[STA 10] STARK L., BOWYER K.W., SIENA S., « Human perceptual categorization of iris texture patterns », *Proceedings of IEEE BTAS*, 2010.

[VAQ 09] VAQUERO D., FERIS R., TRAN D., BROWN L., HAMPAPUR A., TURK M., « Attribute based people search in surveillance environments », *Proceedings of WACV*, 2009.

[WOL 06] WOLF F., POGGIO T., SINHA P., *Bag of words*, CiteSeer, 2006.

[ZEW 04] ZEWAIL R., ELSAFI A., SAEB M., HAMDY N., « Soft and hard biometrics fusion for improved identity verification », *Proceedings of MWSCAS*, vol. 1, p. I-225-8, 2004.

Chapitre 4

# Modélisation, reconstruction et suivi pour la reconnaissance des visages

Avec l'essor des techniques de reconnaissance biométrique, les systèmes de contrôle automatique ont investi de nombreux lieux et équipements ces dernières années (aéroports, locaux sécurisés, etc.). Afin de fluidifier le trafic au niveau de ces systèmes de reconnaissance, il est nécessaire de limiter au maximum les contraintes imposées à l'utilisateur. Pour remplir cet objectif, il est intéressant de procéder à des acquisitions à la volée, sans que l'utilisateur ait besoin de s'arrêter face au capteur.

Dans ce chapitre, nous nous concentrerons sur l'utilisation de la biométrie faciale, et plus spécifiquement sur les problématiques liées à l'acquisition de visages à la volée. Afin de permettre l'authentification au sein de tels systèmes, un certain nombre de problèmes liés à l'estimation de la forme 3D et de la texture des visages sont à résoudre.

Pour aborder les aspects théoriques liés à l'acquisition et à la reconstruction de visages, nous nous plaçons dans le contexte applicatif suivant : un système d'acquisition multi-vues est positionné à l'entrée d'une pièce, dans un couloir ou un parking par exemple, et l'objectif est d'identifier ou d'authentifier la personne observée par le biais de ce dispositif.

Chapitre rédigé par Catherine HEROLD, Vincent DESPIEGEL, Stéphane GENTRIC, Séverine DUBUISSON et Isabelle BLOCH.

## 4.1. Contexte

Les exigences imposées à un système biométrique sont diverses : sa facilité d'utilisation, sa rapidité d'exécution, sa non-intrusivité pour les utilisateurs, son coût et sa fiabilité.

Pour des systèmes destinés au grand public, où le nombre d'utilisateurs est important sans que ceux-ci y soient formés, les trois premiers points sont primordiaux.

Par exemple, dans le cas des passagers dans les aéroports, de nombreuses personnes sont amenées à n'utiliser un système biométrique donné qu'un nombre limité de fois au cours de leur vie. Un moyen efficace d'augmenter l'ergonomie et la fluidité de ce système biométrique est de réduire au maximum les contraintes sur le comportement de l'usager. Si aucune action particulière n'est imposée à l'utilisateur pendant le processus d'acquisition, la marge d'erreur comportementale est très réduite, diminuant ainsi le temps nécessaire.

Parmi les sources d'informations biométriques disponibles (empreintes digitales, iris, visages, veines), toutes ne vérifient pas les mêmes exigences. Les acquisitions classiques d'empreintes digitales ou d'iris imposent une position statique lors de l'acquisition ; elles sont par ailleurs moins bien acceptées par les utilisateurs que la biométrie faciale, qui est la plus naturelle pour l'être humain. Pour le visage, il est aisé d'imaginer un protocole sans contrainte de contact ou d'immobilisation, ce qui en fait une biométrie à la fois plus rapide et beaucoup mieux acceptée, l'utilisateur n'éprouvant pas le besoin de coopérer pendant l'acquisition.

### 4.1.1. *Applications de la reconnaissance faciale*

Ces dernières années, l'essor de la biométrie faciale a été particulièrement important avec des applications telles que :

- le contrôle d'entrée ou les accès sécurisés (identification par rapport à une base de personnes autorisées) ;
- le contrôle aux frontières (authentification avec le passeport) ;
- la délivrance de droits (carte d'électeur, permis de conduire, allocations, etc.) ;
- l'investigation policière.

Dans toutes ces applications, la biométrie faciale peut être utilisée seule ou en complément d'autres biométries.



### 4.1.2. *Authentification à la volée*

Beaucoup de systèmes d'acquisition faciale nécessitent un comportement spécifique de l'utilisateur, tel que l'immobilisation devant une ou plusieurs caméras. Cette contrainte d'arrêt ralentit considérablement l'étape de contrôle d'identité.

La raison principale de cette contrainte est que la majorité des systèmes de biométrie faciale s'appuie sur des comparaisons entre deux vues sous une même pose pour établir un score de correspondance. Les vues de référence enregistrées étant généralement frontales (photographie d'identité), l'objectif du système d'acquisition est de fournir cette vue frontale pour la comparaison.

Dans le cas d'un système où l'utilisateur doit s'arrêter au niveau du capteur, il est assez aisé de disposer directement de ce type de vue. En revanche, lorsque l'acquisition est non contrainte, le visage est perçu sous des poses variées ; la vue frontale doit alors être générée à partir des observations pour vérifier la correspondance : c'est l'étape dite de « frontalisation ».

D'autres méthodes de comparaison sont possibles, comme dans [VET 97] où l'auteur relâche la condition de similarité de pose généralement nécessaire dans les vues à comparer, par le biais de vues de synthèse sous de nouvelles poses. Deux vues peuvent également être comparées par le biais de paramètres de forme 3D et de texture qui sont estimés sur chacune d'entre elles [BLA 03b].

Enfin, il existe des méthodes s'appuyant sur des flux vidéo, qui analysent les dynamiques faciales pour identifier un individu, en complément de l'apparence faciale [MAT 09]. Néanmoins, dans la suite, nous nous limiterons à une comparaison entre deux vues frontales, ce qui correspond à la majorité des scénarios impliquant une photographie d'identité.

Pour obtenir la vue frontale du visage observé, l'idée naturelle est de passer par sa reconstruction tridimensionnelle (en forme et en texture) pour ensuite synthétiser une vue frontale de ce modèle. L'estimation de la pose, de la forme, de la texture et de l'illumination permettant de se ramener à une vue frontale constitue le cœur de ce chapitre.

Afin d'évaluer ces paramètres, de nombreux modules d'acquisition sont disponibles. Nous nous limiterons à des approches qui s'appuient uniquement sur les acquisitions vidéo faites par les caméras du système. D'autres méthodes existent, mais nécessitent un système plus complet (scanners tridimensionnels, capteurs de profondeur [ZOL 11]) ou plus intrusif (marqueurs sur le visage [HUA 11], projection de lumière structurée [ZHA 04], etc.), et ne seront donc pas abordées ici.

Même en ne disposant que d'un système multi-caméras, une grande variété d'informations est disponible pour restituer le visage en 3D : les données de calibration du système, des modèles 3D de visages, etc.

Nous passerons ces informations en revue dans la section 4.2 avant de détailler les approches s'appuyant sur une ou plusieurs vues acquises simultanément dans les sections 4.3 (approches géométriques), 4.4 (approches par modèles) et 4.5 (approches hybrides). Enfin, dans la section 4.6, nous détaillerons les approches intégrant l'aspect temporel avec l'utilisation explicite de la vidéo.

Afin de disposer d'une vue de synthèse du processus, voici un exemple de chaîne d'acquisition du visage « à la volée » (un système d'authentification est présenté dans la figure 4.1), sans interaction spécifique de l'utilisateur, afin d'accélérer l'ensemble du processus d'authentification (ou d'identification).

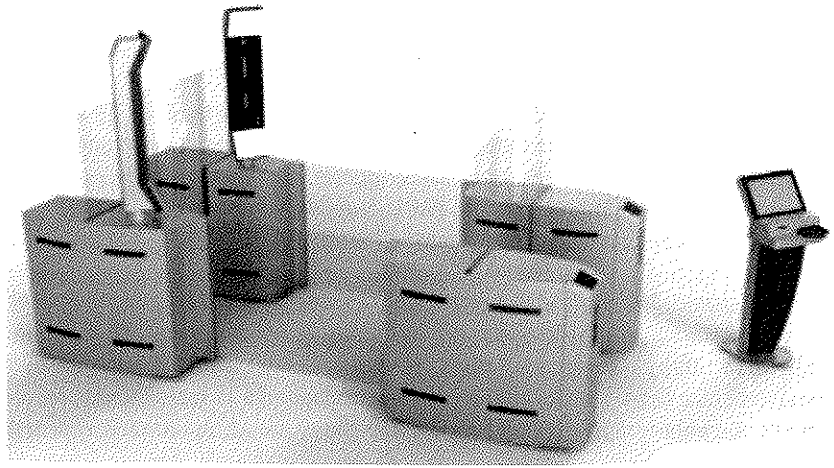


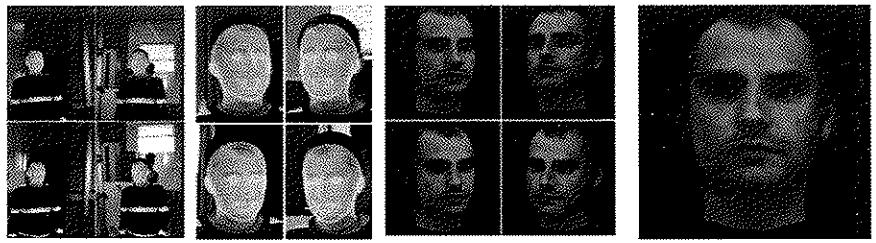
**Figure 4.1.** *Système d'authentification faciale à la volée*

Pendant son déplacement, la tête de l'utilisateur est suivie dans le repère 3D du système et les paramètres du modèle de la tête sont estimés à partir des différentes vues disponibles (figure 4.2a) afin de correspondre au mieux au visage de la personne suivie.

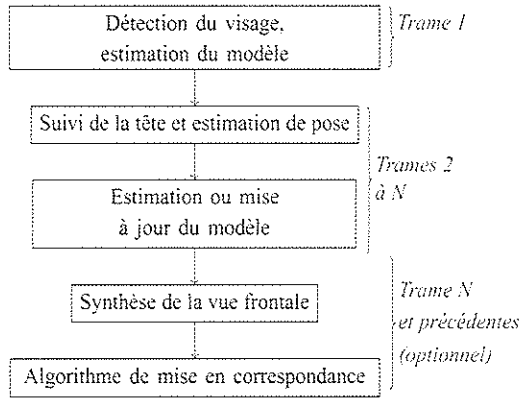
A chaque instant, de nouvelles observations sont disponibles et une modélisation du visage de l'individu est calculée ; de nouvelles vues peuvent alors être générées, en particulier la vue frontale (figure 4.2b), pour être comparée à une (authentification) ou plusieurs (identification) photo(s) d'identité.



L'ensemble du processus est résumé dans la figure 4.2c, et développé plus en détail dans [MOË 10].



