contrôle. Celles-ci sont ensuite analysées par une personne habilitée. Néanmoins, il a été montré [30] que la capacité d'une personne à analyser plusieurs flux vidéo de manière simulta-

> née n'excède pas une vingtaine de minutes. L'utilisation de systèmes de reconnaissance faciale permet de traiter de façon automatique ces séquences vidéo. Les images extraites de celles-ci peuvent ensuite être comparées à une base de données dans le but d'identifier automatiquement les individus.

Ces deux scénarios sont des exemples où seuls des systèmes de reconnaissance 2D de visages sont utilisables. En effet, ici, le type de données utilisées est contraint. Dans le premier cas, le passeport ne permet de stocker qu'une image en deux dimensions. De même, dans le second cas, les images comparées, extraites d'un flux vidéo, sont des captures 2D de l'individu. De plus, l'utilisation d'un système de reconnaissance faciale en 2D permet, au contraire des autres biométries, une utilisation dans des scénarios non-intrusifs et à partir de sources d'acquisition diverses.

Les travaux effectués au cours de cette thèse s'inscrivent dans ce contexte de reconnaissance faciale à partir d'une image en deux dimensions d'un visage.

1.3 Approches et contributions

Les systèmes biométriques de reconnaissance de visage 2D permettent d'atteindre un taux de reconnaissance élevé sur des images acquises dans de bonnes conditions. Malheureusement, les performances obtenues avec les algorithmes actuels de reconnaissance de visage en deux dimensions diminuent fortement lorsque les images sont issues d'acquisitions non contrôlées.

Deux approches peuvent alors être suivies pour adresser ce problème. La première consiste à mettre en place un nouvel algorithme de reconnaissance robuste à ce type de variations. La seconde consiste à proposer un pré-traitement permettant de traiter les changements d'apparence du visage liés aux variations de pose et d'expression en amont d'un algorithme existant de reconnaissance de visage.

Dissocier ainsi la correction de l'expression de l'algorithme de reconnaissance de visage permet une certaine indépendance vis à vis de l'algorithme de reconnaissance utilisé. Cette approche est particulièrement adaptée pour le contexte industriel dans lequel s'inscrit cette thèse (Thèse CIFRE effectuée en collaboration avec Safran Morpho).

En effet, elle permet une capitalisation sur les différents algorithmes existants de Morpho, ceux-ci ayant nécessité plusieurs dizaines d'années-hommes de recherche pour correspondre au mieux aux contraintes des systèmes biométriques déployés à grande échelle. De plus, la reconnaissance de visage étant en perpétuelle évolution, l'approche "pré-traitement", indépendante du reste de la chaine de traitement, pourra être utilisée en amont des futurs algorithmes de reconnaissance faciale.

Nous présenterons donc, dans ce manuscrit, une méthode de pré-traitement permettant d'atteindre l'objectif de cette thèse : Capitaliser sur

les algorithmes existants en proposant une méthode permettant d'améliorer la fiabilité de la reconnaissance en situation non contrôlée.

Plus précisément, nous concentrerons notre travail sur l'amélioration de la robustesse des systèmes actuels aux variations d'expression et de pose. En parallèle, les performances obtenues sur des images contrôlées ne devront pas être dégradées.

FIGURE 1.8 – *Variations d'apparence intra-identité du visage (Base XM2VTSDB [47]). Chaque colonne montre une photographie de la même personne*

Pour traiter cette problématique, nous avons choisi de nous placer dans la perspective d'un pré-traitement pour annuler ces variations. Ce choix nous permet, d'une part, de capitaliser sur les algorithmes existants et d'autre part, de proposer une solution flexible. En effet, notre méthode, indépendante de l'algorithme de reconnaissance de visage utilisé, pourra alors être utilisée en amont des futurs algorithmes.

La méthode que nous proposons est donc un pré-traitement appliqué sur l'image originale en deux dimensions. L'ajustement d'un modèle déformable 3D générateur de visage permet d'approximer la forme 3D du visage à partir de l'image 2D. Les différents paramètres (coefficients d'identité, coefficients d'expression et paramètres d'expression) permettant d'expliquer cette forme ainsi qu'une carte de texture sont déduits de cet ajustement. Une nouvelle vue synthétique frontale et neutre peut alors être générée en utilisant de nouveaux paramètres de pose et d'expression.

La figure 1.9 présente les différentes étapes de ce processus.

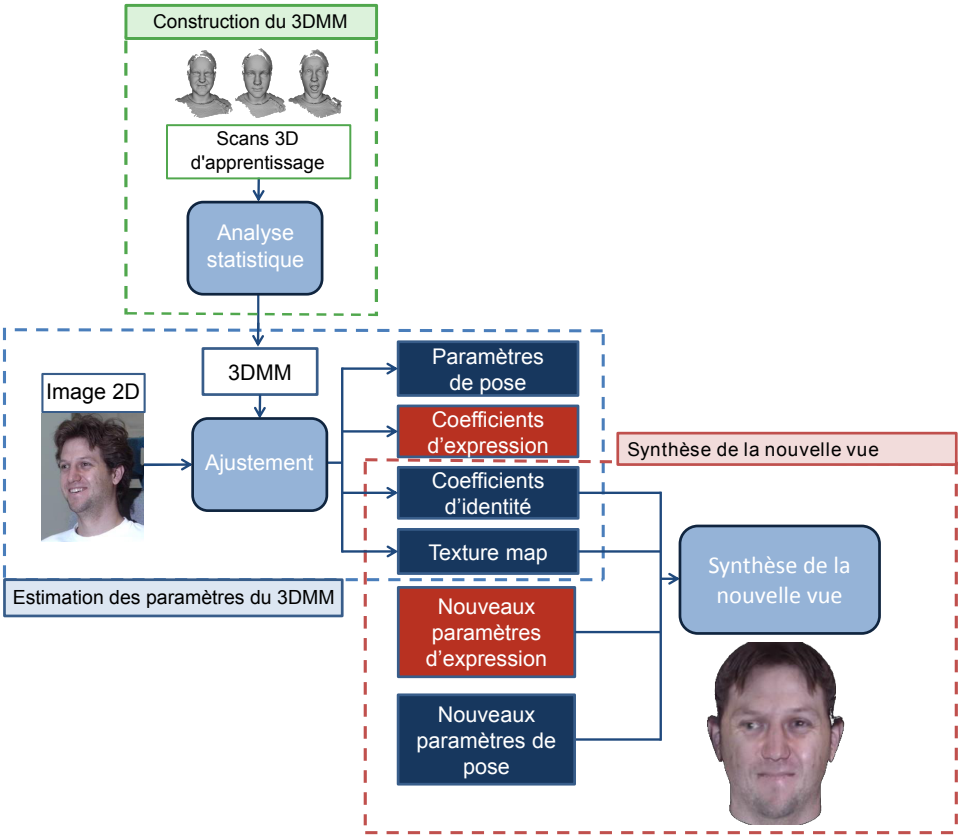


FIGURE 1.9 – *Diagramme fonctionnel de notre méthode*

Nous présenterons au cours de cette thèse, deux méthodes de prétraitement basées sur cette approche. Celles-ci permettent, au contraire des méthodes actuelles de l'état de l'art, de corriger les variations simultanées de pose et d'expression. Ainsi, nos méthodes de *neutralisation de l'expression* et de *transfert de l'expression* constituent les contributions principales de notre thèse.

Nous proposerons ensuite différentes approches permettant d'améliorer les performances biométriques obtenues avec ces méthodes ainsi qu'une extension à la problématique de reconnaissance de visages à partir de flux vidéo.

1.4 Plan de la thèse

Nous commencerons cette thèse par une présentation des principaux travaux traitant de la reconnaissance faciale. Nous porterons une attention particulière aux méthodes de reconnaissance robustes aux variations de pose ainsi que celles développées dans le but d'être robuste aux variations d'expression. Ensuite, nous exposerons l'ensemble des travaux effectués au cours de cette thèse.

Dans une première partie, nous détaillerons l'outil que nous avons utilisé pour représenter un visage (Chapitre 3). Celui-ci, basé sur un modèle déformable 3D, permet de reconstruire un visage en trois dimensions, à partir d'une simple image. Ce modèle permet de représenter n'importe quelle forme de visage comme une combinaison linéaire d'un nombre limité de déformations. De par sa construction, il est capable de dissocier les variations liées à l'identité de celles relatives à l'expression.

Nous présenterons, dans le chapitre 4, les méthodes que nous proposons pour neutraliser l'expression du visage. Celles-ci utilisent le modèle déformable 3D de visage présenté dans le chapitre précédent. Ici, deux méthodes sont présentées : La première effectue une neutralisation de l'expression à partir d'une image 2D, tandis que la deuxième propose de transférer l'expression de l'image de référence vers l'image de test. Nous terminerons ce chapitre par un ensemble d'expérimentations permettant de comparer les performances biométriques obtenues avec ces deux méthodes.

Dans le chapitre 5, nous étendrons la neutralisation d'expression à des scénarios de reconnaissance à partir de séquences vidéos. L'utilisation des informations temporelles apportées par la vidéo permet d'améliorer la qualité de l'approximation du visage par le modèle déformable 3D. Une évaluation expérimentale permettra de valider cette méthode.

L'analyse des résultats obtenus dans le chapitre 4 permet d'identifier un certain nombre de limitations induites par notre méthode. Le chapitre 6 sera donc consacré à la présentation de travaux réalisés en vue d'améliorer le processus de neutralisation de l'expression. Nous proposerons donc, dans cette dernière partie, des solutions pour pallier ces limitations.

Cette thèse sera conclue par un récapitulatif des principales contributions proposées et par un exposé des perspectives et axes de recherche futurs.

1.5 Publications

Les travaux menés au cours de cette thèse ont amené les publications suivantes :

3D-Aided Face Recognition Robust to Expression and Pose Variations
Baptiste Chu, Sami Romdhani, Liming Chen
In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on

3D-Aided Face Recognition from Videos
Baptiste Chu, Sami Romdhani, Liming Chen
In Visual Information Processing, 2014 5th European Workshop on. IEEE

Application of 3D Morphable Model for Faces with Expressions
Baptiste Chu, Sami Romdhani, Liming Chen
In Proceedings of COmpression REprésentation des Signaux Audiovisuels 2013

Etat de l'art

2

Sommaire

Dans ce chapitre, nous présenterons un état de l'art des méthodes de reconnaissances faciales en deux dimensions. Ces méthodes peuvent être classées en deux catégories. Les approches holistiques utilisent une représentation globale des visages tandis que les approches locales proposent de les caractériser par un ensemble de descripteurs locaux. Nous présenterons les principaux travaux relatifs à ces deux approches. Dans un second temps, nous nous concentrerons sur les méthodes traitant de la robustesse aux variations intra-classe du visage et plus précisément sur les variations de pose et les variations d'expression.

L'étude de la reconnaissance automatique de visage a commencé à être étudiée depuis le début des années 1970. Les travaux effectués par Kanade [39] au cours de sa thèse sont considérés comme les premiers sur ce domaine. Depuis, des nombreux travaux de recherche ont eu lieu.

Dans ce chapitre, nous présentons les travaux marquants de ce domaine. Dans un premier temps, nous nous intéresserons aux approches générales de reconnaissance de visages en deux dimensions. Dans un second temps, nous présenterons les méthodes développées en vue de robustifier la reconnaissance de visage aux variations de pose et d'expression.

De manière générale, les systèmes de reconnaissance de visages peuvent être décomposés en quatre étapes :

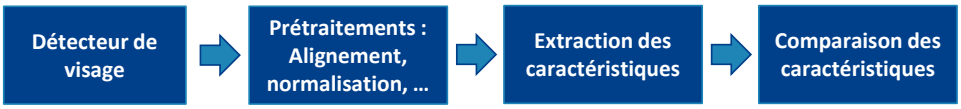


FIGURE 2.1 – *Workflow standard de la reconnaissance de visage*

La première étape consiste à détecter la présence d'un visage dans l'image et d'en extraire sa position le cas échéant [63]. Une fois extraits, un ensemble de pré-traitement est appliqué à chacun de ces visages. Ils permettent d'instaurer un référentiel commun entre eux. Ainsi, un alignement et une normalisation de l'illumination permettent de faciliter leur comparaison. Des caractéristiques faciales sont ensuite extraites de ces visages pré-traités. Ces caractéristiques doivent être suffisamment discriminantes et robustes aux variations intra-identité. L'ensemble des caractéristiques ainsi extraites forment la signature biométrique du visage. Finalement, la comparaison de deux visages est faite à partir de leurs signatures biométriques.

Dans cette partie, nous présentons les principales méthodes d'extraction et de comparaison de caractéristiques faciales de l'état de l'art. Ces méthodes peuvent être classées en deux catégories :

D'un coté, les méthodes holistiques utilisent le visage dans son intégralité pour effectuer la classification. Ces méthodes suivent une approche semblable au processus cognitif permettant aux être humains de reconnaitre des visages. En effet, Sinha *et al.* [57] ont montré que ce processus est basé sur une analyse globale du visage. D'une manière semblable, les méthodes holistiques proposent de représenter les visages par un vecteur de grande dimension contenant les informations de chacun des pixels du visage. Également appelées méthodes globales, elles cherchent ensuite à réduire la dimension de ces vecteurs représentatifs de l'espace des visages. Les méthodes EigenFaces ou FisherFaces, basées sur des procédés standard de réduction de dimensionalité en sont des exemples.

De l'autre coté, se trouvent les méthodes basées sur caractéristiques locales du visage. En opposition aux méthodes holistiques, les visages sont représentés par une collection de vecteurs caractéristiques extraits

dans des régions précises. Les méthodes de cette catégories peuvent, par exemple, être basées sur des descripteurs LBP ou des ondelettes de Gabor.

2.1 Approches holistiques

Les méthodes présentées dans cette partie proposent de représenter chaque visage par un vecteur I de taille $M \times P$. Ce vecteur est constitué de la valeur en niveau de gris de chacun des pixels de l'image de taille $M \times P$ du visage. La principale limitation induite par l'utilisation d'une image dans sa globalité est la dimension importante du vecteur représentant cette image : Une image de taille 128×128 est en effet représentée par un vecteur de taille 16384. L'ensemble des méthodes présentées dans cette partie a donc pour objectif de réduire la dimension de ce vecteur de représentation des images. Nous nous intéresserons plus particulièrement à celles basées sur deux méthodes standard de réduction de dimensionalité : L'Analyse en Composantes Principales et l'Analyse Discriminante Linéaire.

2.1.1 Analyse en Composantes Principales : EigenFaces

Une Analyse en Composantes Principales (ACP) permet de définir, à partir d'un jeu de données d'apprentissage, un sous espace permettant de simultanément conserver l'information discriminante et supprimer les informations secondaires (non informatives). Cette méthode consiste à trouver une nouvelle base de l'espace des données dont tous les vecteurs sont orthogonaux entre eux. Le premier de ces vecteurs correspond à la direction de variance maximale des données d'apprentissage. Les autres composantes sont déterminées par la contrainte d'orthogonalité entre les vecteurs tout en respectant une direction de variance maximum.

Proposée en 1991 par Turk et Pentland [62], l'algorithme EigenFaces est une adaptation de l'ACP à la problématique de la reconnaissance des visages. L'espace de représentation des visages est construit en effectuant une ACP sur un ensemble de N images d'apprentissage.

On note $I = \{I_1, ..., I_N\}$ la collection des N vecteurs de taille $M \times P$ représentant les N images d'apprentissage.

L'image moyenne de cette base d'apprentissage est alors définie par :

$$\bar{I} = \frac{1}{N} \sum_{i=1}^{N} I_i \tag{2.1}$$

Les images d'apprentissage sont alors normalisées en soustrayant l'image moyenne :

$$\Psi_i = I_i - \bar{I}, \forall i \in 1..N \tag{2.2}$$

L'ACP permet de trouver un ensemble de K vecteurs orthogonaux U_i permettant de décrire, au mieux, l'ensemble des visages comme une combinaison linéaire de ces vecteurs :

$$\Psi_i = \sum_{j=1}^{K} \alpha_j U_j, \forall i \in 1..N \tag{2.3}$$

La recherche de ces vecteurs orthogonaux revient à déterminer les vecteurs propres v_i de la matrice de covariance C_x des données d'apprentissage. Cette matrice est définie par :

$$C_x = \sum_{i=1}^{N} \Psi_i \Psi_i^t \tag{2.4}$$

En notant X, la matrice de taille $N \times (M.P)$ contenant les N vecteurs d'apprentissage, la matrice de covariance peut être définie par :

$$C_x = XX^t \tag{2.5}$$

Les vecteurs propres de cette matrice de covariance C_x, correspondant aux composantes principales de visages, sont couramment appelés EigenFaces [49].

FIGURE 2.2 – *Exemples d'EigenFaces [49]*

Une réduction de la dimension de l'espace représentatif des visages peut ensuite être obtenue en ne conservant qu'une sous-partie de ces vecteurs propres : Seuls les N' vecteurs propres associés aux valeurs propres les plus élevées sont conservés pour constituer une base du nouvel espace des visages. Le choix de la valeur de N' peut être fait de manière à conserver un pourcentage donné de l'énergie globale ou bien de manière empirique en utilisant par exemple le critère de Kaiser (seules les vecteurs propres associés à une valeur propre supérieure à 1 sont conservés).

Le nouvel espace de représentation des visages E' est alors l'espace défini par les vecteurs v_i avec $i = 1..N'$. Chaque visage peut alors être caractérisé par sa projection dans cet espace. La comparaison entre deux visages V_1 et V_2 est alors faite en comparant les projections V_1' et V_2' de ses visages dans le nouvel espace E'. Le score de similarité entre ces deux visage est alors égal à la distance $d = \|V_1'V_2'\|$.

2.1.2 Analyse Discriminante Linéaire : FisherFaces

La méthode présentée précédemment traite les changements d'apparence du visage dans leur globalité. En effet, l'analyse en composantes principales est effectuée sur des données d'apprentissage non étiquetées et ne permet donc pas de différencier les variations intra-individus et les variations extra-individus. L'Analyse Discriminante Linéaire permet, à

partir de données d'apprentissage labellisées, de maximiser les variations extra-classe tout en minimisant les variations intra-classe.

L'application de l'Analyse Discriminante Linéaire à la reconnaissance de visage a été proposée par Belhumeur *et al.* [8] en 1997. Pour cela, l'ensemble des visages d'apprentissage sont annotés pour effectuer un apprentissage supervisé (Chacune de ces images est associée à une classe). Une classe est associée à un individu et contient toutes les images relatives à celui-ci. De plus, les classes ainsi définies doivent être composées d'au moins deux images.

L'image moyenne de chacune des classes c_i, notée $\overline{I_{c_i}}$, est définie par :

$$\overline{I_{c_i}} = \frac{1}{N_{c_i}} \sum_{i=1}^{N_{c_i}} I_i \tag{2.6}$$

avec N_{c_i} le nombre d'images relatifs à la classe c_i.

De manière similaire au paragraphe précédent, l'image moyenne des N_c classes d'apprentissage est notée $\bar{I}$. Elle est définie par :

$$\bar{I} = \frac{1}{N_c} \sum_{i=1}^{N_c} N_{c_i} \overline{I_{c_i}} \tag{2.7}$$

Les variations inter-classe (*between class*) S_b et intra-classe (*within class*) S_w sont alors définies ainsi :

$$S_b = \sum_{i=1}^{N_c} N_c \left(\overline{I_{c_i}} - \bar{I}\right) \left(\overline{I_{c_i}} - \bar{I}\right)^t \tag{2.8}$$

$$S_w = \sum_{i=1}^{N_c} \sum_{I \in ci} N_c \left(I - \overline{I_{c_i}}\right) \left(I - \overline{I_{c_i}}\right)^t \tag{2.9}$$

L'analyse discriminante linéaire propose alors de trouver la matrice de projection optimale permettant de maximiser S_b tout en minimisant S_w. Cela revient donc à chercher ζ^m maximisant le critère suivant (appelé critère d'optimisation de Fisher) :

$$\zeta^m = \arg\max_\zeta \left(\frac{\zeta^t S_b \zeta}{\zeta^t S_w \zeta} \right) \tag{2.10}$$

Une fois ce nouvel espace défini, la comparaison entre deux images est faite de manière similaire aux EigenFaces : Le score de similarité entre deux visages est donné par la distance entre leur projection dans ce nouvel espace de représentation des visages.

2.1.3 Conclusion

Les méthodes holistiques prennent en compte la globalité du visage pour effectuer la classification. Les images de taille $n \times m$ sont alors représentées par un vecteur de taille nm contenant les valeurs d'intensité de chacun des pixels composant l'image. Des méthodes de réduction de

dimensionalité (Analyse en composantes principales, analyse discriminante linéaire) peuvent alors être utilisées. Elles permettent de trouver un nouvel espace de représentation des images de plus petite dimension dans lequel l'information discriminante est conservée. Ces méthodes de reconnaissance permettent d'obtenir de bons résultats sur des images acquises dans des conditions similaires. Cependant, ces performances chutent fortement lorsque les images présentent des variations importantes d'illumination, de pose ou d'expression.

Des méthodes locales permettent de robustifier la reconnaissance de visage à ce type de variations. Ces méthodes de représenter les visages par une collection de caractéristiques associées à des sous-parties du visage.

2.2 Approche locales

Comme nous l'avons vu précédemment, l'apparence globale du visage est modifiée par les changements d'illumination, d'expression ou de pose. Les méthodes holistiques ne permettent donc pas une reconnaissance robuste à ces variations. Au contraire, les représentations locales des visages proposent de les caractériser par un ensemble de vecteurs caractéristiques extraits dans des zones précises du visage. Ce type de représentation permet d'obtenir des systèmes plus robustes aux variations intra-classes. En effet, l'utilisation d'une multitude de caractéristiques locales en différents points du visage permet de limiter l'impact des changements d'apparence variations d'apparence intra-identité.

Ces méthodes locales de reconnaissance de visage peuvent être classées en deux catégories : celles utilisant une approche géométrique et celles basées sur l'apparence locale.

2.2.1 Méthodes basées sur une approche géométrique

Ce type de méthode est considéré comme l'un des fondements de la reconnaissance biométrique. En effet, le premier système de reconnaissance d'individu proposé par Bertillon en 1870 était basé sur la mesure de différents paramètres du corps humain. Dans son ouvrage présentant son système, Bertillon [11] précise que la probabilité que deux personnes partagent les mêmes caractéristiques anthropométriques est de 1/286000.

Les premières méthodes automatiques de reconnaissance de visage, proposées au début des années 1970 (Kanade [39]), sont également basées sur cette approche. L'étude d'un ensemble de mesures géométriques (distance ou angle entre certains points du visage) permet de caractériser le visage. Ces points sont choisis de telle sorte à pouvoir être aisément détectés (Centre des yeux, coins de la bouche, pointe du nez, ...). Le visage est alors représenté par un vecteur de distances et d'angles.

Plus récemment, d'autres méthodes de reconnaissance faciale basées sur la position d'un plus grand nombre de points caractéristiques ont été proposées. Brunelli et Poggio [19] ont ainsi décrit un système de

reconnaissance de visage basé sur l'extraction automatique de 35 caractéristiques physiques du visage. Celui-ci est ensuite caractérisé par leur position ou bien par l'angle formé par certains de ces points.

De nombreux travaux ont ensuite proposé une extension de cette méthode. Ils utilisent une représentation des visages par déformation de graphes [42] [69] [61].

Dans ces méthodes, les noeuds des graphes sont définis à partir d'un certain nombre de points caractéristiques et de vecteurs caractéristiques calculés au voisinage de ces points. La distance euclidienne séparant deux de ces points permet de définir la topologie du graphe. La comparaison des deux images I_1 et I_2 se fait en comparant les graphes G_1 et G_2 associés à ces deux images. Cette comparaison est faite à la fois sur la topologie des graphes (distance entre les différents noeuds) et sur les noeuds (similarité entre les vecteurs caractéristiques associés aux noeuds).

En plus d'une meilleure robustesse aux changements d'apparence, ces méthodes nécessitent un coût de stockage négligeable en comparaison aux méthodes holistiques. La qualité de l'extraction de ces caractéristiques géométriques peut toutefois être altérée en cas de fortes variations d'illumination ou de pose.

2.2.2 Méthodes basées sur l'apparence locale

L'ensemble de ces méthodes proposent de caractériser les visages par une collection de descripteurs locaux d'apparence. Ces descripteurs sont calculés sur différentes parties du visage. Celles-ci peuvent être définies par des patchs (avec ou sans chevauchement) ou par des régions informatives. Les caractéristiques locales sont ensuite extraites sur chacune d'elles grâce à des descripteurs standards. Une fois ces caractéristiques extraites, seules les plus informatives sont conservées grâce à l'utilisation d'une ACP ou d'une LDA.

Nous présenterons, dans cette section, deux méthodes d'extraction des caractéristiques couramment utilisées en vision par ordinateur : Les filtres de Gabor et le descripteur Local Binary Pattern.

Filtres de Gabor

De nombreuses méthodes de caractérisation d'images basées sur des filtres de Gabor ont été proposées par la communauté de vision par ordinateur. Ces filtres, dont le fonctionnement peut être rapproché de celui du système visuel humain [35], autorise un paramétrage en fréquence et en orientation.

La première application des filtres de Gabor à la biométrie a été proposé par Daugman en 1993 [25] pour la reconnaissance d'iris. Appliqués à la reconnaissance de visages, ils permettent une extraction efficace des caractéristiques faciales en minimisant l'effet des variations de pose et

d'illumination [76].

Une ondelette de Gabor est définie par :

$$\psi_{u,v}(z) = \frac{\left\|\kappa_{u,v}\right\|^2}{\sigma^2} e^{-\frac{\|\kappa_{u,v}\|^2\|z\|^2}{2\sigma^2}} \left(e^{i\kappa_{u,v}z} - e^{-\frac{\sigma^2}{2}} \right) \tag{2.11}$$

L'orientation du filtre est définie par u et l'échelle donnée par v.
$\kappa_{u,v}$ est le vecteur d'onde. Il est défini ainsi :

$$\kappa_{u,v} = \frac{f_{max}}{\sqrt{2}^v} e^{\frac{i\pi u}{8}} \tag{2.12}$$

La représentation d'une image $I(x,y) = I(z)$ en ondelettes de Gabor est obtenue par la convolution de cette image avec le filtre de Gabor correspondant :

$$G_{u,v} = I(z) * \psi_{u,v}(z) \tag{2.13}$$

L'image est généralement transformée par un ensemble d'ondelettes. Shen *et al.* [56] proposent ainsi d'utiliser 40 ondelettes (Figure 2.3) correspondant à cinq paramètres d'échelles v différents et huit paramètres d'orientations u différentes.

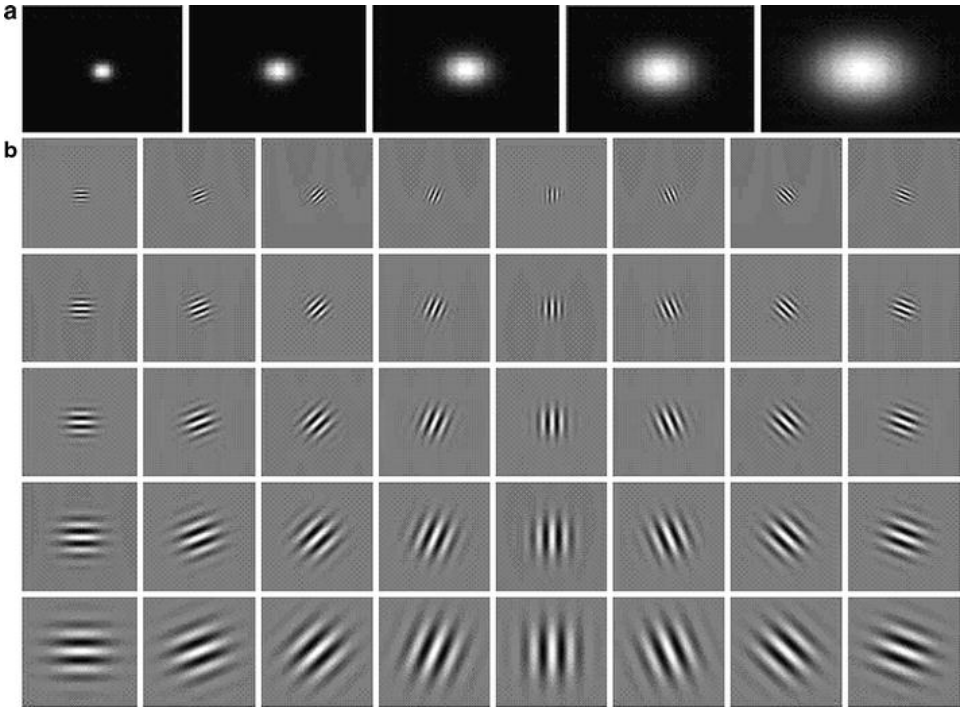


FIGURE 2.3 – *Collection d'ondelettes de Gabor : (a) Intensité à cinq échelles différentes (b) Partie réelle à cinq échelles et huit orientations [56]*

La figure 2.4 montre la réponse en amplitude et en phase de ce banc de Gabor sur une image de visage.

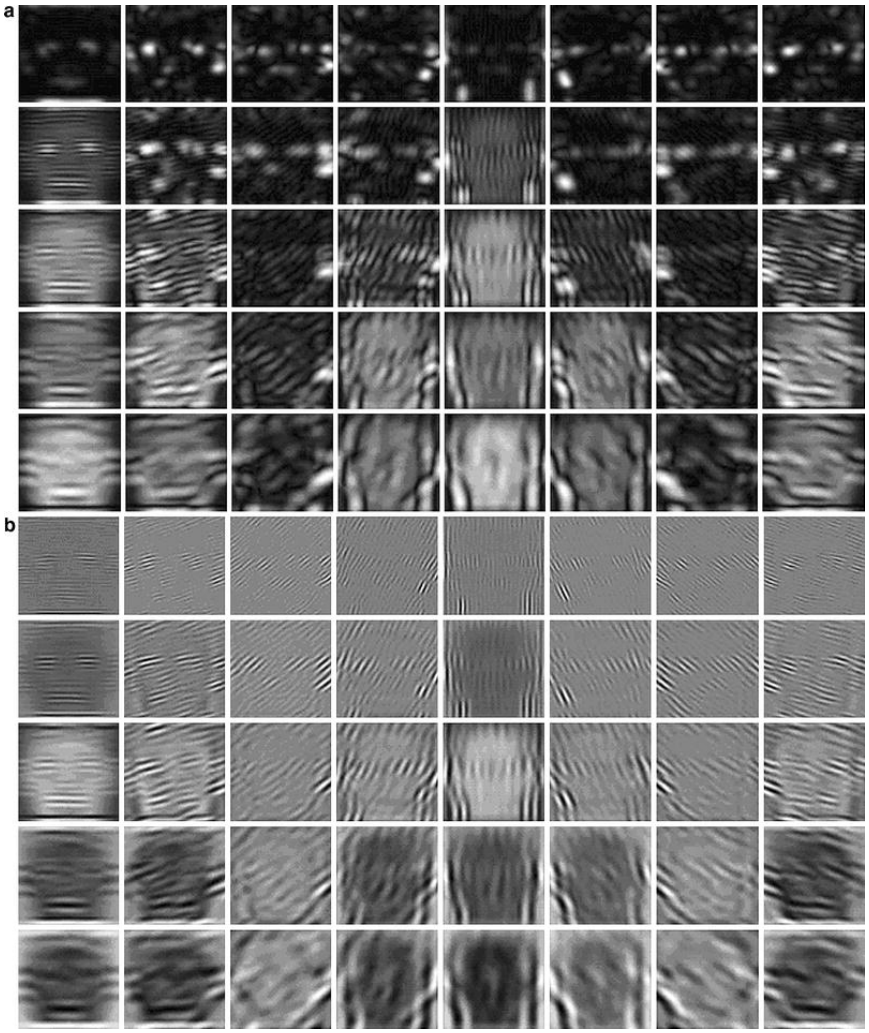


FIGURE 2.4 – *Représentation d'une image de visage dans le domaine de Gabor. Réponse en amplitude (a) et Réponse en phase (b) [56]*

Local Binary Pattern (LBP)

L'opérateur LBP a été proposé en 1996 par Ojala *et al.* [51]. Cet opérateur propose de représenter chaque pixel par un code binaire calculé à partir des 8 pixels voisins. Chacun des pixels voisins se verra représenter par un 1 si sa valeur est supérieure au pixel courant. Dans le cas contraire, ce pixel sera représenté par un 0.

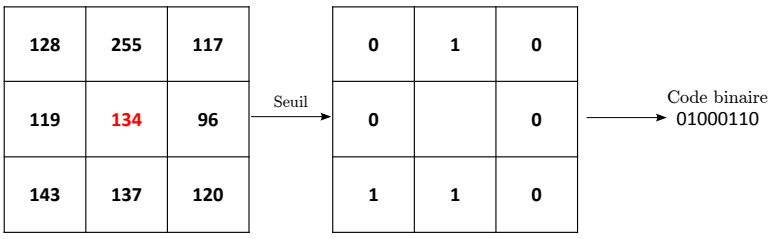


FIGURE 2.5 – *Descripteur LBP : Codage d'un pixel en fonction de la valeur de ses 8 pixels voisins*

De nombreuses extensions ont été proposées par la suite [33]. Oujala

et al. [52] ont notamment proposé une nouvelle définition de voisinage circulaire pour calculer le descripteur des pixels.

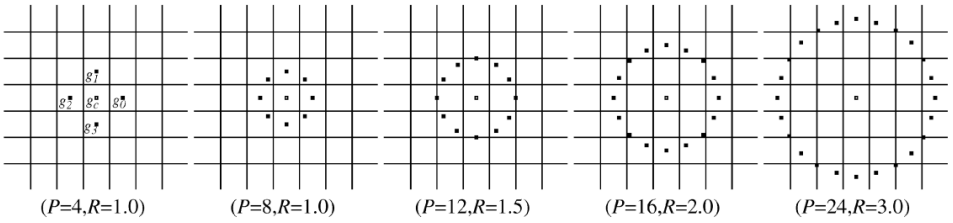


FIGURE 2.6 – *Exemples de voisinages circulaires proposés par Oujala* et al. *[52]*

Les travaux de Ahonen *et al.* [1] ont montré que l'utilisation de descripteurs LBP permettait d'obtenir de bonnes performances en reconnaissance de visage. Dans ces travaux, il est proposé de calculer un certain nombre de descripteurs locaux puis de les combiner pour obtenir un vecteur de description du visage. L'image originale est découpée en différentes régions. Les descripteurs associés à chacune de ces régions sont ensuite concaténés pour former un unique vecteur décrivant l'image.

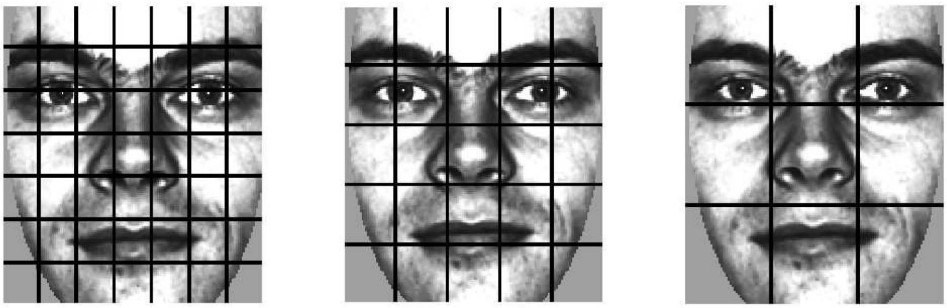


FIGURE 2.7 – *Découpage de l'image originale en* 7×7, 5×5 *et* 3×3 *régions [1]*

Ici encore, l'utilisation de descripteurs locaux permet de robustifier la reconnaissance faciale aux variations modérées de pose ou d'illumination.

2.3 Robustesse aux variations de pose et d'expression

Nous avons présenté, dans la partie précédente, un ensemble de méthodes générales de reconnaissance faciale. Ces méthodes ont été développées dans le but d'obtenir des performances optimales sur des images acquises dans de bonnes conditions. Cependant, un certain nombre de limitations à ces méthodes ont été soulevées. Ces limitations affectant les performances biométriques sont, comme décrit par Bledsoe dès 1966, les variations d'illuminations, de pose et d'expression.

Dans cette thèse, nous nous concentrons sur les deux dernières sources de variations. Nous proposons donc dans cette partie de présenter les principaux travaux traitant des variations de pose et d'expression.

2.3.1 Robustesse aux variations de pose

De nombreux travaux ont eu pour but d'améliorer la robustesse de la reconnaissance aux variations de pose. Celles-ci peuvent être effectuées selon trois axes (Figure 2.8).

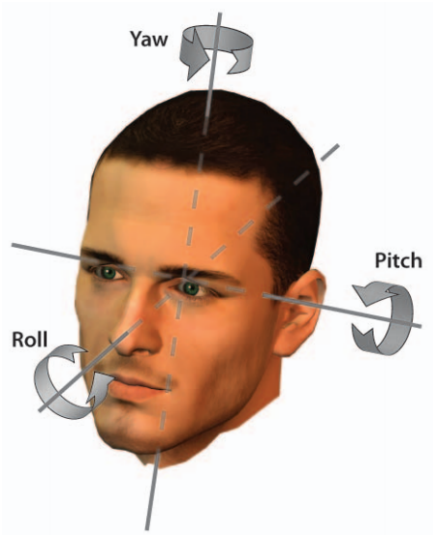


FIGURE 2.8 – *Axes de rotation du visage : Yaw, Pitch et Roll*

Les variations en roll du visage pouvant être corrigées par une rotation dans le plan, seules les variations en yaw et en pitch sont traitées.

L'approche naturelle et vraisemblablement la plus simple consiste à capturer pour chaque individu une collection d'images acquises sous des poses variées. Les premiers travaux traitant de cette problématique ont été effectués par Beymer *et al.* [13]. Dans ce travail, chaque individu est représenté dans la galerie par une quinzaine de photos acquises sous des poses variées dans l'intervalle ± 40 °en yaw et ± 20 °en pitch. La comparaison est ensuite faite en comparant le visage avec l'image de la galerie proposant une pose similaire.

FIGURE 2.9 – *Exemples de vues utilisées dans [13] pour représenter le visage d'un individu*

La méthode proposée par Georghiades *et al.* [29] suit une approche similaire. Les auteurs y présentent une méthode permettant d'extrapoler, à partir d'un nombre limité d'images (sept images sont utilisées dans ce travail), l'apparence du visage obtenue dans des conditions d'acquisition extrême (variations d'illuminations ou de pose). La figure 2.10 montre des exemples d'images synthétiques générées par cette méthode.

FIGURE 2.10 – *Exemples de vues utilisées dans [13] pour représenter le visage d'un individu*

Bien que proposant des résultats intéressants, ces méthode possèdent une limitation importante. En effet, elles nécessitent une coopération de l'utilisateur dont un ensemble d'acquisitions doit être effectué.

D'autres méthodes ont alors été proposées afin d'être utilisées dans des scénarios où seule une image par individu est disponible dans la galerie. L'approche utilisée consiste donc en la génération de nouvelles vues synthétiques à partir d'une unique image. Beymer *et al.* [12] ont notamment proposée une extension de leur précédente méthode leur permettant d'être utilisée dans ce cas. Les auteurs proposent ici de générer à partir d'une seule acquisition I_0, quatorze nouvelles vues synthétiques $I_1 \dots I_{14}$ leur permettant de se rapprocher du contexte de la méthode précédente. Pour chacune de ces nouvelles configurations, un ensemble de transformations T_i est appris des données d'apprentissage $I_{training}$ puis est transféré à l'ensemble des pixels de l'image originale I_0 (Figure 2.11).

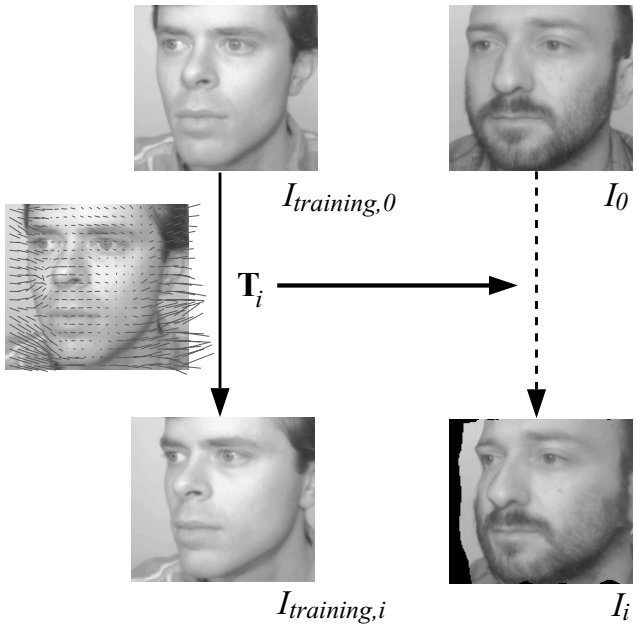


FIGURE 2.11 – *Processus de synthèse de nouvelle vue proposée par Beymer* et al. *[12]*

D'autres méthodes permettent d'effectuer cette synthèse de nouvelles vues grâce à l'utilisation d'un modèle actif d'apparence [21]. L'ensemble de ces méthodes [22] [38] traite alors les problèmes liés à la pose en suivant une approche en deux dimensions. Malgré les résultats encourageants qu'elles présentent, elles ne permettent qu'une correction des poses modérées. Le visage étant un objet tri-dimensionnel, l'utilisation d'un modèle 3D constitue une seconde approche suivie dans de nombreux travaux.

Ainsi, Asthana *et al.* [6] proposent d'effectuer cette correction de pose à l'aide d'un modèle 3D moyen de visage. Une douzaine de points caractéristiques permettent d'ajuster ce modèle à l'image originale. La pose et les informations de texture peuvent alors en être extraites. Une nouvelle vue synthétique peut ensuite être générée en utilisant la forme moyenne de visage, les informations de texture extraites et des paramètres de pose correspondant à une pose frontale (Figure 2.12).

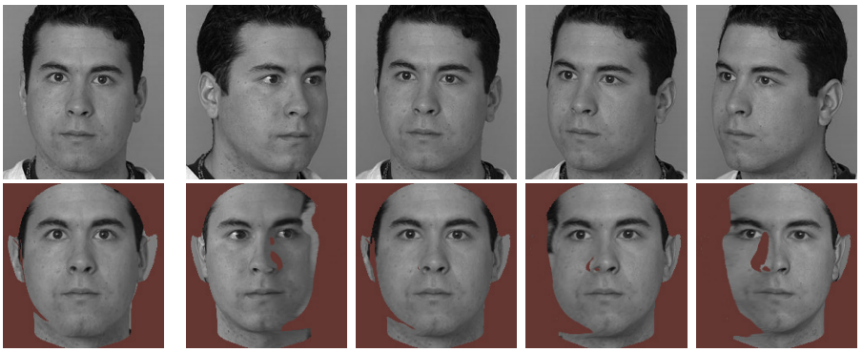


FIGURE 2.12 – *Exemples de corrections de pose effectuées parAsthana* et al. *[6]*

Toutefois, l'utilisation d'un modèle moyen ne permet pas une approximation optimale de la forme 3D du visage. Celle-ci peut être obtenue à

l'aide d'un modèle 3D générateur de visages [16].

Basé sur l'analyse statistique d'un large jeu de données d'apprentissage de visage 3D, ce modèle permet d'approcher la forme du visage de n'importe quel individu à partir d'une combinaison linéaire d'un nombre limité de déformations. A la suite de cette approximation, une nouvelle vue synthétique de l'image originale peut alors être générée en utilisant la forme 3D estimée et de nouveaux paramètres de pose (Blanz *et al.* [15]). L'efficacité de cette méthode, appliquée en tant que pré-traitement d'algorithmes commerciaux de reconnaissance de visage, a été prouvée lors du benchmark effectué par le NIST au début des années 2000 [32]. Lors de cette évaluation, il a été montré que le taux de vérification obtenu lors de la comparaison de visages pris à 45° avec des visages frontaux est passé de 17% à 79% pour un taux de fausses acceptances de 1%.

Une autre catégorie de méthodes propose de suivre une approche statistique permettant de modéliser la façon dont l'apparence du visage change en fonction de la pose. Parmi ceux-ci, Kanade *et al.* [40] définissent une notion de discriminance des zones du visage en fonction de la pose. Pour chaque pose, une analyse de la distribution de la similarité intra-individu et extra-individu de chacune des zones du visage est effectuée. Cet apprentissage est effectué sur une base de 884 images (68 individus acquis sous 13 poses différentes). Cette analyse statistique permet alors de déterminer le pouvoir discriminant de différentes zones pour une pose donnée (Plus les distributions intra-individu et extra-individu sont séparées, plus la zone est discriminante).

La figure 2.13 montre un exemple de carte présentant le pouvoir discriminant des différentes zones du visage en fonction de la pose. Plus une zone est claire, meilleur est son pouvoir discriminant. Cette zone aura donc un poids important dans le calcul de la similarité globale des visages. Au contraire, une zone foncée se verra apportée moins d'importance lors du calcul de similarité.

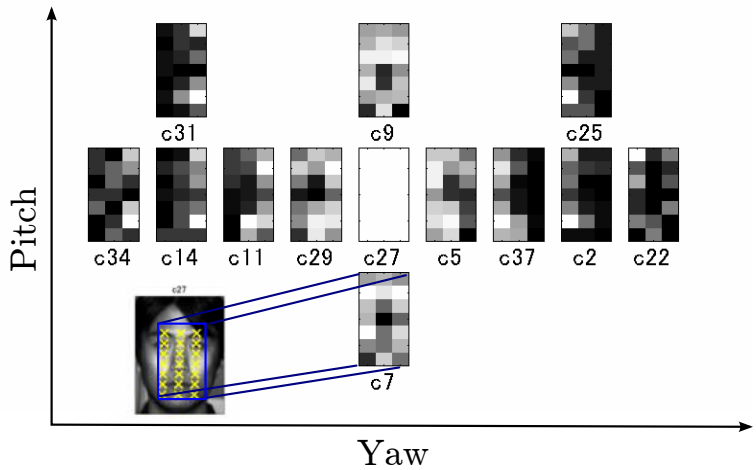


FIGURE 2.13 – *Carte du pouvoir discriminant des zones du visages en fonction de la pose [40]*

Cette présentation des principales méthodes traitant de la problématique de robustesse à la pose, a permis de déterminer plusieurs approches.

La première propose de représenter chaque individu dans la galerie par un ensemble d'images acquises sous différentes poses. Chaque visage à identifier est alors comparé contre l'ensemble de ces images. Cette approche nécessite toutefois d'effectuer plusieurs acquisitions lors de l'enregistrement d'un individu.

La seconde catégorie de méthode propose de s'affranchir de cette contrainte en générant une nouvelle vue de l'image de test avec une pose frontale. Cette correction peut être effectuée en deux ou en trois dimensions. L'utilisation d'un modèle 3D permet une meilleure robustesse aux variations de pose en comparaison des approches bi-dimensionnelles. Toutefois, pour assurer une généricité importante, la construction de ce modèle 3D nécessite un nombre important de scans 3D de visage.

2.3.2 Robustesse aux variations d'expression

La problématique de robustesse aux variations d'expression a fait l'objet d'un nombre plus restreint de travaux. Ces travaux reposent sur une approche commune.

La présence d'expression sur une visage provoque sur celui-ci un ensemble limité de déformations locales d'apparence. Très localisées, ces modifications n'impactent qu'une partie limitée du visage. La comparaison du visage est alors effectuée en traitant ces zones de manière similaire aux zones occultées. Pour rendre robuste la reconnaissance à ces artefacts d'apparence, la plupart des méthodes proposent une comparaison locale durant laquelle certaines zones du visage sont ignorées.

Pour cela, Wei *et al.* [67] suivent une approche semblable à celle du système visuel humain. Celui-ci analyse les visages en se fixant successivement sur des parties précises du visage. Les auteurs proposent ainsi de découper le visage en différentes régions. Une comparaison par patchs (en utilisant la méthode DICW [68]) est alors effectuée entre ces régions et celles extraites des images de galerie. Une décision par un vote à la majorité est alors utilisée pour déterminer l'identité de l'individu.

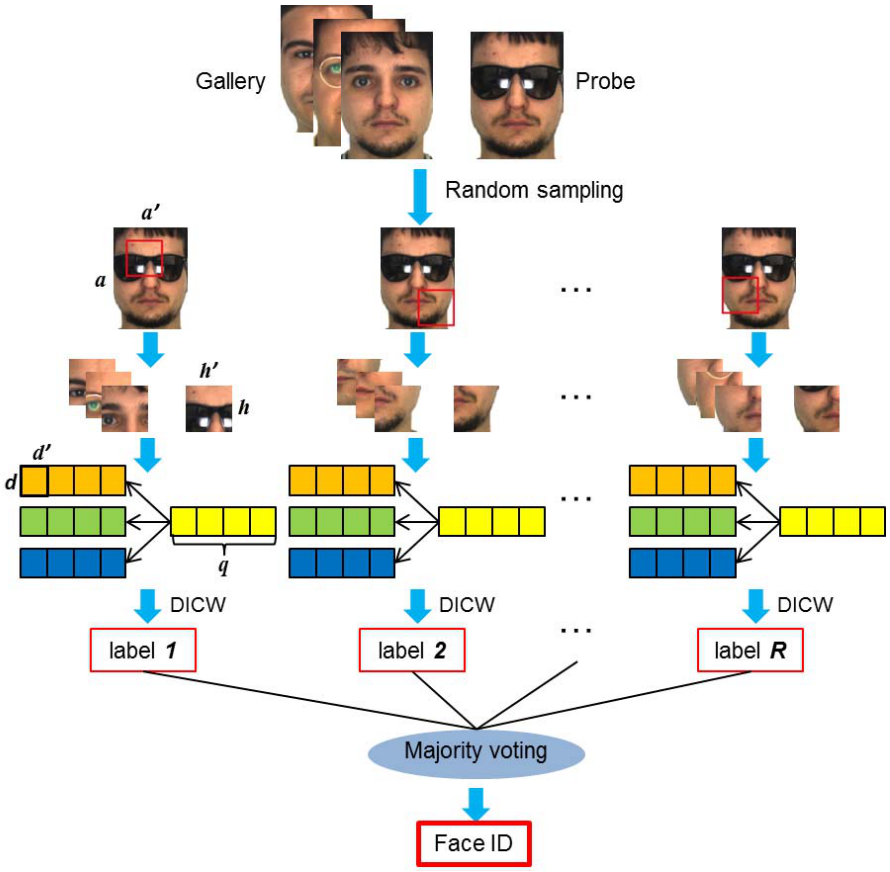


FIGURE 2.14 – *Processus de la méthode proposée par Wei* et al. *[68]*

L'utilisation d'un nombre important de régions permet une grande robustesse aux artefacts locaux provoqués par des occultations ou par des expressions.

Une seconde méthode, proposée par Tan *et al.* [59], suit une approche similaire en introduisant une notion de similarité locale permettant de focaliser le processus de reconnaissance sur les parties semblables du visage. Pour cela, l'image est découpée en blocs de petite taille, chacun d'entre eux étant représenté par un descripteur basique (niveau de gris du bloc). Un seuil de similarité est alors utilisé pour déterminer la ressemblance entre deux blocs. Ainsi, le calcul de similarité globale entre deux images peut être effectué en excluant les parties dont la ressemblance n'est pas suffisante. Une grande robustesse aux occultations et aux variations d'expression peut ainsi être obtenue.

D'autres méthodes de l'état proposent quant à elles de suivre une approche basée sur une représentation parcimonieuse du visage [71]. Ce type de représentation permet alors d'obtenir une représentation compacte du visage en éliminant les parties non informatives du visage afin d'augmenter la robustesse de la reconnaissance aux problèmes d'occultations (Figure 2.15).

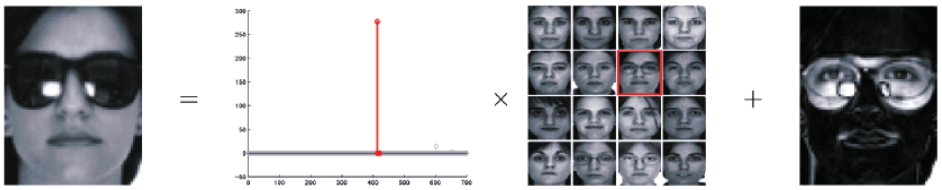


FIGURE 2.15 – *Représentation parcimonieuse du visage proposée par Wright* et al. *[71]*

Les artefacts liés à la présence d'expression pouvant être traités comme des problèmes d'occultation, les auteurs proposent d'étendre leur méthode à cette problématique en présentant des résultats expérimentaux sur des images avec des variations d'expression.

L'ensemble des travaux présentés dans cette section propose des méthodes permettant d'améliorer la robustesse de la reconnaissance faciale aux variations d'expression. Toutes ces méthodes utilisent le fait que les modifications d'apparence liées à la présence d'expression se situent dans des zones particulières du visage. Bien que permettant une bonne robustesse aux variations d'expression, ces méthodes ne permettent pas de traiter des variations simultanées d'expression et de pose.

2.3.3 Variations simultanées de pose et d'expression

Les deux sections précédentes nous ont permis de présenter différents travaux traitant, d'une part, de la robustesse aux variations de pose et, d'autre part, de la robustesse aux variations d'expression. Dans cette thèse, nous concentrerons notre travail sur la problématique de variations simultanées de pose et d'expression. La littérature propose un état de l'art beaucoup plus limité et très récent sur cette problématique.

Une des approches notables relative à cette problématique a été proposée par Berg *et al.* [9]. Dans cette méthode, l'image originale est déformée pour obtenir une nouvelle vue frontale et neutre. Un ensemble de 95 points caractéristiques du visage est utilisé pour effectuer cette déformation. La position de chacun de ces points est ensuite modifiée pour correspondre à une vue frontale et neutre tout en conservant l'identité du visage (Figure 2.17).

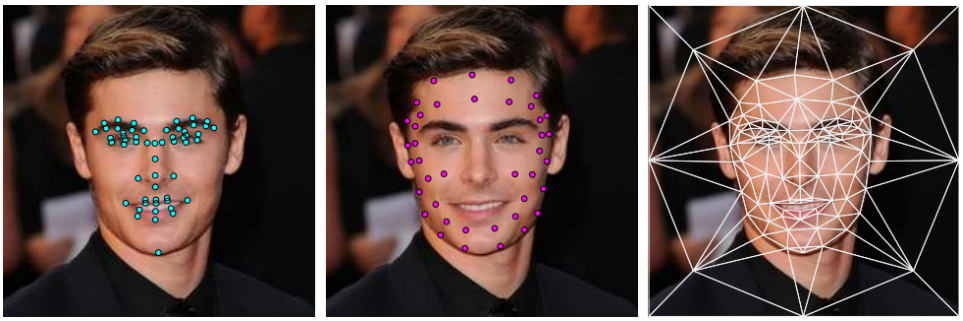


FIGURE 2.16 – *Triangulation du visage à l'aide de 95 points caractéristiques (Berg* et al. *[9])*

Cette méthode repose sur une triangulation de Delaunay de l'image dont la position de chacun des triangles est modifiée, lors de l'alignement,

par une transformation affine vers une position canonique.

Les auteurs proposent d'adapter cette position canonique à chaque individu afin de conserver les différences d'apparence relatives à l'identité de la personne. En effet, utiliser la même forme canonique pour tous les individus entraînerait un "sur-alignement". Pour éviter ce phénomène, une base de référence constituée de 20639 images provenant de 120 individus différents est utilisée pour extraire, pour chaque image à aligner, la triangulation correspondant à un individu moyen sous la même pose et la même expression.

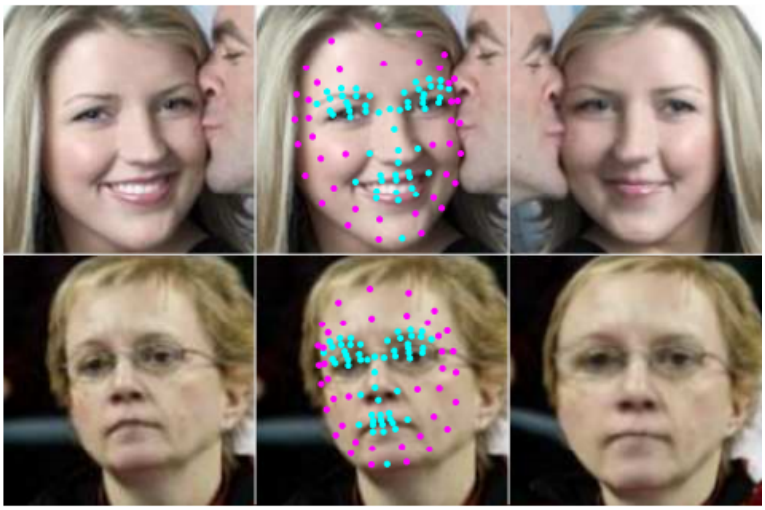


Image originale | Détéction | Image normalisée

FIGURE 2.17 – *Correction simultanée de la pose et de l'expression (Berg* et al. *[9])*

Bien que présentant des résultats intéressants, cette méthode possède une limitation importante. Basée sur une approche bi-dimensionnelle, son application reste limitée aux images présentant une pose modérée.

CONCLUSION DU CHAPITRE

Nous avons présenté au sein de ce chapitre, les principaux travaux traitant de la reconnaissance faciale. L'objectif de cette présentation de l'état de l'art est non pas de lister l'ensemble des méthodes proposées par la communauté, mais de présenter les principaux travaux et d'en déceler les principales limitations.

De nombreuses méthodes générales de reconnaissances faciales (holistiques ou locales) ont été proposées depuis plusieurs décennies. Plus récemment, les travaux présentés au cours de ces dernières ont eu pour objectif d'améliorer les performances obtenues en conditions dégradées.

Différentes méthodes ont ainsi été proposées pour améliorer la robus-

tesse aux variations de pose ou d'expression Ces méthodes peuvent être classées en deux catégories.

La première regroupe les méthodes proposant de traiter de ces problématiques à l'aide de nouveaux algorithmes. Bien que permettant d'obtenir des résultats intéressant, ces méthodes [40] [67] [59] sont généralement proposée dans le but de répondre à une problématique précise.

La seconde catégorie regroupe les méthodes basées sur une approche de type "pré-traitement". Ces méthodes [22] [38] [6] [15] [9] proposent d'augmenter les taux de reconnaissance obtenus avec des algorithmes standards de reconnaissance de visage en améliorant la qualité d'alignement des visages. Indépendantes de l'algorithme de reconnaissance utilisé, ces méthodes permettent une utilisation beaucoup plus pérenne dans le temps.

Cette deuxième approche constitue un réel avantage dans des contextes industriels, tel que celui dans lequel s'est déroulée cette thèse. En effet, cela permet à la fois de capitaliser sur les algorithmes actuels, fruit de plusieurs années-ingénieur de recherche, et d'assurer une solution pérenne dans le temps, en vue d'une utilisation future en amont des prochaines générations d'algorithmes de reconnaissance de visage.

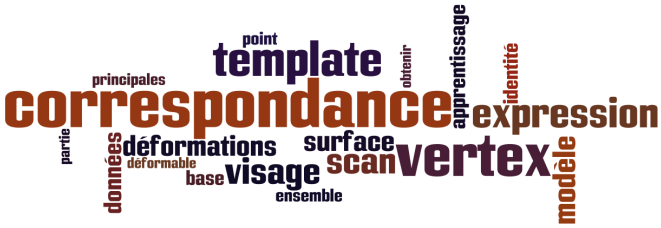
Dans notre thèse, nous traitons le problème des variations simultanées de l'expression et de la pose. Des bases de données telles que CMU Multi-PIE [31] proposent une collection d'images avec des variations combinées d'expression et de pose. Malgré l'existence de telles bases de données, cette problématique n'a été que très peu traitée de l'état de l'art de la reconnaissance de visage en deux dimensions.

Nous plaçons donc notre thèse dans ce contexte en proposant une approche novatrice permettant de robustifier les algorithmes actuels de reconnaissance faciale aux variations simultanées de pose et d'expression. De manière similaire à certains travaux effectués pour corriger les variations de pose [15], notre méthode proposera de générer une nouvelle vue synthétique du visage à l'aide d'un modèle 3D déformable étendu de visage. Ce modèle, étendu aux variations d'expression, permettra de corriger simultanément la pose et l'expression de l'individu. Ainsi, une amélioration significative des performances obtenues dans le scénario étudié pourra être obtenue tout en assurant une stabilité des résultats dans le cas d'images acquises dans de bonnes conditions.

Modèle générateur 3D de visage

3

Sommaire



Dans ce chapitre, nous présenterons en détail le modèle générateur 3D de visages utilisé dans cette thèse. L'état de l'art nous a montré qu'une partie des méthodes permettant de robustifier la reconnaissance de visages aux variations de poses était basée sur un modèle statistique de visages 3D. Une fois ce modèle déformé pour correspondre au mieux à l'image 2D originale, une nouvelle vue frontale est synthétisée. Ce prétraitement permet d'obtenir une nouvelle vue du visage dans des conditions pour lesquelles les performances des algorithmes standards de reconnaissance faciale sont optimales. L'ensemble des étapes permettant la construction de ce modèle déformable de visages seront présentées dans ce chapitre.

3.1 Données d'apprentissage et mise en correspondance

Le visage humain est un objet 3D dont la surface peut être représentée par un mesh. Un mesh est constitué de sommets (appelés vertex dans la suite), d'arêtes et de faces. Ces faces sont généralement des triangles ou des quadrangles. Dans la suite, nous utiliserons des faces triangulaires.

L'estimation de la surface 3D du visage à partir d'une seule image est un problème largement sous-déterminé. L'utilisation d'une connaissance *a priori* est alors nécessaire. Ici, la surface à approcher est celle d'un visage humain. L'utilisation d'un modèle 3D déformable de visage permet d'apporter cette information lors du processus d'approximation de la surface.

Un modèle déformable 3D de visage permet d'approcher la forme du visage de n'importe quel individu à partir d'un nombre limité de déformations. L'approximation de la surface consiste donc à trouver la combinaison linéaire de ces déformations permettant d'obtenir la forme 3D la plus proche de celle du visage présent dans l'image.

Pour garantir que les visages ainsi formés restent réalistes, l'espace des déformations autorisées doit être contraint. Ces contraintes peuvent être définies de différentes façons :

D'un coté, l'utilisation d'un modèle anatomique permet de limiter le déplacement des vertex. Ce type de modèle empêche les combinaisons de déformations entrainant un positionnement des vertex irréalistes. En effet, tous les vertex du visage ne peuvent être déplacés dans n'importe quelle direction. Par exemple, les vertex associés au crâne doivent conserver, à la suite de n'importe quelle combinaison de déformations, une forme ovoïde. L'utilisation d'un modèle anatomique permet d'obtenir des résultats très inintéressants. Ils sont par exemple couramment utilisés dans le domaine de l'animation 3D [72].

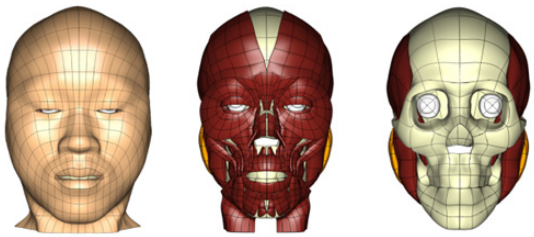


Figure 3.1 – *Modèle anatomique de visage [72]*

D'un autre côté, ces contraintes peuvent être déduites d'une analyse statistique effectuée sur une collection de scans 3D de visage. L'objectif de cette analyse statistique est de caractériser l'ensemble des déformations des vertex. En faisant l'hypothèse d'une distribution normale des visages 3D, une analyse en composantes principales permet de définir l'espace des déformations. Ainsi, cette méthode permet d'obtenir une représentation compacte de l'ensemble des visages : N'importe quel visage humain peut alors être approché par un petit jeu de coefficients. Ces coefficients

définissent alors la combinaison linéaire des déformations appliquées au modèle moyen. La qualité de ce modèle statistique est fortement dépendant de la qualité des données d'apprentissage.

3.1.1 Visages 3D d'apprentissage

Pour garantir un modèle de bonne qualité, il est primordial que le jeu de données d'apprentissage réponde à un certain nombre de critères. Ces données d'apprentissage doivent notamment être d'une grande variété : Pour être capable de générer des visages en contrôlant séparément l'identité et l'expression, le jeu de données d'apprentissage doit être à la fois composé de visages neutres (pour déterminer l'espace des déformations d'identité) et de visages avec expression (pour déterminer l'espace des déformations d'expression).

La constitution de la base d'apprentissage peut être fait de différentes façon : Soit en organisant une campagne d'acquisition de visages 3D ou bien en utilisant des bases de données existantes. Effectuer nous même l'acquisition des scans 3D nous permet de définir précisément les caractéristiques de la base (Nombre d'individus, type d'acquisition, nombre d'expressions, ...). Cependant, la construction d'une telle base est coûteuse (tant en termes de matériel que de temps). Nous avons donc décidé de nous orienter vers une des nombreuses bases de données publiques proposant une collection importante de visages 3D. Parmi elles, nous pouvons citer les bases BU-3DFE [74], Bosphorus [55] ou bien encore D3DFacs [23].

Afin d'obtenir le modèle le plus général possible, la base d'apprentissage doit contenir un nombre important d'individus (avec une diversité importante d'éthnicité). Bien que disposant de scans 3D en haute qualité (en moyenne 30 000 vertex par scan) et d'une grande variété intra-identité (en moyenne 52 séquences acquises par individu), la base D3DFacs [23] n'est composée que de 10 individus. Cette limitation ne lui permet pas d'être utilisée pour l'apprentissage de notre modèle statistique.

FIGURE 3.2 – *Scans 3D de la base Dynamic 3D Facial Action Coding System Database*

Composée de plus de 4000 scans 3D, la base Bosphorus [55] répond parfaitement à la contrainte de généricité du modèle 3D. Cette base de données propose des scans 3D de 105 individus avec jusqu'à 35 acquisitions différentes par individu. La totalité de ces scans a été nettoyée par un

ensemble de pré-traitements. Un détourage de la partie frontale du visage est notamment effectué.

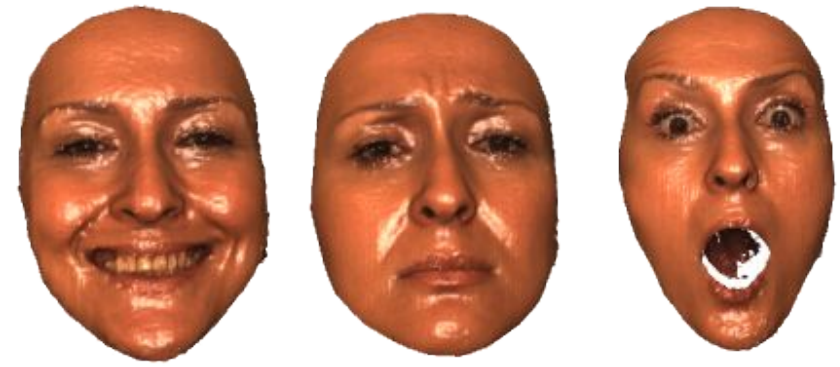


FIGURE 3.3 – *Scans 3D de la base Bosphorus*

Pour obtenir un modèle déformable précis, il convient d'utiliser des scans avec une partie utile la plus grande possible. Les parties du visage supprimées par le détourage des scans de la base Bosphorus sont autant d'informations perdues lors de l'analyse statistique. Nous avons donc choisi de ne pas utiliser cette base pour l'apprentissage du modèle déformable 3D.

La troisième base de données fréquemment utilisée par la communauté est la base BU3D-FE [74]. Cette base est composée de scans 3D issus de 100 individus (56 femmes et 44 hommes) âgés de 18 à 70 ans avec une variété ethnique importante. Le visage de chaque individu a été acquis sous différentes expressions (avec quatre niveaux d'intensité pour chaque expression). Les expressions disponibles sont l'expression neutre et les six expressions prototypiques. Ces six expressions de base sont la joie, la peur, le dégout, la surprise, la colère et la tristesse.

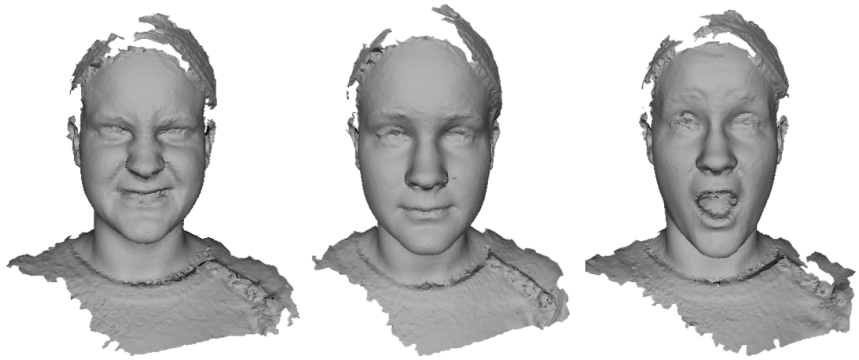


FIGURE 3.4 – *Scans 3D de la bases BU3D-FE*

Nous avons donc choisi d'utiliser cette base de données pour construire la base de données d'apprentissage du modèle déformable. Sept scans par individus ont été conservés : un scan du visage avec l'expression neutre et six scans avec expressions (Pour chacune d'elle, seul le scan avec l'intensité maximale est conservé).

Le jeu de données d'apprentissage est ainsi composé de 700 scans 3D.

3.1.2 Mise en correspondance des visages 3D

Le modèle 3D déformable de visage est une représentation de l'espace des visages humains. Il permet de générer n'importe quel visage à partir d'une combinaison linéaire d'un nombre limité de déformations. Une analyse en composantes principales du jeu de données d'apprentissage permet de calculer l'espace de ces déformations. En acceptant l'hypothèse de normalité des visages 3D, cette analyse statistique permet de déterminer la distribution des données. Ainsi, le modèle obtenu sera *générique* (La forme de n'importe quel visage pourra être approchée par ce modèle) et *restrictif* (N'importe quelle forme générée par ce modèle peut être un visage rencontré dans la réalité).

Pour que cette notion de déformation puisse avoir un sens, l'ensemble des objets 3D d'apprentissage doivent partager une définition commune de l'indexation de leurs sommets. Cette stabilité entre les sommets des données d'apprentissage est obtenue par une opération de mise en correspondance de surface 3D.

Le mise en correspondance de deux objets 3D est basée sur le principe suivant. L'un des deux objets, dont la topologie géométrique sera celle de référence, appelé gabarit ou *template* et noté T, est déformé afin de correspondre au mieux à la surface 3D notée S du second objet 3D. A l'issue de cette opération, le nouvel objet 3D obtenu possède les caractéristiques suivantes :

- Sa surface 3D est quasiment identique à S
- L'indexation de ces sommets est identique à celle ci T.

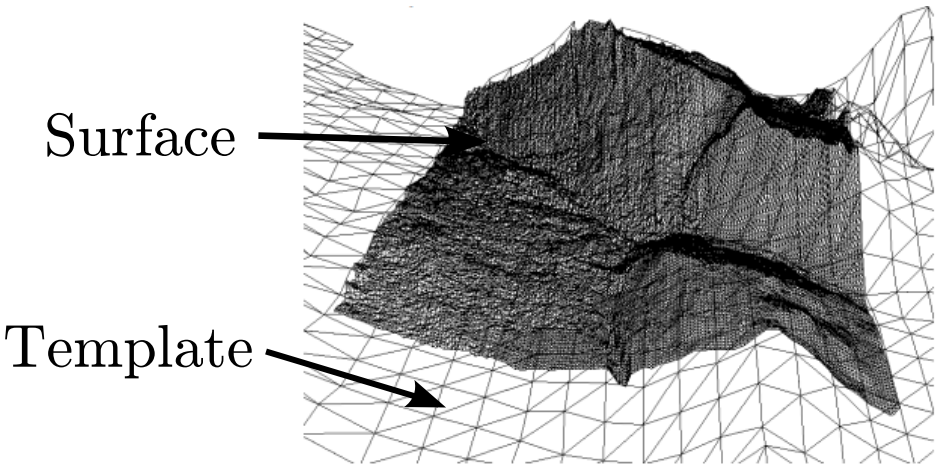


FIGURE 3.5 – *Exemple de mise en correspondance d'une carte élévation [37]*

Au cours de ce processus, chaque sommet du template est déplacé pour correspondre au mieux à la surface S. De plus, une contrainte de régularisation doit être utilisée pour minimiser l'ensemble de ces déformations et ainsi conserver l'aspect du maillage du template.

Lors de ce processus de mise en correspondance, chaque vertex v_i du template est déformé par une transformation notée T_i. La partie affine de

cette transformation est représentée par la matrice A_i de taille 3×3, tandis que la translation est représentée par t_i :

$$T_i : \begin{cases} \mathbb{R}^3 \rightarrow \mathbb{R}^3 \\ v_i \mapsto A_i v_i + t_i \end{cases} \quad (3.1)$$

L'utilisation de cette représentation affine des transformations permettra le calcul de l'énergie de régularisation. Cette énergie, basée sur la minimisation du laplacien des transformations, sera présentée dans un prochain paragraphe.

Energie d'erreur aux données

L'objectif du processus de mise en correspondance est donc de chercher l'ensemble des transformations T_i permettant de minimiser l'erreur de correspondance entre les sommets du template et la surface 3D. Cette erreur, notée E_{data}, est définie comme la somme des distances au carré de chaque sommet du template au point de la surface le plus proche :

$$E_{data} = \sum_{i=1}^{n_{ver}} dist^2 \left(T_i \left(v_i \right), S \right) \quad (3.2)$$

avec :

- v_i : Le vertex d'indice i du template.
- n_{ver} : Le nombre de vertex du template T.
- T_i : La transformation appliquée au vertex v_i du template.
- $dist(v, S)$: La fonction permettant de déterminer la distance entre le vertex v et la surface S.

La fonction $dist(v, S)$ permet de calculer la distance entre le vertex v et la surface S. Traditionnellement, ce type de fonction calcule la distance entre le vertex et le point le plus proche de la surface. Dans le cas d'une surface maillée, cette recherche est faite pour chacune des facettes et seule celle permettant une distance minimale est conservée.

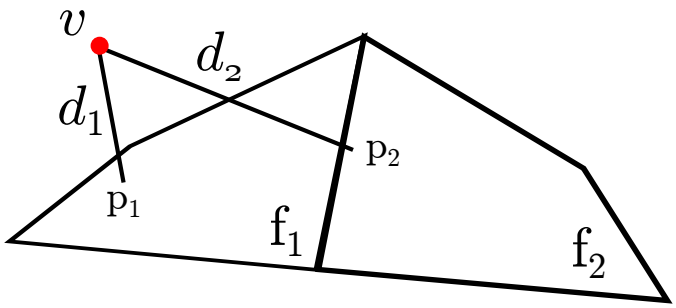


FIGURE 3.6 – *Recherche de correspondance point-surface : La distance d_1 du vertex v à la facette est f_1 est inférieure à d_2 (distance à la facette f_2). La correspondance est donc établie entre v et p_1.*

Dans notre contexte de mise en correspondance de visage 3D, il convient de tenir compte d'un certain nombre de contraintes afin d'empêcher des associations erronées entre un vertex v du template et un point

de la surface S.
Dans un premier temps, l'association n'est permise que si la distance d'association est inférieure à un seuil d_{thres}. Cette condition permet d'éviter une mise en correspondance entre deux parties différentes du visage. Pour robustifier cette étape d'association, une seconde contrainte basée sur la cohérence entre les normales des deux points associés est exigée. Cette contrainte permet par exemple d'empêcher la mise en correspondance d'un sommet de la narine droite avec un point de la narine gauche même si la distance d'appariement est inférieur au seuil d_{thres}.