*a)* *b)*




*c)*

**Figure 4.2.** *Processus global et suivi d'authentification. Source* [HER 11]
*a) Estimation des paramètres du modèle ; b) vue frontale ;*
*c) chaîne de traitement global*

## 4.2. Ensemble des informations disponibles

A partir d'un ensemble de vidéos synchronisées, beaucoup d'informations peuvent être exploitées pour reconstituer une vue frontale du visage observé.

Nous distinguons ici deux types de données :

– le premier est lié aux propriétés du système d'acquisition ;

– le deuxième à la nature de l'objet à reconstruire : le visage.

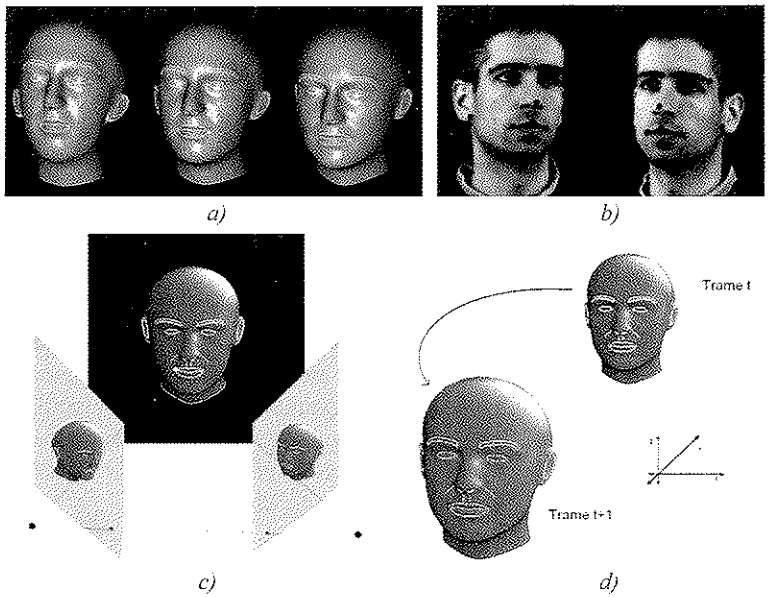



**Figure 4.3.** *Données disponibles pour estimer le visage à partir des acquisitions video : a) données sur le visage – forme ; b) données sur le visage – texture ; c) données du système – contraintes géométriques ; d) suivi – contraintes temporelles*

### 4.2.1. *Informations liées au système d'acquisition*

En bénéficiant d'un système multi-caméras, il est possible de s'appuyer sur un ensemble de vues synchronisées pour mettre en correspondance des points 2D et estimer le point 3D associé. En outre, si la calibration du système est connue, les contraintes épipolaires permettent d'améliorer le couplage des points inter-vues. Bon nombre de méthodes ont été développées pour estimer les paramètres de calibration d'une ou plusieurs caméras, avec ou sans mire [HAR 04, ZHA 00]. Les contraintes géométriques induites par la calibration permettent alors de reconstituer la forme d'un objet. Une autre solution, proposée par certains algorithmes, est d'estimer conjointement la calibration du système et la position des points mis en correspondance.

En outre, si le système est installé dans un environnement contrôlé (position et direction des lumières), l'ombrage de l'objet peut également être utilisé pour le reconstituer, par des techniques dites de *Shape from Shading* (forme à partir de l'ombrage) [ZHA 99]. Bien qu'il ne soit pas toujours possible de maîtriser l'environnement lumineux, ni de connaître les propriétés des sources de lumière, ces méthodes sont tout de même applicables avec l'ajout d'hypothèses sur certaines propriétés de forme de l'objet.



Enfin, l'aspect temporel peut également être exploité pour reconstruire le visage observé. Il est tout d'abord bénéfique d'exploiter la cohérence des positions et poses estimées entre les instants successifs dans un processus de suivi. Par ailleurs, la reconstruction peut être faite à partir de différentes vues d'un même flux vidéo, en optimisant conjointement la forme et la pose de l'objet, à partir d'une corrélation temporelle. Cette technique, parfois appelée *Structure from Motion*, a déjà été utilisée pour de nombreuses applications : reconstruction d'environnements urbains filmés à bord d'un véhicule, de bâtiments [POL 04], d'objets observés par une *webcam* mobile [NEW 10], etc. Nous verrons à la fin de ce chapitre comment exploiter le flux vidéo pour consolider la reconstruction du visage.

### 4.2.2. *Caractéristiques du visage*

Toutes les techniques précédemment citées s'appuient sur des données du système, et ne prennent en compte aucune information *a priori* sur les propriétés de l'objet à reconstruire. Or ici, nous nous intéressons spécifiquement à la reconstruction 3D de visages, information qui peut être intégrée dans le processus. C'est le cas des approches par modèles qui s'appuient sur les caractéristiques spécifiques des visages. Celles-ci peuvent être séparées en deux catégories : d'une part les informations de texture ou de couleur, et d'autre part les informations de forme.

Dans la première catégorie, on peut différencier les descripteurs globaux, qui spécifient les propriétés d'un visage dans son ensemble, et les descripteurs locaux, qui, à l'inverse, décrivent localement certaines sous-parties ou points caractéristiques du visage (comme ceux de la norme MPEG-4 FACE [PAN 03]). Les ondelettes de Haar ou les filtres de Gabor sont deux exemples de descripteurs couramment utilisés pour caractériser le visage ou ses parties. Ces descripteurs sont à la base des algorithmes de détection qui permettent de déterminer les positions des visages ou de leurs points d'intérêt dans une image. Ceux-ci sont généralement appris à partir d'une base d'apprentissage contenant les positions 2D d'intérêt [VIO 04]. Une autre information souvent utilisée pour caractériser le visage est sa couleur de peau. En effet, des modèles de couleur peuvent être appris pour décrire la teinte chair et utilisés comme détecteurs de visages [YAN 98]. Une revue globale des méthodes de détection de visages est donnée dans [ZHA 10]. Enfin, la réflectance, qui est une caractéristique plus physique du visage, peut également être rattachée à la couleur. Cette propriété explicite la lumière renvoyée par un point de la surface et peut être reliée à l'intensité perçue dans une image en ce point projeté.

La forme spécifique des visages peut également être caractérisée par le biais de distances types (entre les points caractéristiques par exemple), de répartition des points en 2D ou 3D, ou d'un maillage de la surface. Rattachées à cette information

de forme, les silhouettes issues de la projection de la forme dans le repère de l'image sont des informations riches pour restituer la forme 3D d'un visage.

Tant la forme que la texture des visages peuvent être apprises pour construire des modèles de cette classe. Cependant, malgré la généricité de la classe des visages en termes d'apparence et de forme, il faut noter la grande variabilité intra-classe des individus, qui nous permet de différencier un individu d'un autre. C'est cette différence qu'il s'agit d'exploiter lors d'algorithmes d'identification et d'authentification. Des modèles proposent à la fois la prise en compte des aspects génériques et individuels des visages. Ceux-ci sont obtenus par le biais d'un apprentissage dont sont extraits un modèle moyen (2D ou 3D) ainsi que des déformations, auxquelles est associée une probabilité de réalisation. Elles caractérisent soit la forme, soit la texture de la classe de visages, voire les deux de manière conjointe.

Une telle approche par modèle a plusieurs avantages. Tout d'abord, la prise en compte d'une information *a priori* sur la forme et/ou la texture contraint l'espace des solutions, et permet de régulariser la solution en cas de données bruitées. Par ailleurs, la connaissance d'un modèle de texture et de forme associé est riche en information pour l'estimation du visage. En effet, elle informe sur les zones d'intérêt (points caractéristiques, gradients particuliers, silhouette) et permet de calculer une similitude avec les observations en ces zones afin d'optimiser les paramètres.

## 4.3. Approches géométriques pour la reconstruction

De nombreux algorithmes ont été développés pour reconstruire un objet à partir d'un ensemble d'acquisitions. Dans le cas d'images de visages, ceux s'appuyant sur la stéréovision (de manière plus générale sur des acquisitions multi-vues) ou le *Shape from Shading* sont les plus utilisés.

### 4.3.1. *Stéréovision multi-vues*

Le premier type d'algorithmes s'appuie sur un ensemble de vues synchronisées de l'objet sous différents angles et prend en compte les contraintes de stéréovision issues de la calibration du système. Le principe de base est le suivant : à partir d'un ensemble de points d'intérêt détectés sur chaque vue, des appariements de points sont effectués (éventuellement contraints par les droites épipolaires issues des données de calibration). On en déduit ensuite les positions 3D associées pour remonter à l'information tridimensionnelle de l'objet. Les points non texturés sont ensuite reconstruits par interpolation à partir de l'ensemble épars calculé dans l'étape précédente, ou en utilisant à nouveau les contraintes épipolaires.

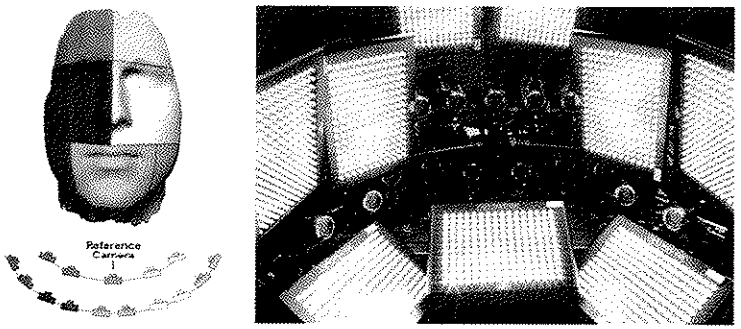


Une évaluation détaillée est donnée dans [SEI 06], où les auteurs catégorisent les différents algorithmes en fonction de leur initialisation, de leur méthodologie et des informations *a priori* utilisées.

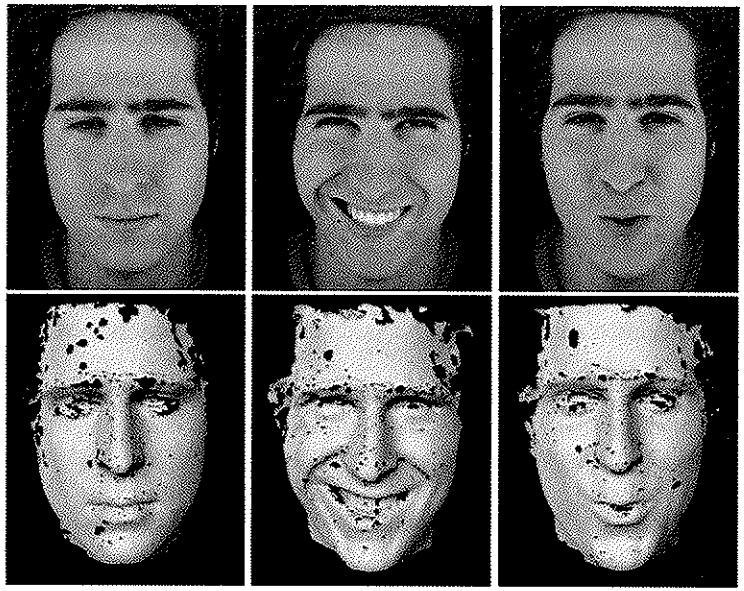
Ces méthodes imposent plusieurs contraintes. Tout d'abord, il est primordial de bénéficier d'un nombre important d'appariements, et ce sur toute la surface du visage, pour obtenir une reconstruction valide en tout point. Il est donc nécessaire d'avoir des vues prises sous des angles proches pour vérifier cette condition, sans quoi un point n'est pas forcément visible dans les différentes images. D'autre part, les paramètres de calibration doivent être connus avec précision pour effectuer une triangulation correcte des points mis en correspondance. Certaines méthodes proposent toutefois d'estimer la forme d'un objet en 3D à partir d'un ensemble de vues acquises quand les paramètres extrinsèques de calibration ne sont pas (ou seulement partiellement) connus [DAL 09, POL 04]. La procédure peut alors se rapprocher de problématiques d'estimation de la forme à partir du mouvement (*Structure From Motion*), que nous détaillerons dans la section 4.6.3.