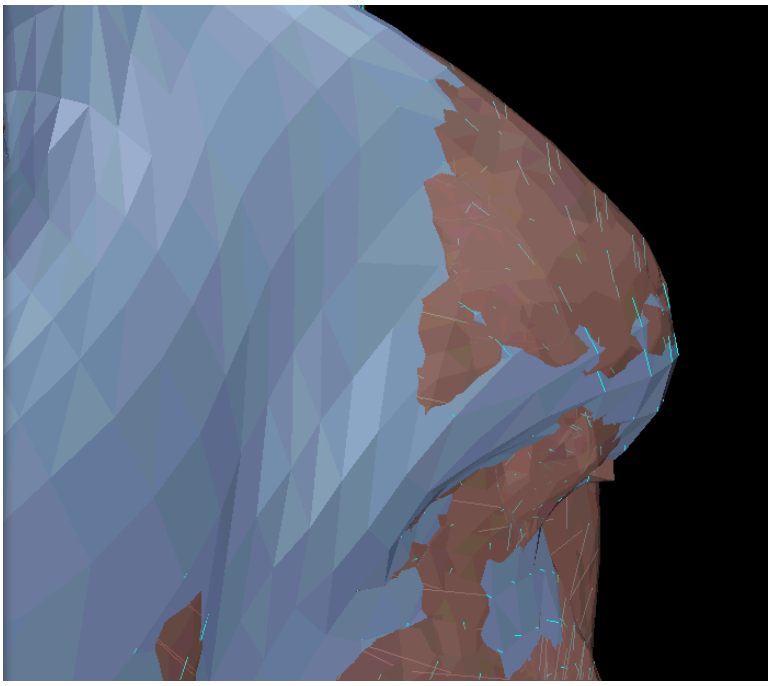


FIGURE 3.7 – *Les lignes turquoise représentent les appariements entre les sommets du template (En bleu) et la surface du scan (En rose)*

L'énergie d'erreur aux données (Equation 3.2) est alors calculée pour chacune des correspondances obtenues par ce processus d'appariement entre les sommets du template et la surface 3D. Malgré l'utilisation de ces contraintes, la qualité de ce processus n'est pas optimale. Celle-ci peut alors être améliorée en utilisant de nouvelles associations.

Une annotation manuelle des données d'apprentissage permet de définir la position exacte d'un certain nombre de points facilement identifiables [53]. Ceux-ci sont associés à un ensemble de sommets du template dont l'index est connu. Ainsi, cette vérité terrain permet d'assurer une initialisation correcte du processus de mise en correspondance. La figure 3.8 montre ces correspondances : Les points annotés sur le scan 3D sont en rouge tandis que leur correspondance sur le template sont représentées par des points violets.

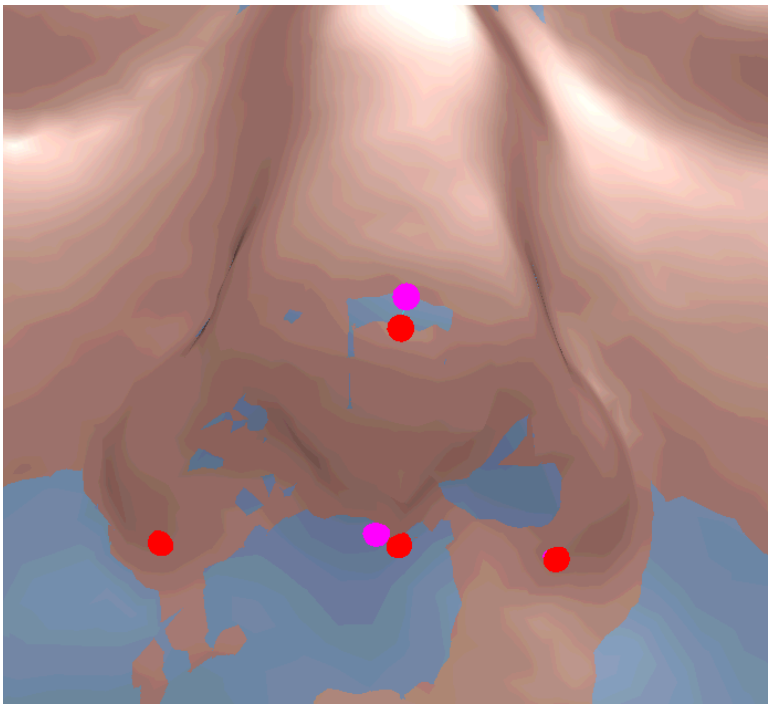


FIGURE 3.8 – *Exemple de correspondances entre points caractéristiques*

Dans ce travail, le processus de mise en correspondance est utilisé pour recaler des scans 3D de visage avec expression. La zone de la bouche est la partie du visage la plus fortement modifiée par l'expression. Il est donc nécessaire d'apporter une attention particulière à cette zone au cours du processus de recalage.

Nous proposons donc d'ajouter de nouvelles contraintes d'appariement de type point à courbe. De manière similaire aux points caractéristiques, une annotation manuelle du contour des lèvres est effectuée sur chacune des données d'apprentissage. Pour cela, l'utilisateur annote un ensemble de points régulièrement espacés sur celui-ci. Une approximation du contour par une courbe de Bézier en 3D peut alors être effectuée en utilisant ces annotations comme points de contrôle.

Lors de la mise en correspondance, chacun des vertex du template situé sur le contour des lèvres est ensuite associé à un point de la courbe 3D de Bézier annotée. Ce point est défini comme étant l'intersection entre le plan tangent au contour de la bouche du template et la courbe de Bézier annotée (Figure 3.9).

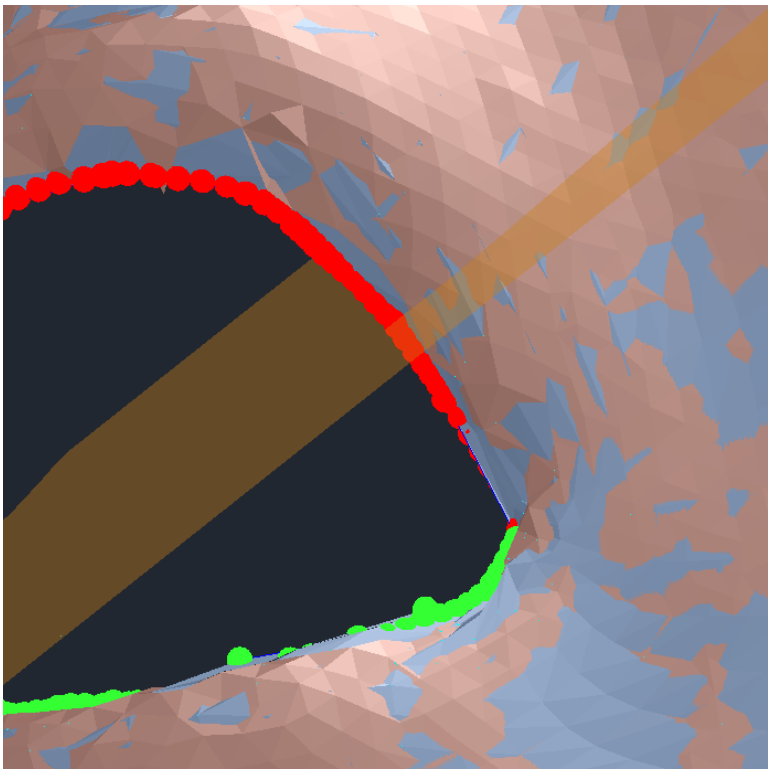


FIGURE 3.9 – *Correspondances des vertex de la bouche avec, en jaune, le plan tangent au contour de la bouche*

L'ensemble de ces nouveaux appariements de points sont ajoutés dans la fonction d'énergie d'erreur aux données E_{data} :

$$E_{data} = \sum_{i=1}^{n_{ver}} dist^2\left(T_i\left(v_i\right), S\right) + \sum_{v_i \in v_{mouth}} dist^2_{curve}\left(T_i\left(v_i\right), Bezier_{mouth}\right) + \sum_{v_i \in Fp} \left\| T_i\left(v_i\right) - Fp_i^{annot} \right\|^2 \tag{3.3}$$

avec :

- $dist_{curve}(v, C)$: La distance du vertex v à la courbe C.
- Fp : Ensemble des points caractéristiques du template
- Fp_i^{annot} : Coordonnées 3D de l'annotation manuelle du i^{me} point caractéristique.

Energie de régularisation

La seule utilisation des énergies présentées précédemment ne permet pas de résoudre le problème de mise en correspondance. En effet, il existe une infinité de transformations permettant de déformer un vertex 3D vers un autre. Le système est donc largement sous-déterminé. De nouvelles contraintes de régularisation doivent donc être rajoutées afin de contraindre le système. Les contraintes de régularisation couramment utilisées dans la littérature sont basées sur une conservation de la surface ou du volume. La grande variabilité dans la forme 3D des visages entre

différents individus rend impossible l'utilisation de ce type de contrainte.

Nous proposons donc d'utiliser ici, une contrainte permettant d'assurer une continuité entre les déformations. Allen *et al.* [3] proposent une énergie permettant de limiter les variations entre deux transformations voisines. Cette énergie est définie ainsi :

$$E_{regularisation} = \sum_{i,j|(v_i,v_j)\in Edges(T)} \left\|\mathbf{T}_i - \mathbf{T}_j\right\|^2 \tag{3.4}$$

Ainsi, les déformations appliquées à deux sommets voisins ne présentent pas de variations trop importantes. Une variante a été proposée par Amberg [4] lors de ses travaux de mise en correspondance de visage 3D. Il propose de minimiser le laplacien des transformations sur la surface :

$$E_{smoothness} = \int_S \left\|\nabla^2 T(v)\right\| \tag{3.5}$$

où $T(v)$ est la transformation appliquée au vertex 3D v.

Cette méthode permet de conserver la forme globale du template et ainsi de s'assurer que le résultat de la mise en correspondance conserve un aspect de visage 3D.

Cette contrainte est d'autant plus importante qu'une grande partie des sommets du template n'est pas associée à un point de la surface 3D. En effet, les données d'apprentissage ne contiennent pas de scans complets du visage : Seules les données de la partie frontale du visage sont acquises (Figure 3.10).

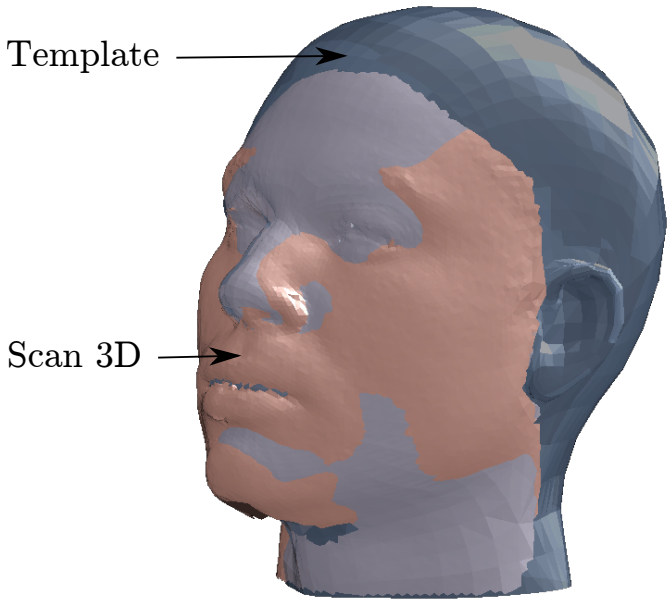


FIGURE 3.10 – *Recalage de scans incomplets : Le scan 3D est représenté en brun tandis que le template est affiché en bleu.*

De plus, pour garantir des déformations réalistes de cette zone, nous proposons d'ajouter un *a priori* rigide de forme en complétant le scan 3D

par une coque à l'arrière du scan 3D. L'ajout de cette coque est visible dans la figure 3.14.

L'ajout de l'énergie $E_{smoothness}$ (Equation 3.5) dans le processus de minimisation permet de conserver l'aspect général du visage, en particulier dans la partie antérieure du visage.

L'énergie globale à minimiser devient alors :

$$E(v) = E_{data}(v) + \alpha E_{smoothness}(v) \tag{3.6}$$

$$E_{data}(v) = \|Cv - c\|^2 \tag{3.7}$$

$$E_{smoothness}(v) = \left\| D \begin{bmatrix} \mathrm{v} \\ \mathrm{n(v)} \end{bmatrix} \right\|^2 \tag{3.8}$$

avec :

— v et $n(v)$ les vertex et les normales aux facettes du template.
— c précise les coordonnées 3D des points de la surface S en correspondance avec les sommets du template.
— C et D sont des matrices parcimonieuses dépendantes uniquement du template initial et de sa topologie.

Les normales $n(v)$ du template sont nécessaires lors du calcul de (3.8), entrainant la non linéarité de l'équation (3.6). Afin de rendre ce problème solvable linéairement, Amberg [4] propose d'introduire un nouveau jeu de données $\overline{n}$ relatif aux normales et une nouvelle fonction de coût permettant de rendre $\overline{n}$ et les normales à l'itération précédente $n(v^{t-1})$ aussi proches que possible. La nouvelle énergie à minimiser devient donc :

$$E(v) = \alpha E_{data}(v) + \alpha E_{smoothness}(v) + \beta E_{normal}(v, \overline{n}) \tag{3.9}$$

$$E_{data}(v) = \|Cv - c\|^2 \tag{3.10}$$

$$E_{smoothness}(v) = \left\| D \begin{bmatrix} v \\ \overline{n} \end{bmatrix} \right\|^2 \tag{3.11}$$

$$E_{normal}(v) = \left\| \overline{n} - n(v^{t-1}) \right\|^2 \tag{3.12}$$

Cette approximation permet de rendre l'équation (3.9) solvable linéairement. Une minimisation par algorithme de Gauss-Newton peut alors être utilisée pour résoudre ce problème efficacement.

L'ensemble du processus de mise en correspondance présenté dans cette partie est un **recalage flexible** : Chaque sommet v_i du template est déformé selon une transformation T_i qui lui est propre. Cette étape nécessite un premier alignement grossier entre le template et l'objet 3D : le **recalage rigide**. Au cours de cette étape, une unique transformation T est appliquée à tous les vertex du scan. Le but de cette étape est de trouver les paramètres de mise à l'échelle, de translation et de rotation permettant de faire correspondre au mieux le template et le scan 3D. La recherche de cette transformation est faite à partir des annotations manuelles de

certains points caractéristiques du visage.

La figure 3.11 illustre ces deux étapes de recalage rigide et flexible. L'entrelacement des deux objets 3D (Scan 3D et template) à l'issue du recalage flexible montre la précision du processus de mise en correspondance.

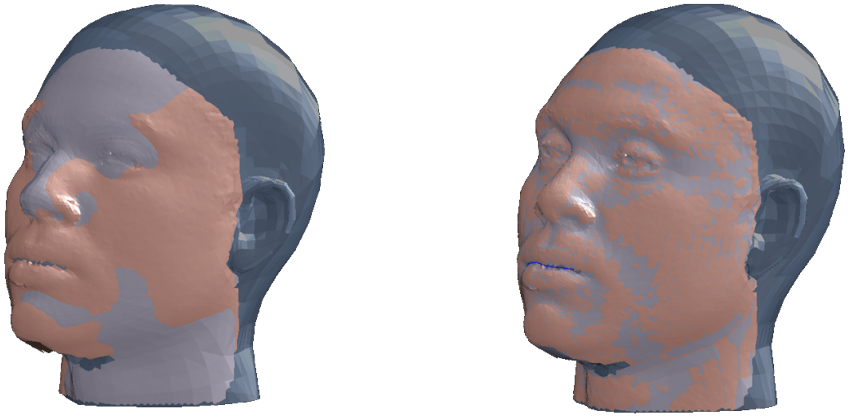


FIGURE 3.11 – *Recalage rigide (à gauche) et recalage flexible (à droite)*

Etapes de mise en correspondance

Le base d'apprentissage, présentée au début de ce chapitre, est composée de scans 3D provenant de 100 individus. Pour chacun d'entre eux, sept scans (un avec l'expression neutre et six autres avec les expressions prototypiques) sont acquis. Etant donnée cette structure, nous proposons un processus de recalage en deux étapes.

Dans un premier temps, l'ensemble des scans neutres est mis en correspondance avec un template générique :

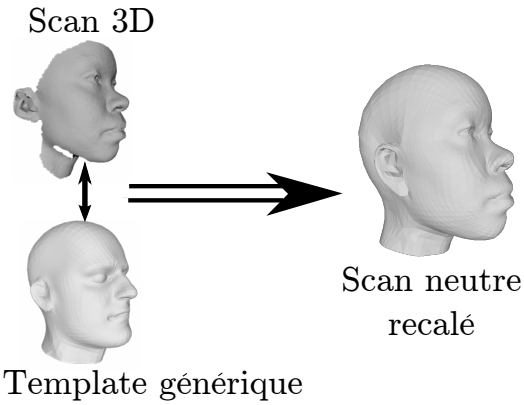


FIGURE 3.12 – *Mise en correspondance des scans neutres*

A l'issue de cette étape, une collection de scans neutres recalés est obtenue. Ceux-ci sont alors utilisés lors de la deuxième étape où chacun des scans avec expression est mis en correspondance avec le scan neutre recalé associé au même individu.

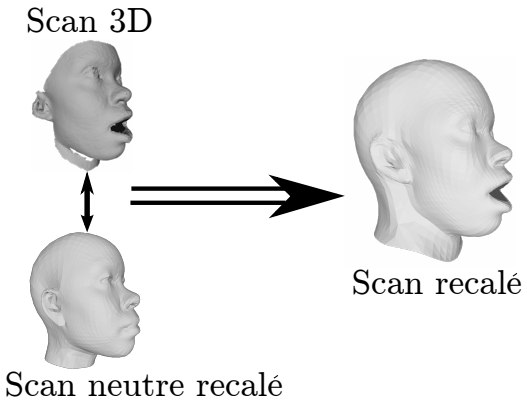


FIGURE 3.13 – *Mise en correspondance des scans avec expression*

3.1.3 Résultats de mise en correspondance

Dans cette partie, nous présentons les résultats de mise en correspondance des données d'apprentissage. La figure 3.14 montre des exemples de recalage du template (en bleu) sur des différents scans 3D complétés par l'*a priori* de forme sur la partie antérieure du visage (en rouge).

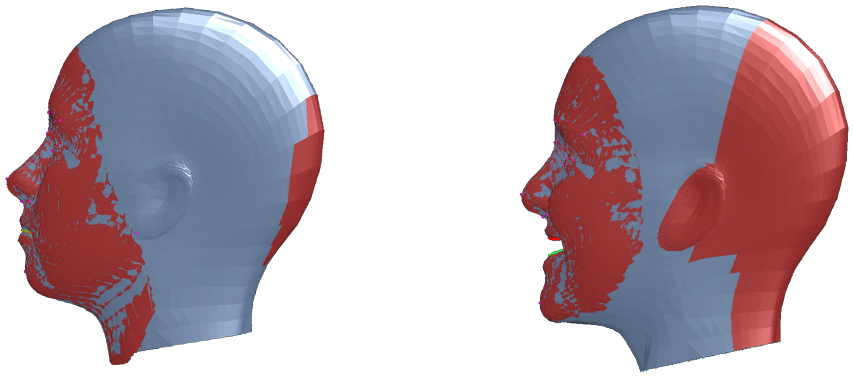


FIGURE 3.14 – *Résultats du processus de mise en correspondance*

La qualité de la mise en correspondance peut être évaluée de différentes façons. Dans un premier temps, l'erreur de mise en correspondance des points caractéristiques et l'erreur de mise en correspondance entre les sommets du template et la surface du scan 3D peuvent être utilisées comme métrique de qualité. Le tableau 3.1 présente la valeur de ces métriques obtenues sur les scans recalés présentés dans la figure 3.14.

Scans d'entrée	Erreur points caractéristiques	Erreur point/surface
Scan neutre	0.017%	0.061%
Scan expression dégout	0.016%	0.089%
Scan expression surprise	0.007%	0.052%

TABLE 3.1 – *Erreur de mise en correspondance donnée en pourcentage de la distance inter-oculaire*

Ces résultats montrent que le processus décrit dans ce chapitre permet une mise en correspondance de haute qualité. En effet, l'erreur moyenne

de recalage très faible montre un recalage quasi-parfait du template et la surface du scan d'entrée tant pour les scans neutres que pour les scans avec expression. L'utilisation du contour de la bouche pour ajouter de nouvelles correspondances permet cette grande robustesse aux expressions. En effet, la qualité du recalage au niveau de la bouche est très largement améliorée par cette contrainte d'ajustement courbe à courbe.

Une analyse visuelle permet également de valider le processus de mise en correspondance. En plus de regarder la différence entre le template et la surface 3D, l'analyse de caricatures peut permettre de déceler des artefacts de recalage tels que des problèmes de correspondance (par exemple, la mise en correspondance d'un vertex du nez avec la lèvre).

Les caricatures sont construites en exagérant le déplacement des vertex du template lors de la mise en correspondance. Les vertex v_{caric} de la caricature sont obtenus par l'équation suivante :

$$v_{caric} = v_{template} + coeff_{caric}(v_{registered} - v_{template}) \tag{3.13}$$

avec :

- $v_{registered}$ les vertex du template mis en correspondance avec le scan.
- $v_{template}$ les vertex du template à l'état initial.
- $coeff_{caric}$ le coefficient de caricature.

Différents types de caricatures peuvent être obtenus en faisant varier la valeur du coefficient $coeff_{caric}$:

- $coeff_{caric} = 0$: Le visage 3D obtenu est celui du template original.
- $coeff_{caric} = 1$: Le visage 3D obtenu est celui du template déformé sur le scan.
- $coeff_{caric} = -1$: Dans cette configuration, l'ensemble des vertex sont déplacés dans la direction opposée au recalage. Le visage obtenu est appelé "anti-face".
- $coeff_{caric} = 1$: Ici, le déplacement des vertex est exagéré. Chacun des vertex est déplacé du double par rapport au template original. Le visage ainsi obtenu est appelé "caricature"

Les caricatures obtenues dans la figure 3.15 ne montrent pas d'apparition d'artefacts pouvant par exemple être liés à des problèmes d'appariement.

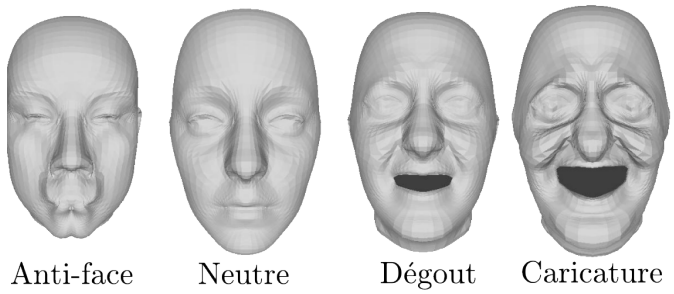


FIGURE 3.15 – *Caricatures du scan dégout mis en correspondance*

Les différents critères de qualité, utilisés dans cette partie, nous prouvent l'efficacité du processus de mise en correspondance.

Ce processus de mise en correspondance est appliqué à l'ensemble des scans 3D. Cela permet d'obtenir un jeu de visages 3D d'apprentissage recalés composé de 700 scans (7 scans avec variation d'expression pour chacune des 100 identités de la base BU-3DFE). Ces données d'apprentissage peuvent alors être utilisées pour déterminer l'espace des visages humains 3D grâce à une analyse en composantes principales.

3.2 Analyse statistique

La section précédente nous a permis d'obtenir un ensemble de 700 scans 3D partageant une définition commune des vertex. L'espace des visages 3D peut alors être estimé à l'aide d'une analyse en composantes principales.

Chaque visage 3D, noté $S_{i,e}$ est représenté par un vecteur $(X_1, Y_1, Z_1, X_2, \ldots, Y_{n_{vert}}, Z_{n_{vert}})$ de taille $3.n_{vert} \times 1$ où e indique l'expression (0 pour l'expression neutre et 1-6 pour les expressions) et $i \in \{1, \ldots, n_{id}\}$ fait référence à l'index de l'individu.

L'objectif de cette étape est de construire un modèle déformable de visages permettant de faire varier indépendamment l'identité et l'expression. Le jeu de données d'apprentissage est divisé en deux. Une analyse en composantes principales (ACP) est ensuite effectuée sur chacun des sous-sets [5].

Le premier sous-set est composé des scans neutres $S_{i,0}$. La forme moyenne de ces visages, notée $\overline{S}$, est définie par :

$$\overline{S} = \sum_{i=1}^{n_{id}} S_{i,0} \tag{3.14}$$

Une ACP sur les scans normalisés $S_{i,0} - \overline{S}$ de ce set permet alors d'extraire les déformations principales d'identité. Ces déformations sont stockées dans une matrice A_{id} de taille $n_{vert} \times (n_{id} - 1)$. Similairement au travail proposé par Blanz *et al.* [16], de nouveaux visages 3D neutres peuvent ensuite être générés en modifiant le vecteur de coefficients d'identité α_{id} de taille $(n_{id} - 1) \times 1$:

$$S = \overline{S} + A_{id}\alpha_{id} \tag{3.15}$$

Le second sous-set permet d'extraire les déformations principales liées aux variations d'expression. Ce set est composé, pour chacun des individus, des offsets entre les 6 scans avec expression et le scan avec l'expression neutre. Ces offsets sont notés par :

$$\Delta S_{id,exp} = S_{id,exp} - S_{id,0} \text{ pour exp } \in \{1 \ldots 6\} \tag{3.16}$$

L'ACP est effectuée sur l'ensemble des déformations d'expressions $\Delta S_{id,exp}$. Elle permet d'obtenir une matrice A_{exp}, de taille $n_{vert} \times (6n_{exp} - 1)$

contenant les déformations principales relatives à l'expression. Les déformations faciales dues à l'expression, ΔS_{exp}, peuvent alors être générées par :

$$\Delta S_{exp} = A_{exp}\alpha_{exp} \tag{3.17}$$

Les équations (3.15) et (3.17) peuvent être combinées pour obtenir le modèle déformable 3D de visages avec variations d'identité et d'expression :

$$S = \overline{S} + \underbrace{A_{id}\alpha_{id}}_{identite} + \underbrace{A_{exp}\alpha_{exp}}_{expression} \tag{3.18}$$

La formulation suivante sera adoptée dans la suite :

$$S = \overline{S} + \begin{bmatrix} A_{id} & A_{exp} \end{bmatrix} \begin{bmatrix} \alpha_{id} \\ \alpha_{exp} \end{bmatrix} \tag{3.19}$$

La figure 3.16 montre des exemples de visages 3D générés par ce modèle déformable. L'abscisse montre une variation des paramètres d'expression tandis que l'axe des ordonnées montrent l'influence des paramètres d'identité.

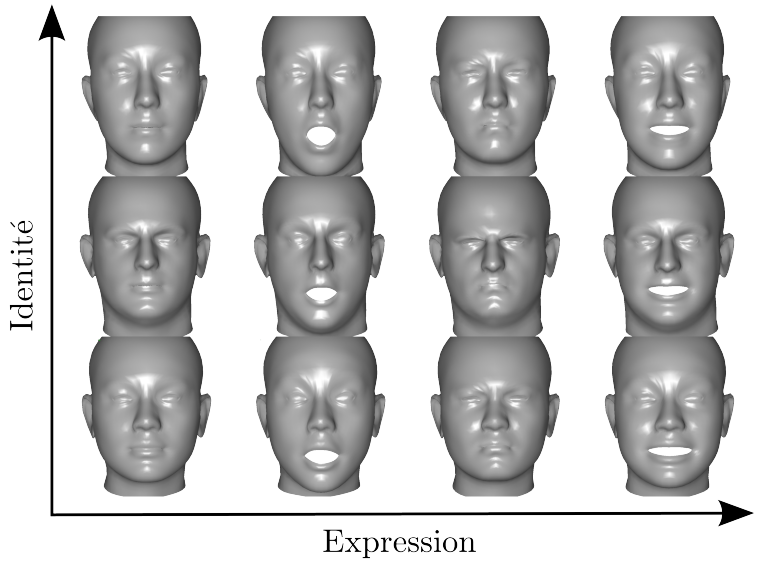


FIGURE 3.16 – *Exemples de visages 3D générés*

A l'issue de ces deux analyses en composantes principales, 99 déformations d'identité et 599 déformations d'expression sont extraites. Les figures 3.17 et 3.18 présentent l'énergie cumulée conservée par les déformations principales d'identité et d'expression. Cette énergie est définie comme étant la variance cumulée associée à chacune des composantes principales d'identité et d'expression.

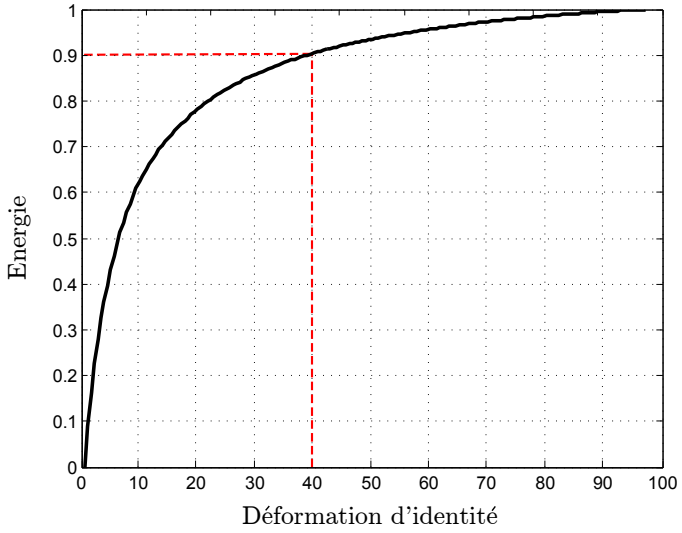


FIGURE 3.17 – *Energie cumulée en fonction des déformations d'identité*

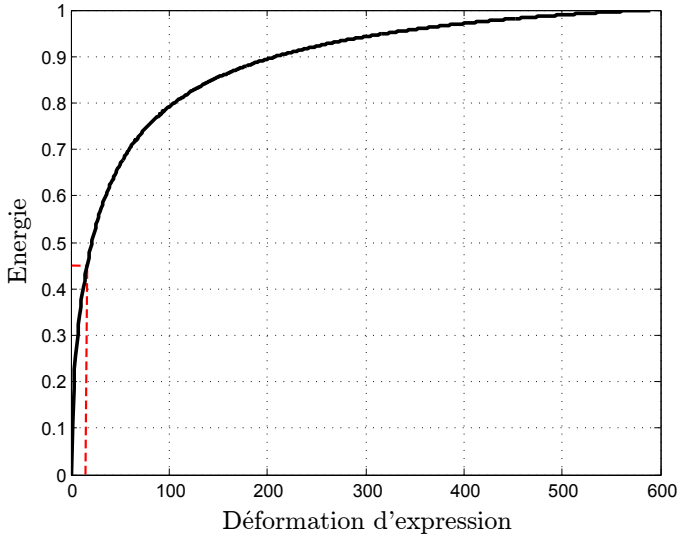


FIGURE 3.18 – *Energie cumulée en fonction des déformations d'expression*

Dans le but d'obtenir un modèle plus compact, seules 40 déformations d'identité et 8 déformations d'expression sont conservées. Cette réduction de dimensionalité permet de conserver 90% de l'énergie des déformations d'identité et 45% de l'énergie des déformations d'expression. Bien que paraissant importante, cette réduction permet un bon compromis entre diminution du temps de calcul nécessaire à l'ajustement du modèle 3D sur une image 2D et précision de cet ajustement. Nous aborderons cette notion d'ajustement dans le chapitre suivant.

Conclusion du chapitre

Ce chapitre a présenté l'ensemble des méthodes mises en place pour l'obtention d'un modèle 3D déformable de visage avec variations d'identité et d'expression. La construction d'un 3DMM standard repose sur une analyse en composantes principales d'un jeu de données d'apprentissage. Nous proposons, dans le but de dissocier les variations liées à l'identité

de celles liées à l'expression, de construire notre modèle déformable sur deux jeux de données. Le premier, pour extraire les données d'identité, est constitué de scans neutres. Le deuxième est une collection de déformations 3D de visages liées à des variations d'expressions. En amont de ces deux analyses en composantes principales, une étape de mise en correspondance a été réalisée afin d'obtenir une indexation commune de l'ensemble des sommets des scans d'apprentissage. La méthode de mise en correspondance de visage 3D proposée par Amberg [4] a été étendue, dans le but d'améliorer la robustesse de cette étape aux scans avec de fortes expressions.

Correction de l'expression

4

Sommaire

Dans ce chapitre, nous présenterons les différentes stratégies mises en place pour effectuer une neutralisation des expressions du visage. Le modèle déformable 3D de visage, dont la présentation a été effectuée dans le chapitre précédent, est l'outil clef de ces stratégies. L'ajustement du modèle 3D déformable de visage sur l'image 2D d'entrée permet d'extraire différentes informations (Coefficients d'identité, coefficients d'expression, paramètres de pose, carte de texture). Des méthodes de l'état de l'art proposent de corriger les problèmes de pose à travers la génération d'une vue synthétique à l'aide de ces informations et de nouveaux paramètres de pose correspondant à une vue frontale. Dans ce chapitre, nous proposerons d'étendre cette approche à la correction des problèmes liés aux variations simultanées de pose et d'expression.

4.1 Ajustement du modèle 3D

Le modèle déformable 3D, présenté dans le chapitre précédent, propose une représentation de l'espace des visages 3D. La forme 3D de n'importe quel visage peut alors être approchée par une forme moyenne de visage sur laquelle un nombre limité de déformations est appliqué. L'analyse en composantes principales effectuée sur le jeu de données d'apprentissage a permis de définir l'ensemble de ces déformations. Dans cette section, nous présentons le processus d'ajustement du modèle 3D sur une image 2D. Cette opération consiste à calculer les coefficients associés à chacune des déformations, permettant au modèle 3D ainsi déformé d'approcher au mieux la forme du visage présent dans l'image 2D.

4.1.1 Estimation des paramètres

Le modèle de forme présenté dans le chapitre précédent permet d'approximer la forme 3D de n'importe quel visage par un jeu de coefficient α. Similairement [16], un modèle de texture peut être construit afin d'approximer la texture de n'importe quel visage par un jeu de coefficients β.

L'objectif de l'ajustement du modèle déformable est de trouver les paramètres de forme α et des paramètres de texture β du modèle 3D ainsi que les paramètres s, R et t de pose permettant d'approcher au mieux le visage présent dans l'image 2D originale I_{input}. L'approximation de la forme S du visage, composée de N_{ver} sommets, est alors donnée par :

$$S = s.R.\left(\overline{S} + \sum_i \alpha_i A_i\right) + T \tag{4.1}$$

avec

$\overline{S}$: Matrice de taille $3 \times N_{ver}$ représentant la forme moyenne de visage.

α_i : i^{eme} coefficient du vecteur de forme α.

A_i : Matrice de taille $3 \times N_{ver}$ relative à la i^{eme} déformation.

s : Facteur d'échelle.

R : Matrice de rotation de taille 3×3.

T : Matrice de taille $3 \times N_{ver}$ constituée de la concaténation du vecteur de translation t de taille 3×1.

Les paramètres optimaux du modèle 3D sont obtenus par la minimisation d'une énergie globale E. Celle-ci est fonction de deux énergies E_{pixels} et E_{reg} :

$$E = E_{pixels} + \lambda E_{reg} \tag{4.2}$$

- Le premier terme E_{pixels} est relatif à l'attache aux données. Cette énergie est basée sur la distance entre les pixels de l'image originale I_{input} et ceux de I_{render}, l'image artificielle générée à partir de la forme 3D

générée par les coefficients α de forme, les coefficients β de texture et les paramètres de pose s, R et t [16] :

$$E_{pixels} = \sum_{pixels} \left\| I_{input}(x,y) - I_{render}(x,y) \right\|^2 \tag{4.3}$$

- Le second terme E_{reg} est une énergie de régularisation. Il permet, lors de l'ajustement du modèle 3D, de tenir compte de l'information *a priori* de distribution gaussienne des coefficients de forme α :

$$E_{reg} = \sum_{i=1}^{n_{coeff}} \frac{\alpha_i^2}{\sigma_i^2} \tag{4.4}$$

L'énergie globale E peut alors être minimisée de manière itérative par une méthode standard, telle qu'une descente de gradient. Ce processus de minimisation est très sensible aux problèmes de minima locaux (Figure 4.1).

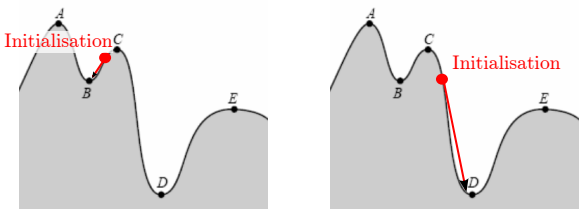


FIGURE 4.1 – *Influence de l'initialisation sur les problèmes de minima locaux*

Pour réduire ces risques de tomber dans un minimum local lors de la minimisation, Romdhani *et al.* [54] proposent d'effectuer une minimisation en plusieurs étapes. Les premières étapes utilisent de nouvelles contraintes d'attache aux données dans le but de se rapprocher du minimum local lors de l'initialisation du processus.

Ces nouvelles contraintes utilisent des informations de texture discriminantes. Ainsi, la cohérence entre certains contours du visage (contours internes ou silhouette) est utilisée lors du processus d'ajustement.

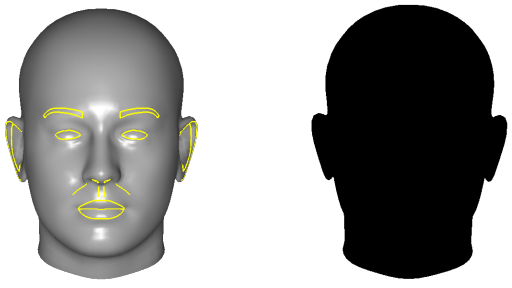


FIGURE 4.2 – *Contours internes et silhouette du visage issus du 3DMM*

La position d'un certain nombre de points caractéristiques du visage peut également être utilisée pour initialiser le processus de minimisation. Ces points doivent avoir une définition sémantique précise et pouvoir être détectés d'une manière précise. Cette annotation peut être fait manuellement par un opérateur ou bien de manière automatique [65]. Les énergies

relatives à ces données sont notées E_{edge}, E_{sil} et E_{fp}.

La fonction de minimisation devient donc :

$$E = \lambda_{pixels}E_{pixels} + \lambda_{edge}E_{edge} + \lambda_{sil}E_{sil} + \lambda_{fp}E_{fp} + \lambda_{reg}E_{reg} \tag{4.5}$$

avec λ_{pixels}, λ_{edge}, λ_{sil}, λ_{fp} et λ_{reg} les poids associés aux énergies correspondantes.

La minimisation de l'énergie E peut être réalisée en plusieurs étapes en faisant varier les poids associés aux différents termes.

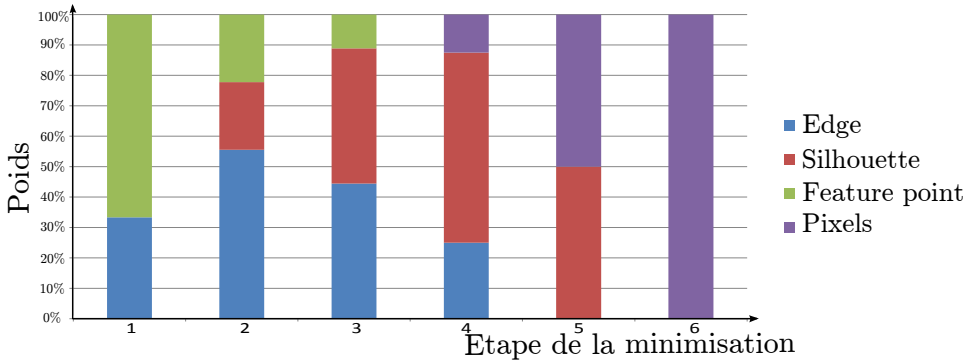


FIGURE 4.3 – *Variation des poids des énergies en fonction des étapes*

Dans cette configuration, les premières étapes utilisent les informations de points caractéristiques et de contours internes pour initialiser le processus de minimisation (Etape 1, 2, 3 et 4) en se rapprochant du minimum local. Dans un second temps (Etape 4, 5 et 6), les informations de données pixelliques sont utilisées pour affiner l'estimation de la forme du modèle 3D. Au cours de ce processus, les étapes sont réalisées successivement jusqu'à convergence.

4.1.2 Résultat de l'ajustement du modèle 3D

L'énergie globale est minimisée par l'algorithme de Levenberg-Marquart [44]. Cette méthode de minimisation de fonction est dérivée de la descente de gradient et de l'algorithme de Gauss-Newton. Cet algorithme est particulièrement adapté à notre problème en bénéficiant des avantages de ces deux méthodes. En effet, il permet de trouver précisément le minimum d'une fonction tout en assurant une convergence rapide vers ce minimum même si la position initial est éloignée de celui-ci.

Il est difficile de proposer une mesure permettant d'évaluer la qualité du recalage du modèle 3D sur une image 2D. Une telle mesure doit être capable de mesurer l'écart entre le visage original et le résultat de l'ajustement du modèle 3D (Forme 3D du visage + paramètres de pose). Il est alors nécessaire de posséder la vérité terrain. Malgré les récents progrès des dispositifs de capture 3D, l'acquisition de cette vérité est délicate : La synchronisation doit être exacte avec le système d'acquisition 2D et les paramètres de pose du visage doivent être parfaitement estimés.

Devant cette difficulté à acquérir cette vérité terrain, nous proposons ici une évaluation visuelle du processus d'ajustement du modèle 3D à partir de la superposition de l'image 2D d'origine et de la projection du modèle 3D déformé. Nous proposerons, dans la suite, une évaluation complète de notre chaine globale de pré-traitement à travers des performances biométriques.

La figure 4.4 montre la superposition de l'image originale et de la projection du modèle 3D déformé.

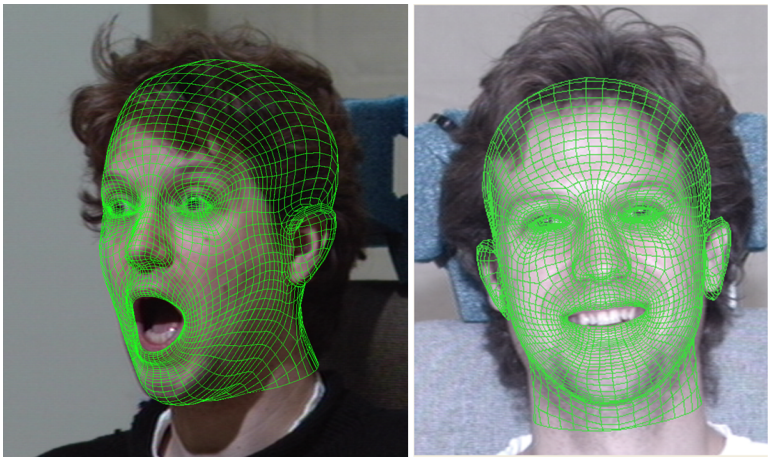


FIGURE 4.4 – *Résultat du recalage du modèle 3D déformable sur des images présentant des variations de pose et d'expression*

Il apparait assez clairement qu'une bonne correspondance entre l'image originale et le modèle 3D est obtenue. La fidélité du recalage du modèle 3D sur des images présentant des variations d'expression est démontrée sur ces deux images ainsi qu'une bonne robustesse aux variations de pose (Image de gauche).

L'objectif de cette opération de recalage est double. D'une part, elle permet l'approximation de la forme du visage en dissociant les déformations liées à l'identité de l'individu de celles liées à son expression. La connaissance de cette information de forme permettra dans la suite de notre méthode une estimation réaliste de la forme 3D du visage de l'individu avec une pose frontale et une expression neutre. D'autre part, elle permet l'extraction de l'information de texture liée à cette personne. La carte de texture ainsi extraite est le deuxième élément nécessaire à la génération d'une nouvelle vue synthétique.

La carte de texture (ou *texture map*) est une image 2D contenant l'ensemble des informations pixelliques relatives à un objet 3D. Chaque vertex v du modèle 3D est alors associé à un pixel p de coordonnées (u, v) de la carte de texture (Figure 4.5).

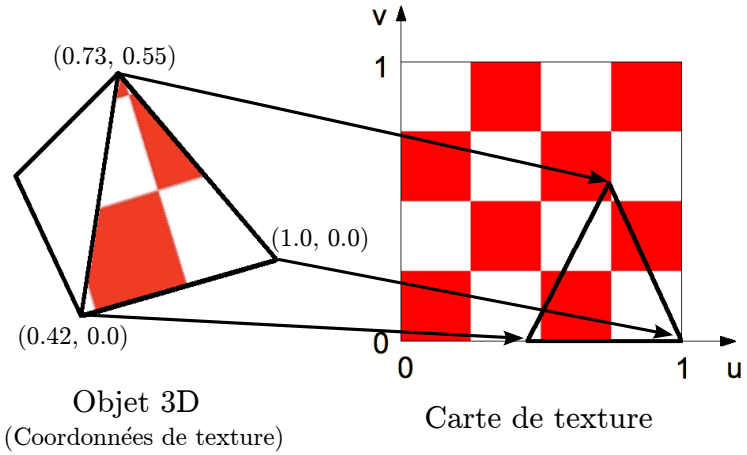


FIGURE 4.5 – *Coordonnées de texture d'un modèle 3D*

A l'issue du processus d'ajustement du modèle 3D, chaque sommet visible du modèle 3D est associé à un point de l'image d'origine. La carte de texture associée au visage présent dans l'image peut alors être reconstruite (Figure 4.6).

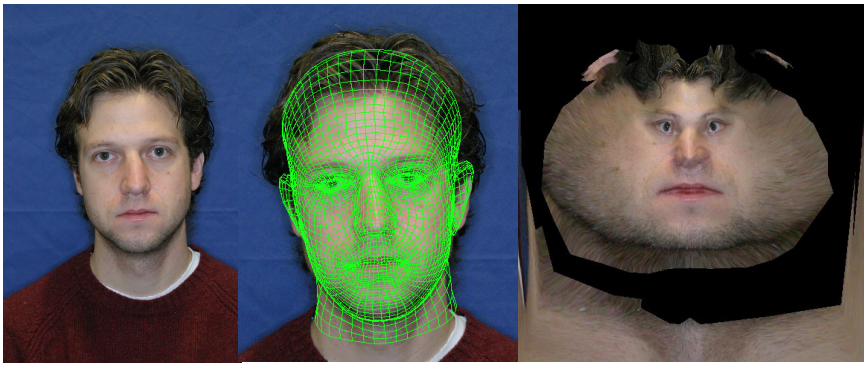


FIGURE 4.6 – *Image originale, ajustement du modèle déformable 3D et extraction de la texture map*

Dans le cas d'image avec des variations de pose importantes, l'information de texture associée à un nombre plus ou moins important de pixels de la carte de texture ne peut être extraite. Plus la pose sera importante, plus nombreux seront ces pixels. Différentes techniques peuvent alors être utilisées pour compléter la carte de texture. Les pixels manquants peuvent par exemple être remplacés par ceux correspondant à un visage moyen. Une autre possibilité est (en acceptant l'hypothèse de symétrie du visage) de les remplacer par leur symétrique dans la carte de texture (Figure 4.7).

FIGURE 4.7 – *Correction de la texture map par symétrie*

A l'issue du recalage du modèle 3D sur l'image 2D, les informations de forme et de texture relatives à l'individu sont extraites. Une nouvelle vue synthétique du visage peut alors être générée.

Génération de la nouvelle vue synthétique

Cette étape de synthèse de vue est l'étape clef de certaines méthodes proposées pour augmenter la robustesse des algorithmes de reconnaissance faciale aux problème de pose. En effet, une bonne correspondance entre les différentes images à comparer est nécessaire à l'obtention d'un système de reconnaissance faciale robuste. Cette correction de la pose du visage peut notamment être effectuée en trois dimensions (Blanz *et al.* [15]) en générant une nouvelle vue synthétique du visage présent dans l'image originale. Cette synthèse de vue est effectuée à partir du modèle 3D (forme et carte de texture) estimé de l'individu.

De nouveaux paramètres de pose peuvent alors être utilisés pour synthétiser cette nouvelle vue de l'individu (Figure 4.8).

Dans ce manuscrit, un modèle 3D déformable étendu avec des variations d'expression est utilisé. En vue de neutraliser le visage, des modifications des coefficients d'expression peuvent également être effectuées. Nous présentons ces nouvelles stratégies d'ajustement du modèle 3D et de génération de vues synthétiques dans la suite (Sections 4.2 et 4.3).

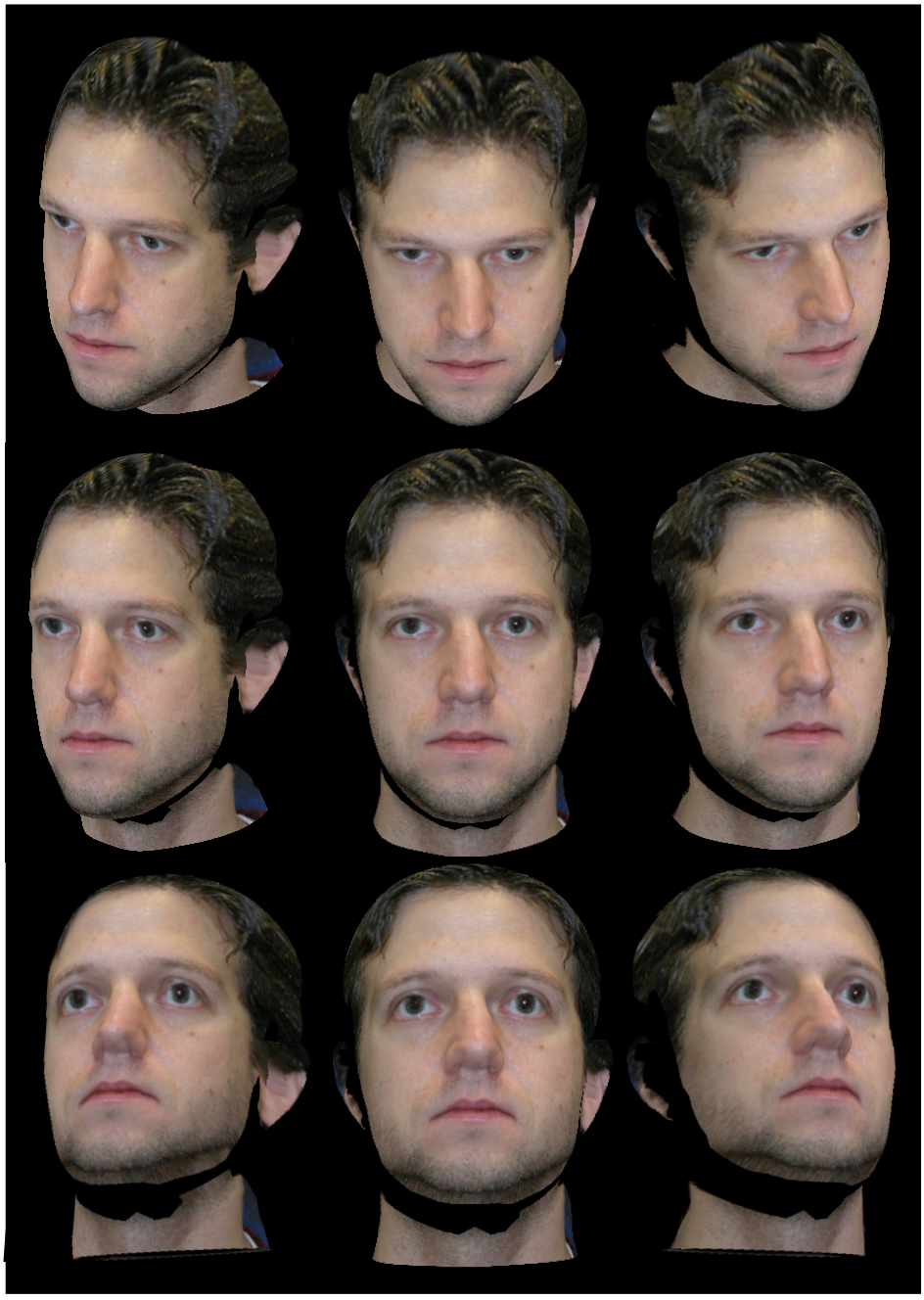


FIGURE 4.8 – *Exemples de vue synthétiques générées à partir du modèle 3D*

4.1.3 Régularisation des paramètres d'expression

Le processus d'ajustement du modèle 3D présenté dans la section précédente est applicable au modèle 3D déformable standard (sans variations d'expression). Dans le cas d'un 3DMM avec variations d'expression, l'énergie de régularisation de coefficients de forme peut être séparée en deux afin de dissocier la régularisation des paramètres d'identité de ceux liés aux paramètres d'expression.

La nouvelle énergie de régularisation devient alors :

$$E_{reg} = \lambda_{reg}^{id} E_{reg}^{id} + \lambda_{reg}^{exp} E_{reg}^{exp} \tag{4.6}$$

Cette décomposition de l'énergie de régularisation permet de modifier le poids des informations d'*a priori* d'identité (*a priori* de forme moyenne) et d'expression (*a priori* d'expression neutre). Les scalaires λ_{reg}^{id} et λ_{reg}^{exp} permettent respectivement d'ajuster l'attache à l'*a priori* d'identité et d'expression.

Pour illustrer cette séparation de l'énergie de régularisation, nous présentons dans la figure 4.9, les résultats d'une expérience montrant l'effet d'une variation du poids λ_{reg}^{id} pour un λ_{reg}^{exp} constant sur la norme du vecteur de coefficients d'identité α_{id} et celui des coefficients d'expression α_{exp} estimés lors de l'ajustement du modèle 3D. Pour cette expérience, le modèle 3D a été ajusté sur un ensemble de 250 images.

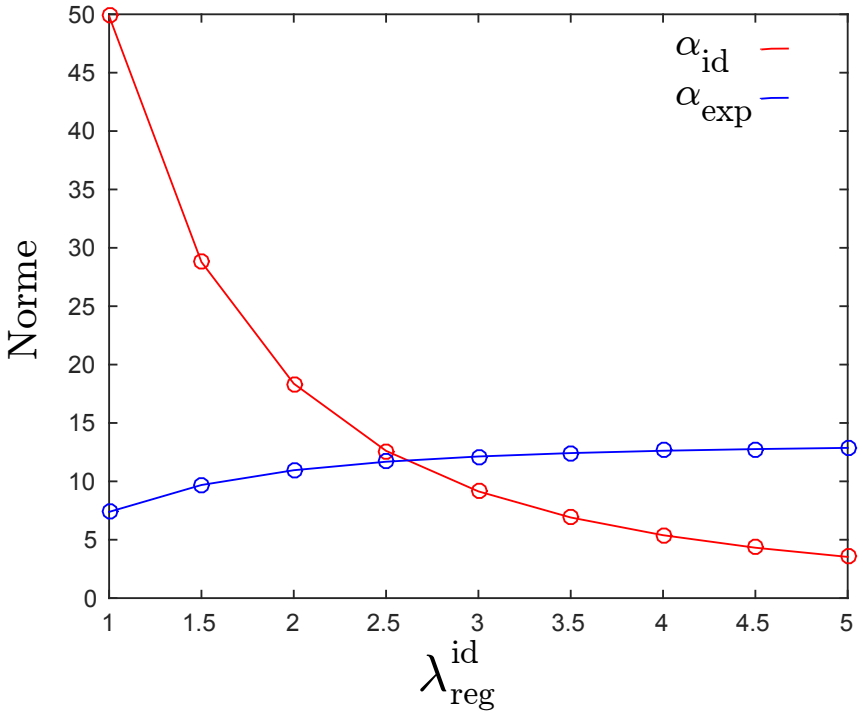


FIGURE 4.9 – *Influence des énergies de régularisation d'identité et d'expression*

L'augmentation du poids λ_{reg}^{id} de l'énergie de régularisation E_{reg}^{id} implique une diminution de la norme des coefficients d'identité. Au

contraire, les coefficients d'expression ne sont que peu impactés. Seule une légère augmentation de l'amplitude de ces coefficients est observée : Une partie des déformations du visage ne peut plus être expliquée par les déformations d'identité à cause de l'augmentation du poids λ_{reg}^{id}. La séparation entre la partie identité et la partie expression n'étant pas parfaite, ces déformations sont alors expliquées par la partie expression du modèle déformable.

La figure 4.10 montre l'influence des paramètres λ_{reg}^{id} et λ_{reg}^{exp} sur la qualité de l'ajustement du modèle 3D. Cette figure présente le résultat de cet ajustement sur une image 2D pour différentes valeurs λ_{reg}^{id} et λ_{reg}^{exp}. Pour plus de clarté, seuls la silhouette du modèle 3D (en bleu) et certains contours internes à celui-ci (en vert) sont affichés.

L'influence des paramètres de régularisation y est clairement visible. En effet, un λ_{reg}^{exp} faible permet une amplitude plus importante des coefficients d'expression et ainsi de l'ouverture de la bouche sur le modèle ajusté.

Dans cette même zone du visage, les déformations relatives à l'identité permettent de contrôler l'épaisseur des lèvres. Ainsi, pour un λ_{reg}^{exp} constant, l'utilisation d'un λ_{reg}^{id} faible tentera d'expliquer la forme du visage par ces déformations d'identité. Ainsi, la forme sera expliquée par une augmentation de l'épaisseur des lèvres, au détriment de l'ouverture de la bouche. Ce phénomène peut s'expliquer ainsi : Les données utilisées lors de l'ajustement du modèle 3D sont essentiellement situées sur le contour extérieur des lèvres. Une valeur faible du poids accordé à l'énergie de régularisation des paramètres d'identité tendra alors à expliquer l'écartement entre les contours extérieurs des lèvres par la présence de lèvres épaisses.

Au contraire, une valeur élevée de λ_{reg}^{id} diminuera l'amplitude des déformations d'identité. La forme 3D sera donc plus proche de la forme moyenne de visage. Celle-ci possédant des lèvres relativement fines, l'écartement entre les contours extérieurs des lèvres sera expliqué par les déformations d'expression et donc par l'ouverture de la bouche.

Nous aborderons, plus en détail dans la section 4.4.3, la problématique d'ajustement de ces paramètres λ_{reg}^{id} et λ_{reg}^{exp}. En effet, leurs valeurs ont un impact important sur les performances biométriques et sont fortement dépendant du type d'image à comparer.

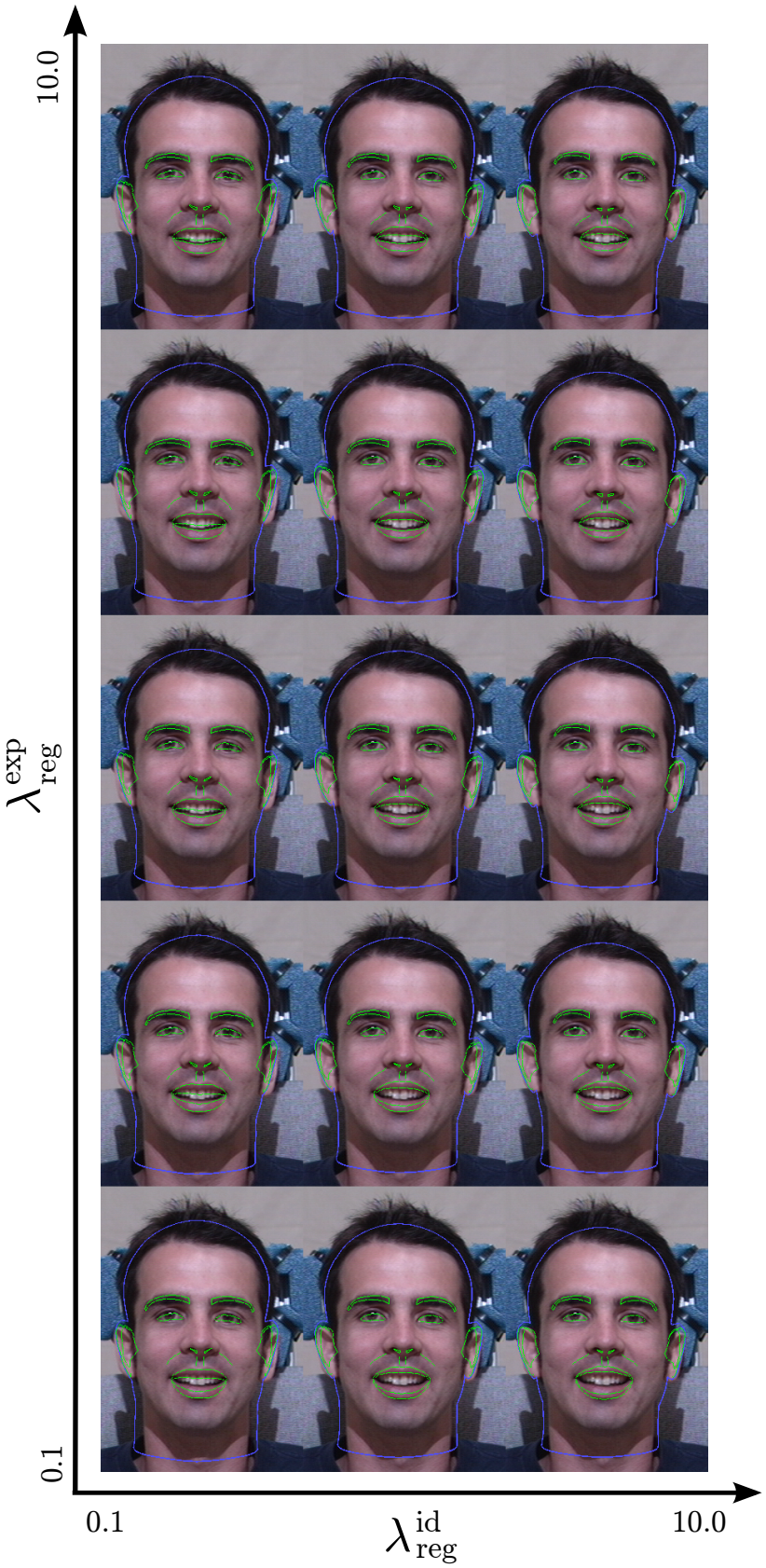


FIGURE 4.10 – *Influence des énergies de régularisation d'identité et d'expression sur l'ajustement du modèle 3D*

4.2 Neutralisation de l'expression

Les variations intra-identité de l'apparence du visage constituent l'une des limitations principales de la reconnaissance de visage en comparaison d'autres biométries telles que la reconnaissance d'empreintes digitales ou d'iris. Ces variations d'apparence peuvent être liées à différents phénomènes tels que les variations de pose ou les variations d'expression.

Les systèmes actuels de reconnaissance de visage atteignent un niveau de performances optimal lorsque les images à comparer sont frontales et neutres. Dans cette section, nous présentons une méthode de pré-traitement des images permettant ensuite d'utiliser les algorithmes de reconnaissance faciale dans des conditions optimales. Notre contribution se place donc au coeur de l'étape "Ajustement du modèle déformable 3D / Génération d'une vue synthétique" du pré-traitement.

Dans le but de générer une nouvelle vue frontale, Blanz *et al.* [15] propose d'extraire les paramètres du modèle déformable (Carte de texture, coefficients d'identité et paramètres de pose) à partir d'une image 2D puis de générer une nouvelle vue synthétique avec de nouveaux paramètres de pose. Ici, nous proposons d'utiliser le modèle déformable 3D de visage étendu (Chapitre 3). Celui-ci permet de dissocier les déformations liées à l'identité de celles liées à l'expression. Ainsi, le recalage de ce modèle 3D sur une image permet l'extraction d'un vecteur de coefficients relatifs aux déformations d'expression (appelé plus simplement coefficients d'expression dans la suite) en plus des paramètres obtenus avec un 3DMM standard (coefficients d'identité et paramètres de pose).

Cette séparation des déformations intra-identité et extra-identité nous permet ensuite de générer de nouvelles images synthétiques où les variations intra-classe du visage sont annulées. Nous proposons donc de mettre en place un processus durant lequel les informations d'identité extraites de l'image originale sont conservées tandis que de nouveaux paramètres de pose et coefficients d'expression sont utilisés. Ainsi, toutes les images générées ne comportent plus de variations intra-identité. La figure 4.11 décrit ce processus.

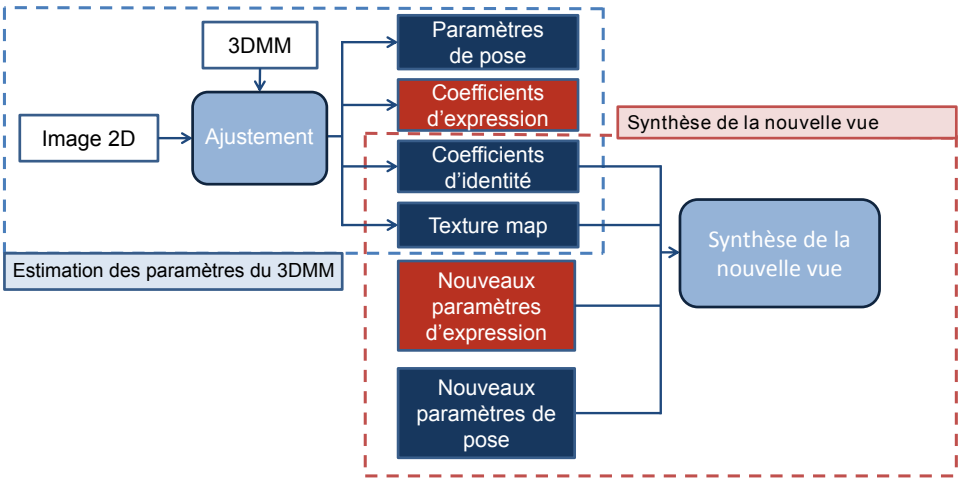


Figure 4.11 – *Processus d'ajustement du modèle 3D et de synthèse d'une nouvelle vue*

Les algorithmes de reconnaissance présentant des performances optimales sur des images sans expression, nous proposons d'utiliser lors de la synthèse des paramètres d'expression correspondant à une expression neutre.

De part sa construction, le modèle déformable 3D de visages génère des images sans expression lorsque l'ensemble des coefficients relatifs à l'expression sont nuls. La première méthode que l'on présente ici, appelée **Neutralisation d'expression**, est basée sur cette observation.

Dans cette méthode, l'ajustement du modèle 3D étendu sur l'image 2D de test (appelée *probe*) permet d'obtenir les coefficients d'identité α_{id} et d'expression α_{exp}^{probe} et les paramètres de pose. Dans un second temps, une nouvelle vue est synthétisée en utilisant ces paramètres d'identité, des paramètres d'expression neutre α_{exp}^{neutre} et des paramètres de pose correspondant à une pose frontale. Cette nouvelle vue est ensuite comparée avec l'image de référence (appelée *galerie*) à l'aide d'un algorithme standard de reconnaissance de visage. La figure 4.12 présente ce processus de neutralisation de l'expression.

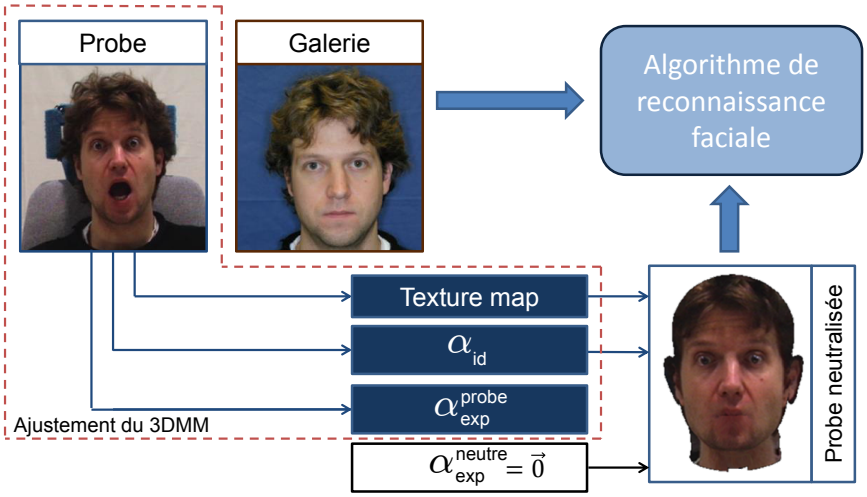


FIGURE 4.12 – *Processus de neutralisation de l'expression*

Ce nouveau processus de synthèse d'image n'induit pas de surcoût en terme de temps de calcul par rapport à un processus standard de frontalisation. En effet, l'ensemble des étapes composant ce pré-traitement est identique à la frontalisation. Seuls les paramètres d'expression utilisés lors de la synthèse sont modifiés :

Lors d'une frontalisation standard, les paramètres d'expression α_{exp}^{probe} utilisés sont ceux issus de l'ajustement du modèle 3D sur l'image de probe tandis que la méthode de neutralisation de l'expression que l'on présente ici est basée sur l'utilisation d'un jeu de coefficients d'expression neutre α_{exp}^{neutre}.

Scénario	α_{exp} utilisé lors de la synthèse
Frontalisation standard	α_{exp}^{probe}
Neutralisation d'expression	$\alpha_{exp}^{neutre} = \vec{0}$

TABLE 4.1 – *Paramètres d'expression utilisés lors de la synthèse*

Le principal avantage de cette méthode est sa capacité à être effectuée indépendamment sur chacune des images. En considérant un système d'authentification de visages regroupant plusieurs centaines de milliers d'images, l'ensemble des pré-traitements peuvent être effectués une seule fois par image, en amont du stockage en base de données. Cette possibilité de fonctionnement offline de cette méthode constitue un réel atout en vue d'une intégration dans des systèmes déployés à grande échelle.

La figure 4.13 montre des exemples de résultats obtenus par notre méthode de neutralisation de l'expression sur des images présentant des variations simultanées de pose et d'expression.

FIGURE 4.13 – *Image de test (Images de gauche), images obtenues par une frontalisation standard (Images du milieu) et images obtenues par notre méthode de correction simultanée de la pose et de l'expression (Image de droite)*

La figure 4.14 montre d'autres exemples d'images obtenues par notre méthode de neutralisation de l'expression.

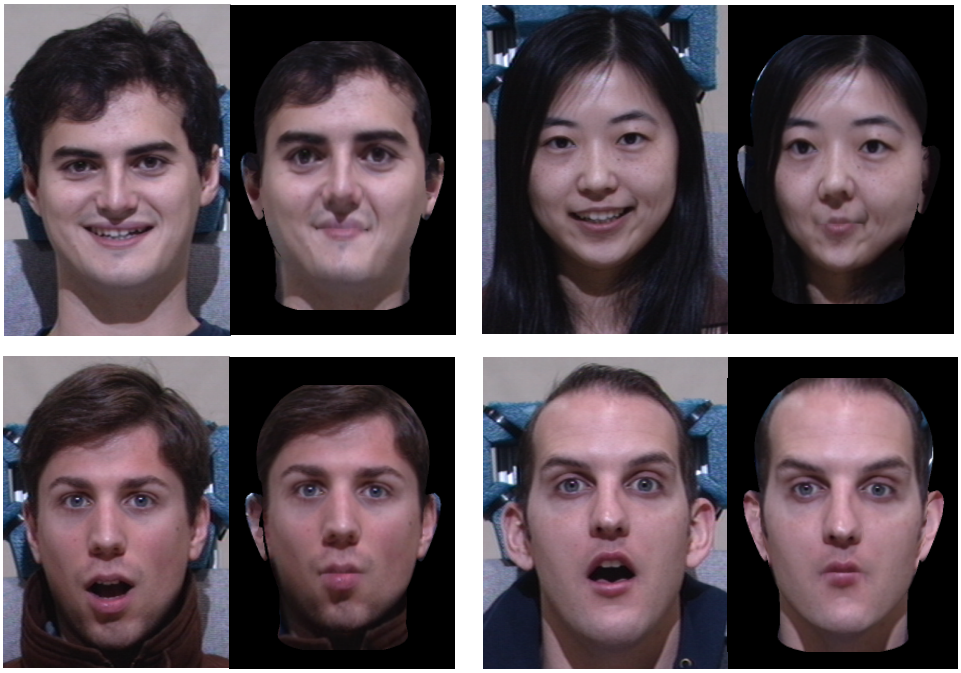


FIGURE 4.14 – *Exemples d'images obtenues par notre méthode de neutralisation de l'expression*

Bien que proposant des résultats intéressants, cette méthode possède certaines limitations. La plus importante est la difficulté à séparer les déformations du visage liées à l'identité de celles relatives aux variations d'expression. En effet, lors de l'ajustement du modèle déformable, un certain nombre de déformations relatives à l'identité peuvent être assignées à l'expression et inversement. Cette difficulté à séparer les déformations inter-identité et intra-identité nous amène à proposer une nouvelle méthode.

4.3 Transfert de l'expression

Dans cette seconde méthode, nous proposons d'améliorer la qualité de la séparation des déformations d'identité et des déformations d'expression dans un scénario de vérification ("Suis-je la personne que je prétends être ?"). Dans ce contexte 1 : 1, les deux images originales de probe et de galerie sont disponibles lors de la comparaison et peuvent ainsi être prétraitées simultanément. Cette méthode repose sur l'hypothèse suivante :

Les deux images présentées à l'algorithme de comparaison sont supposées provenir de la même personne. Ainsi, la différence entre leurs apparences, sous la même pose et des conditions d'illuminations semblables, provient principalement des variations d'expression.

Ainsi, en utilisant l'information d'identité *a priori* identique dans l'image de probe et dans l'image de galerie, le modèle 3D peut être simultanément ajusté sur les deux images. Cette opération permet d'estimer plus précisément les déformations d'identité communes aux deux images. Les variations de forme entre les deux images seront, quant à elles, expli-

quées par les déformations d'expression.

La figure 4.15 présente de manière plus détaillé cette méthode.

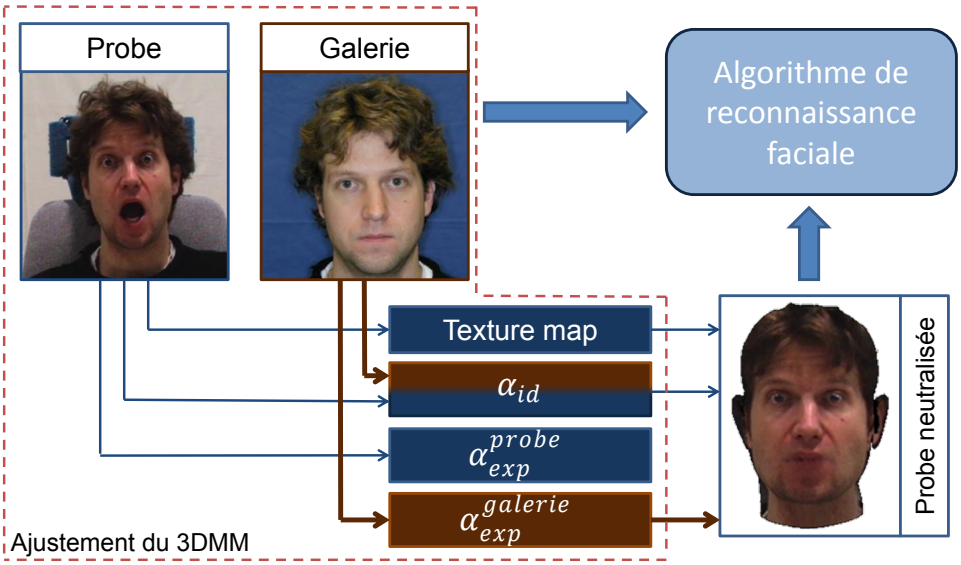


FIGURE 4.15 – *Processus de transfert d'expression*

Dans un premier temps, l'ajustement du modèle 3D permet d'extraire les informations de forme suivante :

- Un jeu de coefficients d'identité α_{id} commun aux deux images.
- Un jeu de coefficients d'expression α_{exp}^{probe} pour l'image de probe.
- Un jeu de coefficients d'expression $\alpha_{exp}^{galerie}$ pour l'image de galerie.