L'utilisation d'acquisitions multi-vues pour la reconstruction de visages a été proposée à plusieurs reprises [BRA 10, LIN 10, BEE 10], avec différents nombres de vues et qualités de capteurs selon les systèmes. Avec l'essor des appareils photographiques à haute résolution, les reconstructions s'appuient de plus en plus sur des systèmes multi-vues de très bonne qualité, approchant la précision obtenue avec des méthodes d'acquisition active (scanner laser, lumière projetée). Bien que sans marqueurs, ces systèmes sont parfois contraignants s'ils impliquent plusieurs capteurs et un système d'éclairage particulier (la figure 4.4a illustre le système proposé dans [BRA 10]). Cependant, ils permettent une reconstruction très fidèle du visage observé (figure 4.4b). En effet, grâce à la résolution élevée des images, des appariements sont faits sur des détails mésoscopiques (pores de la peau, rides), ce qui permet d'obtenir un nuage de points dense (de l'ordre de 8 à 10 millions de points pour un visage) et un maillage final très précis.

D'autres méthodes de reconstruction ont également été proposées à partir d'un unique système binoculaire à haute résolution [BEE 10], en exploitant les détails très fins du visage comme précédemment. Pour limiter le coût du système et le temps d'exécution, certaines méthodes s'appuient sur des images moins résolues, au prix d'une qualité de reconstruction moins précise. Lin *et al.* [LIN 10] utilisent cinq vues du visage à pose fortement variable pour reconstruire le visage, en faisant appel à un algorithme d'ajustement de faisceaux et à la programmation dynamique. L'utilisation de l'information de silhouette et de vues de profil permet d'améliorer la reconstruction, en particulier au niveau du nez, mais reste en-deçà des méthodes précédentes.




*a)*



*b)*

**Figure 4.4.** *Reconstruction multi-vues proposée dans* [BRA 10] :
*a) couverture des caméras et système d'acquisition ; b) trois exemples de reconstructions (indépendantes) pendant une séquence*

### 4.3.2. *Forme à partir de l'ombrage*

Les techniques d'estimation de forme à partir de l'ombrage (*Shape from Shading*) [ZHA 99] consistent à estimer la géométrie d'un objet à partir d'une ou plusieurs vues de celui-ci, grâce à l'information d'ombrage. Cette donnée caractérise la variation d'intensité observée dans une image, entre deux points d'une surface



aux propriétés identiques, ou alors d'un même point observé avec deux illuminations différentes dans deux vues. Comme l'intensité observée dépend de l'orientation de la surface associée, des informations de forme de l'objet peuvent être déduites de l'ombrage. Il est nécessaire pour cela de bénéficier d'une modélisation du système optique, mais aussi de l'illumination de la scène et des propriétés de réflectance de l'objet à reconstruire. Une hypothèse classiquement faite pour les méthodes de *Shape from Shading* est de considérer l'objet à reconstruire comme lambertien, c'est-à-dire que la lumière reflétée par un point de sa surface est la même dans toutes les directions. D'autres modélisations plus réalistes existent, comme le modèle d'illumination de Phong. Celui-ci prend en compte non seulement la composante ambiante, la réflexion diffuse (modèle lambertien), mais aussi la réflexion spéculaire, qui caractérise une direction de réflexion privilégiée. L'intensité $I$ d'un point est alors donnée par la somme de ces trois termes :

$$I = \underbrace{k_a I_a}_{\text{composante ambiante}} + \underbrace{k_d I_d \cos\theta}_{\text{composante diffuse}} + \underbrace{k_s I_d \left(\cos\alpha\right)^{\upsilon}}_{\text{composante spéculaire}} \qquad [4.1]$$

– $I_a$, $I_d$ sont respectivement les intensités des lumières ambiante et directionnelle ;

– $k_a$, $k_d$, $k_s$ sont les coefficients de réflexion ambiante, diffuse et spéculaire ;

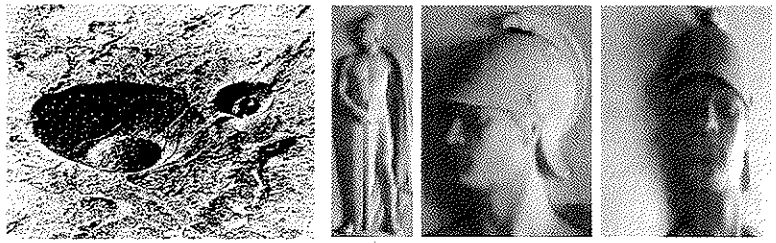
– $\theta$ est l'angle entre la normale au point considéré et la direction de la lumière directionnelle ;

– $\alpha$ est l'angle entre les directions de réflexion et de vue, et $\upsilon$ le coefficient de brillance du point considéré.

Ce modèle permet d'affiner l'estimation de la forme du fait d'une modélisation plus réaliste. Certains auteurs ont proposé de mesurer spécifiquement la réflectance du visage [MAR 99] par l'apprentissage de la fonction de réflectance bidirectionnelle BRDF (*Bidirectional Reflectance Distribution Function*), qui modélise la réflexion de la lumière en un point d'une surface.

Une hypothèse nécessaire pour utiliser ces techniques avec une seule image est de bénéficier d'une information *a priori*, la position de la source lumineuse par exemple. En effet, sans cela, le problème du *Shape from Shading* est mal posé et il n'est pas possible d'inférer directement une surface de manière unique à partir d'une image. Différentes ambiguïtés ont été relevées dans la littérature, telles que celle du *cratère* [PEN 89] ou du *bas-relief* [BEL 97]. La première est illustrée par la figure 4.5a et montre l'ambiguïté existant si l'éclairage doit être conjointement estimé avec la surface. Ici, l'éclairage peut être perçu comme venant de haut (vue d'un cratère) ou de bas (vue d'un volcan à l'envers) : la surface et l'éclairage ne peuvent donc pas être déterminés de manière unique.

La figure 4.5b montre un exemple d'ambiguïté dite du bas-relief, où l'estimation du relief du visage estimé en regardant l'image centrale est erronée. En effet, celui-ci est, en réalité, beaucoup plus écrasé (photo de droite). Différentes surfaces tridimensionnelles, associées à des sources de lumière adaptées, peuvent donc donner lieu à une même image après projection.



*a)* *b)*

**Figure 4.5.** *Exemples d'ambiguïtés en* Shape from Shading *[BRA 10] : a) cratère. Source [PEN 89] ; b) bas-relief. Source [BEL 97]*

Initialement, la méthode de *Shape from Shading* a été développée pour estimer la forme d'un objet à partir de plusieurs vues sous une pose fixe et différentes illuminations connues [WOO 89]. Des évolutions ont été proposées pour relâcher ces conditions, et ne nécessitent plus de connaître les paramètres d'illumination [BAS 07, WU 11].

Récemment, une méthode de *Shape from Shading* sans contrainte de pose ni de source lumineuse a été proposée, permettant ainsi d'utiliser un grand nombre d'acquisitions pour reconstruire le visage [KEM 11b]. Pour s'affranchir des changements de forme dans l'ensemble des images utilisées (typiquement, les expressions utilisées), la notion de forme canonique, définie comme la forme *localement similaire* au plus grand nombre de photographies possibles, est introduite.

A la différence des techniques de stéréovision qui reconstruisent un objet par interpolation à partir d'un ensemble épars de points 3D, les méthodes de *Shape from Shading* estiment la normale en chaque pixel de l'image, et apportent donc plus de précision à la reconstruction. Cela est particulièrement le cas pour des surfaces peu texturées, comme les joues du visage par exemple où très peu de points d'intérêt sont détectés. Néanmoins, ces deux méthodes peuvent s'avérer complémentaires, en initialisant un maillage par stéréovision multi-vues avant de le raffiner par des techniques de *Shape from Shading* [WU 11].



Dans les techniques précédemment citées, aucune hypothèse sur l'objet à reconstruire n'est utilisée, l'avantage est donc de pouvoir reconstituer un objet quelconque. Or les visages présentent des zones très peu texturées, en particulier sur les joues ou le front. Pour certaines de ces techniques, il est donc difficile d'en déduire la forme 3D. De ce fait, comme la forme et la texture des visages peuvent être modélisées, il est intéressant d'utiliser ces informations *a priori* pour déterminer la solution. La prise en compte d'une modélisation de la classe des visages permet d'une part de réduire l'espace de recherche à un sous-espace adapté, et d'autre part de régulariser la solution.

## 4.4. Approches par modèles pour la reconstruction

### 4.4.1. *Modélisation du visage*

Beaucoup de travaux ont été consacrés à la modélisation des visages en deux et trois dimensions. Nous faisons ici un rapide tour d'horizon des modèles les plus courants, pour finir par une description plus détaillée du modèle 3D le plus usité, à savoir le *3D Morphable Model* (*3DMM*) [BLA 99].

Le choix de la modélisation de la classe des visages est contraint par le type d'informations à traiter (images, capteur de profondeur, centrale inertielle, etc.) et par l'application pour laquelle le modèle est utilisé. En effet, pour des applications d'interaction homme-machine ou de vidéo-conférence, l'information importante sera contenue dans les mouvements du visage (expressions, paroles). Un modèle générique commun à tous les individus est donc suffisant. Il est cependant nécessaire de lui associer un modèle de déformation, attaché à des expressions par exemple, comme c'est le cas de *GRETA* [PAS 01].

Au contraire, dans le cas où la reconstruction du visage s'inscrit dans le cadre d'une application de reconnaissance faciale, il est nécessaire de bénéficier d'un modèle déformable – la déformation s'entend ici au sens des spécificités de chaque individu. Notons que certains modèles combinent un modèle d'identité et d'expressions [BLA 03a], offrant ainsi une grande souplesse d'utilisation, mais nécessitant aussi des algorithmes plus performants pour estimer l'ensemble des paramètres d'identité et d'expressions.