De manière similaire à la méthode de neutralisation de l'expression, l'information de texture est extraite de l'image de test.

Une fois ces différentes informations extraites, la génération de la vue synthétique est effectuée en utilisant les coefficients d'identité α_{id}, les coefficients d'expression $\alpha_{exp}^{galerie}$ issus de l'image de galerie et la carte de texture extraite de l'image de probe. Ainsi synthétisée, la nouvelle vue de l'image de probe possède une expression proche de celle de l'image de galerie. Ainsi, les variations d'apparence intra-classes du visage sont minimisées et l'algorithme de reconnaissance faciale est utilisé dans son contexte optimal.

Le point clef de cette méthode est l'ajustement simultané du modèle 3D sur les deux images. Ainsi, une meilleure séparation entre les déformations liées à l'identité et celles relatives à l'expression est obtenue.

Cette méthode possède toutefois une contrainte non négligeable. Pour pouvoir être utilisée, elle nécessite que les deux images originales soient disponibles lors de l'ajustement du modèle 3D. Cette contrainte n'a pas d'impact dans un scénario de vérification où un couple d'images est fourni au système afin d'obtenir un score de similarité permettant de valider ou non l'identité de la personne. Au contraire, cette contrainte est limitante dans des contextes d'identification $1 : N$, où une image est comparée

contre une base complète d'images. Dans ce type de scénario, les vecteurs caractéristiques sont généralement extraits lors de l'enregistrement de l'image dans la base. L'ensemble du processus *pré-traitement / extraction des vecteurs caractéristiques* peut ainsi être effectué hors-ligne. La méthode de neutralisation de l'expression, effectuée indépendamment sur chaque image, autorise cette approche d'extraction hors-ligne des caractéristiques tandis que la méthode de transfert d'expression ne le permet pas. Ce pré-traitement, devant alors être effectué pour chaque couple d'images, impacte fortement le temps de calcul dans ce type de scénario.

Dans ce cas, un transfert d'expression ne peut donc pas être utilisé dans ce cas au contraire de la neutralisation d'expression qui peut être effectuée indépendamment pour chaque image lors de son enregistrement en base.

La figure 4.16 propose une comparaison visuelle des deux méthodes proposées dans les sections 4.2 et 4.3.

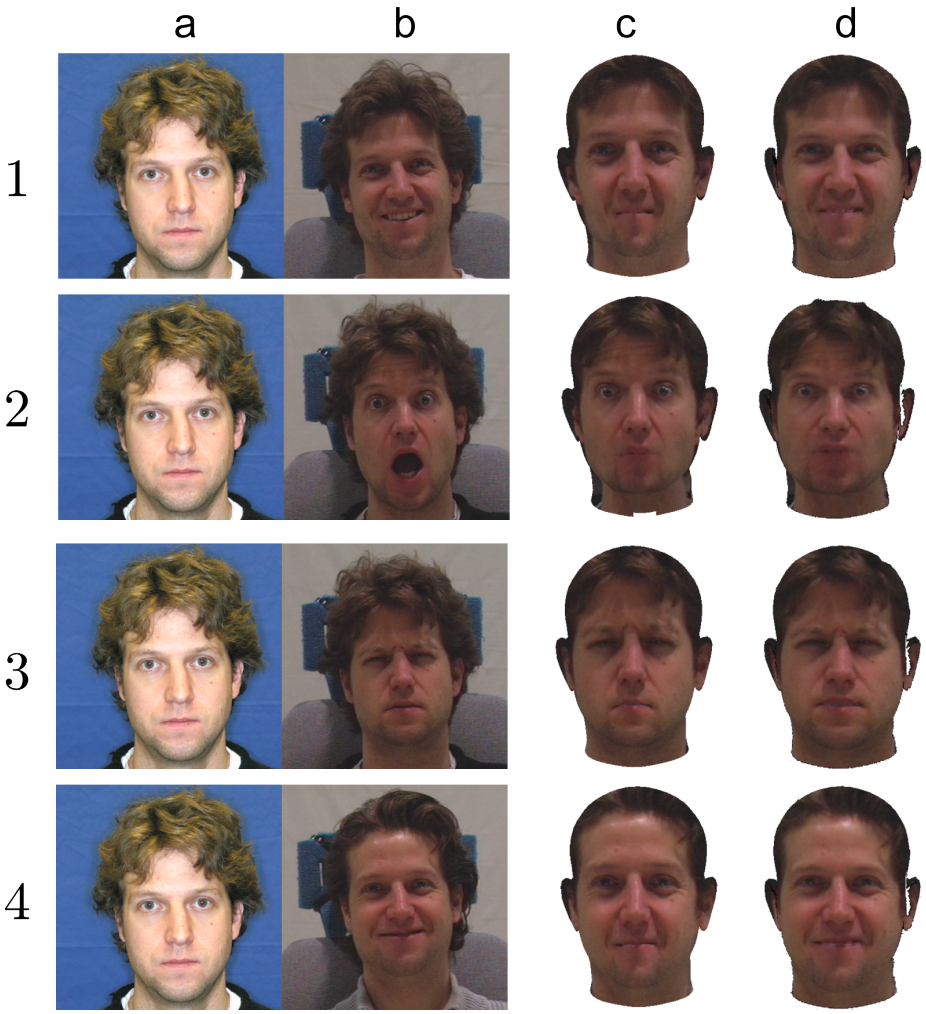


FIGURE 4.16 – *Comparaison de la neutralisation de l'expression (c) et du transfert d'expression (d) pour un couple d'image de galerie (a) et d'image de probe (b)*

A première vue, les images synthétisées par la neutralisation (Colonne c) ou par le transfert d'expression (Colonne d) semblent être identiques. Une analyse plus approfondie de ces images nous permet de détec-

ter un certain nombre de différences, en particulier dans la zone de la bouche. Ce phénomène peut notamment être observé sur la troisième et la quatrième ligne où le transfert d'expression permet un rendu plus réaliste.

Nous présentons, dans la section suivante, un ensemble de résultats biométriques permettant de quantifier l'apport de ces méthodes lorsqu'elles sont utilisées en tant que pré-traitement d'algorithmes standard de reconnaissance faciale.

4.4 Reconnaissance de visages

Dans la section précédente, les deux méthodes de correction de l'expression proposées ont été évaluées à travers une analyse visuelle. L'objectif *in-fine* de ces méthodes étant de robustifier les algorithmes actuels de reconnaissance de visages, nous procédons ici une évaluation des performances biométriques obtenues avec nos méthodes.

Nous proposons donc d'évaluer dans un premier temps, nos deux méthodes sur certaines bases de données de l'état de l'art. Ces bases, couramment utilisées dans les travaux traitant de ces problématiques, font partie des "protocoles standards" d'évaluation des méthodes proposant une robustesse aux variations d'expression. De plus, notre méthode, basée sur un modèle déformable 3D, permet une correction simultanée de la pose et de l'expression. A notre connaissance, il n'existe pas, dans l'état de l'art, de protocole expérimental relatif à cette problématique. Nous proposons donc un nouveau protocole de tests permettant d'évaluer les performances biométriques en présence de variations simultanées de pose et d'expression (Section 4.4.2).

Dans la seconde partie de cette section, nous présentons les performances obtenues dans des scénarios réalistes. Le but de cette partie est d'évaluer nos travaux en vue d'une intégration ultérieure au sein de systèmes complets. Il convient en effet de s'assurer que leur intégration permet également d'améliorer les résultats obtenus également sur des bases représentatives de situations réelles.

4.4.1 Performances obtenues sur des bases avec variations d'expressions

Nous présentons, dans cette partie, les performances obtenues à l'aide des méthodes décrites précédemment sur les bases CMU Multi-PIE [31] et AR Face Database [45].

Pour les expériences présentées dans cette partie, les images frontales neutres (appelées *mug-shot*) sont utilisées en tant qu'images de galerie. Les images avec des variations d'expressions, d'illumination et de pose composent le jeu d'images de tests.

Les différentes méthodes sont comparées en termes de taux de reconnaissance au rang 1. Ce taux constitue une métrique couramment utilisée dans les scénarios 1 : N. Dans ce type de scénario, une image est comparée contre une base de données composée d'images provenant de différents individus.

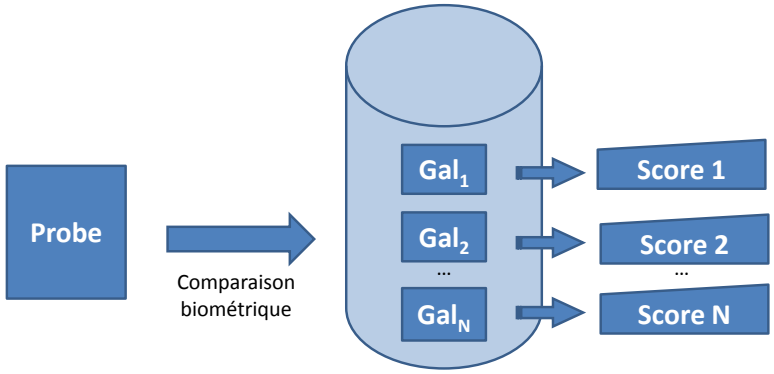


FIGURE 4.17 – *Scénario 1 :N : Comparaison d'une image contre une base de données*

Le taux de reconnaissance au rang 1 indique le pourcentage d'images de tests (*probes*) dont l'image de galerie avec le score de similarité le plus élevé provient du même individu.

Dans cette section, nous souhaitons évaluer la capacité des méthodes de neutralisation de l'expression et de transfert de l'expression à augmenter la robustesse des algorithmes de reconnaissances de visages aux variations d'expression. Pour limiter les dégradations de performances liées au bruit apporté par les détecteurs automatiques, nous effectuons une annotation manuelle des images. Ces annotations permettent une initialisation correcte du processus d'ajustement du modèle 3D.

Les méthodes que l'on a proposées précédemment permettent de pré-traiter les images en amont de leur utilisation dans un algorithme standard de reconnaissance de visages. Nous proposons donc de comparer différentes configurations de pré-traitement :

Configuration standard : Cette configuration présente les résultats obtenus avec l'algorithme commercial de reconnaissance de visage utilisé au cours de cette thèse.

Configuration frontalisation standard : Dans cette configuration, le seul pré-traitement appliqué est une frontalisation standard de l'image d'entrée. Cette frontalisation est effectuée à partir d'un modèle déformable 3D de visages sans variations d'expression, comme proposé par Blanz *et al.* dans [15].

Configuration neutralisation d'expression : A travers cette configuration, nous évaluons l'apport de notre méthode de neutralisation de l'expression. Chaque image est frontalisée et neutralisée avant d'être transmise à l'algorithme de reconnaissance de visage.

Configuration transfert d'expression : Dans cette configuration, nous testons notre méthode de transfert d'expression. Le pré-traitement

est appliqué sur chaque couple comparé : Les deux images sont frontalisées et l'image de test est rendue en utilisant l'expression extraite de l'image de galerie.

Nous avons montré, dans la section 4.1, que les scalaires λ_{reg}^{id} et λ_{reg}^{exp} permettaient la pondération des énergies de régularisation des paramètres d'identité et d'expression. Pour définir leur valeur, des tests ont été effectués sur des images, propriétaires Morpho, présentant des variations de pose et d'expression. Pour limiter les problèmes liés à un sur-apprentissage, nous nous sommes assurés que les individus présents sur ces images étaient différents de ceux présents dans les bases de tests. Ces tests nous ont conduit à utiliser, pour les configurations "neutralisation d'expression" et "transfert d'expression", les valeurs suivantes : $\lambda_{reg}^{id} = 0.1$ et $\lambda_{reg}^{exp} = 0.1$

Nous comparons ces trois configurations à travers les performances biométriques obtenues lorsqu'elles sont utilisées en amont d'un algorithme Morpho de reconnaissance faciale. Nous testons nos méthodes sur des bases de données utilisées régulièrement pour évaluer des travaux traitant des problématiques relatives à l'expression. Nous utilisons donc les bases CMU Multi-PIE [31] et AR Face Database [45] sur lesquels différentes méthodes proposant de robustifier la reconnaissance faciale aux problèmes de variations d'expression ont été évaluées.

CMU Multi-PIE

La base CMU Multi-PIE [31] est une collection de plus de 750 000 images comprenant des variations de pose, d'illumination et d'expression de 337 individus. Chaque sujet est photographié sous 15 poses et 19 illuminations différentes en effectuant différentes expressions (Neutre, sourire, surprise, yeux plissés, cri). Cette grande diversité (Figure 4.18) a permis à cette base de données d'être très largement utilisée dans de nombreux articles traitant de reconnaissance de visage.

FIGURE 4.18 – *Exemple d'images de la base CMU-Multi-PIE*

Pour effectuer l'évaluation de nos méthodes en les comparant à l'état de l'art, nous suivons le protocole expérimental proposé par Yang *et al.* [73]. Une collection d'images neutres et frontales acquises dans de bonnes conditions d'illumination compose la base d'images de galerie tandis que quatre sous-parties de la base de données composent les jeux de données de tests. Dans ces expérimentations, toutes ces images de

tests sont frontales. Elles sont acquises sous différentes illuminations pour rendre la reconnaissance plus délicate.

Chacun de ces jeux de données est relatif à une expression particulière :
- Smi-S1 : Expression *Sourire* acquise lors la première session

FIGURE 4.19 – *Exemples d'images de tests du jeu de données Smi-S1*

- Sqi-S2 : Expression *Yeux plissé* acquise lors de la deuxième session

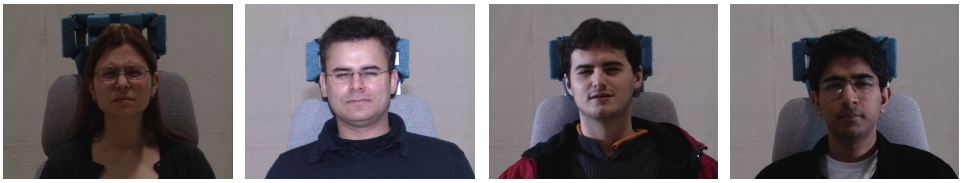


FIGURE 4.20 – *Exemples d'images de tests du jeu de données Sqi-S2*

- Sur-S2 : Expression *Surprise* acquise lors de la deuxième session

FIGURE 4.21 – *Exemples d'images de tests du jeu de données Sur-S2*

- Smi-S3 : Expression *Sourire* acquise lors de la troisième session

FIGURE 4.22 – *Exemples d'images de tests du jeu de données Smi-S3*

Le deux méthodes proposées précédemment peuvent être utilisées de manière indépendante de l'algorithme de reconnaissance faciale utilisé. Nous proposons donc, dans un premier de visualiser, l'apport de ces méthodes sur un algorithme standard de reconnaissance de visage de type *HOG+LDA*.

Les descripteurs HOG, introduits par Lowe *et al.* [43], sont des descripteurs couramment utilisés par la communauté de vision par ordinateur. Ils présentent en effet un certain nombre d'avantage. La représentation par histogramme permet notamment une invariance aux légers défauts

d'alignement. L'utilisation de descripteurs HOG à la problématique de reconnaissance de visage a été proposée par Albiol *et al.* [2].

Dans cette expérience, nous proposons de représenter la zone d'intérêt du visage par un ensemble de descripteurs HOG calculés sur un ensemble de cellules, de taille 4×4, couvrant l'ensemble de la zone. Une réduction de dimensionalité par LDA est ensuite appliquée afin de ne conserver que les parties informatives des descripteurs.

Nous comparons ainsi, dans le tableau 4.2, les taux de reconnaissance au rang 1 obtenus.

	Sur-S2	Sqi-S2	Smi-S1	Smi-S3
HOG+LDA standard	33.3%	39.5%	43.8%	42.0%
HOG+LDA+Frontalisation	33.4%	39.2%	44.0%	42.1%
HOG+LDA+Neutralisation	34.9%	39.1%	43.8%	42.3%
HOG+LDA+Transfert d'expression	**37.2%**	**39.6%**	**48.3%**	**46.4%**

TABLE 4.2 – *Comparaison du taux de reconnaissance au rang 1 obtenu sur les différents jeux de données*

Cette expérience nous montre l'apport de nos méthodes de pré-traitement sur le taux de reconnaissance lorsqu'elles sont utilisées en amont d'un algorithme standard de type *HOG+LDA*. La deuxième ligne (*HOG+LDA+Frontalisation*) présente les résultats obtenus, lorsqu'une étape de frontalisation est effectuée en pré-traitement. Il apparait assez clairement que ce type de pré-traitement ne permet pas d'améliorer les performances sur ces jeux de données. Sur ces images, les visages étant acquis à une pose frontale, les variations d'apparence liées à la pose sont minimes. Les descripteurs HOG utilisés dans cette expérience étant robustes aux légers défauts d'alignement, la correction de la pose par frontalisation n'apporte pas de gain significatif du taux de reconnaissance.

De même, il apparait, à la vue de la troisième ligne de ce tableau, que la méthode de neutralisation de l'expression n'a qu'un impact limité sur les performances biométriques obtenues avec l'algorithme *HOG+LDA*. La difficulté à séparer, avec cette méthode, la partie relative à l'identité de celle relative à l'expression, ne permet pas de corriger parfaitement la totalité des défauts d'alignement des visages liés à la présence d'expression. Lors de cette expérience, des descripteurs HOG sont utilisés. Ceux-ci étant robustes aux défauts modérés d'alignement (tels que ceux corrigés par la méthode de neutralisation de l'expression), aucune amélioration significative du taux de reconnaissance n'est observée sur les images avec une expression d'intensité faible. Sur l'expression Surprise (*Sur-S2*), la neutralisation de l'expression permet une amélioration plus importante du taux de reconnaissance.

Notre deuxième méthode, le transfert d'expression, permet d'améliorer, sur tous les jeux de test, le taux de reconnaissance biométrique. L'ajustement simultané du modèle 3D sur l'image de probe et l'image de galerie, permet de mieux dissocier les parties identité et expression du

modèle 3D et ainsi un meilleur alignement des images après la génération de la nouvelle vue synthétique. Appliqué en amont d'un algorithme *HOG+LDA*, notre méthode de transfert de l'expression permet d'améliorer le taux de reconnaissance de visage obtenu.

L'objectif *in-fine* étant d'améliorer les performances obtenues avec les algorithmes commerciaux Morpho de reconnaissance de visage, nous proposons de tester l'apport de nos méthodes de correction de l'expression en amont d'un de ces algorithmes.

La tableau 4.3 présente les taux de reconnaissances au rang 1 obtenus sur chacun de ces jeux de données par différentes méthodes de l'état de l'art (méthodes de reconnaissance de visage robustes aux variations d'expression) et par nos trois configurations testées.

	Sur-S2	Sqi-S2	Smi-S1	Smi-S3
SRC [71]	51.4%	58.1%	93.7%	60.3%
LLC [66]	52.3%	64.0%	95.6%	62.5%
RRC_L_2[46]	59.2%	58.1%	96.1%	70.2%
RRC_L_1[46]	68.8%	65.8%	**97.8%**	76.0%
Algorithme commercial	84.0%	88.7%	94.8%	91.4%
Frontalisation	83.7%	89.4%	94.6%	91.5%
Neutralisation	89.4%	87.0%	94.2%	92.5%
Transfert d'expression	**99.1%**	**95.9%**	**97.8%**	**98.6%**

TABLE 4.3 – *Comparaison du taux de reconnaissance au rang 1 obtenu sur les différents jeux de données avec les configurations testées*

Nous pouvons remarquer que notre méthode de neutralisation de l'expression améliore les performances obtenues sur les expressions les plus fortes : Sourire (Smi-S3) et Surprise (Sur-S2). Sur les deux autres jeux de données, une légère diminution des performances est observée, notamment sur l'expression yeux plissés (Sqi-S2) par rapport à la frontalisation standard. Dans cette expression, la déformation principale est relative à la fermeture des yeux. Il est difficile pour l'algorithme d'ajustement de modèle 3D d'associer ces déformations à la partie identité (Individus aux yeux plissés) ou à la partie expression (Individus aux yeux fermés). Cette difficulté à séparer correctement ces déformations entraine donc une dégradation des performances.

La dernière ligne du tableau montre les performances obtenues avec la méthode de transfert de l'expression. Elle permet d'améliorer les performances de l'algorithme de reconnaissance faciale obtenues sur chacun des jeux de données. L'ajustement simultané du modèle 3D sur l'image de test et l'image de galerie permet une meilleure séparation entre la partie identité et la partie expression du modèle déformable. En effet, les déformations communes aux deux images ont, dans ce cas, tendance a être affectées à la partie identité et les déformations présentes uniquement sur l'une des deux images à la partie expression.

La tableau suivant (Tableau 4.4) montre la moyenne et l'écart type des taux de reconnaissance au rang 1 obtenus sur les différentes expressions.

	Moyenne	Ecart type
SRC [71]	65.9%	18.9
LLC [66]	68.6%	18.7
RRC_L_2[46]	70.9%	17.7
RRC_L_1[46]	77.1%	14.4
Algorithme commercial	89.7%	4.6
Frontalisation	89.8%	4.6
Neutralisation	90.8%	3.2
Transfert d'expression	**97.9%**	1.4

TABLE 4.4 – *Moyenne et écart type des taux de reconnaissances obtenus sur les différents jeux de données*

Ces résultats montrent que nos méthodes permettent d'améliorer les performances biométriques en améliorant la robustesse de l'algorithme de reconnaissance faciale aux variations d'expression. La diminution de l'écart type démontre l'apport de nos méthodes sur la stabilité du taux de reconnaissance obtenu sur des images variées.

AR Face database

La base AR Face database contient plus 4000 images frontales issues de 126 individus. Pour chaque individu, les acquisitions ont été effectuées lors de deux sessions espacées de plusieurs semaines. Lors de ces deux sessions, 13 scénarios différents ont été enregistrés proposant des variations d'expressions (Neutre, sourire, colère, ...), d'illuminations (illumination provenant de la gauche, de la droite, ...) ainsi que des occultations (lunettes de soleil, écharpe, ...).

FIGURE 4.23 – *Exemple d'images de la base CMU-Multi-PIE*

Cette grande variété de condition a permis à cette base de données d'être couramment utilisée dans la littérature. Afin de se comparer nos méthodes avec l'état de l'art, nous utilisons des données de test similaires à celles proposées dans [67].

Les images utilisées proviennent de 100 individus différents (50 hommes et 50 femmes). La galerie est composée des images acquises avec une expression neutre lors de la première session. Comme dans l'expérimentation précédente, plusieurs sets d'images de tests sont utilisés, chacun d'entre eux étant relatif à une expression particulière (Sourire, Colère, Cri) acquise lors des deux sessions (Figures 4.24, 4.25 et 4.26).

FIGURE 4.24 – *Exemples d'images de tests avec l'expression Sourire*

FIGURE 4.25 – *Exemples d'images de tests avec l'expression Colère*

FIGURE 4.26 – *Exemples d'images de tests avec l'expression Cri*

Les tableaux 4.5 et 4.6 présentent les résultats obtenus avec les trois configurations présentées précédemment ainsi que ceux obtenus avec six autres méthodes de l'état de l'art.

	Session 1		
	Sourire	Colère	Cri
SRC [71]	98.0%	89.0%	55.0%
PD [59]	**100.0%**	97.0%	93.0%
SOM [60]	**100.0%**	98.0%	88.0%
DICW [68]	**100.0%**	99.0%	84.0%
CTSDP [28]	**100.0%**	**100.0%**	95.5%
FS [67]	**100.0%**	**100.0%**	91.4%
Algorithme commercial	**100.0%**	**100.0%**	94.0%
Frontalisation	**100.0%**	**100.0%**	94.0%
Neutralisation	**100.0%**	**100.0%**	96.0%
Transfert d'expression	**100.0%**	**100.0%**	**97.0%**

TABLE 4.5 – *Taux de reconnaissance au rang 1 obtenus sur les différents jeux de données de tests de la base AR database (Session 1)*

	Session 2		
	Sourire	Colère	Cri
SRC [71]	79.0%	78.0%	31.0%
PD [59]	88.0%	86.0%	63.0%
SOM [60]	88.0%	90.0%	64.0%
DICW [68]	91.0%	92.0%	44.0%
CTSDP [28]	98.2%	**99.1%**	**86.4%**
FS [67]	94.5%	**98.0%**	58.6%
Algorithme commercial	**99.0%**	**99.0%**	73.0%
Frontalisation	**99.0%**	**99.0%**	76.0%
Neutralisation	**99.0%**	**99.0%**	85.0%
Transfert d'expression	**99.0%**	**99.0%**	82.0%

TABLE 4.6 – *Taux de reconnaissance au rang 1 obtenus sur les différents jeux de données de tests de la base AR database (Session 2)*

L'expression Cri apparait clairement comme l'expression la plus difficile, le taux de reconnaissance sur les autres sous-bases étant entre 99% et 100%. Cette expression peut se caractériser par une bouche largement ouverte et des yeux fermés. Une fois de plus, les deux méthodes que l'on propose permettent d'améliorer largement l'identification. En particulier, les résultats obtenus sur l'expression Cri sont supérieurs à ceux obtenus avec les méthodes de l'état de l'art proposant de traiter les problèmes d'expression par une approche locale. Sur le jeu de données où l'apparence du visage est également modifiée par un phénomène de vieillissement (L'apparence de certains individus est largement modifiée entre la session 1 et la session 2), nos méthodes obtiennent des résultats comparables à la méthode CTSDP [28] également basée sur une approche de déformation de l'image.

Les expérimentations conduites dans cette section ont montré l'apport de nos méthodes lorsqu'elles sont utilisées en tant que pré-traitement d'un algorithme standard de reconnaissance de visage. Dans la prochaine section, nous proposons de combiner ces deux méthodes afin de bénéficier de leurs avantages respectifs (Temps de calcul faible pour l'une et performances optimales pour l'autre).

Combinaison des deux stratégies

Les expériences présentées précédemment ont montré que la neutralisation et le transfert d'expression augmentaient la robustesse aux expressions des algorithmes standards de reconnaissance faciale.

L'ensemble de ces tests ont été effectués dans un contexte de comparaison 1 : N où une image appelée probe est comparée contre un ensemble de N images. Un score de similarité est calculé pour chacun des N couples image de probe / image de galerie.

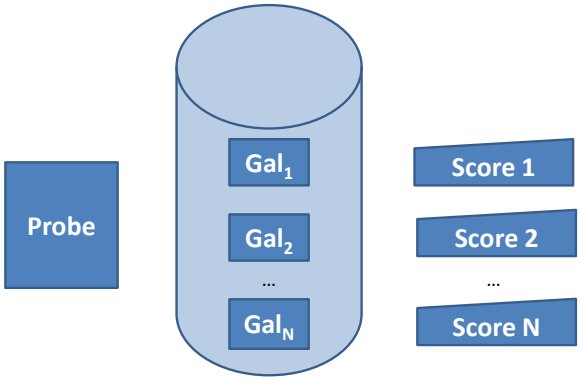


FIGURE 4.27 – *Comparaison* 1 : *N ou authentification*

Il a été montré au cours de ces expérimentations que les performances obtenues sont meilleures lorsque l'expression de la galerie est transférée vers l'image de test. Cependant, cette stratégie impose un temps de pré-traitement plus important.

En effet, en supposant la comparaison d'une image contre une base de N images, la méthode de neutralisation coûte $\tau_{ajustement} + \tau_{rendu}$ tandis que la méthode de transfert d'expression coûte $N \times \left(\tau_{ajustement} + \tau_{rendu}\right)$ avec :

- $\tau_{ajustement}$: Le temps de calcul nécessaire à l'ajustement du modèle 3D déformable de visage sur l'image d'entrée.
- τ_{rendu} : Le temps nécessaire à la génération de l'image synthétique.

Dans le but de limiter le temps de pré-traitement tout en maximisant le taux de reconnaissance, nous proposons de combiner les deux méthodes en utilisant une stratégie en deux étapes :

Tout d'abord, une première passe est effectuée en utilisant la méthode de neutralisation d'expression. Cette première passe permet d'effectuer un premier filtre sur les N images de la galerie : Seule une partie des images de la galerie (celles avec le score de similarité le plus élevé) est conservée pour la seconde passe.

Lors de la seconde passe, un transfert d'expression entre chacune des images de la galerie et l'image de probe est effectué. Grâce à la première passe, le nombre de transfert d'expression à effectuer est beaucoup plus faible, ce qui permet un net gain en terme de temps de calcul.

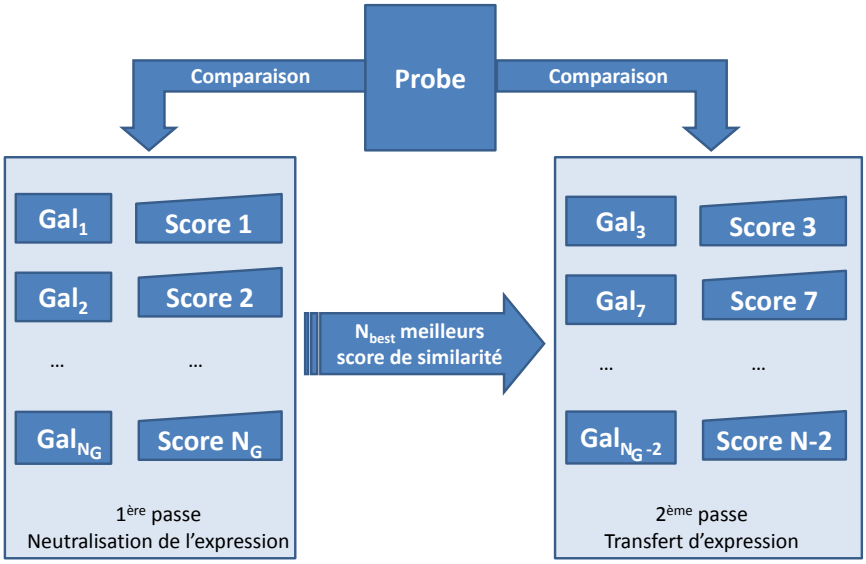


FIGURE 4.28 – *Comparaison double-passe*

En reprenant les notations précédentes, le temps de pré-traitement devient alors $(\tau_{ajustement} + \tau_{rendu}) + N_{best}(\tau_{ajustement} + \tau_{rendu})$.

Le tableau suivant présente les résultats obtenus avec cette stratégie multi-passe. Dans cette expérience, seules les 10 premières matchantes sont conservées à l'issue de la première passe.

	Sur-S2	Sqi-S2	Smi-S1	Smi-S3
Algorithme commercial	84.0%	88.7%	94.8%	91.4%
Frontalisation	83.7%	89.4%	94.6%	91.5%
Neutralisation	89.4%	87.0%	94.2.%	92.5%
Transfert d'expression	99.1%	95.9%	97.8%	98.6%
Double passe	**95.8%**	**92.1%**	**96.5%**	**95.1%**

TABLE 4.7 – *Résultats obtenus avec la stratégie double passe*

4.4.2 Variations simultanées de l'expression et de la pose

Grâce à l'utilisation d'un modèle 3D déformable de visage, nos méthodes permettent de corriger simultanément l'expression et la pose du visage. A notre connaissance, il n'existe pas dans l'état de l'art de travaux proposant des performances biométriques sur des images présentant des variations conjointes de pose et d'expression. Nous proposons donc ici un nouveau protocole de test permettant d'évaluer la robustesse de la reconnaissance faciale aux variations simultanées de pose et d'expression.

La base CMU MultiPIE [31] propose une grande collection d'images acquises sous des poses et des expressions différentes avec une illumination ambiante contrôlée. Pour chaque expression (Sourire Session 1, Surprise Session 2, Yeux fermés Session 2 et Sourire Session 3), trois jeux de données sont utilisés. Chacun d'eux est relatif à une pose particulière : 0°, 15°et 30°(Figure 4.29).

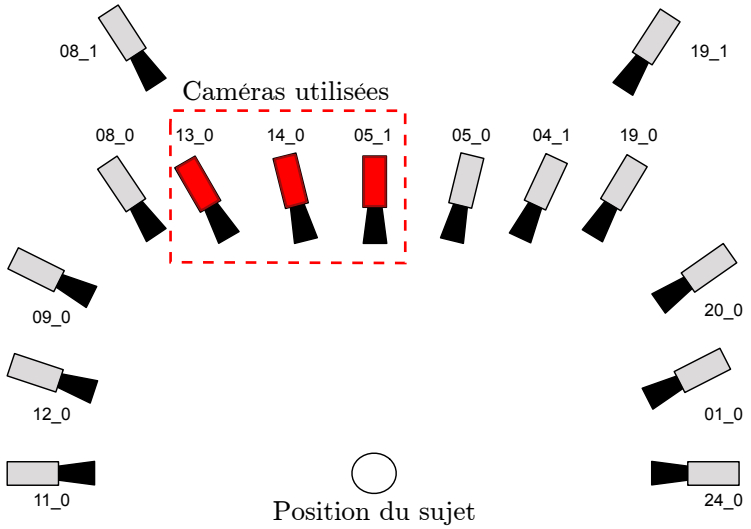


FIGURE 4.29 – *Positionnement des caméras utilisées au cours de cette expérience*

Les tableaux 4.8, 4.9 et 4.10 présentent les performances biométriques obtenues sur ces différents jeux de données.

Caméra 05_1 (0 °)				
Configuration	Sur-S2	Sqi-S2	Smi-S1	Smi-S3
Algorithme commercial	85.6%	88.8%	94.2%	89.8%
Frontalisation	85.0%	89.7%	94.3%	89.7%
Neutralisation	93.5%	87.7%	94.8%	89.4%
Transfert d'expression	**99.5%**	**96.6%**	**99.2%**	**97.8%**

TABLE 4.8 – *Résultats biométriques (taux de reconnaissance) avec une acquisition à 0 °*

Caméra 14_0 (15 °)				
Configuration	Sur-S2	Sqi-S2	Smi-S1	Smi-S3
Algorithme commercial	53.4%	60.7%	87.3%	80.2%
Frontalisation	66.5%	79.5%	89.5%	79.5%
Neutralisation	86.7%	75.1%	90.7%	85.8%
Transfert d'expression	**95.6%**	**89.2%**	**97.2%**	**98.7%**

TABLE 4.9 – *Résultats biométriques (taux de reconnaissance) avec une acquisition à 15 °*

Caméra 13_0 (30 °)				
Configuration	Sur-S2	Sqi-S2	Smi-S1	Smi-S3
Algorithme commercial	10.0%	18.4%	32.9%	25.9%
Frontalisation	46.6%	58.7%	73.3%	58.7%
Neutralisation	62.0%	52.2%	73.5%	66.1%
Transfert d'expression	**82.3%**	**71.9%**	**90.8%**	**86.1%**

TABLE 4.10 – *Résultats biométriques (taux de reconnaissance) avec une acquisition à 30 °*

Ces tableaux montrent que la précision de la reconnaissance de visage est significativement améliorée lorsque les images sont à la fois corrigées en pose et en expression tant avec la neutralisation de l'expression qu'avec

le transfert d'expression. Cette amélioration est d'autant plus grande que le scénario difficile. Ainsi, le taux d'identification est amélioré de plus de 76% lorsqu'un transfert d'expression est effectué sur les images acquises à une pose de 30°avec l'expression surprise.

L'ensemble des expérimentations présentées ici ont permis de prouver l'apport de nos méthodes de correction de l'expression. De plus, basées sur l'utilisation d'un modèle 3D déformable de visage, elles permettent d'effectuer simultanément une correction de la pose et une correction de l'expression. Cette capacité à traiter de manière conjointe ces deux problématiques majeures de la reconnaissance faciale constituent une réelle contribution à l'état de l'art. Nous proposons, dans la prochaine partie, d'évaluer nos méthodes sur des bases de données standard de visages présentant des caractéristiques variées représentatives de certains cas réels d'utilisation.

4.4.3 Performances sur des bases de données *in-situ*

Nous avons pu montrer, à travers les expériences menées précédemment, le gain obtenu lorsque nos méthodes sont utilisées en pré-traitement d'un algorithme standard de reconnaissance faciale sur des bases de données spécifiques à notre problématique de robustesse aux variations d'expression. L'objectif *in-fine* étant une utilisation dans des systèmes réels, il convient également d'évaluer ces méthodes sur d'autres types de données.

En effet, les expérimentations présentées précédemment ne permettent pas de quantifier l'apport de nos travaux dans des contextes d'utilisation réels. Les bases de données utilisées ayant été spécifiquement choisies pour prouver la robustesse aux expressions des méthodes que l'on propose, leurs caractéristiques sont assez éloignées des images rencontrées dans des situations réelles. Nous proposons ici de quantifier l'apport de nos travaux sur de nouvelles bases de données.

La méthode de transfert d'expression entrainant un surcoût important en terme de temps de calcul, sa mise en application dans des cas réels est délicate. Seule la neutralisation d'expression sera donc évaluer dans cette section.

De plus, l'objectif de cette section étant d'évaluer nos méthodes dans des contextes d'application réaliste, aucune annotation manuelle n'est effectuée. Les caractéristiques nécessaires à l'initialisation du processus d'ajustement du modèle 3D sont extraites de manière automatique [65].

Présentation des bases de données utilisées

Les domaines d'application de la reconnaissance faciale peuvent être séparés en trois catégories :

- Les applications en vidéo-protection : Celles-ci représentent sans aucun doute le contexte d'application de la reconnaissance faciale le

plus difficile à l'heure actuelle. En effet, les images représentatives de ces scénarios sont acquises dans des situations non contrôlées. Une forte variation de pose, d'expression et d'illumination est donc observée sur ce type d'images. De plus, les systèmes d'acquisitions ne permettent en général qu'une capture en basse résolution.

- Les applications policières : Ici, l'objectif est de comparer des images acquises lors de l'arrestation d'une personne. Les images constituant cette base sont frontales avec une bonne résolution. Un phénomène de vieillissement peut être observé (plusieurs années peuvent s'écouler entre deux acquisitions) ainsi que certaines variations d'expression.

- Les applications de contrôle aux frontières : Dans ce type de scénario, l'identité du voyageur est validée en comparant son visage à celui contenu dans son passeport. Une acquisition du passager est effectuée lors de son passage au sein du sas d'authentification. Afin d'améliorer l'ergonomie du système, aucune contrainte n'est imposée à l'individu. La base représentative de ce type de scénario est donc composée d'une galerie comportant des photos de type passeport (Images frontales, sans expression et acquises dans de bonnes conditions d'illuminations) et d'images de tests obtenues dans des conditions similaires à celles du sas (Poses variées, présence d'expression sur une partie des images).

Le tableau suivant résume les caractéristiques principales de ces bases de données ainsi que leurs identifiants dans la suite du manuscrit :

	Base V	Base P	Base F
Scénario	Vidéo-protection	Police	Contrôle aux frontières
Images de probe	Basse résolution avec variations d'expression et de pose	Images frontales avec variations d'expression et vieillissement	Images de bonne résolution acquises dans un contexte non contrôlé
Images de galerie	Basse résolution avec variations d'expression et de pose	Images frontales avec variations d'expression et vieillissement	Image frontale et neutre

TABLE 4.11 – *Caractéristiques des bases représentatives utilisées*

Protocole expérimental

Nous présentons ici, les performances biométriques obtenues dans un contexte d'authentification (également appelé matching 1 : 1 ou matching

de couples). En 1 : 1, l'image de probe est comparée à l'image de galerie par l'algorithme de reconnaissance faciale. La décision est ensuite effectuée en comparant le score de similarité obtenu S_{sim} à une valeur seuil S_{thres} :

- Si le score S_{sim} est supérieur à un seuil S_{thres}, les deux images sont considérées comme issues du même individu.
- Si le score S_{sim} est inférieur à un seuil S_{thres}, les deux images sont considérées comme issues de deux individus différents.

Le comportement du système est déterminé par la valeur du seuil S_{thres}. Dans le cas d'un système d'accès à une zone hautement sensible, ce seuil sera élevé : Le risque d'accepter une personne à tort sera faible au risque de refuser plus facilement une personne autorisée. Au contraire, abaisser la valeur du seuil permet de rendre le système plus permissif en limitant le nombre de faux refus. De ce cas, le risque d'accepter à tort une personne est augmenté.

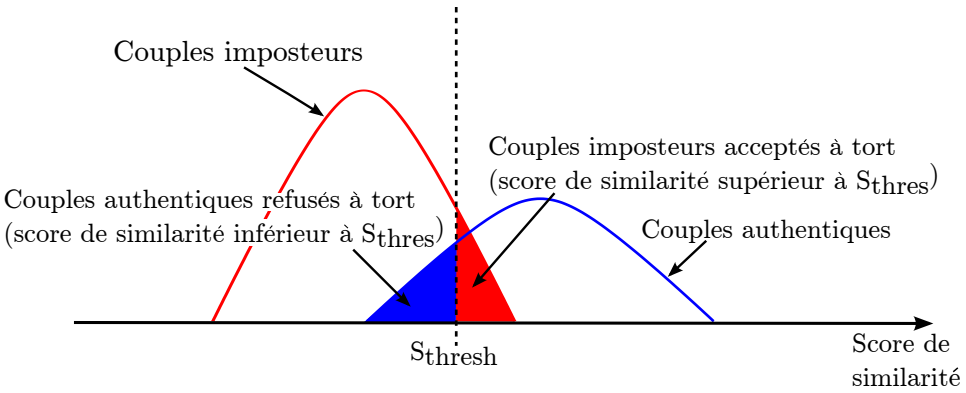


FIGURE 4.30 – *Distributions des scores de similarité des couples authentiques et des couples imposteurs*

Les résultats sont affichés sous forme de courbe ROC (*Receiver Operating Characteristic*). Cette courbe montre le taux de faux refus (FRR) en fonction du taux de fausses acceptances (FAR).

Ajustement des poids des fonctions de régularisation

Nous avons vu, dans le chapitre 4, que l'énergie minimisée lors de l'ajustement du modèle 3D est fonction de plusieurs énergies (Equation 4.5). Les énergies E_{reg}^{exp} et E_{reg}^{id} (pondérés respectivement par λ_{reg}^{exp} et λ_{reg}^{id}) permettent notamment la régularisation des coefficients d'identité et des coefficients d'expression.

Un λ_{reg}^{exp} faible tendra, en effet, à augmenter la norme des coefficients d'expression et donc à augmenter la capacité du modèle 3D à s'ajuster sur un visage avec expression. Au contraire, une valeur trop élevée aura pour

conséquence de diminuer la flexibilité du modèle 3D en terme d'expression.

Le choix de la valeur optimale de λ_{reg}^{exp} doit être un compromis entre la flexibilité du modèle 3D en termes de variation d'expression et la robustesse du processus d'ajustement. Cette valeur optimale dépend fortement du type d'images utilisées.

En effet, un poids λ_{reg}^{exp} faible tendra à relâcher la contrainte sur la norme des coefficients α_{exp}. Ainsi, une meilleure approximation de la forme 3D d'un visage avec expression pourra être obtenue. Au contraire, sur des images sans expression, la diminution de ce paramètre tendra à dégrader les performances biométriques. En effet, la séparation entre la partie identité et la partie expression de notre modèle 3D n'étant pas parfaite, l'augmentation de la flexibilité de ce modèle en terme d'expression entraine l'attribution à tort de certaines déformations du visage relatives à l'identité par la partie expression du modèle 3D.

Nous avons donc choisi d'ajuster la valeur de ce paramètre sur un ensemble de bases variées afin d'obtenir un bon compromis entre performances sur des images avec expression et performances sur des images sans expression. Les figures 4.31 et 4.32 montrent les performances obtenues avec différentes valeurs de λ_{reg}^{exp}.

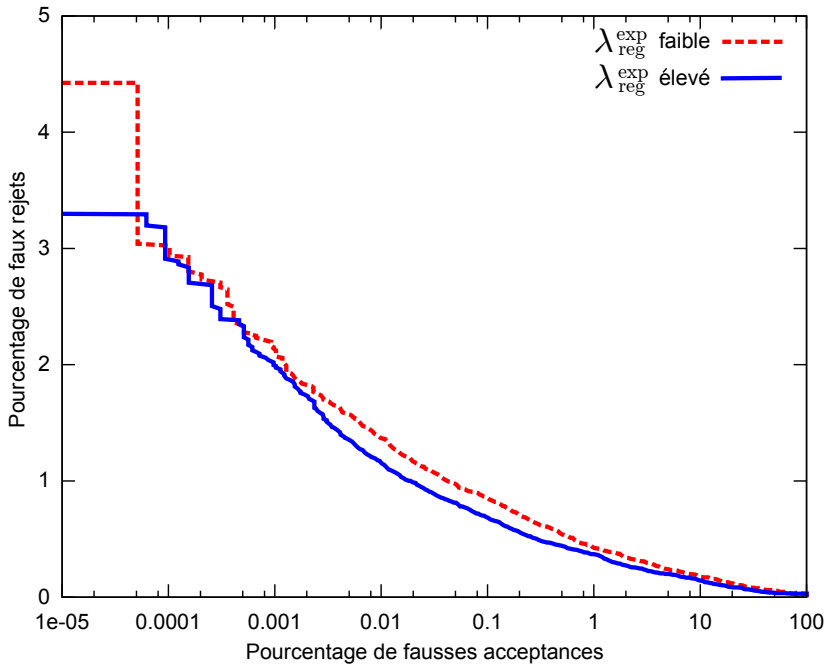


FIGURE 4.31 – *Influence de l'énergie de régularisation des paramètres d'expression sur une base sans variations d'expression*

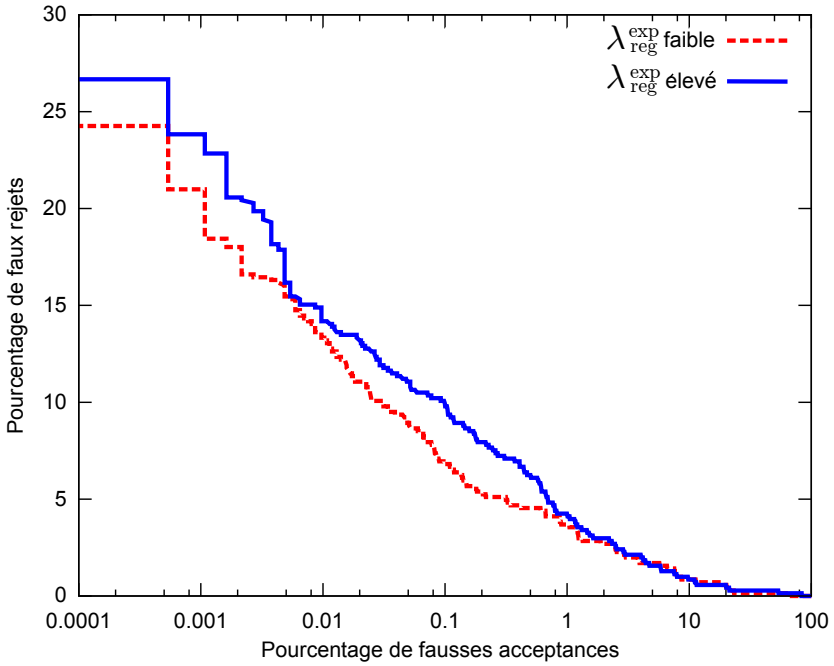


FIGURE 4.32 – *Influence de l'énergie de régularisation des paramètres d'expression sur une base avec variations d'expression*

A travers ces deux courbes, nous pouvons confirmer l'analyse faite précédemment sur l'influence du poids λ_{reg}^{exp}. Lorsque ce poids augmente, la liberté accordée aux paramètres d'expression est diminuée. Cela a pour effet de dégrader les performances obtenues sur les images avec variations d'expression. Au contraire, son augmentation permet une amélioration des performances sur des bases plus générales où les problèmes relatifs à l'expression ne concernent qu'une partie limitée des images.

Nous proposons une nouvelle expérience permettant de visualiser l'impact de ces poids λ_{reg}^{id} et λ_{reg}^{exp} en termes de taux de faux rejets.

Ainsi, les figures 4.34 et 4.33 montrent les taux de faux rejets obtenus sur différents types d'image en fonction des valeurs de λ_{reg}^{id} et λ_{reg}^{exp} pour un taux de fausses acceptances fixé à 10^{-3}.

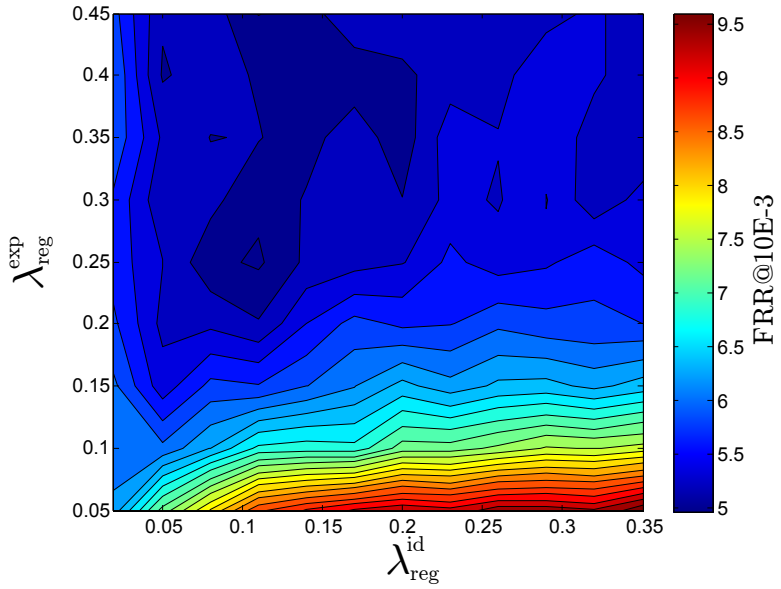


FIGURE 4.33 – *Influence de λ^{id}_{reg} et λ^{exp}_{reg} sur des images neutres*

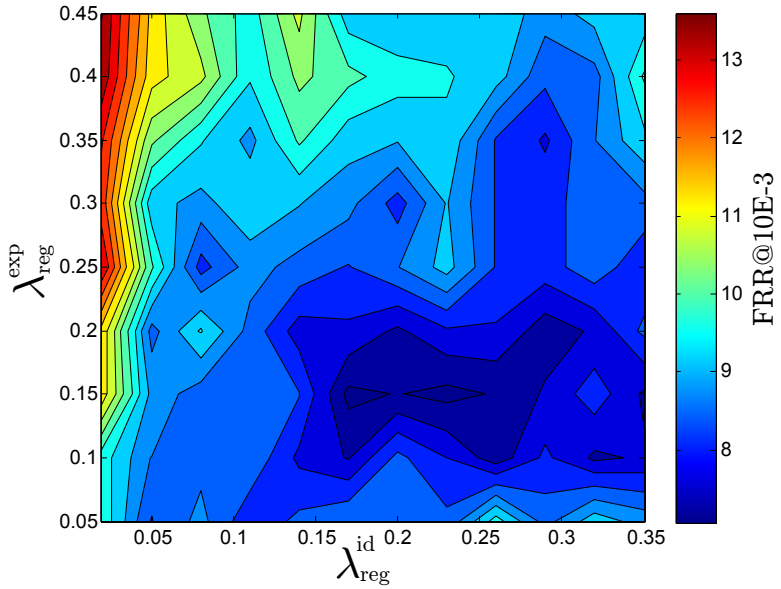


FIGURE 4.34 – *Influence de λ^{id}_{reg} et λ^{exp}_{reg} sur des images avec expression*

Sur ces graphiques, il apparait clairement que l'influence de ces poids est fortement dépendant du type d'image. En effet, sur des images neutres (Figure 4.33), la valeur optimale peut être obtenue pour $\lambda^{exp}_{reg} = 0.3$ et $\lambda^{id}_{reg} = 0.1$. Sur des images avec expression (Figure 4.34), les valeurs optimales sont $\lambda^{exp}_{reg} = 0.15$ et $\lambda^{id}_{reg} = 0.22$. Les minimums se trouvant, dans ces deux configurations, dans des zones éloignées, il est impossible de définir un jeu de paramètres permettant de maximiser les performances biométriques obtenues sur tous les types d'image.

L'objectif *in-fine* de cette thèse étant d'améliorer les performances biométriques obtenues dans des systèmes réels, nous proposons de définir la valeur de ces paramètres de manière empirique sur des bases de données

représentatives des scénarios réels.

Ainsi, nous utilisons un λ_{reg}^{exp} égal à 0.3. Cette valeur, plus élevée que dans les expériences précédentes (où $\lambda_{reg}^{exp} = 0.1$), permet ainsi des performances optimales sur les bases usuelles au détriment d'une légère dégradation sur les bases spécifiques aux problèmes d'expression.

Résultats biométriques

Base V Dans ce paragraphe, nous présentons les résultats obtenus dans un contexte de vidéo-protection. La base de données utilisée ici est composée de plus de 10 000 images (5 000 individus).

Ici, la difficulté ne se limite pas uniquement aux variations d'expression. En effet, sur ce type d'image les variations d'apparence liées au caractère non coopératif du sujet sont multiples : variations de la pose, de l'illumination et de l'expression. De plus, la résolution médiocre des acquisitions impacte fortement les performances biométriques. La figure 4.35 montre les courbes ROC obtenues sur cette base avec d'une part, une frontalisation standard et d'autre part, notre méthode de neutralisation de l'expression.

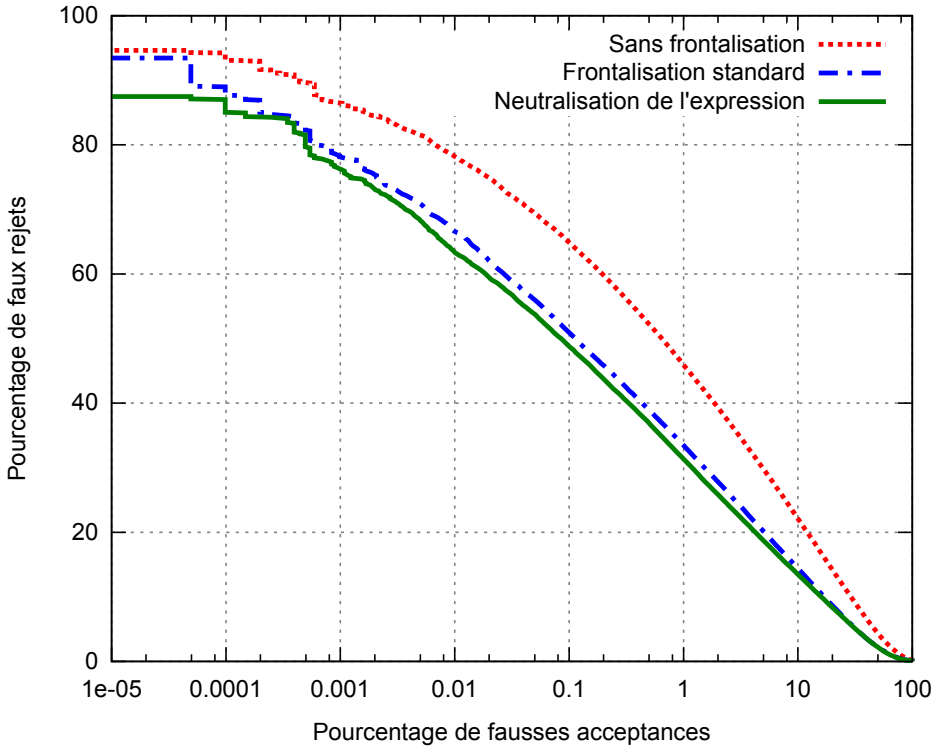


FIGURE 4.35 – *Courbes ROC sur la base V*

Bien que les difficultés de cette base soient multiples, une amélioration des performances est observée lorsque notre méthode de neutralisation de l'expression est utilisée lors du pré-traitement des images. Cette amélioration relative est de l'ordre de 5%. En se plaçant à un taux de fonctionnement de 1 fausse acceptance toutes les 10 000 requêtes, cette amélioration représente, sur cette base, 8 000 requêtes refusées à tort en moins.

Base F Cette base est composée d'environ 5 000 images de probe et 5 000 images de référence issues de 5 000 individus. Elle permet de quantifier l'apport de notre méthode dans un scénario de type "sas de contrôle aux frontières". Les images de référence ont des caractéristiques semblables aux photos contenues dans les passeports. Elles sont donc frontales avec une expression neutre et acquises sous un éclairage contrôlé. Les images de probe sont quant à elles acquises lors du passage dans le sas. Elles sont donc susceptibles de contenir certains problèmes de pose (la personne ne regarde pas forcément la caméra) ou d'expression (l'individu peut par exemple être en train de parler). Les courbes ROC obtenues avec les deux méthodes comparées sont affichées dans la figure 4.36.

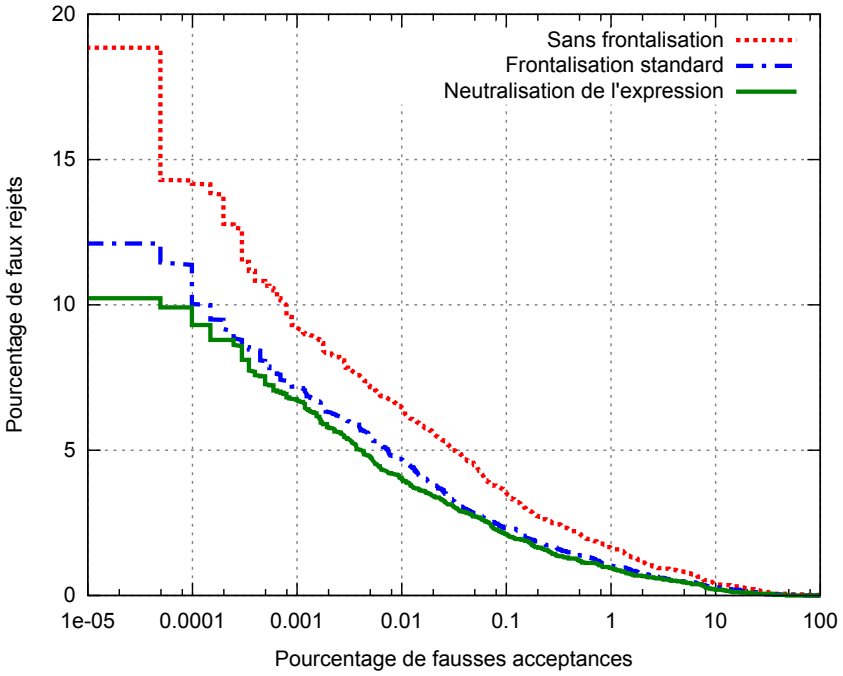


FIGURE 4.36 – *Courbes ROC sur la base F*

Les images de probe étant acquises dans un contexte non contraint, les problèmes de pose sont la difficulté principale. Ainsi, l'utilisation d'une frontalisation standard permet d'obtenir un système performant. Une fois ce type de problème corrigé, la présence d'expression devient le facteur limitant. L'utilisation de notre méthode de neutralisation de l'expression permet alors d'améliorer encore les performances obtenues. La correction de l'expression dans les images de probe permet un gain relatif de l'ordre de 10% sur cette base de données. La neutralisation de l'expression ici la vérification de 30 personnes supplémentaires pour un taux de fausses acceptances de 0.01%.

Base P Cette base de données, permettant d'évaluer notre méthode pour des applications policières, est composée de plus de 150 000 images (issues d'environ 100 000 individus). Ces images sont de bonne qualité avec une pose généralement frontale. Les principales difficultés de cette base sont les problèmes de vieillissement et la présence d'expression. La figure 4.37 montre les performances obtenues avec une frontalisation standard et avec notre méthode sur cette base de données.

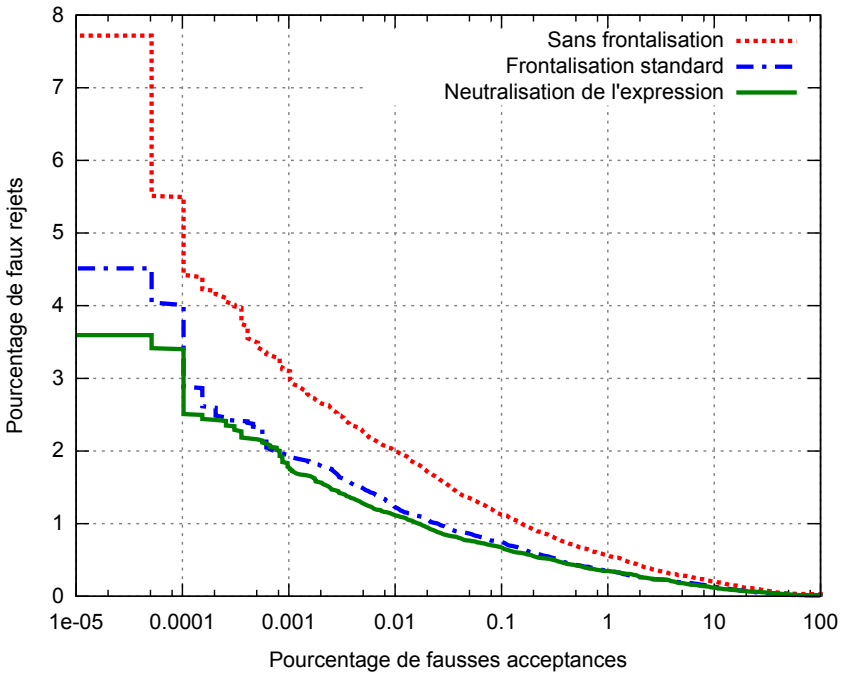


FIGURE 4.37 – *Courbes ROC sur la base P*

Une fois encore, la neutralisation de l'expression permet une amélioration des performances biométriques dans un contexte d'application réaliste. L'amélioration relative observée est de l'ordre des 10%. Sur cette base de données, l'amélioration représente la vérification d'une cinquantaine de personnes supplémentaires pour un taux de fonctionnement de 1 fausse acceptance toutes les 10 000 requêtes. Dans cette expérience où le nombre d'images contenues dans la base de données est très important, la correction de l'expression permet notamment de diminuer le risque de fausses acceptances causées par la présence d'une même expression dans deux images. Ce risque est d'autant plus élevé que la taille de la base est grande.

Conclusion du chapitre

Dans ce chapitre, nous avons présenté notre processus de neutralisation de l'expression du visage, en vue d'améliorer les performances obtenues avec un algorithme standard de reconnaissance faciale.

L'ajustement du modèle 3D déformable de visage, présenté dans le chapitre précédent, nous permet d'extraire les coefficients d'identité, les coefficients d'expression, les paramètres de pose et la carte de texture

associée à l'image d'entrée.

Une nouvelle vue synthétique est ensuite générée avec de nouveaux coefficients d'expression et paramètres de pose. Ceux-ci correspondent à une expression neutre et une pose frontale.

Pour valider les méthodes proposées, plusieurs expérimentations sont faites. Elles ont permis de prouver l'apport de nos méthode en présentant des taux d'identification supérieurs à ceux obtenus par les autres méthodes de l'état de l'art traitant de cette problématique. De plus, un nouveau protocole permettant d'évaluer la robustesse de la reconnaissance faciale aux variations simultanées de pose et d'expression a été proposé.

Finalement, les performances obtenues sur des bases de données représentatives de cas réels d'utilisation de la biométrie faciale nous permet d'envisager l'utilisation de ces méthodes dans des systèmes déployés à grande échelle. De plus, notre méthode de neutralisation de l'expression ne nécessite pas de temps de calcul supplémentaire en comparaison d'une frontalisation standard. Ainsi, l'ensemble des étapes nécessaires à la comparaison de visages (Détection du visage, Frontalisation, Correction de l'expression, Extraction des caractéristiques, Comparaison des caractéristiques) peut être effectué en moins d'une seconde.

Applications aux vidéos

5

Sommaire

Dans ce chapitre, nous nous concentrerons sur une problématique de plus en plus importante : La reconnaissance de visage à partir d'un flux vidéo.

Dans ce scénario, nous cherchons à identifier un individu présent dans une vidéo parmi une base de visage 2D. Traditionnellement, la reconnaissance depuis une vidéo se fait par l'extraction d'une image de la vidéo. Cette image est ensuite comparée contre la base d'images 2D avec un algorithme standard de reconnaissance faciale. Cette image, couramment appelée *best image*, est choisie selon un certain nombre de critères de qualité de l'image (Résolution de l'image, pose du visage, ...).

Dans ce chapitre, nous proposons, à partir d'une séquence vidéo, de générer une nouvelle image frontalisée et neutralisée en tenant compte des informations temporelles apportées par la vidéo. Une évaluation expérimentale permettra ensuite de valider cette méthode en tant que pré-traitement d'un algorithme standard de reconnaissance faciale.

De par son caractère non intrusif, la reconnaissance faciale est aujourd'hui une biométrie couramment utilisée. En parallèle, le développement des systèmes de vidéo-protection offre de nouvelles perspectives d'application de la reconnaissance de visage. La problématique de reconnaitre un visage à partir d'une vidéo devient un enjeu majeur. Il devient nécessaire de concevoir des systèmes acceptant des vidéos en entrée.

Nous proposons ici d'étendre la méthode de neutralisation présentée précédemment pour une utilisation à partir de séquences vidéo : L'objectif de ce chapitre est donc, de présenter une méthode permettant, à partir d'une séquence vidéo, de générer une nouvelle vue frontale et neutre qui puisse ensuite être comparée avec une image 2D de référence. Pour optimiser les performances obtenues avec cette méthode, nous proposons d'utiliser les informations temporelles de la séquence vidéo pour améliorer la qualité de l'ajustement du modèle 3D.

L'utilisation de flux vidéo au sein de systèmes biométriques est une technologie relativement jeune en comparaison de la reconnaissance image contre image. Les premiers travaux effectués sur la problématique de reconnaissance vidéo contre image ont naturellement été effectués en étendant les solutions existantes au cas de la vidéo. Ainsi, l'état de l'art propose un certain nombre de méthodes [27] [50] [70] basées sur l'extraction d'une unique image de la vidéo. Cette extraction est faite selon certains critères de qualité (Résolution ou pose du visage par exemple).

Ce type de méthode permet alors de se ramener à un scénario connu et maitrisé. Le principal avantage de ces méthodes est donc la réutilisation d'approches déjà optimisées dans le cas de reconnaissance d'images 2D et ainsi bénéficier de l'expérience acquise sur la reconnaissance d'images 2D. Cette stratégie a cependant un inconvénient important : Les informations temporelles données par la vidéo sont perdues lors de l'extraction de la *best image*.

Nous proposons donc d'utiliser cette information au cours de l'ajustement du modèle 3D pour améliorer l'extraction des coefficients d'expression et des paramètres de pose.

5.1 Informations temporelles

Dans cette section, nous nous plaçons dans le cas où le modèle déformable est ajusté sur un ensemble de N images consécutives. Nous appellerons *tracklet*, la collection d'images consécutives associées à un individu. Cette tracklet est obtenue par un algorithme de suivi de visage [18].

Comme nous l'avons vu dans le chapitre précédent, l'ajustement du modèle 3D sur une image consiste à déterminer les paramètres d'identité, d'expression et de pose permettant de minimiser la distance entre la projection du modèle 3D ainsi déformé et l'image d'entrée. Nous proposons, dans ce chapitre, d'étendre celui-ci aux vidéos.

Les paramètres du modèle 3D sont alors calculés pour correspondre au mieux à l'ensemble des images de la tracklet. Une partie des paramètres du modèle 3D, relatifs aux variations inter-identité, doit être constante tout au long de la séquence. En effet, une tracklet est le résultat du suivi d'un individu dans une vidéo. L'identité de la personne présente dans la tracklet est donc constante dans le temps. Ainsi, un unique jeu de coefficients d'identité est calculé pour l'ensemble des images. La carte de texture est quant à elle extraite à partir d'une seule image choisie selon un critère de résolution optimale du visage.

Au contraire, d'autres paramètres du modèle 3D sont associés aux variations intra-identité. Les paramètres de pose et les paramètres d'expression varient tout au long de la séquence plus ou moins rapidement en fonction de la vidéo. Il est donc nécessaire de calculer un jeu de coefficients d'expression et un jeu de paramètres de pose pour chacune des images de la tracklet.

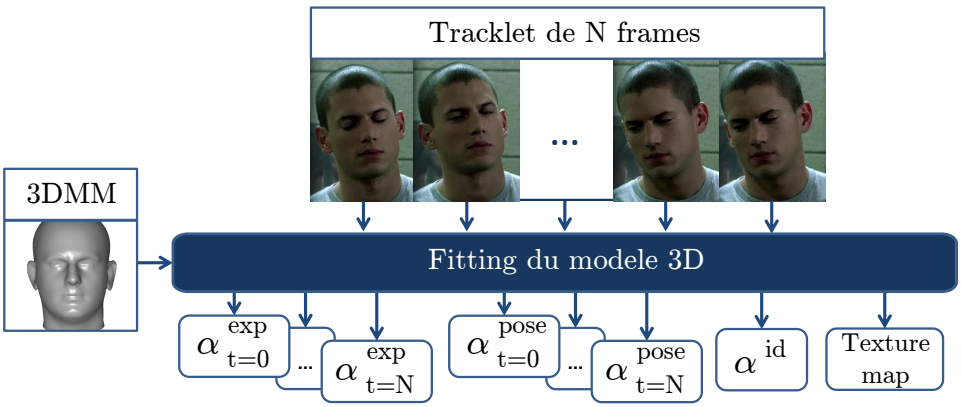


FIGURE 5.1 – *Fitting du modèle 3D sur une tracklet*

L'utilisation de plusieurs images permet d'obtenir un ajustement du modèle 3D plus précis. En effet, cet ajustement simultané sur plusieurs images permet une meilleure séparation entre les déformations liées à l'identité et celles liées à l'expression. Les parties constantes du visage auront ainsi tendance à être attribuées aux déformations d'identité tandis que les déformations non constantes à travers la vidéo seront associées aux déformations d'expression.

De plus, l'utilisation d'images supplémentaires permet de mieux contraindre le problème d'ajustement du modèle 3D.

La méthode présentée dans la figure 5.1 ne tient compte que partiellement des informations temporelles. En effet, les coefficients d'expression et les paramètres de pose sont calculés indépendamment sur chaque image.

Nous proposons donc d'ajouter une nouvelle fonction de coût lors du fitting. Le but de cette fonction de coût est d'assurer une continuité dans la variation des déformations d'expression au cours de la tracklet.

Pour limiter cette variation, nous proposons de minimiser la dérivée seconde de ces coefficients. La fonction de coût associée est alors définie par :

$$c_t = \left(\frac{d^2 \alpha_t}{dt^2} \right)^2 \tag{5.1}$$

Cette nouvelle fonction de coût est approximée par différences finies :

$$c_t \approx \left(\frac{(\alpha_t - 2\alpha_{t-1} + \alpha_{t-2})}{\Delta t^2} \right) \tag{5.2}$$

L'énergie globale à minimiser devient donc :

$$E = E_{data} + \lambda_{id} E_{reg}^{id} + \lambda_{exp} E_{reg}^{exp} + \lambda_{smooth} E_{smooth} \tag{5.3}$$

avec

$$E_{smooth} = \sum_{i=1}^{n_{coeff}^{forme}} \sum_{t=3}^{N_{frames}} \frac{\alpha_{i,t} - 2\alpha_{i,t-1} + \alpha_{i,t-2}}{\Delta t^2} \tag{5.4}$$

Le graphique suivant (Figure 5.2) montre l'impact de cette nouvelle contrainte de cohérence temporelle sur le premier coefficient d'expression.

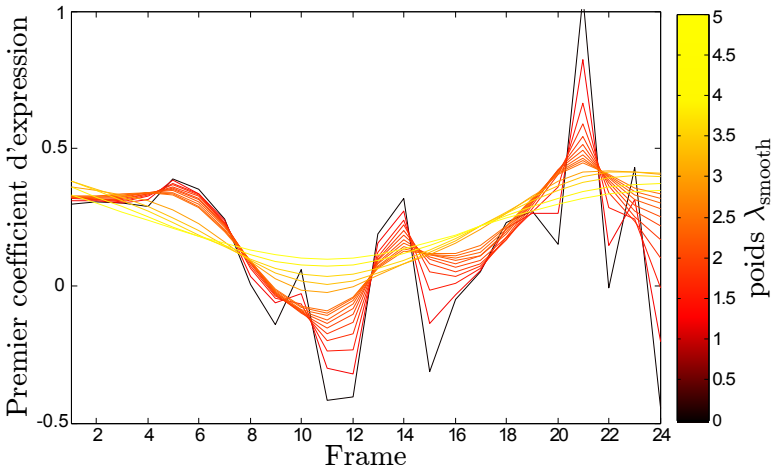


FIGURE 5.2 – *Impact de la contrainte de cohérence temporelle*

Il montre l'évolution temporelle de la valeur du premier coefficient d'expression pour différents poids λ_{smooth} de la fonction de lissage temporel.

Pour obtenir ce graphique, le modèle 3D a été ajusté sur chacune des 24 images composant une tracklet relative à un individu. Plusieurs ajustements ont été effectués avec des valeurs différentes de λ_{smooth}. La courbe noire de cette figure (correspondant à un poids λ_{smooth} nul) montre le résultat obtenu lorsqu'aucune contrainte de lissage temporelle n'est utilisée. A l'opposé, la courbe jaune, montre le résultat obtenu lorsque la contrainte de cohérence temporelle est la plus élevée.

L'ajustement de ce paramètre α_{smooth} est effectué en fonction des caractéristiques de la vidéo utilisée. Un *frame rate* élevé entrainera une valeur α_{smooth} élevée.

5.2 Evaluations expérimentales

5.2.1 Protocole de test

Un protocole d'évaluation de la reconnaissance faciale à partir de vidéos issues de la série *Prison Break* a été proposé par Biaud *et al.* dans [14]. Nous proposons d'évaluer notre méthode en utilisant ce même protocole de test dont nous rappelons le principe ici.

Une méthode de suivi de personne est appliquée sur chacune des vidéos afin de la découper en un certain nombre de tracklets, chaque tracklet correspondant à un individu.

Dans [14], un vecteur caractéristique de visage est extrait de chaque tracklet de la vidéo avant d'être comparé à une collection d'images de galerie. Cette base de galerie est composée d'images des acteurs principaux présents dans la vidéo ainsi que d'un grand nombre d'images de bruit (personnes qui ne sont pas présentes dans la vidéo). Cette base est ainsi composée de 174 images d'acteurs et de 1821 images de bruits issues de la base LFW [34].

La métrique V1, proposée par Biaud *et al.* , permet d'évaluer les performances de différents algorithmes de reconnaissance faciale. Mise au point pour être utiliser sur des données non étiquetées, cette métrique propose de considérer une tracklet comme correctement identifiée si les deux conditions suivantes sont vérifiées :

— Le premier template matché est un des 174 templates d'acteurs.
— Au moins 3 des 8 premiers template matchés sont issus du même individu que le premier template matchant.

Nous proposons donc d'utiliser cette métrique pour évaluer notre méthode. De manière similaire aux expérimentations des chapitres précédents, nous comparons deux configurations de pré-traitements appliquées à un algorithme standard de reconnaissance faciale.