#### 4.4.1.1. *Modélisation 2D du visage*

Les premiers modèles de visage apparus dans les années 1990 étaient des représentations bidimensionnelles. On peut citer, entre autres :

– les *eigenfaces* [TUR 91], qui sont les vecteurs principaux issus d'une analyse en composantes principales (ACP) sur une base de visages en vue frontale. L'ACP a pour but de capturer la variabilité de l'ensemble d'apprentissage et de la retranscrire

dans la base par ordre d'importance. Cet ensemble de vecteurs, dits *eigenfaces*, définit une base dans laquelle on peut exprimer un visage comme leur combinaison linéaire ;

– les *Labelled Graphs* [WIS 97], qui définissent le visage comme un graphe étiqueté. A chaque nœud du graphe est associé un vecteur concaténant des réponses à des ondelettes de Gabor autour du point du visage correspondant. Chaque arête est quant à elle associée à une distance caractérisant l'éloignement de deux points ;

– les modèles actifs de forme, ou *Active Shape Models* [COO 95], qui caractérisent de manière statistique la distribution de formes de visages (en 2D). L'ajustement du modèle (en termes de pose et de déformations) avec une image d'entrée se fait de manière récursive, par mise en correspondance du modèle avec les contours ou les points d'intérêt observés, puis par une mise à jour de la pose et des paramètres de forme ;

– les modèles actifs d'apparence, ou *Active Appearance Models* (AAM) [EDW 98], qui prennent en compte la texture en plus du modèle statistique de forme. L'estimation du modèle se fait en minimisant la différence entre la texture observée dans l'image d'entrée et celle synthétisée à partir des paramètres de forme et de texture estimés.

La majorité des algorithmes qui estiment les paramètres d'un de ces modèles étant donnée une image nécessite d'avoir une vue frontale ou quasi-frontale du visage. Sans cette hypothèse, ces modèles ne permettent pas d'estimer les paramètres du visage observé.

#### 4.4.1.2. *Modélisation 3D du visage*

Etant données les caractéristiques du système d'acquisition présenté dans la section 4.1.2, il est nécessaire de gérer des images de visages sous des poses non-frontales. En effet, du fait du positionnement des caméras (par exemple, sur des montants de portes, ou dans un coin de pièce), la pose sous laquelle le visage est perçu peut être très variée. Pour traiter cette problématique, il est naturel de travailler avec un modèle de visage en 3D. Ainsi, l'estimation conjointe de la pose et des paramètres du modèle permet ensuite de procéder à la frontalisation. En outre, l'utilisation d'un modèle 3D permet de traiter les problèmes d'auto-occultations et d'ombres si l'on intègre les sources lumineuses dans les paramètres à estimer.

Un modèle simple de visage 3D appelé *Candide* a été proposé en 1987 et se présente sous la forme d'un maillage caractérisant la partie frontale du visage [RYD 87]. Ce maillage a été modifié pour s'accorder à la norme MPEG4 et des unités d'actions y ont ensuite été ajoutées pour caractériser des expressions (modèle *Candide-3* [AHL 01]). Cependant, ce modèle ne permet pas de caractériser la variabilité inter-individus de la classe des visages, ce qui a entraîné la construction d'autres modèles, tels que le *3D Morphable Model* (*3DMM*). L'article fondateur du



*3DMM* [BLA 99] proposé par Blanz *et al.* est à l'origine de nombreux travaux sur la modélisation tridimensionnelle du visage. L'apport principal de cet article est l'introduction d'un modèle statistique de visage en termes de forme et de texture, à partir d'un ensemble de $M$ acquisitions 3D de visages, recalées de manière dense.

Chaque visage y est décrit par sa forme : $S = \{(X_1, Y_1, Z_1),..., (X_N, Y_N, Z_N)\}$ composée de $N$ points 3D, et par sa texture $T = \{(R_1,G_1,B_1),...,(R_N, G_N, B_N)\}$.

A partir des $M$ visages $\{(S_i,T_i),\ i \in \{1,...,M\}\}$ auxquels on retire la moyenne $(\overline{S},\overline{T})$, une ACP est effectuée de manière indépendante sur la forme et sur la texture en utilisant les matrices de covariance $C_S$ et $C_T$. Les principaux axes de déformation de forme et de texture sont respectivement caractérisés par les vecteurs propres $s_i$ et $t_i$.

Un visage $(S,T)$ issu de cette modélisation est décrit par :

$$S = \overline{S} + \sum_{i=1}^{M-1} \alpha_i s_i, \quad T = \overline{T} + \sum_{i=1}^{M-1} \beta_i t_i \qquad [4.2]$$

où $\alpha = (\alpha_1,..., \alpha_{M-1})$ est un vecteur à valeurs réelles distribué avec une probabilité :

$$p(\alpha) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{M-1}\left(\frac{\alpha_i}{\sigma_{S,i}}\right)^2\right\} \qquad [4.3]$$

où les $\sigma_{S,i}$ sont les valeurs propres de la matrice de covariance de forme $C_S$. La probabilité du vecteur des coefficients de texture $\beta = (\beta_1,..., \beta_{M-1})$ s'exprime de manière similaire. La figure 4.6 illustre l'influence de la variation des paramètres de forme $\alpha$ sur la forme globale du visage pour une texture donnée. Chaque visage est généré avec la même texture, et la projection est appliquée avec des paramètres de calibration identique. Il y a deux principaux avantages à définir le visage par un modèle de type *3DMM* :

– le nombre d'inconnues à définir pour caractériser la forme et la texture est très largement réduit. En effet, au lieu de définir indépendamment des milliers de points 3D et la couleur qui leur est associée, l'ACP limite la définition de texture et de forme à un ensemble restreint de paramètres qui pondèrent les vecteurs propres ;

– la définition d'un nouveau visage comme combinaison des vecteurs propres retenus à la suite de l'ACP utilise une connaissance *a priori* forte déduite de la base d'apprentissage de visages. Ainsi, cette connaissance permet de créer des visages cohérents du fait de l'attache au modèle.

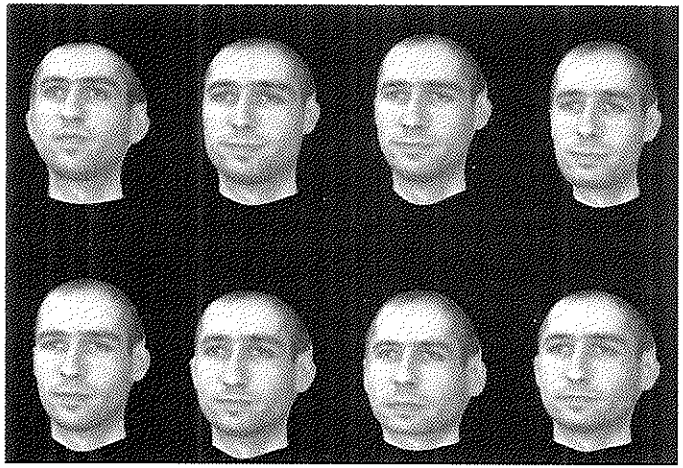


**Figure 4.6.** *Variation de la projection d'un visage pour différents paramètres de forme, à pose et texture données*

Un point à évaluer avec un modèle construit à partir d'une base d'apprentissage est sa capacité à caractériser le visage d'un individu quelconque. Celui-ci ne pourra en effet pas être parfaitement reconstruit par les vecteurs propres issus de l'ACP. Il s'agit donc dans ce cas de trouver les paramètres $\{(\alpha_i, \beta_i), i = 1,..., M-1\}$ tels que la distance du visage considéré à l'espace $V$ des visages défini par le *3DMM* soit minimale (selon une distance à définir). La solution est donc la projection du visage réel dans $V$.

Un intermédiaire entre le modèle actif d'apparence et le *3DMM* a été proposé par Xiao *et al.* [XIA 04] pour caractériser les visages. Ce modèle, qui permet de caractériser autant de formes que le *3DMM*, ne gère cependant pas les problèmes d'occultations (l'information 3D n'étant pas explicite). Par ailleurs, il est moins densément défini que le *3DMM*, et peut donc être limitatif pour des applications de reconnaissance faciale. En revanche, l'avantage de ce modèle est sa vitesse d'ajustement de pose et de déformations étant donnée une image, qui est similaire à celle d'un *AAM* classique, et bien plus grande que celle des méthodes d'estimation du *3DMM* que nous passons en revue à présent.

### 4.4.2. *Estimation des paramètres du modèle*

Dans cette partie, nous verrons différentes méthodes proposées pour estimer les paramètres $\{(\alpha_i, \beta_i), i = 1,..., M-1\}$ (équation 4.2) d'un visage observé à partir d'une image.



#### 4.4.2.1. *Estimation conjointe de forme et de texture*

Différents critères peuvent être utilisés pour estimer la forme tridimensionnelle du visage paramétrée par les coefficients $\alpha_i$ (équation 4.2), ainsi que la texture qui lui est associée. L'article [BLA 99] propose une méthode d'estimation des paramètres de visage $(\alpha_i, \beta)$ conjointement à des paramètres d'illumination de la scène ainsi que de calibration (concaténés dans le vecteur $p$ pour des commodités de lecture). Cette procédure est effectuée en minimisant l'énergie globale $E$ composée d'un terme d'attache aux données $E_I$ et d'un terme de régularisation $E_M$. Le premier s'exprime par :

$$E_I = \sum_{x,y} \left\| I_{obs}(x,y) - I_{render}(x,y,\alpha,\beta,p) \right\|, \qquad [4.4]$$

où $(x, y)$ caractérise la position d'un pixel, $I_{obs}(x, y)$ sa valeur dans l'image d'entrée, et $I_{render}(x, y, \alpha, \beta, p)$ celle dans l'image synthétisée étant données les valeurs courantes des paramètres. Le terme de régularisation $E_M$ intègre l'hypothèse de distribution normale des paramètres de forme et de texture :

$$E_M = \sum_{i=1}^{M-1} \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_{i=1}^{M-1} \frac{\beta_i^2}{\sigma_{T,i}^2} \qquad [4.5.]$$

Une minimisation de l'énergie $E = E_I + E_M$ par descente de gradient stochastique est proposée dans [BLA 99], afin d'être robuste à des minima locaux et d'augmenter la vitesse d'exécution de l'algorithme.

Romdhani *et al.* [ROM 02] proposent une méthode itérative d'estimation des paramètres, en exploitant la linéarité des équations quand les grandeurs non estimées sont fixées. La méthode s'appuie sur le calcul du flot optique entre l'image synthétisée avec les paramètres courants et l'image d'entrée. Cet algorithme nécessite de connaître la direction de la lumière ainsi qu'une pose approximative pour initialiser l'algorithme d'optimisation. Il aboutit à des résultats similaires à ceux de [BLA 99], mais avec un temps d'exécution divisé par cinq. Cependant, il ne prend pas en compte l'ombrage pour estimer la forme, contrairement au gradient stochastique. La méthode *Inverse Compositional Image Alignement* proposée dans [ROM 03] s'inspire d'une méthode d'ajustement du visage initialement proposée en 2D pour l'*AAM* dans [BAK 01] et s'appuie sur la projection inverse du modèle de forme. Pour augmenter le rayon de convergence des méthodes précédentes, Romdhani *et al.* [ROM 05] proposent de multiplier les critères de vraisemblance à prendre en compte pour aligner un modèle sur l'image considérée. Ainsi, le risque de glisser vers des minima locaux durant la procédure d'optimisation est réduit, tout en augmentant la qualité du modèle estimé.