Configuration A Un modèle 3D déformable standard (sans déformations d'expression) est utilisé lors du pré-traitement des tracklets. L'ajustement de ce modèle 3D sur l'ensemble de la tracklet permet le calcul d'un unique jeu de coefficients d'identité. Une vue synthétique peut ensuite être générée en utilisant ces paramètres d'identité et des paramètres de pose correspondant à une pose frontale. L'information de texture utilisée lors du rendu est extraite d'une unique image de la vidéo. Cette image est choisie selon certains critères de qualité (Résolution du visage, pose apparente du visage, ...).

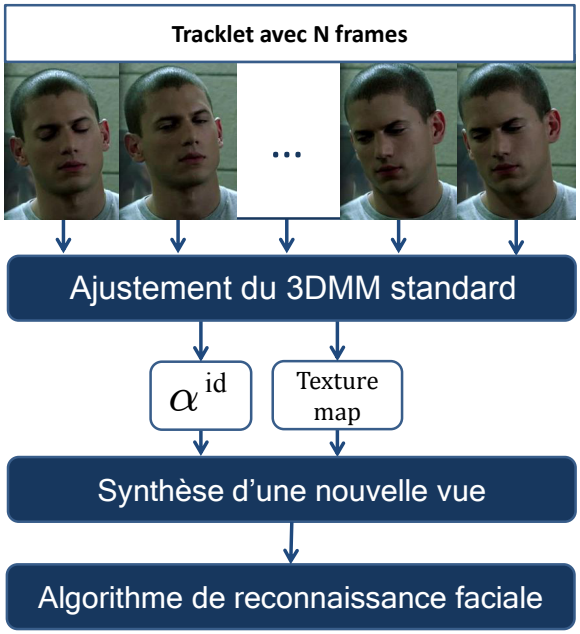


FIGURE 5.3 – *Configuration A (Baseline)*

Configuration B Dans cette configuration, nous évaluons le processus proposé dans ce chapitre permettant l'ajustement du modèle 3D étendu sur une vidéo.

Le modèle déformable 3D de visage étendu est ajusté sur les n images de la tracklet afin d'extraire un unique jeu de coefficients d'identité et n jeux de coefficients d'expression. Une nouvelle vue synthétique est ensuite générée en utilisant les coefficients d'identité extraits, des paramètres de pose permettant une vue frontale et des coefficients d'expression correspondant à une expression neutre. L'information de texture utilisée lors de cette synthèse est extraite, de manière similaire à la configuration A, à partir d'une unique image de la vidéo.

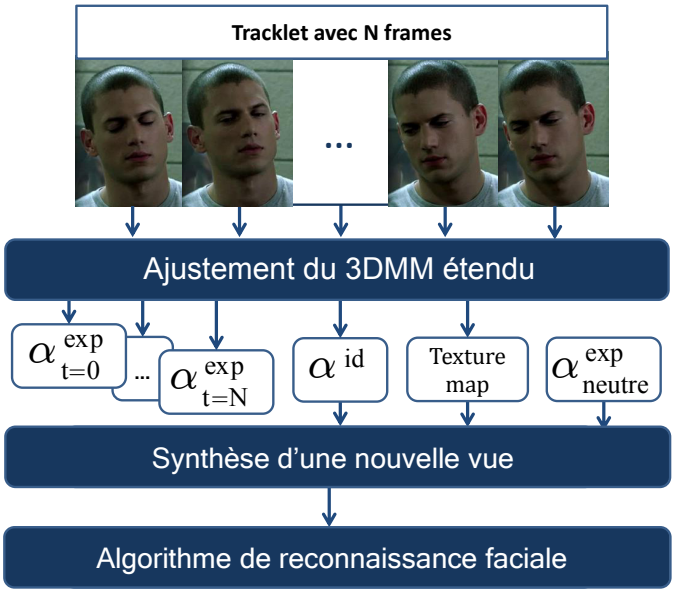


FIGURE 5.4 – *Configuration B (Méthode proposée)*

Nous présentons dans le graphique suivant le taux de tracklets vali-

dées selon la métrique V1 avec les configurations A et B.

Afin d'évaluer l'influence de la taille de la fenêtre glissante *i.e.,* le nombre d'images consécutives utilisées lors de l'ajustement du modèle 3D déformable, seules les tracklets d'au moins 11 images sont conservées. 362 tracklets composent donc le jeu de test.

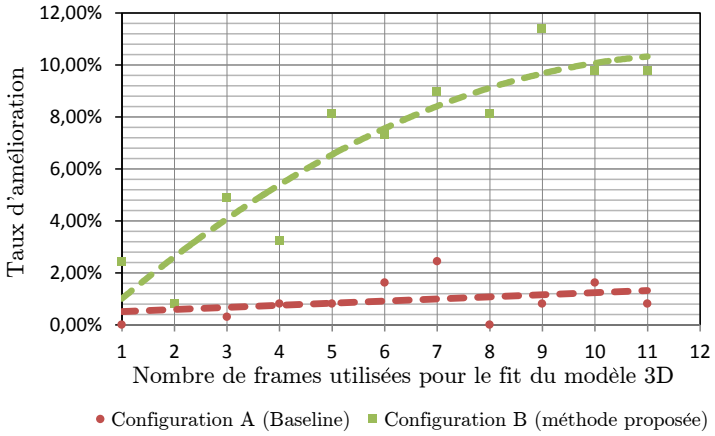


FIGURE 5.5 – *Amélioration du nombre de tracklets validées selon la métrique V1 avec les configurations A et B*

Le graphique montre clairement une amélioration des performances lorsque plusieurs images sont utilisées pour ajuster le modèle 3D.

Lorsque le modèle 3D utilisé ne contient que des déformations d'identité (Configuration A), une légère amélioration (de l'ordre de 1.5%) est observée lorsque 10 images supplémentaires sont utilisées pour ajuster le modèle 3D. Cette amélioration peut s'expliquer par une compensation d'un certain nombre de défauts pouvant apparaitre lors de l'initialisation de l'étape d'ajustement du modèle 3D.

Cette amélioration relative atteint les 10% lorsque le modèle 3D étendu avec des variations d'expression est utilisé. Cette amélioration peut être expliquer à travers différentes pistes :
Similairement à la configuration A, la compensation des outliers apparus lors de la détection des informations de textures permettant l'initialisation de l'ajustement du modèle 3D explique une partie du gain.

Une amélioration est également observée lorsqu'une seule image est utilisée pour estimer la forme 3D du visage. Ce gain s'explique par la neutralisation de l'expression.

Enfin, la troisième cause d'amélioration est une meilleure séparation entre les déformations liées à l'identité de celles relatives à l'expression. En effet, une séparation plus précise des déformations du visage peut être obtenue lorsque plusieurs images sont utilisées pour ajuster le 3DMM. Une partie des déformations du visage (déformations extra-identité) sont constantes tout au long de la tracklet. La méthode proposée dans ce chapitre impose de ne calculer qu'un seul jeu de coefficients d'identité pour

toute les images de la vidéo tandis que N jeux de coefficients d'expression sont calculés (un pour chaque image). Au contraire, le visage d'un individu peut varier à travers la tracklet. Ces déformations intra-identité sont alors expliquées par l'expression.

Grâce à ces différents points, une meilleure approximation de la forme 3D du visage peut être obtenue. Celle-ci permet la génération d'une nouvelle vue synthétique de meilleure qualité.

Conclusion du chapitre

Dans ce chapitre, nous avons étendu la méthode de neutralisation présentée dans le chapitre précédent aux vidéos. L'ajustement du modèle 3D sur une tracklet de N images aboutit à un unique jeu de coefficients d'identité et à N jeux de coefficients d'expression. Ainsi, la variation de la forme du visage à travers la vidéo est expliquée par les paramètres d'expression tandis que la partie constante de la forme du visage est expliquée par les coefficients d'identité.

A travers l'expérimentation effectuée sur des séquences vidéos extraites de la série Prison Break, nous avons pu montrer un gain de l'ordre de 10% des performances obtenues avec un algorithme standard de reconnaissance faciale lorsque cette méthode est utilisée en tant que prétraitement.

6 Améliorations de la méthode de neutralisation de l'expression

Sommaire

Dans ce chapitre, nous présenterons un certain nombre de pistes d'amélioration qui ont été suivies lors de cette thèse afin d'améliorer les performances obtenues avec la méthode de neutralisation des expressions.

Des analyses approfondies des résultats obtenus et des cas d'erreurs observés (Image refusée à tort ou acceptée à tort) nous ont permis de définir certaines limitations de notre méthode.

En premier lieu, les poids des différentes fonctions de coût minimisées durant le fitting ont été ajustées, dans le chapitre 4.4, avec l'objectif d'assurer une non régression sur des bases de données sans variation d'expression. Il apparait que ces paramètres limitent l'amplitude des déformations d'expression autorisées. Cela induit donc, en contre partie, une dégradation des performances obtenues sur des bases avec expressions.

En second lieu, certaines images neutralisées conservent un résidu d'expression. En effet, l'une des conséquences de la présence d'expression est l'apparition de rides dynamiques. Ces rides entrainent l'apparition de forts gradients sur le visage. Or, les algorithmes de reconnaissance faciale sont très sensibles à de tels gradients. Ainsi, la présence de rides d'expression provoque une dégradation des performances biométriques.

6.1 Ajustement de la prior d'expression en fonction d'une mesure de bouche ouverte

L'approximation de la forme 3D du visage, présentée dans le chapitre 4.1, s'appuie sur des connaissances *a priori* d'identité et d'expression. L'énergie de régularisation des coefficients d'identité et celle des coefficients d'expression permettent d'utiliser cette information lors de l'ajustement du modèle 3D.

Nous avons pu voir précédemment que le poids associé à l'énergie E_{reg}^{exp} conditionnait fortement la qualité du pré-traitement proposé. Pour obtenir des performances optimales sur des images caractéristiques de situations réelles, la valeur du poids λ_{reg}^{exp} a été augmentée. Bien que permettant une amélioration significative des performances sur l'ensemble des bases testées dans le chapitre précédent, l'augmentation de la valeur de λ_{reg}^{exp} diminue la précision de l'ajustement du modèle 3D sur des images avec une expression importante (Figure 6.1).

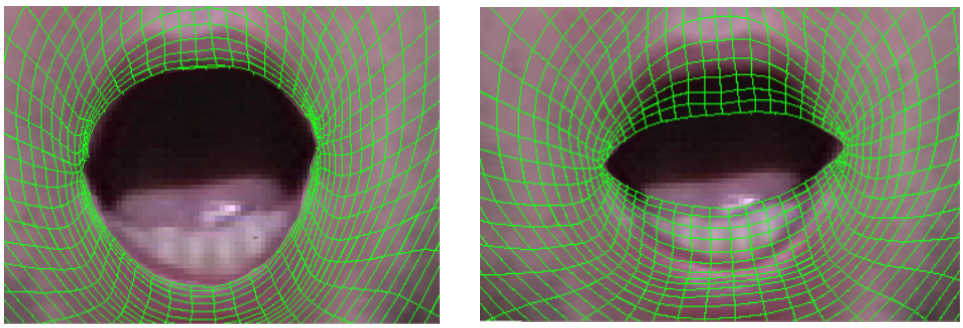


Figure 6.1 – *Résultat de l'ajustement de modèle avec un poids de l'énergie de régularisation des coefficients d'expression faible (à gauche) et élevée (à droite)*

Cette diminution de précision de l'ajustement du modèle peut s'expliquer par la définition de l'énergie de régularisation des coefficients d'expression. En effet, l'énergie de régularisation relative aux paramètres d'expression minimisés durant l'ajustement du modèle 3D est définie par :

$$E_{reg}^{exp} = \sum_{i=1}^{N} \frac{\alpha_i^2}{\sigma_i^2} \tag{6.1}$$

Cette énergie tend à régulariser tous les coefficients d'expression vers zéro. Ainsi, l'hypothèse faite lors de la définition de cette énergie est la présence d'une expression neutre dans l'image de probe. L'observation des résultats d'ajustement du modèle 3D (Figure 6.1) nous montre que l'hypothèse faite ici est trop forte dans certains cas. En effet, elle impose une contrainte trop forte limitant l'amplitude des déformations d'expression autorisées lors de l'ajustement du modèle 3D. Deux approches peuvent alors être suivies pour relâcher cette contrainte sur les paramètres d'expression.

Dans un premier cas, une diminution du poids associé à cette énergie de régularisation des coefficients d'expression peut permettre un meilleur ajustement du modèle 3D. La diminution de ce poids $\alpha_{reg}^{expression}$ induit

une variation plus importante des coefficients d'expression et donc une meilleure approximation de la forme du visage en présence d'expression. La contre-partie est une perte de précision sur les images sans expression (Section 4.4.3). De plus, la diminution de $\alpha_{reg}^{expression}$ ne modifie pas l'*a priori* d'expression neutre dans l'image de probe. Nous proposons donc ici de suivre une seconde approche en modifiant de manière dynamique l'*a priori* d'expression utilisée lors de l'ajustement du modèle 3D.

La modification de la prior d'expression en fonction de l'image d'entrée permet d'homogénéiser la qualité sur l'ajustement du modèle 3D. L'objectif est donc d'obtenir un niveau de performance optimale à la fois sur des images sans expression et sur des images avec expression.

Pour cela, nous proposons d'ajuster la valeur a priori du coefficient d'expression du modèle 3D relatif à la principale source de déformation intra-classe de la forme du visage : L'ouverture de la bouche.

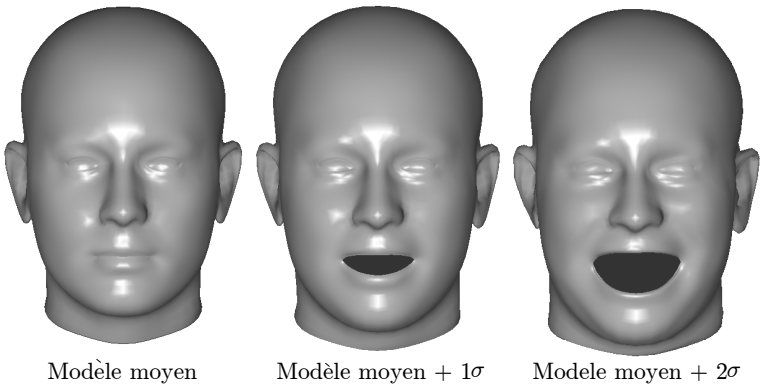


FIGURE 6.2 – *Première déformation du modèle déformable 3D de visage*

Celle-ci, étant la déformation principale relative à l'expression, est associée à la première déformation d'expression de notre modèle 3D (Figure 6.2). L'ajustement de la prior d'expression sera donc fait en modifiant la valeur cible de régularisation du premier coefficient d'expression. Cette valeur est déterminée en fonction d'un détecteur d'ouverture de bouche [75].

Ce détecteur permet d'indiquer l'amplitude de l'ouverture de la bouche à travers un scalaire $\alpha_{mouth\,open}$. Une bouche grande ouverte dans l'image I entrainera un $\alpha_{mouth\,open}$ faible tandis qu'une bouche fermée conduira à un $\alpha_{mouth\,open}$ elevé (Figure 6.3).

$\alpha_{mouth\,open} = 0.3$ $\alpha_{mouth\,open} = 9.3$

FIGURE 6.3 – *Mesure $\alpha_{mouth\,open}$ sur différentes images*

Le graphique suivant 6.4 montre la corrélation entre la valeur du premier coefficient d'expression et la mesure d'ouverture de bouche sur une collection de plus de 5000 images. Au cours de cette expérience, le processus d'ajustement du 3DMM a été guidé par un certain nombre d'annotations manuelles. Celles-ci permettent, grâce à une diminution du poids associé à l'information a priori de forme, une meilleure approximation de la forme 3D du visage.

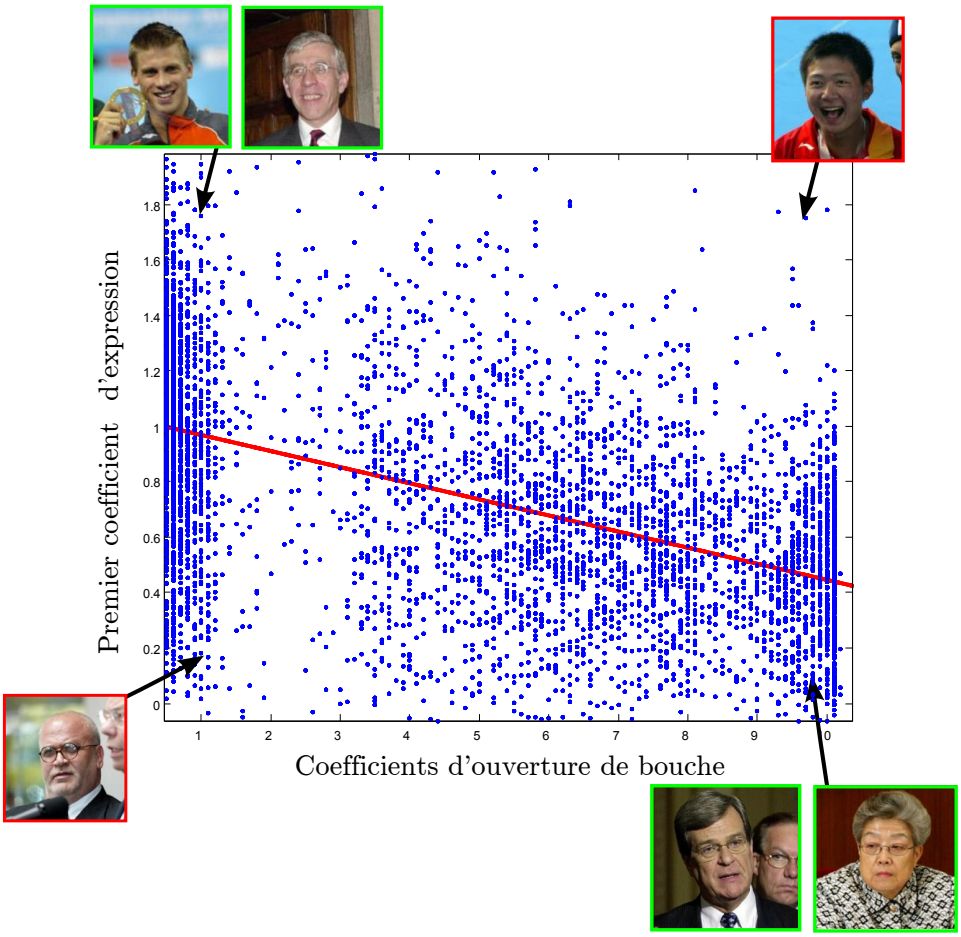


FIGURE 6.4 – *Valeur du premier coefficient d'expression en fonction de la mesure d'ouverture de bouche*

Ce graphique nous montre une corrélation entre la mesure d'ouverture de bouche et la valeur du premier coefficient d'expression du modèle 3D ajusté. Cependant, sur ce graphique, un certain nombre de points est éloigné de la droite de régression. L'approximation de la forme 3D, ayant été guidée par des annotations manuelles, la confiance accordée à l'estimation du premier coefficient d'expression est relativement élevée. Au contraire, le détecteur d'ouverture de bouche utilisé au cours de cette expérience provoque un certain de nombre de classifications erronées. La figure 6.4, où des images correctement classifiées (encadrées en vert) et d'autres dont la classification est erronée (encadrées en rouge) sont affichées, illustre ce problème.

Malgré ces imprécisions dans le détecteur d'ouverture de bouche, nous proposons d'utiliser la corrélation extraite de ce graphique pour ajuster de manière dynamique la valeur a priori de ce premier coefficient d'expression.

Pour intégrer cette modification de prior d'expression, une nouvelle étape est donc ajoutée dans le processus d'ajustement du modèle 3D. Premièrement, une mesure de l'ouverture de la bouche est effectuée sur l'image d'entrée. La fonction de corrélation déduite de la figure 6.4 permet alors de calculer la nouvelle valeur a priori du premier coefficient d'expression $\alpha^{prior}_{expression}$ à partir de $\alpha_{mouth\,open}$. La figure 6.5 montre le nouveau diagramme fonctionnel du processus d'ajustement du modèle 3D.

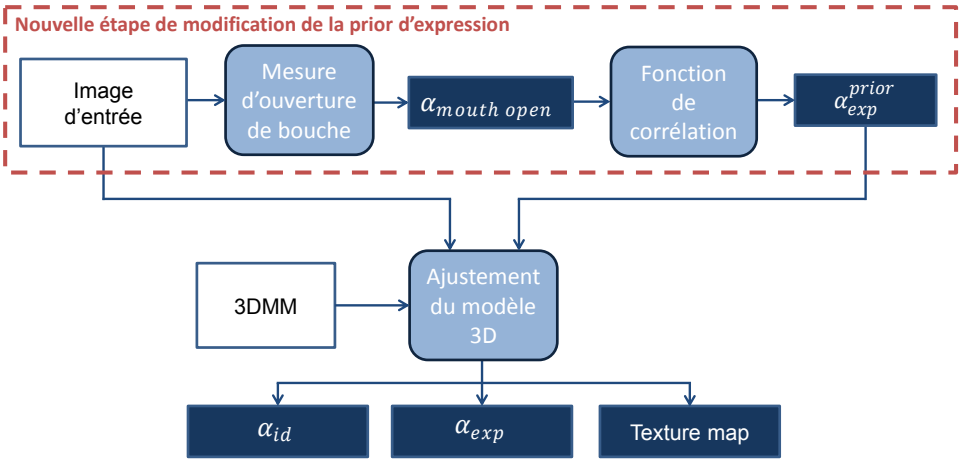


FIGURE 6.5 – *Modification de la prior d'expression lors de l'ajustement du modèle 3D*

La nouvelle fonction de régularisation des coefficients d'expression est alors :

$$E^{exp}_{reg} = \frac{\left(\alpha_1 - \alpha^{prior}_{exp}\right)^2}{\sigma_1^2} + \sum_{i=2}^{N} \frac{\alpha_i^2}{\sigma_i^2} \tag{6.2}$$

Cette modification dynamique de la prior d'expression permet d'améliorer la qualité d'ajustement du modèle 3D dans le cas d'images avec

expression tout en limitant l'impact sur les images sans expression.

En effet, sur les images avec une expression importante (comme par exemple lorsque l'individu sourit), la valeur de $\alpha_{mouth\,open}$ est faible. La nouvelle prior d'expression utilisée est alors proche de 1. Cela entraine l'utilisation d'une forme a priori du modèle 3D proche de celle présentée dans la figure 6.2 (Modèle moyen + 1σ).

La figure 6.6 montre des exemples de neutralisation de l'expression standard (à gauche) et en utilisant notre méthode de modification de la prior d'expression (à droite).

FIGURE 6.6 – *Exemples d'images rendues sans modification de la prior d'expression (à gauche) et avec modification de la prior d'expression (à droite)*

Ces exemples permettent de visualiser l'apport de cette méthode de modification de la prior d'expression sur des images avec une expression. La neutralisation des images sans expression n'est, pour sa part, que très peu impactée par cette méthode. Cependant, un risque inhérent à une mesure incorrecte de $\alpha_{mouthopen}$ existe. En effet, une incohérence dans la mesure de cet indicateur conduirait à l'utilisation d'une prior non adaptée lors de l'ajustement du modèle 3D.

6.2 Rides d'expression

Nous présentons dans cette section, d'autres travaux effectués dans le but d'améliorer le processus de neutralisation des expressions.

Comme nous avons pu le voir à travers les chapitres précédents, la neutralisation de l'expression repose sur une modification de la forme 3D du visage. Cette correction de la forme du visage permet de modifier la position spatiale de chacun des points du visage en les replaçant à la position qu'ils occupent lorsque tous les muscles de visage sont relâchés (ce qui correspond à l'expression neutre).

L'information de texture est, quant à elle, extraite depuis l'image d'entrée en utilisant les informations du modèle 3D déformé sur cette image : L'extraction de la carte de texture peut alors être comparée à un remapping de l'image d'entrée dans l'espace des cartes de texture.

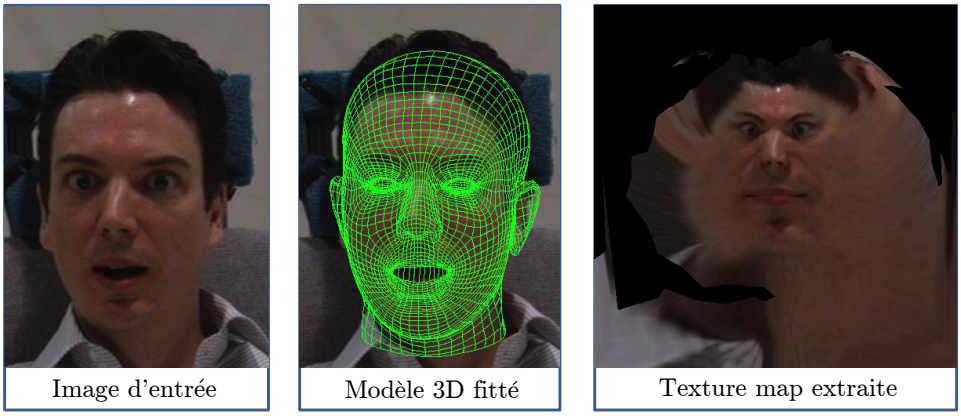


FIGURE 6.7 – *Extraction de la texture map depuis l'image d'entrée*

Cette carte de texture extraite est ensuite utilisée lors de la synthèse de la nouvelle vue sans aucune modification.

Cette technique permet d'obtenir, dans la majorité des cas, des résultats réalistes tout en limitant le temps de calcul. Cependant, cette synthèse de vue peut contenir des résidus d'expression dans certains cas. En effet, l'apparence du visage peut être modifiée par la présence de rides liées à l'expression.

Ces rides, également appelées rides dynamiques, sont provoquées par la contraction des muscles peauciers. La majorité de ces rides se situe dans la région du sillon naso-génien.

FIGURE 6.8 – *Rides d'expression dans la région du sillon naso-génien*

La figure 6.8 montre des exemples d'images présentant des rides dans la région du sillon naso-génien.

Dans le processus actuel de neutralisation de l'expression, ces modifications d'apparence de la texture sont conservées lors de l'extraction de la carte de texture. La nouvelle image synthétisée à partir celle-ci comporte également ces artefacts d'expression. Les résultats de notre méthode de neutralisation de l'expression appliquée aux images précédentes sont présentés dans la figure 6.9.

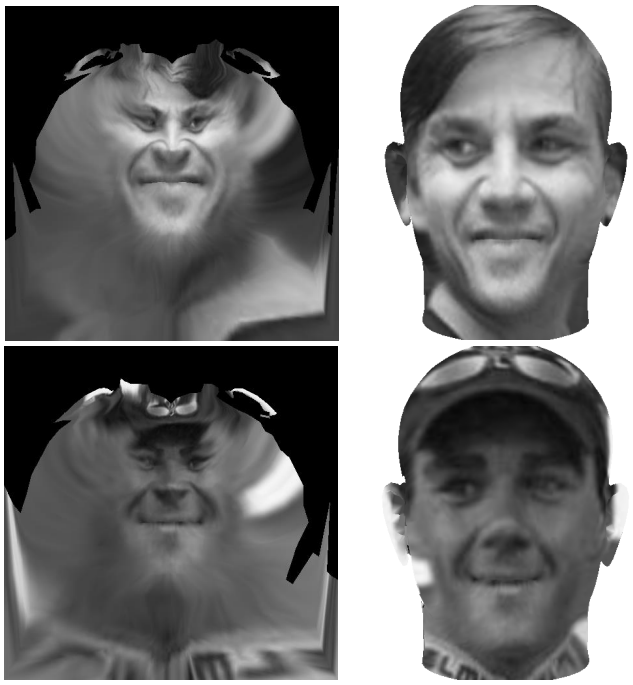


FIGURE 6.9 – *Cartes de texture extraites et images neutres synthétisées*

Les artefacts liés aux rides d'expression apparaissent clairement dans la carte de texture extraite et dans la nouvelle vue neutre synthétisée. Les conséquences de ces résidus d'expression dans les images neutralisées sont multiples.

D'une part, ils empêchent d'avoir un rendu réaliste des images neutralisées. Naturellement, ce type de rides ne peut apparaître lorsque la bouche est fermée. C'est la présence de ces deux caractéristiques incompatibles sur la même image qui empêche de percevoir l'image comme réaliste.

D'autre part, ces rides déstabilisent les algorithmes de reconnaissance de visage qui les considèrent comme une information discriminante du visage. En effet, dans la majorité des cas, la zone du sillon naso-génien ne comporte pas ou peu d'information permettant de discriminer les individus. La présence de forts gradients dans ces zones habituellement non discriminantes entraine donc une dégradation des performances de reconnaissance. La présence de rides dans des zones similaires du visage entraine par exemple une augmentation du score de similarité lorsque les deux images comparées comportent des rides similaires.

Il est donc nécessaire d'effectuer une correction de la carte de texture pour éviter cette dégradation des performances. Cette correction de la texture map peut être intégrée dans différentes étapes de la chaine (Figure 6.10).

FIGURE 6.10 – *Extraction et utilisation de la carte de texture pour synthétiser une nouvelle vue*

La correction de la texture map peut ainsi être effectuée soit lors de son extraction depuis l'image d'entrée (1) soit lors de la génération de la nouvelle vue synthétique (2).

Au moment de son extraction depuis l'image d'origine, la carte de texture peut être corrigée pour ne contenir que l'information de couleur intrinsèque du pixel. La présence de rides entraine un phénomène d'ombrage. La couleur apparente de cette zone est donc modifiée. Il est alors nécessaire d'utiliser un modèle d'illumination. En considérant un modèle basique, l'illumination de chaque point du modèle 3D peut être calculée en fonction de :

- La normale à la surface
- Le point de vue de l'observateur
- La couleur intrinsèque de ce point
- La source d'illumination de la scène

Pour ce faire, un modèle 3D plus précis doit être utilisé (le maillage du modèle 3D doit être suffisamment fin pour permettre une bonne approximation géométrique des rides). Le modèle déformable proposé par Vlasic *et al.* [64] propose une telle représentation. Il permet de générer des formes de visages détaillées. La figure 6.11 montre des exemples de visages 3D générées avec ce modèle.

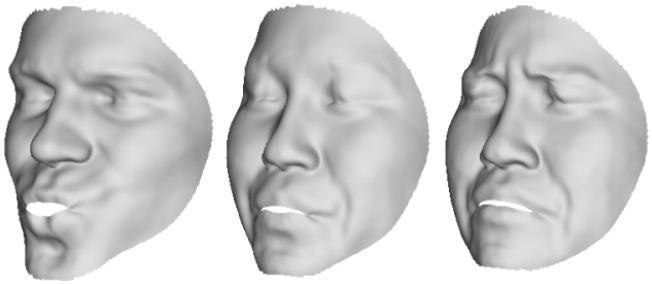


FIGURE 6.11 – *Exemples de visages 3D générées par Vlasic et al. [64]*

L'information géométrique apportée par un modèle 3D de plus haute résolution permet une meilleure approximation de la surface du visage. Pour notre problématique, l'utilisation d'un maillage plus fin permet notamment une estimation plus fine des normales situées dans les zones sujettes à des déformations lorsque des rides apparaissent.

Surface du visage au repos Apparition de rides sur la surface du visage

FIGURE 6.12 – *Pertubations des normales au visage liées à l'apparition de rides*

L'approximation de ces normales peut alors nous permettre d'estimer la variation d'illumination apportée par la déformation de la surface. Cette information peut ensuite être prise en compte lors de l'extraction de la carte de texture pour obtenir la couleur intrinsèque du point et non sa couleur apparente liée aux modifications des normales à la surface en ce point du visage.

Le principal inconvénient de cette méthode est une forte augmentation du temps de pré-traitement. En effet, l'utilisation d'un modèle plus défini entraine une augmentation importante du temps d'ajustement du modèle 3D. De plus, la correction de l'information colorimétrique pour chacun des pixels de la carte de texture est très couteux. Ces contraintes nous empêchent donc d'envisager cette méthode pour un contexte temps réel.

Nous avons donc choisi de nous orienter vers une deuxième stratégie en proposant de modifier la carte de texture après son extraction pour supprimer les artefacts causés par la présence de rides.

Cette opération sera effectuée en deux étapes. Dans un premier temps, nous proposons une méthode de détection de présence de rides. Une fois cette détection de rides effectuée, nous appliquerons une méthode d'inpainting pour reconstruire la partie du visage associée à ces rides.

6.2.1 Détection de la présence de rides

Nous souhaitons, au cours de cette étape, déterminer si la carte de texture contient des artefacts liés à la présence de rides d'expression dans l'image d'origine. Travailler directement sur la carte de texture présente un avantage important : La correspondance établie entre toutes ces données permet de faciliter la localisation des zones probables d'apparition de rides. Ces zones peuvent être multiples : Zone du sillon naso-génien, zone frontale...

Les rides du sillon naso-génien sont les plus impactantes sur les algorithmes de reconnaissance faciale. Nous avons donc choisi de nous concentrer sur cette catégorie dans la suite.

Notre méthode de détection est basée sur une approximation basique de ces rides (Figure 6.13).

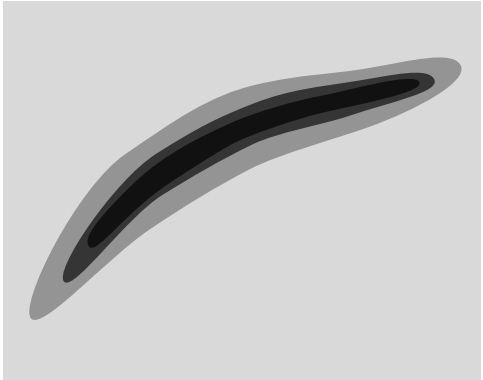


FIGURE 6.13 – *Approximation d'une zone de ride*

Comme nous le montre cette figure, une ride peut être représentée par un ensemble de zones concentriques dont l'intensité lumineuse décroit en fonction de sa proximité au centre. Nous proposons donc d'approcher la zone susceptible de contenir des rides par ce type de modèle avant de conclure en fonction de la qualité de cette approximation. Le processus de détection de rides est donc basé sur les étapes suivantes :

- Recadrage de la zone d'intérêt des rides.
- Partitionnement de cette zone en trois parties distinctes.
- Classification basée sur la variabilité des centroïdes des zones.

La première étape (Recadrage de la zone d'intérêt) est immédiate grâce à la correspondance entre l'ensemble des cartes de textures. La position des zones d'intérêt est donc constante pour toutes les cartes de textures. Une annotation manuelle de ces rides sur de nombreuses images permet de définir cette zone (Figure 6.14).

Une fois la zone d'intérêt extraite, un partitionnement en trois zones est effectué en se plaçant dans l'espace colorimétrique HSV (Hue, Saturation, Value). Dans cet espace, la *value* est définie par $V = \frac{R+G+B}{3}$. Ce canal représente donc l'intensité lumineuse des pixels.

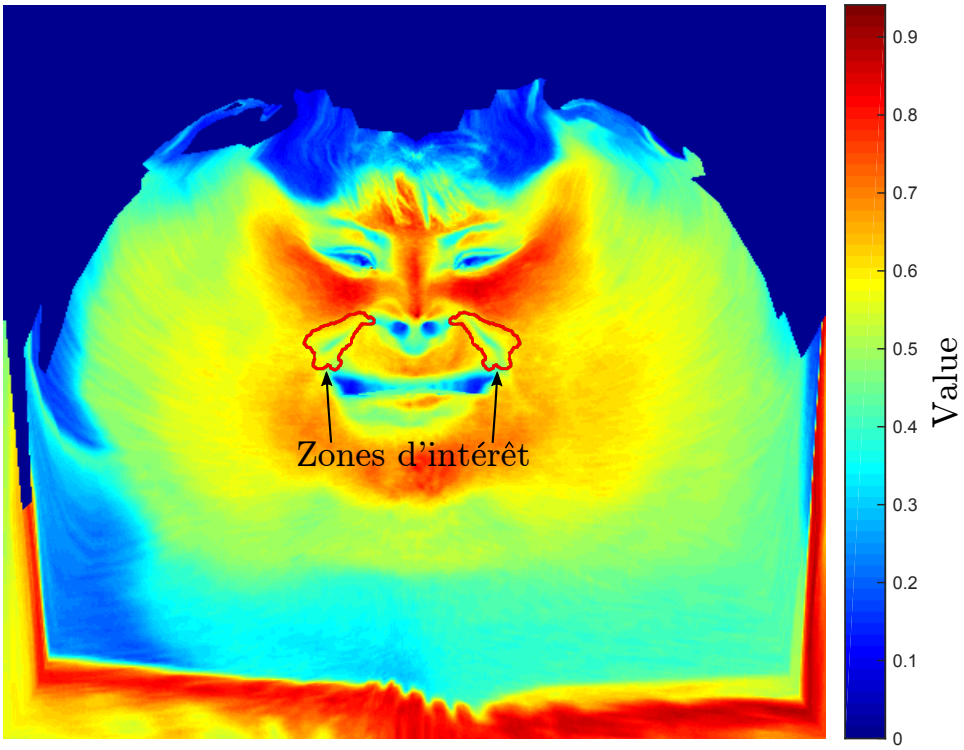


FIGURE 6.14 – *Canal Value d'une texture map*

La particularité des rides étant une modification de l'intensité des pixels, le partitionnement de la zone d'intérêt sera donc effectué sur le canal *Value* de l'espace HSV (Figure 6.14).

Le partitionnement des N_p pixels de la zone d'intérêt est effectué à partir de l'histogramme de distribution des pixels selon leur value. Ces pixels sont partitionnés en trois classes.

Les classes C_1, C_2 et C_3 sont définies ainsi :

$$C_1 = [value_{min}; value_1] \; C_2 =]value_1; value_2] \; C_3 =]value_2; value_{max}]$$

tel que :

$$value_{min} < value_1 < value_2 < value_{max}$$

et

$$card\{C_1\} = card\{C_2\} = card\{C_3\} = \frac{N_p}{3}$$

La figure 6.15 montre le résultat de ce partitionnement sur l'histogramme des values des pixels de la zone d'intérêt.