En plus du terme d'attache aux données pixelliques et des informations *a priori* de forme et de texture, les auteurs prennent en compte la position des contours et des reflets spéculaires dans l'image. Comme précédemment, cette procédure propose un compromis entre la fidélité avec les observations (qui peuvent présenter du bruit ou contenir des mauvaises données) et le modèle *a priori*. En outre, la direction de la lumière n'est plus requise en entrée mais elle est également estimée par l'algorithme. Les critères intégrés dans ce dernier algorithme imposent des prétraitements (extraction des contours, génération de cartes de distances) sur l'image parfois bruitée, et impliquent de paramétrer justement la pondération entre les critères.

4.4.2.2. *Estimation des paramètres de forme et extraction de texture*

Il est possible de n'utiliser que l'aspect géométrique du *3DMM* pour reconstruire les visages en 3D. En effet, l'objectif étant *in fine* de valider l'identité d'une personne, il est important de bénéficier de la texture la plus fidèle possible. Plutôt que d'exprimer celle-ci en fonction d'une base d'apprentissage, elle peut donc directement être extraite des observations une fois l'estimation de pose et de forme effectuée.

Etant donnée la variabilité des textures des individus (couleur de peau, présence de cicatrices, etc.), il faudrait un très grand nombre d'individus dans la base pour assurer que tout individu considéré dans la population soit assez proche d'au moins une solution dans l'espace défini par le *3DMM*. Par ailleurs, comme moins de paramètres sont à déterminer, les temps de calcul associés à leur estimation sont réduits. Dans cette partie, nous nous concentrons donc sur des algorithmes dans lesquels la forme seule du modèle est estimée, avec la possibilité d'extraire la texture des observations dans un deuxième temps.

En s'affranchissant de la texture, il est possible d'employer des critères moins complexes que ceux vus dans la section 4.4.2.1.

Par exemple, la restitution des paramètres d'un modèle 3D de type *3DMM* à partir d'un ensemble de points 2D détectés sur les images est envisageable. Il s'agit alors de résoudre le problème inverse de détermination de la pose et des paramètres $\alpha_i$, $i = 1,\ldots, M-1$, tels que les points caractéristiques du modèle ainsi estimé $S = \bar{S} + \sum_{i=1}^{M-1} \alpha_i s_i$ se projettent le plus près possible des positions détectées.

Du fait d'une connaissance statistique du modèle, il est possible d'établir une fonction de coût construite à partir de deux termes [BLA 04, FAG 08] :

– un terme d'attache aux données, qui est la distance entre les reprojections des points 3D du modèle estimé et les points détectés ;



– un terme de régularisation, issu de la construction du *3DMM*, à savoir la norme de Mahalanobis du vecteur des coefficients de déformation (à rapprocher de l'équation 4.5). Celui-ci est pondéré par un facteur $\eta$, qui règle l'impact de l'information *a priori* vis-à-vis de l'attache aux données, ainsi que l'illustre la figure 4.7.

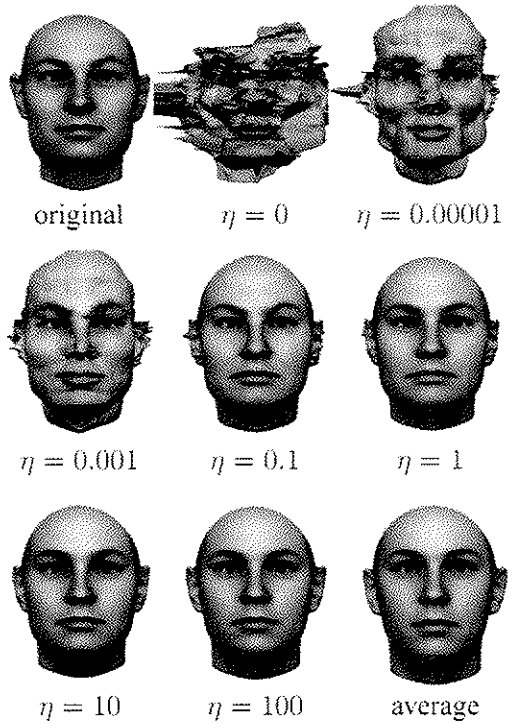



**Figure 4.7.** *Variation de la reconstruction en fonction du coefficient de régularisation $\eta$. Source* [BLA 04]

Ainsi, les déformations induites par des détections imprécises sont régularisées par le modèle appris. Hormis les points détectés, il est aussi possible d'ajouter des informations complémentaires, telles que des directions tangentes aux contours du visage (figure 4.8).

Cependant, l'énergie proposée n'est pas robuste aux détections aberrantes, qui sont prises en compte dans l'erreur de projection avec une norme euclidienne. Des solutions ont été développées pour mieux gérer les erreurs de détection des points caractéristiques, en relâchant cette contrainte d'attache aux données.

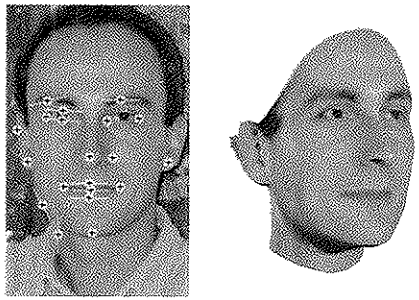


**Figure 4.8.** *Annotation de l'image d'entrée et estimation du modèle associé Source* [BLA 04]

Dans [BRE 10] par exemple, après quelques itérations, l'énergie à minimiser ne prend plus en compte la distance 2D entre les points projetés du modèle et les détections, mais s'appuie sur un score de correspondance (par ZNCC, *Zero-Normalized Cross Correlation*) d'imagettes autour de ces points. Plus l'image de synthèse correspond à l'image d'entrée, meilleurs sont les scores de correspondance et donc l'estimation des paramètres. Ainsi, le modèle ne s'adapte pas directement aux points détectés mais cherche un compromis entre la configuration globale des points d'intérêt et l'erreur de reconstruction.

Il faut noter que même en cas de points correctement détectés, la position des points caractéristiques varie peu pour des visages échantillonnés selon le *3DMM*. En effet, les zones du visage à forte variance de forme ne sont pas toutes situées à proximité des points saillants du visage, et il est donc difficile de capter les déformations par l'unique information de ces points caractéristiques. L'ajout d'autres critères, tels que la proximité de certaines arêtes du modèle (lèvres, yeux) ou des silhouettes projetées avec les gradients détectés dans l'image permet d'améliorer l'estimation des paramètres, mais augmente simultanément la complexité et le temps d'exécution de l'algorithme.

Une méthode totalement automatique pour générer une vue frontale à partir de n'importe quelle image d'entrée est proposée dans [AST 11]. Au lieu d'utiliser le modèle de forme du *3DMM*, les auteurs apprennent plusieurs modèles d'apparence, appelés *View Active Appearance Models*, pour différents intervalles de pose. Les modèles actifs les plus appropriés sont ajustés aux observations, et on ne garde que celui minimisant l'erreur résiduelle de forme et de texture. Une estimation de pose précise est ensuite évaluée pour ce modèle avec les paramètres estimés, par régression avec des machines à vecteurs de support. Enfin, la texture de l'image d'entrée est extraite avec un modèle moyen de forme 3D avant de générer la vue frontale correspondante.



La forme spécifique de chaque visage n'est donc pas utilisée pour la frontalisation, car on se rapporte au modèle moyen pour faire cette étape. Des erreurs peuvent donc exister sur l'extraction de texture, lorsque la forme du visage observé diffère trop de celle du modèle moyen. En revanche, le gain de temps par rapport à une méthode d'ajustement d'un modèle 3D complet est notable. Les résultats de quelques vues frontales ainsi générées sont illustrés à la figure 4.9.

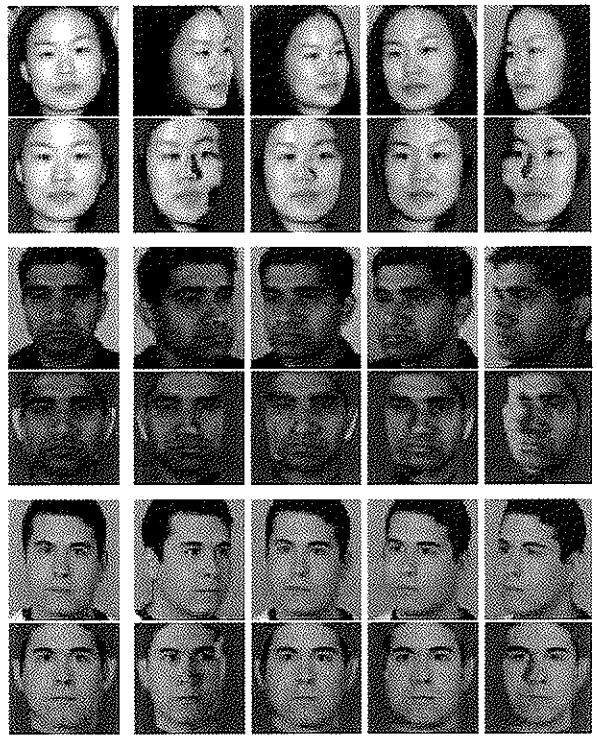


**Figure 4.9.** *Quelques exemples de visages en vue frontale générés à partir de* View Active Appearance Models [AST 11]

Une fois les paramètres de forme et de pose estimés avec une des méthodes décrites ici, la texture est extraite étant donnés la forme du modèle 3D et un modèle de projection (orthographique, ou en perspective). Le modèle complet est ensuite utilisé pour générer la vue frontale. Avec ces méthodes d'ajustement géométrique pur, l'illumination n'est pas prise en compte, et la texture n'est donc pas corrigée en cas d'ombres ou de reflets spéculaires. Il faut donc veiller à contrôler l'environnement lumineux dans l'espace d'acquisition pour limiter ces effets. Par ailleurs, un problème de validité de la texture extraite peut être posé par les

accessoires en relief tels que les lunettes. En effet, celles-ci sont considérées comme étant directement posées sur le visage lors de l'extraction, alors qu'il faudrait les modéliser spatialement pour extraire séparément la texture des lunettes d'une part, et celle du visage d'autre part. Sans cela, en changeant la pose pour synthétiser la vue frontale, la texture des lunettes peut être reprojetée dans des zones incorrectes. Une solution possible est de détecter la présence de tels objets et de les supprimer dans les images d'entrée (par des algorithmes d'*inpainting* par exemple), pour n'en extraire ensuite que la texture du visage.

Un des inconvénients des approches par modèles est, par construction, leur dépendance à l'ensemble d'apprentissage. Il faut veiller à varier les représentants de la classe lors de l'apprentissage pour couvrir au mieux l'ensemble des visages existants (barbe, lunettes, etc.). Il est par exemple difficile de recréer un visage avec des cicatrices particulières en l'absence de marques similaires dans la base d'apprentissage. L'avantage de méthodes qui ne s'appuient pas sur un modèle (section 4.3) est de pouvoir reconstituer des formes particulières de visage ou des accessoires portés par l'individu, tels que des lunettes, un chapeau, ou un foulard, sans être contraint par une information *a priori*. Cela évite le problème de plaquage de texture aberrant sur le visage qui a été explicité ci-dessus.

## 4.5. Approches hybrides

Les méthodes précédemment exposées (géométriques d'une part, à partir de modèles d'autre part), peuvent être employées simultanément. Le problème à résoudre possède ainsi plus de contraintes et d'informations en entrée, ce qui permet de lever certaines incertitudes. Cependant, étant donnée la quantité d'informations à prendre en compte (ou à estimer), les fonctions associées sont plus complexes, induisant généralement un temps de résolution plus important.

Une première fusion de méthodes possibles est l'utilisation conjointe de la stéréovision et d'un modèle *a priori* de visage. A partir d'un système calibré, l'estimation de forme peut ainsi être ainsi effectuée grâce à l'information de silhouette extraite dans chaque vue [JIN 03, LEE 03], consolidée en 3D.

L'avantage de cette caractéristique est sa facilité d'extraction sous des poses variées, sans compter qu'il s'agit d'une information très importante pour estimer les paramètres de forme d'un visage. La calibration permet également de calculer les positions 3D de points caractéristiques tels que les yeux et la bouche en multi-vues et d'en déduire la pose 3D ainsi que l'échelle du visage. Le modèle de forme peut ensuite être déformé à partir de points mis en correspondance entre les différentes vues [IVA 07]. Enfin, la méthode d'estimation de forme et de texture mono-vue présentée dans [ROM 05] a également été étendue au cas multi-vues [AMB 07].



Un critère de cohérence stéréoscopique est alors intégré dans la fonction de coût et l'optimisation permet d'estimer non seulement les paramètres du modèle, mais aussi les paramètres de calibration des caméras (qui ne nécessitent donc pas d'être connus contrairement aux méthodes précédentes).

A partir de plusieurs caméras, l'information de profondeur du visage est conservée, contrairement au cas de l'optimisation sur une image avec projection orthographique. L'utilisation de plusieurs vues permet en outre de résoudre des problèmes d'occultation existant avec une seule image et aboutit à des reconstructions plus précises et complètes (figure 4.10).

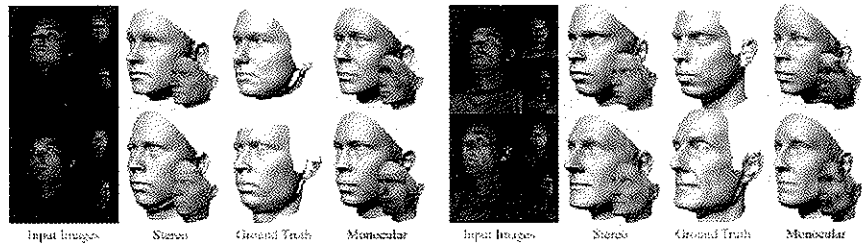



**Figure 4.10.** *Images d'entrée, reconstruction associée avec une méthode multi-vues, forme réelle et reconstruction avec la vue frontale uniquement. Les visages au bas des reconstructions indiquent l'erreur associée : plus la zone est sombre, plus l'erreur est importante. Source* [AMB 07].

Des méthodes associant l'approche par *Shape from Shading* et l'utilisation de modèles ont également été proposées récemment. Comme dans [ROM 05], une méthode d'optimisation de type Levenberg-Macquardt est utilisée, mais l'énergie à optimiser intègre alors les contraintes issues du modèle de réflectance de Blinn-Phong. Les paramètres de forme du *3DMM* ainsi que l'albédo de la surface peuvent ainsi être conjointement estimés [PAT 09]. Pour se défaire des limitations d'un modèle de forme, Kemelmacher-Shlizerman et Basri [KEM 11a] s'appuient uniquement sur un modèle moyen pour estimer la pose et des sources lumineuses dans un premier temps. Celui-ci est ensuite déformé pour s'accorder avec les observations d'une unique image, en optimisant la profondeur des points du modèle. Cette méthode peut s'avérer avantageuse car elle ne nécessite pas l'apprentissage d'un modèle déformable de forme qui implique la mise en correspondance dense d'un grand nombre d'acquisitions 3D, et elle permet de recréer des formes non présentes dans le modèle.

La force des méthodes présentées dans cette partie réside dans l'exploitation conjointe des informations *a priori* sur la classe d'objets à reconstruire, à savoir les

visages, et l'utilisation d'informations photométriques ou géométriques liées au système. Les hypothèses *a priori* permettent d'accélérer l'initialisation, et les méthodes sans modèles permettent de reconstituer une forme et une texture avec une précision supérieure à celle possible avec les contraintes d'un modèle.

### 4.6. Intégration de l'aspect temporel

Dans les sections 4.3, 4.4 et 4.5, la reconstruction du visage est effectuée à partir d'une ou plusieurs images acquises simultanément. Or, de plus en plus de systèmes intègrent désormais des capteurs vidéo, comme le sas d'authentification présenté au début de ce chapitre (figure 4.1). Il est alors intéressant d'exploiter l'information temporelle, pour guider l'estimation de pose (section 4.6.1), pour multiplier les scores de vérification (section 4.6.2) ou encore pour consolider le modèle en cours d'estimation (section 4.6.3).

#### 4.6.1. *Suivi de visage*

Avant d'estimer la forme et la texture d'un visage, il faut tout d'abord déterminer (au moins approximativement) sa position, son orientation, et, selon les algorithmes utilisés, détecter certains points caractéristiques. La plupart des algorithmes cités précédemment s'appuient en effet sur ces points, et la qualité de la reconstruction dépend alors du nombre de détections et de leur précision. Dans un environnement non contraint, le visage n'est pas toujours vu frontalement, et la détection du visage et de ses points caractéristiques peut s'avérer difficile et ainsi aboutir à des points aberrants ou imprécis, ainsi qu'à des non-détections. De plus, l'utilisation de détecteurs sur toute l'image est une opération coûteuse, surtout si un détecteur différent est employé pour chaque point sémantique recherché. Si l'on dispose de flux vidéo, il est alors intéressant d'intégrer un filtrage temporel pour guider la détection du visage et des points.

De nombreuses méthodes ont été proposées depuis plus d'une vingtaine d'années pour résoudre cette problématique dans le cadre d'une séquence vidéo. Le suivi de pose peut se décliner en plusieurs cas : suivi de la position 2D (et éventuellement de l'orientation) à partir d'une caméra ou le suivi de la pose 3D à partir d'une ou plusieurs caméras. Une vue d'ensemble des méthodes d'estimation de pose est proposée dans [MUR 09], traitant le cas de l'estimation de pose sur une seule image et dans les flux vidéo. Dans cette section, nous nous intéressons en particulier aux méthodes fondées sur le suivi, qui tirent ainsi bénéfice de l'aspect séquentiel des images et de l'historique associé.



Plusieurs approches s'appuient sur des méthodes de flot optique pour estimer récursivement la pose du visage, mais elles sont généralement contraintes par l'hypothèse d'invariance de luminosité, et nécessitent un taux de rafraîchissement élevé. Ces méthodes peuvent être combinées à l'utilisation de caractéristiques d'un modèle moyen pour faire le suivi et relâcher ces conditions [MAL 00]. D'autres méthodes s'appuient sur l'information issue de détecteurs de visages [YAN 06] ou de points caractéristiques [COM 03] pour évaluer la pose [ZHU 04]. Cependant, en cas de grandes variations d'apparence de l'objet à suivre (ici, le visage change de pose du fait de la variation de position vis-à-vis des capteurs), l'apprentissage d'un détecteur robuste de visages ou de points peut s'avérer difficile. Il est alors préférable d'employer des approches ne s'appuyant pas sur des informations de détection.

Le filtre de Kalman [KAL 60], le filtre de Kalman étendu et le filtre particulaire [DOU 00] sont différentes déclinaisons de la théorie bayésienne appliquée à des problématiques de filtrage. Le filtre de Kalman permet d'estimer récursivement un état $X_t$ à l'instant $t$ (par exemple la position de l'objet d'intérêt dans l'image $I_t$) et l'erreur associée sous la forme d'une matrice de covariance $C_t$, à partir des observations courantes $y_t$ et des valeurs $X_{t-1}$ et $C_{t-1}$ calculées à l'instant précédent. Pour appliquer le filtre de Kalman, des hypothèses gaussiennes et de linéarité sont requises sur les fonctions et les bruits impliqués dans le processus. Le filtrage particulaire permet de relâcher ces conditions en s'appuyant sur une approximation de la densité de probabilité de l'état $X_t$ par un ensemble de particules, chacune d'entre elles représentant une hypothèse de l'état. Chaque particule est associée à un poids qui caractérise sa cohérence avec les observations, et qui est mis à jour à chaque nouvelle trame.

Les travaux qui s'appuient sur un filtre particulaire pour suivre la pose d'un visage se différencient notamment par les critères utilisés pour évaluer la vraisemblance des particules. Classiquement, un critère de couleur est utilisé pour l'estimer [PER 02], ce qui est limitant en cas de variations d'illumination et/ou de pose, car l'apparence du visage change beaucoup dans ces cas. Dans [OKA 05], les particules sont évaluées à partir de vraisemblances locales en certains points caractéristiques, ce qui augmente la robustesse aux variations de pose. Kobayashi *et al.* [KOB 06] proposent un critère original de vraisemblance en intégrant des classifieurs faibles (utilisant des réponses de filtres de Haar) fusionnés par Adaboost dans le filtre. Pour être robustes aux variations de pose, des classifieurs peuvent être appris pour différents intervalles d'orientation. Ainsi, le choix du classifieur informe sur l'intervalle de la pose, sans toutefois l'estimer avec précision. Ba *et al.* proposent de coupler le processus de suivi et d'estimation de pose [BA 04] pour estimer conjointement la position et l'orientation avec précision. Les particules possèdent un état dit mixte, caractérisé par la position 2D du visage dans l'image d'une part, et par son orientation d'autre part. Afin d'évaluer la vraisemblance des particules,

un apprentissage est fait en amont pour caractériser les réponses de visages à des filtres gaussiens et de Gabor pour différentes poses. Pour une particule donnée, la réponse observée dans l'image est alors comparée à la réponse attendue étant donné son état, ce qui permet d'évaluer sa vraisemblance. En faisant simultanément le suivi et l'estimation de pose, le nombre de particules à utiliser pour assurer un suivi robuste augmente exponentiellement avec la dimension de l'espace de recherche. Le temps de calcul associé à l'étape de filtrage est donc plus important, mais permet d'obtenir simultanément une estimation de la position et de l'orientation du visage et offre de la robustesse aux variations de pose.

Un autre moyen d'augmenter la robustesse aux changements d'illumination et de pose sans modèle 3D est de mettre à jour des caractéristiques d'apparence de l'objet à suivre [ROS 08, OKA 05]. Néanmoins, en adaptant les descripteurs aux observations les plus récentes, les méthodes de mise à jour souffrent potentiellement du problème de dérive (souvent appelé *drift*), c'est-à-dire d'insertion de mauvaises caractéristiques de l'objet dans le modèle d'apparence. Ce biais entraîne une accumulation d'erreurs sur le suivi pour finalement perdre l'objet. Pour limiter cet effet, des contraintes sur la mise à jour du modèle peuvent être imposées, en contrôlant par exemple la différence entre les anciennes et nouvelles caractéristiques [KIM 08].