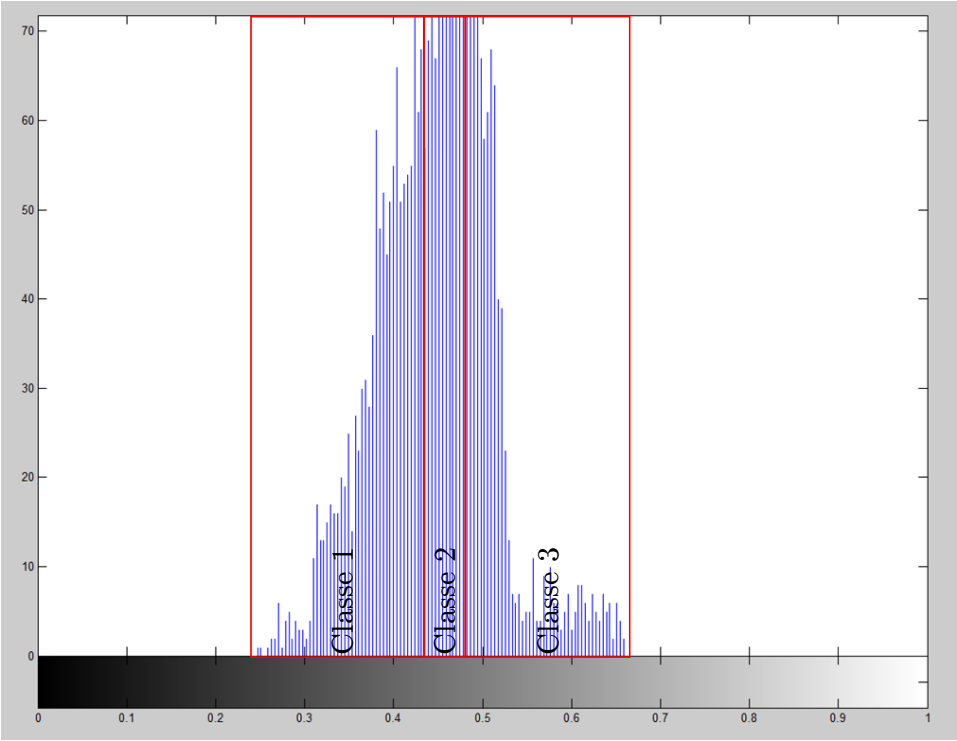


FIGURE 6.15 – *Classification à partir de l'histogramme (Canal Value)*

Cette classification nous permet ainsi de découper la zone d'intérêt en trois classes. La figure 6.16 montre le résultat de cette classification sur une image positive (présence de rides d'expression). Il apparait clairement que les trois zones obtenues sont englobantes.

FIGURE 6.16 – *Canal Value de la zone d'intérêt (à gauche) et résultat du clustering en trois classes de cette zone (à droite)*

La figure 6.17 montre les résultats de la classification sur des images négatives (sans rides d'expressions).

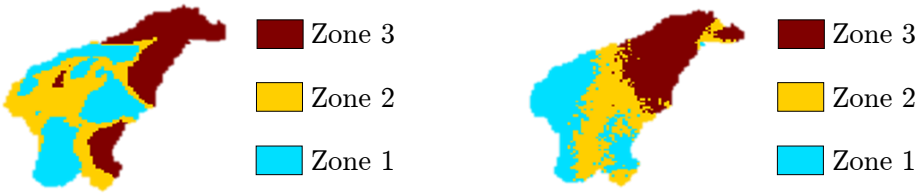


FIGURE 6.17 – *Résultat de la classification sur des images sans rides*

Le résultat de la classification est beaucoup moins structuré sur des exemples négatifs. Pour déterminer la concentricité de ces zones et ainsi déterminer la présence ou non de rides d'expression, nous proposons d'analyser la position des centroïdes de ces zones. Le centroïde d'une région est définie comme étant la position moyenne de l'ensemble des points de la zone.

Dans le cas de zones concentriques, les différents centroïdes sont proches les uns des autres (Figure 6.18). Au contraire, leur position est beaucoup plus chaotique dans le cas de zones non-concentriques (Figure 6.19).

FIGURE 6.18 – *Position des centroïdes de chaque classe (Exemples positifs)*

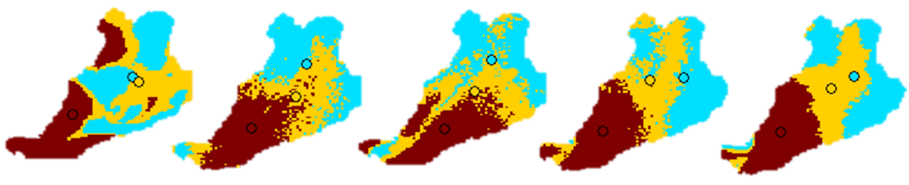


FIGURE 6.19 – *Position des centroïdes de chaque classe (Exemples négatifs)*

Nous proposons donc d'effectuer la classification de la zone en fonction des variabilités en X et en Y de la position des centroïdes des classes. La figure 6.20 nous montre l'écart type de la position des centroïdes obtenue sur des cas positifs et sur des cas négatifs.

Un SVM linéaire peut alors être appris sur ces données puis utilisé pour déterminer si des rides sont présentes dans la carte de texture. Appliquée sur un ensemble de 72 images, la classification permet d'obtenir un taux d'identification correcte supérieur à 91% .

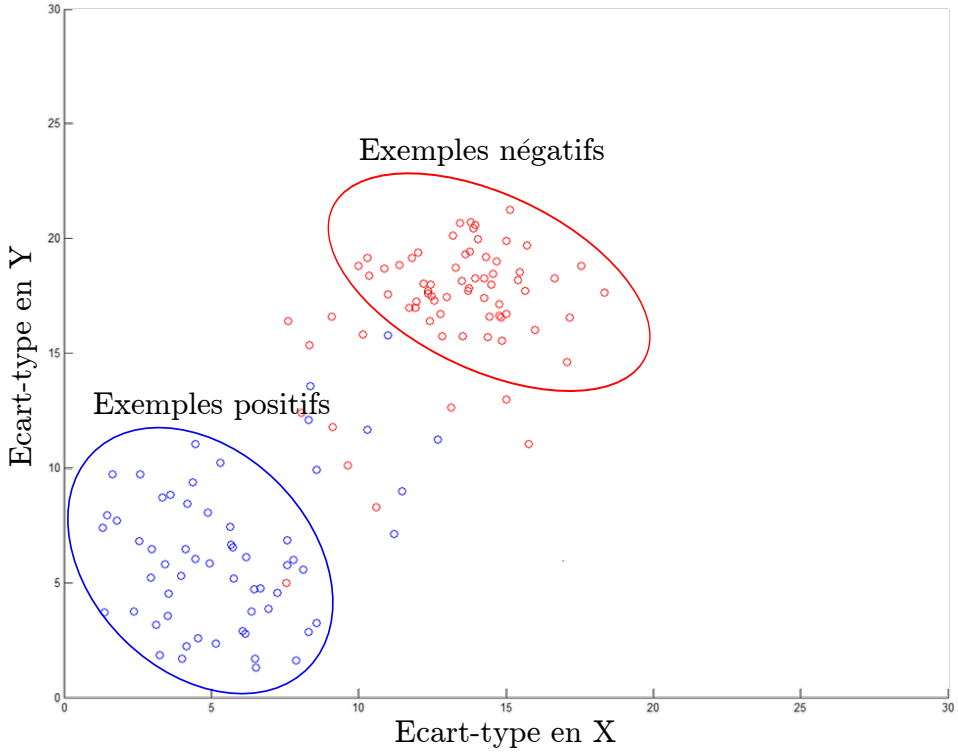


FIGURE 6.20 – *Variabilité en X et Y des centroïdes des classes)*

6.2.2 Suppression des rides

L'étape précédente nous a permis de déterminer si des rides d'expression étaient présentes ou non dans la carte de texture extraite. Une correction de l'information de texture peut alors être effectuée dans les zones où des rides ont été détectées.

La problématique de correction d'artefacts a été traitée dans de nombreux travaux de l'état de l'art. Sobiecki *et al.* [58] proposent notamment d'utiliser un pré-traitement basé sur une méthode d'inpainting pour supprimer ces artefacts. Dans ce travail, les auteurs proposent de supprimer les lunettes des images originales (Figure 6.21).

FIGURE 6.21 – *Correction de lunettes par inpainting [58]*

Dans cette section, nous proposons de suivre une approche similaire pour corriger les rides d'expressions.

Le processus complet de correction des rides d'expression peut alors être résumé par le schéma suivant :

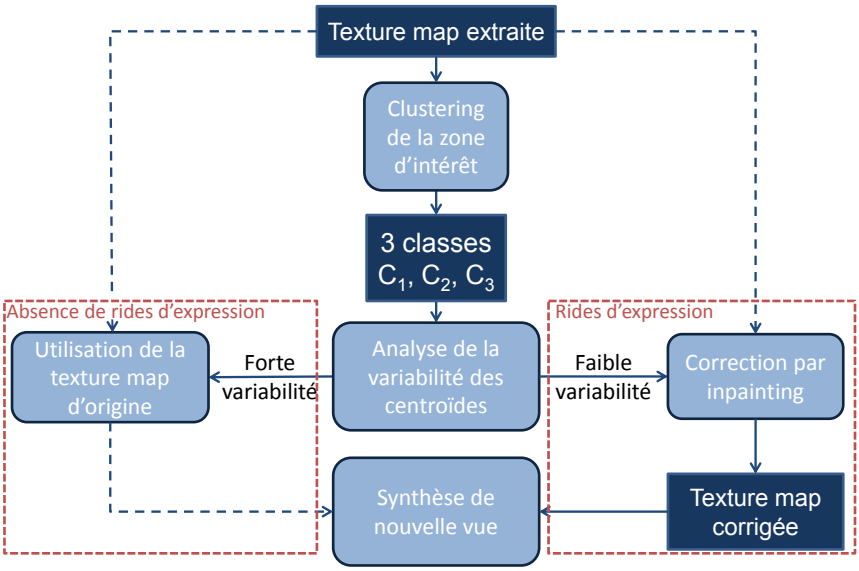


FIGURE 6.22 – *Diagramme fonctionnel du processus de correction des rides d'expression*

Les méthodes d'inpainting existantes dans la littérature permettent de reconstruire des zones de pixels manquants d'une image à partir des pixels voisins de cette région. Les techniques d'inpainting peuvent être classées en deux catégories principales :

- Les "diffusion-based", basées des équations aux dérivées partielles [10]
- Les "patch-based", basées sur des méthodes de correspondance de patchs [7]

Toutes les méthodes d'inpainting nécessitent la définition d'un masque décrivant la zone à reconstruire. Compte tenu de l'information a priori de la localisation de la zone des rides (grâce à la correspondance entre les cartes de texture), nous proposons d'utiliser un masque binaire unique pour toutes les cartes de texture. L'utilisation d'un masque unique permet un gain de temps indispensable pour une utilisation en temps réel de notre méthode.

Des exemples de correction des rides d'expression sont montrés dans la figure 6.23.

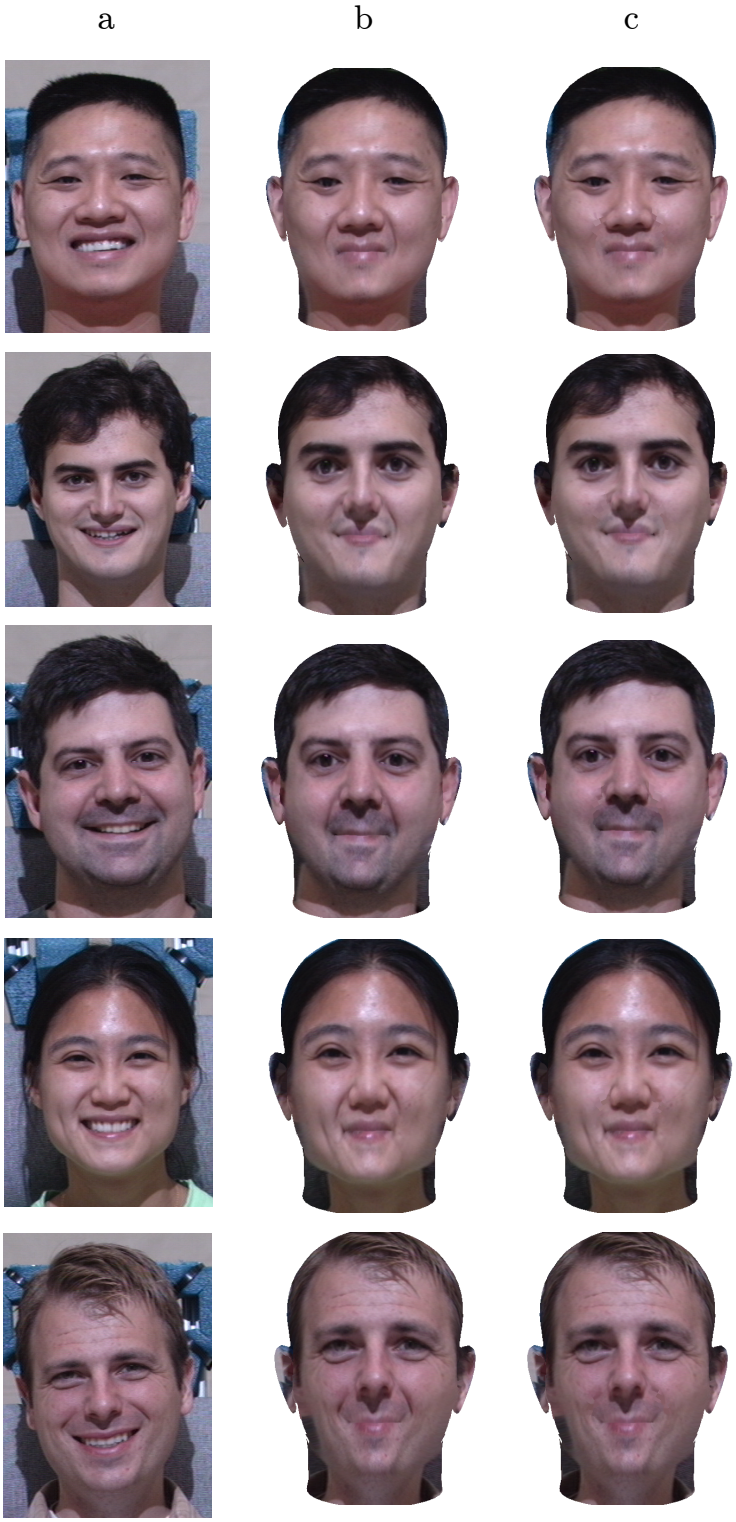


FIGURE 6.23 – *(a) Image originale, (b) Image obtenue par la neutralisation de l'expression, (c) Image obtenue après la correction des rides d'expression*

La méthode de correction des rides proposées dans cette section permet la génération d'image de meilleure qualité.

Pour quantifier l'apport de notre méthode de correction de rides, nous proposons le protocole d'évaluation expérimentale suivant :

Une base de test composée de 65 images de tests et 2085 images de référence est utilisée. Les images de la base de test ont été choisies afin de correspondre aux problématiques étudiées dans ce chapitre (Variations d'expression avec apparition de rides dynamiques).

Les résultats obtenus sur cette base de données sont proposés sous la forme de courbe de rang. Cette représentation permet d'évaluer les performances d'un système biométrique dans un contexte de comparaison 1 : N en visualisant le taux d'identification obtenu pour un rang donné. Par exemple, un taux d'identification de 90% au rang 5 signifie que pour 90% des images de tests, l'image de référence associée possède l'un des cinq scores de similarité les plus élevés.

La figure 6.24 présente les courbes en rang obtenues en utilisant d'une part le processus de neutralisation de l'expression standard et d'autre part, ce même processus complété par notre méthode de correction des rides d'expression.

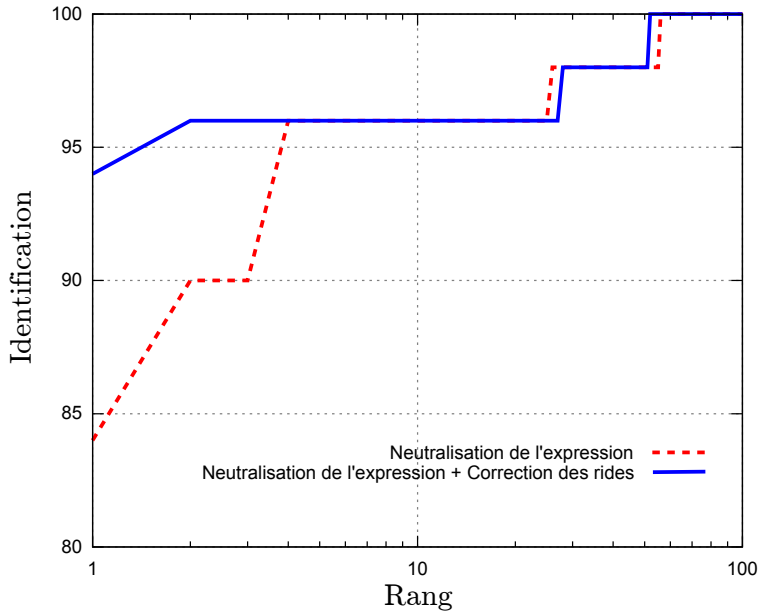


FIGURE 6.24 – *Résultat en rang*

L'amélioration apportée par la méthode de correction de rides permet d'améliorer de manière significative le taux d'identification obtenu dans les premiers rangs. Ainsi, le taux d'identification au rang 1 (Pourcentage d'images de tests dont l'image associée a le score de similarité le plus élevé) passe de 84% à 94%.

CONCLUSION DU CHAPITRE

A travers ce chapitre, nous avons présenté différentes approches permettant d'améliorer notre méthode de neutralisation de l'expression. Les

deux processus proposés ici traitent chacun de l'une des deux étapes principales de la méthode de neutralisation de l'expression.

Dans un premier temps, nous avons proposé d'améliorer la qualité de l'ajustement du modèle 3D. Ce processus présenté dans le chapitre 4 utilise l'information d'expression neutre a priori dans l'image. La conséquence principale de cette hypothèse est une approximation imprécise de la forme 3D de la région de la bouche. Nous avons donc proposé de modifier cette prior d'expression en s'appuyant sur un détecteur de bouche ouverte. Cette méthode permet d'améliorer la qualité de l'ajustement du modèle déformable sur des images avec une bouche ouverte tout en conservant une bonne approximation de la forme 3D sur des images sans expression.

Dans un second temps, nous nous sommes intéressés à une amélioration de la qualité de la synthèse de la nouvelle vue. En effet, certains des résultats obtenus par la neutralisation de l'expression conservent des artefacts liés à la présence d'expression dans l'image d'origine après neutralisation. Ces artefacts sont principalement dus à la présence de rides d'expression sur l'image originale. Tout d'abord, une détection de la présence de rides d'expression est effectuée. Si cette détection est positive, un algorithme d'inpainting est appliqué sur la zone de rides pour corriger cet artefact lors de la synthèse de la nouvelle vue. L'évaluation expérimentale conduite a permis de prouver l'efficacité de cette méthode.

Conclusion

Dans cette thèse, nous avons traité l'une des problématiques majeures de la reconnaissance faciale. La maturité de cette technologie lui permet de proposer un taux de reconnaissance élevé lorsqu'elle est appliquée à des images acquises dans de bonnes conditions (Images frontales acquises sous une illumination contrôlée et une expression neutre). Ces performances diminuent fortement lorsque les images sont issues d'acquisitions non contrôlées.

Nous avons donc axé notre travail sur ces problématiques. Plus précisément, nous nous sommes concentrés sur l'amélioration de la robustesse des systèmes actuels de reconnaissance faciale aux variations d'expression et de pose. En effet, il n'existe dans l'état de l'art que très peu de méthodes traitant de cette problématique de variations simultanées d'expression et de pose. Nous avons ainsi, durant cette thèse, proposé différentes approches de pré-traitement permettant de capitaliser sur les algorithmes existants. Basées sur la synthèse d'une nouvelle vue synthétique, ces méthodes permettent de placer ensuite les algorithmes standards de reconnaissance faciale dans leur condition optimale.

Pour accomplir cette tâche, un modèle 3D déformable de visage étendu est utilisé. La construction d'un tel modèle repose sur une analyse en composantes principales effectuée sur une large collection de scans 3D de visage. Ceux-ci sont représentatifs de la variété des visages tant en termes d'identité que d'expression. Pour permettre l'analyse statistique, l'ensemble des données a été mis en correspondance selon un processus robuste aux variations d'expression. Effectuées séparément sur les scans neutres et sur les scans avec expression, les analyses en composantes principales ont permis l'extraction des déformations d'identité et d'expression. Ainsi, la forme de n'importe quel visage peut être approchée par une combinaison linéaire de ces déformations. L'ajustement de ce modèle sur une image 2D permet ensuite d'extraire les coefficients d'identité et d'expression ainsi que les paramètres de pose et les informations de texture relatives au visage. Une fois cet ajustement effectué, une nouvelle vue synthétique peut être générée à partir de ces informations.

Pour obtenir des performances optimales de l'algorithme de reconnaissance, la méthode de neutralisation de l'expression propose de rendre l'image sous une expression neutre et une pose frontale. Une seconde méthode, conçue pour être utilisée dans un contexte de vérification (Comparaison 1:1), a été proposée. Celle-ci est basée sur un ajustement simultané du modèle 3D sur les deux images comparées. Une meilleure séparation entre les déformations d'identité et d'expression peut ainsi être

obtenue. Les nombreuses expérimentations conduites ont prouvé l'apport de ces méthodes lorsqu'elles sont appliquées en pré-traitement d'un algorithme standard de reconnaissance faciale. Par ailleurs, un nouveau protocole d'évaluation de la robustesse aux variations simultanées de pose et d'expression a été présenté.

La reconnaissance de visages à partir de vidéos est aujourd'hui une autre problématique majeure de la reconnaissance faciale. Nous avons donc choisi d'étendre notre méthode de neutralisation de l'expression à ce type de scénario. Dans cette approche, le modèle 3D est ajusté sur l'ensemble des images de la vidéo. L'ajout d'une contrainte de lissage des coefficients d'expression tout au long de la trajectoire permet notamment de robustifier l'approximation de la forme 3D du visage.

Finalement, deux processus d'amélioration de notre méthode de correction de l'expression ont été présentés. Le premier propose de modifier la prior d'expression utilisée lors de l'ajustement du modèle 3D en fonction d'un indicateur d'ouverture de bouche tandis que le second permet une correction des rides d'expression. Ces deux méthodes ont été développées dans le but d'améliorer les performances obtenues dans certaines configurations tout en conservant un comportement similaire sur autres images. Le choix de ces améliorations a été effectué à la suite d'une analyse approfondie des résultats obtenus par la méthode de neutralisation de l'expression. D'autres perspectives d'évolution peuvent être envisagées.

Perspectives

Dans cette thèse, nous avons choisi de construire notre modèle déformable à partir de deux analyses en composantes principales. Nous avons donc fait l'hypothèse que les déformations liées à l'identité et celles relatives à l'expression étaient indépendantes. Pour s'affranchir de cette hypothèse forte, un modèle déformable de visage multilinéaire peut être utilisé. Celui-ci permet, à l'aide d'une représentation tensorielle de l'espace des formes 3D de visage, de considérer les interactions entre identité et expression. Cependant, la construction de ce type de modèle nécessite un très grand nombre de données d'apprentissage pour apprendre ces interactions.

Parallèlement, nous souhaitons améliorer le processus d'ajustement du modèle 3D. Pour contraindre celui-ci et limiter les risques d'approximation aberrante, une énergie de régularisation est utilisée. Celle-ci est basée sur une connaissance a priori de l'objet estimé. Malgré l'utilisation de cette contrainte, le risque d'approcher la forme du visage par une combinaison irréaliste de déformation existe. Pour limiter ce risque, l'espace des déformations et leurs combinaisons doit être contraint. Nous souhaitons en particulier restreindre ces déformations à une enveloppe convexe dont la structure est apprise sur une large collection de scans 3D de visages.

Bibliographie

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns : Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12) :2037–2041, 2006. (Cité pages viii et 23.)

[2] Alberto Albiol, David Monzo, Antoine Martin, Jorge Sastre, and Antonio Albiol. Face recognition using hog–ebgm. *Pattern Recognition Letters*, 29(10) :1537–1543, 2008. (Cité page 75.)

[3] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes : reconstruction and parameterization from range scans. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 587–594. ACM, 2003. (Cité page 43.)

[4] Brian Amberg. *Editing faces in videos*. PhD thesis, University of Basel, 2011. (Cité pages 43, 44 et 51.)

[5] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3d face recognition with a morphable model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008. (Cité page 48.)

[6] Akshay Asthana, Tim K Marks, Michael J Jones, Kinh H Tieu, and M Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 937–944. IEEE, 2011. (Cité pages viii, 26 et 32.)

[7] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan Goldman. Patchmatch : A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 28(3) :24, 2009. (Cité page 120.)

[8] Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. fisherfaces : Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7) :711–720, 1997. (Cité page 18.)

[9] Thomas Berg and Peter N Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *BMVC*, volume 2, page 7. Citeseer, 2012. (Cité pages viii, 30, 31 et 32.)

[10] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000. (Cité page 120.)

[11] A. Bertillon. *Ethnographie moderne : Les races sauvages. Les peuples de l'Afrique, les peuples de l'Amerique, les peuples de l'Océanie, quelques peuples de l'Asie ...* Bibl. de la nature. G. Masson, 1882. (Cité page 19.)

[12] David Beymer and Tomaso Poggio. Face recognition from one example view. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 500–507. IEEE, 1995. (Cité pages viii, 25 et 26.)

[13] David J Beymer. Face recognition under varying pose. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 756–761. IEEE, 1994. (Cité pages viii, 24 et 25.)

[14] Valentin Biaud, Vincent Despiegel, Catherine Herold, Olivier Beiler, and Stéphane Gentric. Semi-supervised evaluation of face recognition in videos. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, page 1. ACM, 2013. (Cité page 100.)

[15] Volker Blanz, Patrick Grother, P Jonathon Phillips, and Thomas Vetter. Face recognition based on frontal views generated from non-frontal images. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 454–461. IEEE, 2005. (Cité pages 27, 32, 60, 65 et 72.)

[16] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. (Cité pages 27, 48, 55 et 56.)

[17] Kevin W Bowyer, Kyong Chang, and Patrick Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 101(1) :1–15, 2006. (Cité page 7.)

[18] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1515–1522. IEEE, 2009. (Cité page 97.)

[19] Roberto Brunelli and Tomaso Poggio. Face recognition : Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10) :1042–1052, 1993. (Cité page 19.)

[20] CNIL. *Biometrie : des dispositifs sensibles soumis a autorisation de la CNIL*, 07 Avril 2011. (Cité page 5.)

[21] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *Computer VisionâECCVâ98*, pages 484–498. Springer, 1998. (Cité page 26.)

[22] Timothy F Cootes, Gavin V Wheeler, Kevin N Walker, and Christopher J Taylor. View-based active appearance models. *Image and vision computing*, 20(9) :657–664, 2002. (Cité pages 26 et 32.)

[23] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2296–2303. IEEE, 2011. (Cité page 36.)

[24] John Daugman. How iris recognition works. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1) :21–30, 2004. (Cité page 6.)

[25] John G Daugman. High confidence visual recognition of persons by a test of statistical independence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(11) :1148–1161, 1993. (Cité page 20.)

[26] Hassen Drira, Ben Amor Boulbaba, Daoudi Mohamed, Srivastava Anuj, et al. Pose and expression-invariant 3d face recognition using elastic radial curves. In *Proceeding of British machine vision conference*, pages 1–11, 2010. (Cité page 7.)

[27] Xiufeng Gao, Stan Z Li, Rong Liu, and Peiren Zhang. Standardization of face image sample quality. In *Advances in Biometrics*, pages 242–251. Springer, 2007. (Cité page 97.)

[28] Tobias Gass, Leonid Pishchulin, Philippe Dreuw, and Hermann Ney. Warp that smile on your face : optimal and smooth deformations for face recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 456–463. IEEE, 2011. (Cité pages 78 et 79.)

[29] Athinodoros S. Georghiades, Peter N. Belhumeur, and David Kriegman. From few to many : Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6) :643–660, 2001. (Cité page 25.)

[30] Valérie Gouaillier. La vidéosurveillance intelligente : promesses et défis. *Technopôle Défense et Sécurité, Québec*, 2009. (Cité page 8.)

[31] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5) :807–813, 2010. (Cité pages vii, 7, 32, 71, 73 et 81.)

[32] Patrick J Grother, George W Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709 :106, 2010. (Cité pages 6 et 27.)

[33] Di Huang, Caifeng Shan, Mohsen Ardebilian, and Liming Chen. Facial image analysis based on local binary patterns : A survey. *IEEE Trans. Sys., Man, and Cyber.,-Part C*, 41(6) :765–781, 2011. (Cité page 22.)

[34] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild : A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. (Cité page 100.)

[35] David H Hubel and Torsten N Wiesel. Ferrier lecture : Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, pages 1–59, 1977. (Cité page 20.)

[36] Information technology â Biometric data interchange formats â Part 5 : Face image data, 2005. (Cité page 6.)

[37] Andrew Edie Johnson and Martial Hebert. Surface registration by matching oriented points. In *3-D Digital Imaging and Modeling, 1997. Proceedings., International Conference on Recent Advances in*, pages 121–128. IEEE, 1997. (Cité pages viii et 38.)

[38] Fatih Kahraman, Binnur Kurt, and Muhittin Gokmen. Robust face alignment for illumination and pose invariant face recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007. (Cité pages 26 et 32.)

[39] Takeo Kanade. Picture processing system by computer complex and recognition of human faces. In *doctoral dissertation, Kyoto University*. November 1973. (Cité pages 6, 15 et 19.)

[40] Takeo Kanade and Akihiko Yamada. Multi-subregion based probabilistic approach toward pose-invariant face recognition. In *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*, volume 2, pages 954–959. IEEE, 2003. (Cité pages viii, 27 et 32.)

[41] Ingo Kennerknecht, Thomas Grueter, Brigitte Welling, Sebastian Wentzek, Juergen Horst, Steve Edwards, and Martina Grueter. First report of prevalence of non-syndromic hereditary prosopagnosia (hpa). *American Journal of Medical Genetics Part A*, 140(15) :1617–1622, 2006. (Cité page 3.)

[42] Martin Lades, Jan C Vorbruggen, Joachim Buhmann, Jörg Lange, Christoph von der Malsburg, Rolf P Wurtz, and Wolfgang Konen. Distortion invariant object recognition in the dynamic link architecture. *Computers, IEEE Transactions on*, 42(3) :300–311, 1993. (Cité page 20.)

[43] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110, 2004. (Cité page 74.)

[44] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2) :431–441, 1963. (Cité page 57.)

[45] Aleix M Martinez. The ar face database. *CVC Technical Report*, 24, 1998. (Cité pages 71 et 73.)

[46] Yang Meng, Zhang Lei, Yang Jian, and David Zhang. Regularized robust coding for face recognition. *IEEE Transactions on Image Processing*, 2013. (Cité pages 76 et 77.)

[47] K Messer, J Matas, J Kittler, J Luettin, and G Maitre. Xm2vtsdb : The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999. (Cité pages vii et 10.)

[48] Naoto Miura, Akio Nagasaka, and Takafumi Miyatake. Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification. *Machine Vision and Applications*, 15(4) :194–203, 2004. (Cité page 6.)

[49] Baback Moghaddam, Wasiuddin Wahid, and Alex Pentland. Beyond eigenfaces : Probabilistic matching for face recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 30–35. IEEE, 1998. (Cité pages vii et 17.)

[50] Kamal Nasrollahi and Thomas B Moeslund. Face quality assessment system in video sequences. In *Biometrics and Identity Management*, pages 10–18. Springer, 2008. (Cité page 97.)

[51] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59, 1996. (Cité page 22.)

[52] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7) :971–987, 2002. (Cité pages viii et 23.)

[53] Jörn Ostermann. Animation of synthetic faces in mpeg-4. In *Computer Animation 98. Proceedings*, pages 49–55. IEEE, 1998. (Cité page 40.)

[54] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 986–993. IEEE, 2005. (Cité page 56.)

[55] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008. (Cité page 36.)

[56] Linlin Shen and Li Bai. A review on gabor wavelets for face recognition. *Pattern analysis and applications*, 9(2-3) :273–292, 2006. (Cité pages viii, 21 et 22.)

[57] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans : Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11) :1948–1962, 2006. (Cité page 15.)

[58] André Sobiecki, Alexandru Telea, Gilson A Giraldi, Luiz A Pereira Neves, and Carlos Eduardo Thomaz. Low-cost automatic inpainting for artifact suppression in facial images. In *VISAPP (1)*, pages 41–50, 2013. (Cité pages x, 119 et 120.)

[59] Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Jun Liu. Face recognition under occlusions and variant expressions with partial similarity. *Information Forensics and Security, IEEE Transactions on*, 4(2) :217–230, 2009. (Cité pages 29, 32, 78 et 79.)

[60] Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. Recognizing partially occluded, expression variant faces from single training image per person with som and soft k-nn ensemble. *Neural Networks, IEEE Transactions on*, 16(4) :875–886, 2005. (Cité pages 78 et 79.)

[61] Anastasios Tefas, Constantine Kotropoulos, and Ioannis Pitas. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(7) :735–746, 2001. (Cité page 20.)

[62] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991. (Cité page 16.)

[63] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001. (Cité page 15.)

[64] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433. ACM, 2005. (Cité pages x et 114.)

[65] Danijela Vukadinovic and Maja Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1692–1698. IEEE, 2005. (Cité pages 56 et 83.)

[66] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010. (Cité pages 76 et 77.)

[67] Xingjie Wei and Chang-Tsun Li. Fixation and saccade based face recognition from single image per person with various occlusions and expressions. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 70–75. IEEE, 2013. (Cité pages 28, 32, 77, 78 et 79.)

[68] Xingjie Wei, Chang-Tsun Li, and Yongjian Hu. Face recognition with occlusion using dynamic image-to-class warping (dicw). In *FG13*. (Cité pages viii, 28, 29, 78 et 79.)

[69] Laurenz Wiskott, J-M Fellous, N Kuiger, and Christoph Von Der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7) :775–779, 1997. (Cité page 20.)

[70] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 74–81. IEEE, 2011. (Cité page 97.)

[71] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2) :210–227, 2009. (Cité pages viii, 29, 30, 76, 77, 78 et 79.)

[72] Tim Wu, Peter Hunter, and Kumar Mithraratne. Simulating and validating facial expressions using an anatomically accurate biomechanical model derived from mri data. In *Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications. SciTePress : Setubal, Portugal*, pages 267–272, 2013. (Cité pages viii et 35.)

[73] Meng Yang, D Zhang, and Jian Yang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625–632. IEEE, 2011. (Cité page 73.)

[74] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE, 2008. (Cité pages 36 et 37.)

[75] Liang Zhang. Estimation of the mouth features using deformable templates. In *Image Processing, 1997. Proceedings., International Conference on*, volume 3, pages 328–331. IEEE, 1997. (Cité page 108.)

[76] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition : A literature survey. *Acm Computing Surveys (CSUR)*, 35(4) :399–458, 2003. (Cité pages 7 et 21.)

Titre Neutralisation des expressions faciales pour améliorer la reconnaissance du visage

Résumé Les variations de pose et d'expression constituent des limitations importantes à la reconnaissance de visages en deux dimensions. Dans cette thèse, nous proposons d'augmenter la robustesse des algorithmes de reconnaissances faciales aux changements de pose et d'expression. Pour cela, nous proposons d'utiliser un modèle 3D déformable de visage permettant d'isoler les déformations d'identité de celles relatives à l'expression. Plus précisément, étant donné une image de probe avec expression, une nouvelle vue synthétique du visage est générée avec une pose frontale et une expression neutre. Nous présentons deux méthodes de correction de l'expression. La première est basée sur une connaissance a priori dans le but de changer l'expression de l'image vers une expression neutre. La seconde méthode, conçue pour les scénarios de vérification, est basée sur le transfert de l'expression de l'image de référence vers l'image de probe. De nombreuses expérimentations ont montré une amélioration significative des performances et ainsi valider l'apport de nos méthodes. Nous proposons ensuite une extension de ces méthodes pour traiter de la problématique émergente de reconnaissance de visage à partir d'un flux vidéo. Pour finir, nous présentons différents travaux permettant d'améliorer les performances obtenues dans des cas spécifiques et ainsi améliorer les performances générales obtenues grâce à notre méthode.

Mots-clés Reconnaissance de visage, Expression, Pose, Modèle déformable 3D, Neutralisation, video

Title Cancelling Facial Expressions for Reliable 2D Face Recognition

Abstract Expression and pose variations are major challenges for reliable face recognition (FR) in 2D. In this thesis, we aim to endow state of the art face recognition SDKs with robustness to simultaneous facial expression variations and pose changes by using an extended 3D Morphable Model (3DMM) which isolates identity variations from those due to facial expressions. Specifically, given a probe with expression, a novel view of the face is generated where the pose is rectified and the expression neutralized. We present two methods of expression neutralization. The first one uses prior knowledge to infer the neutral expression from an input image. The second method, specifically designed for verification, is based on the transfer of the gallery face expression to the probe. Experiments using rectified and neutralized view with a standard commercial FR SDK on two 2D face databases show significant performance improvement and demonstrates the effectiveness of the proposed approach. Then, we aim to endow the state of the art FR SDKs with the capabilities to recognize faces in videos. Finally, we present different methods for improving biometric performances for specific cases.

Keywords Face Recognition, Expression, Pose, 3D Morphable Model, Neutralization, video