Le suivi d'un objet avec pose variable peut aussi être amélioré avec un modèle 3D explicite, bénéficiant ainsi de l'apparence de l'objet sous n'importe quelle pose. Cette connaissance peut être intégrée dans une approche par filtre particulaire, où la vraisemblance est alors calculée en comparant les observations aux vues synthétisées à partir des états des particules [HER 11, BRO 12].

Hors du contexte du filtrage particulaire, des méthodes d'optimisation de type Gauss-Newton s'appuient également sur cette donnée pour effectuer le suivi, en optimisant les paramètres de pose qui explicitent la projection du modèle sur l'image observée [MUÑ 09].

Les processus de suivi détaillés dans cette partie peuvent être vus comme une étape préliminaire aux algorithmes d'estimation de forme et de texture. En effet, la plupart des méthodes d'ajustement du modèle nécessitent une initialisation de pose et/ou des positions de points caractéristiques dans l'image. Le résultat du suivi fournit la première information, qui permet ensuite de limiter les zones de recherche des détecteurs de points, réduisant le temps de traitement pour une trame. D'autre part, les processus de filtrage sur la pose permettent de vérifier la cohérence temporelle des positions successives, et de détecter des valeurs de pose aberrantes en cas d'échec de l'algorithme.



### 4.6.2. *Approche statique à partir de flux vidéo*

Disposant de flux vidéo plutôt que d'une seule image pour reconnaître une personne, un premier moyen d'exploiter la séquence est d'appliquer à chaque trame les processus de reconstruction précédemment présentés. Ainsi, on ne fait plus une seule comparaison, mais autant que de trames (et donc de vues frontales) disponibles. Même si des occultations sont présentes temporairement, ou si le visage est mal estimé à une trame donnée, d'autres seront potentiellement valides et donc utilisables. Pour optimiser cette méthode en temps d'exécution, il faut cependant déterminer des règles qui permettent de filtrer les mauvaises trames et établir une stratégie de décision étant donné l'ensemble des scores obtenus par les différentes mises en correspondance entre la photographie de référence et les vues frontales synthétisées.

La sélection des trames à traiter pour estimer le modèle de visage peut être faite par le biais de différents critères. La confiance des détecteurs, la résolution du visage ou encore sa pose sont couramment utilisées dans ce but [SAT 00, VIL 10]. En plus de ces critères estimés sur une image, il est également possible de prendre en compte leur variation temporelle entre deux trames pour une meilleure définition de la qualité. Tous ces critères doivent ensuite être fusionnés pour améliorer le processus de sélection de trames. Hormis les règles de fusion par moyenne, produit, ou sélection du minimum/maximum, d'autres méthodes ont été proposées, telles que l'utilisation de séparateur à vaste marge (*Support Vector Machine*), de $k$ plus proches voisins, ou encore une fusion par maximisation de l'aire sous la courbe ROC (*Receiver Operating Characteristics*), qui fournit les meilleurs résultats d'après l'étude comparative proposée dans [VIL 10].

Une fois la sélection de trames effectuée, les visages reconstruits et les vues frontales générées, il s'agit d'établir un résultat de reconnaissance pour la séquence par fusion des différents scores de mise en correspondance. Pour étendre les méthodes classiques de comparaison entre images au cas de la vidéo (ou un sous-ensemble de ses trames), une distance doit être définie entre l'ensemble des images de la requête (base ou unique photo d'identité) d'une part, et le flux vidéo d'autre part. Une définition possible est la plus petite distance calculée sur l'ensemble des paires possibles (établie par exemple dans l'espace des *eigenfaces* [SAT 00]). Dans le cas particulier de l'identification (où un individu doit être sélectionné parmi une liste donnée), des critères spécifiques peuvent être utilisés pour pondérer les résultats obtenus sur une trame, à savoir :

– une distance au modèle, qui caractérise la distance à la classe du visage de l'individu le plus proche. Celle-ci a pour but d'éliminer des visages détectés avec une pose ou une illumination non représentée dans les classes apprises ;

– une distance à la deuxième plus proche classe, pour vérifier la validité de la classification. Ce critère s'appuie sur le fait que si la classe sélectionnée est correcte, elle doit présenter des scores de mise en correspondance bien meilleurs qu'avec la deuxième meilleure classe.

En intégrant ces différents critères dans la fusion de scores, les résultats de classification sont à la fois meilleurs que ceux obtenus indépendamment sur chaque trame, et par somme sur tous les scores [STA 07].

Les stratégies présentées dans cette partie permettent d'exploiter au mieux les informations biométriques disponibles dans une séquence, mais en considérant le processus de reconstruction indépendamment sur chaque trame. Selon le nombre de capteurs, la pose du visage observé, ou les algorithmes employés, la reconstruction du visage associée à une trame n'est pas toujours complète (en cas d'auto-occultations) ni précise (bruit sur les images, caractéristiques du visage insuffisantes, etc.). L'objet de la prochaine partie sera de voir comment exploiter simultanément les différentes trames d'une séquence pour améliorer la qualité de la reconstruction.

### 4.6.3. *Consolidation temporelle à partir de flux vidéo*

Comme nous l'avons relevé précédemment, le problème de reconstruction d'un visage peut être mal posé selon les caractéristiques utilisées pour procéder au recalage et à l'estimation du visage. L'utilisation de plusieurs vues permet tout d'abord de lever l'ambiguïté de profondeur issue de la projection dans le cas d'une image unique. Par ailleurs, comme la pose du visage varie au cours de la séquence, des zones occultées deviennent visibles, permettant ainsi de compléter l'estimation de forme et/ou de texture. Voyons donc à présent comment fonctionnent les méthodes exploitant les séquences vidéo pour consolider la reconstruction.

Citons tout d'abord les méthodes de consolidation 2D, qui ne nécessitent pas de modèle de forme tridimensionnelle. Ainsi, Hu *et al.* [HU 09] proposent une reconstruction incrémentale de la vue frontale d'un visage en accumulant les régions correspondant aux zones visibles à chaque instant. Une distorsion est appliquée aux observations pour recaler les textures extraites sur un modèle moyen de vue frontale. Ainsi, les zones observées pendant la séquence permettent de reconstruire un visage plus complet que sur une seule trame. L'avantage de telles méthodes est leur rapidité et leur indépendance vis-à-vis de modèles complexes. Néanmoins, plus l'angle de pose est important, plus il sera difficile de trouver les paramètres de similitude de la déformation.

L'utilisation de modèles 3D offre une plus grande robustesse aux variations de pose observées dans les flux vidéo. Un modèle de forme 3D peut être estimé à partir



d'un ensemble de trames issues d'une vidéo, en exploitant l'information de silhouette [SAI 07], de points caractéristiques détectés [FAG 08], ou de points saillants mis en correspondance [FID 07] dans les trames de la vidéo. La pose dans chaque trame peut être obtenue en utilisant un marqueur spécifique sur le visage [SAI 07] ou estimée par des méthodes dites de structure à partir du mouvement (*Structure from Motion*) s'appuyant sur un ensemble épars de points du visage mis en correspondance, éventuellement contraints par un modèle générique [FID 07]. Comme dans le cas de certaines méthodes hybrides présentées dans la section 4.5, ce modèle peut ensuite être abandonné pour obtenir une restitution précise des formes du visage. Les méthodes de *Structure From Motion* s'appuient classiquement sur une mise en correspondance de points particuliers sur les différentes vues, ce qui contraint la variation de pose entre les trames pour des questions de visibilité des points. L'utilisation d'un modèle de visage crée un espace intermédiaire pour lier toutes les observations, et il n'est alors plus nécessaire d'apparier les détections entre les vues, celles-ci étant rattachées au modèle. Pour estimer au mieux à la forme observée, il est cependant nécessaire de disposer d'un nombre conséquent de points (46 dans [FAG 08]) répartis sur l'ensemble du visage. Sans cela, plusieurs jeux de paramètres peuvent vérifier la correspondance, sans avoir une validation dense de la similitude de forme (ce qui rejoint certaines problématiques relevées dans la section 4.4.2.2). Cette méthode, qui a l'avantage d'être rapide, nécessite toutefois un nombre important de points caractéristiques en entrée, qu'il n'est pas toujours possible de détecter selon la pose du visage dans les images. Comme précédemment, un compromis entre les critères utilisés (liés la précision de la reconstruction) et la vitesse d'exécution est donc à trouver pour remplir les conditions de précision et de rapidité imposées par le système. Une approche probabiliste peut également être envisagée pour estimer les paramètres de forme à partir d'une séquence vidéo, par exemple par un filtre particulaire [HER 12].

Une difficulté supplémentaire des séquences vidéo par rapport au cas de l'acquisition multi-vues synchronisée est due aux variations que peut présenter un visage entre deux instants (clignement d'œil, pincement de bouche, etc.). Certaines méthodes exploitent un modèle de déformation d'expressions (de type *Candide* par exemple) pour estimer les déformations faciales et en déduire les expressions [DOR 05, OKA 05, MUÑ 09]. Bien sûr, pour imposer le moins de contraintes possibles à l'utilisateur, un système optimal permettrait d'estimer la forme tout en étant robuste à des variations d'expressions. Des méthodes ont déjà été proposées dans ce but [AMB 08], mais n'exploitent pas directement d'images (ou de flux vidéo). L'optimisation d'un double modèle de forme et d'expressions est faite à partir d'une information géométrique issue d'un scanner 3D, et n'exploite ni l'intensité des images, ni une cohérence temporelle sur les expressions. Les potentialités de telles méthodes sur des données purement images restent donc à évaluer.

## 4.7. Conclusion

Au fil de ce chapitre, nous avons passé en revue les différentes méthodes existantes pour reconstruire un visage en 3D. Du fait des contraintes applicatives d'un système biométrique, nous nous sommes concentrés sur des méthodes dites passives, s'appuyant uniquement sur des acquisitions vidéo. La reconstruction est ensuite utilisée pour générer la vue frontale associée à des fins de reconnaissance faciale.

Des méthodes génériques telles que la stéréovision ou le *Shape from Shading* peuvent être utilisées, mais l'utilisation d'hypothèses *a priori* sur les formes et les apparences permet d'améliorer la qualité des reconstructions, en particulier en environnement non contrôlé (pose non frontale, illumination variée). Certains auteurs ont proposé de bénéficier des avantages des deux types de méthodes, et proposent des résultats très convaincants en mêlant modèle 3D et stéréovision par exemple.

Enfin, l'utilisation de la vidéo permet, d'une part, d'imposer des contraintes temporelles sur les poses de visages estimées pour accélérer l'initialisation, et d'autre part de bénéficier de plusieurs estimations de visages, voire de consolider temporellement l'estimation des paramètres du modèle. Cela est utile en particulier lorsqu'une seule caméra est disponible, pour améliorer la qualité de l'estimation de la forme 3D et compléter la texture au cours de l'acquisition. Dès lors qu'un flux vidéo est exploité, il faut prendre en compte l'existence de dynamiques du visage, liées aux expressions, aux mouvements des yeux, etc. La robustesse de la reconstruction de visages aux expressions est, dans une image unique ou un flux vidéo, un axe de recherche très actif actuellement.

## 4.8. Bibliographie

[AHL 01] AHLBERG J., CANDIDE-3 - An Updated Parameterised Face, Rapport n°LiTHISY-R-2326, Departement of Electrical Engineering, Linköping University, Suède, janvier 2001.

[AMB 07] AMBERG B., BLAKE A., FITZGIBBON A.W., ROMDHANI S., VETTER T., « Reconstructing High Quality Face-Surfaces using Model Based Stereo », *International Conference on Computer Vision - Rio de Janeiro*, p. 1-8, 2007.

[AMB 08] AMBERG B., KNOTHE R., VETTER T., « Expression invariant 3D face recognition with a Morphable Model », *IEEE International Conference on Automatic Face and Gesture Recognition - Amsterdam*, p. 1-6, 2008.

[AST 11] ASTHANA A., MARKS T.K., JONES M.J., TIEU K.H., RONITH M., « Fully Automatic Pose-Invariant Face Recognition *via* 3D Pose Normalization », *International Conference on Computer Vision - Barcelona*, p. 937-944, 2011.



[BA 04] BA S., ODOBEZ J., « A Probabilistic Framework for Joint Head Tracking and Pose Estimation », *International Conference on Pattern Recognition*, vol. 4, p. 264-267, 2004.

[BAK 01] BAKER S., MATTHEWS I., « Equivalence and Efficiency of Image Alignment Algorithms », *IEEE Conference on Computer Vision and Pattern Recognition - Kauai, HI*, p. 1090-1097, 2001.

[BAS 07] BASRI R., JACOBS D., KEMELMACHER I., « Photometric Stereo with General, Unknown Lighting », *International Journal of Computer Vision*, vol. 72, p. 239-257, Kluwer Academic Publishers, Londres, mai 2007.

[BEE 10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M., « High-Quality Single-Shot Capture of Facial Geometry », *ACM Transactions on Graphics (SIGGRAPH)*, vol. 29, n°4, p. 40 :1, 40 :9, Los Angeles, Etats-Unis, 2010.

[BEL 97] BELHUMEUR P.N., KRIEGMAN D.J., YUILLE A.L., « The Bas-Relief Ambiguity », *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1060-1066, 1997.

[BLA 03a] BLANZ V., BASSO C., POGGIO T., VETTER T., « Reanimating Faces in Images and Video », *Computer Graphics Forum, Eurographics*, vol. 22, n°3, p. 641-650, 2003.

[BLA 03b] BLANZ V., VETTER T., « Face Recognition based on Fitting a 3D Morphable Model », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, p. 1063-1074, 2003.

[BLA 04] BLANZ V., MEHL A., VETTER T., SEIDEL H.-P., « A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data », *3D Data Processing, Visualization, and Transmission*, p. 293-300, 2004.

[BLA 99] BLANZ V., VETTER T., « A Morphable Model for the Synthesis of 3D Faces », *SIGGRAPH*, p. 187-194, 1999.

[BRA 10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A., « High Resolution Passive Facial Performance Capture », *ACM Transactions on Graphics (SIGGRAPH) - Los Angeles*, vol. 29, n°4, p. 41:1, 41:10, 2010.

[BRE 10] BREUER P., BLANZ V., « Self-Adapting Feature Layers », *European Conference on Computer Vision - Heraklion*, p. 299-312, 2010.

[BRO 12] BROWN J.A., CAPSON D.W., « A Framework for 3D Model-Based Visual Tracking Using a GPU-Accelerated Particle Filter », *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, p. 68-80, 2012.

[COM 03] COMANICIU D., RAMESH V., MEER P., « Kernel-based object tracking », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, n°5, p. 564-577, 2003.

[COO 95] COOTES T.F., TAYLOR C.J., COOPER D.H., GRAHAM J., « Active Shape Models - Their Training and Application », *Computer Vision and Image Understanding*, vol. 61, p. 38-59, 1995.

[DAL 09] DALALYAN A., KERIVEN R., « L1-Penalized Robust Estimation for a Class of Inverse Problems Arising in Multiview Geometry », *Conference on Neural Information Processing Systems*, p. 441-449, Vancouver, Canada, 2009.

[DOR 05] DORNAIKA F., DAVOINE F., « Simultaneous Facial Action Tracking and Expression Recognition Using a Particle Filter », *International Conference on Computer Vision - Beijing*, p. 1733-1738, 2005.

[DOU 00] DOUCET A., GODSILL S., ANDRIEU C., « On Sequential Monte Carlo Sampling Methods for Bayesian Filtering », *Statistics And Computing*, vol. 10, n°3, p. 197-208, 2000.

[EDW 98] EDWARDS G.J., TAYLOR C.J., COOTES T.F., « Interpreting Face Images Using Active Appearance Models », *IEEE International Conference on Automatic Face and Gesture Recognition - Nara*, p. 300-305, 1998.

[FAG 08] FAGGIAN N., PAPLINSKI A.P., SHERRAH J., « 3D Morphable Model fitting from multiple views », *IEEE International Conference on Automatic Face and Gesture Recognition - Amsterdam*, p. 1-6, 2008.

[FID 07] FIDALEO D., MEDIONI G., « Model-assisted 3D face reconstruction from video », *International Conference on Analysis and Modeling of Faces and Gestures*, p. 124-138, Springer-Verlag, Heidelberg, 2007.

[HAR 04] HARTLEY R.I., ZISSERMAN A., *Multiple View Geometry in Computer Vision*, Cambridge University Press, Londres, 2004 (2e édition).

[HER 11] HEROLD C., GENTRIC S., MOËNNE LOCCOZ N., « Suivi de la pose 3D du visage en environnement multi-caméras avec un modèle tridimensionnel individualisé », *ORASIS*, Praz-sur-Arly, France, 2011.

[HER 12] HEROLD C., DESPIEGEL V., GENTRIC S., DUBUISSON S., BLOCH I., « Head Shape Estimation using a Particle Filter including Unknown Static Parameters », *International Conference on Computer Vision Theory and Applications - Rome*, p.284-293, 2012.

[HU 09] HU C., HARGUESS J., AGGARWAL J.K., « Patch-Based Face Recognition from Video », *IEEE International Conference on Image Processing - Le Caire*, p. 3285-3288, 2009.

[HUA 11] HUANG H., CHAI J., TONG X., WU H.-T., « Leveraging Motion Capture and 3D Scanning for High-Fidelity Facial Performance Acquisition », *ACM Transactions on Graphics (SIGGRAPH) - Vancouver*, vol. 30, n°4, p. 74 :1-74:10, 2011.

[IVA 07] IVALDI W., Synthèse de vue frontale et modélisation 3D de visages par vision multi-caméras, Thèse de doctorat, ISIR/université Pierre et Marie Curie, Paris, 2007.

[JIN 03] JINHO B.M., LEE J., PFISTER H.,MACHIRAJU R., « Model-Based 3D Face Capture with Shape-from-Silhouettes », *IEEE International Workshop on Analysis and Modeling of Faces and Gestures - Nice*, p. 20-27, 2003.

[KAL 60] KALMAN R.E., « A New Approach to Linear Filtering and Prediction Problems », *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, Series D, p. 35-45, 1960.



[KEM 11a] KEMELMACHER-SHLIZERMAN I., BASRI R., « 3D Face Reconstruction from a Single Image Using a Single Reference Face Shape », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, p. 394-405, 2011.

[KEM 11b] KEMELMACHER-SHLIZERMAN I., SEITZ S.M., « Face Reconstruction in the Wild », *International Conference on Computer Vision*, Barcelone, Espagne, 2011.

[KIM 08] KIM M., KUMAR S., PAVLOVIC V., ROWLEY H.A., « Face Tracking and Recognition with Visual Constraints in Real-World Videos », *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage (AL), Etats Unis, 2008.

[KOB 06] KOBAYASHI Y., SUGIMURA D., SATO Y., HIRASAWA K., SUZUKI N., KAGE H., AKIHIRO SUGIMOTO, « 3D Head Tracking using the Particle Filter with Cascaded Classifiers », *British Machine Vision Conference - Edinburgh*, p. 1-10, 200

[LEE 03] LEE J., MOGHADDAM B., PFISTER H., MACHIRAJU R., « Silhouette-Based 3D Face Shape Recovery », *Graphics Interface*, p. 21-30, Halifax, Nova Scotia, Canada, 2003.

[LIN 10] LIN Y., MEDIONI G.G., CHOI J., « Accurate 3D Face Reconstruction from Weakly Calibrated Wide Baseline Images with Profile Contours », *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1490-1497, San Francisco, Etats-Unis, 2010.

[MAL 00] MALCIU M., PRÊTEUX F., « A Robust Model-Based Approach for 3D Head Tracking in Video Sequences », *IEEE International Conference on Automatic Face and Gesture Recognition*, p. 169-174, 2000.

[MAR 99] MARSCHNER S., WESTIN S., LAFORTUNE E., TORRANCE K., GREENBERG D., « Image-Based BRDF Measurement Including Human Skin », *Eurographics Workshop on Rendering*, p. 131-144, 1999.

[MAT 09] MATTA F., DUGELAY J.-L., « Person Recognition using Facial Video Information : A State of the Art », *Journal of Visual Languages*, vol. 20, p. 180-187, 2009.

[MOË 10] MOËNNE-LOCCOZ N., ROQUEMAUREL B.D., ROMDHANI S., GENTRIC S., « Reconstruction à la volée de portraits frontaux par modélisation 3D des visages », *Revue Electronique Francophone d'Informatique Graphique*, vol. 4, p. 14-19, 2010.

[MUÑ 09] MUÑOZ E., BUENAPOSADA J.M., BAUMELA L., « A Direct Approach for Efficiently Tracking with 3D Morphable Models », *International Conference on Computer Vision*, p. 1615-1622, 2009.

[MUR 09] MURPHY-CHUTORIAN E., TRIVEDI M.M., « Head Pose Estimation in Computer Vision : A Survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence – Kyoto*, vol. 31, p. 607-626, 2009.

[NEW 10] NEWCOMBE R.A., DAVISON A.J., « Live Dense Reconstruction with a Single Moving Camera », *IEEE Conference on Computer Vision and Pattern Recognition - San Francisco*, p. 1498-1505, 2010.

[OKA 05] OKA K., SATO Y., « Real-Time Modeling of Face Deformation for 3D Head Pose Estimation », *IEEE International Conference on Automatic Face and Gesture Recognition - Beijing*, p. 308-320, 2005.

[PAN 03] PANDZIC I.S., FORCHHEIMER R. (dir.), *MPEG-4 Facial Animation : The Standard, Implementation and Applications*, John Wiley & Sons, New York, Etat-Unis, 2003.

[PAS 01] PASQUARIELLO S., PELACHAUD C., « Greta : A Simple Facial Animation Engine », *Conference on Soft Computing in Industrial Applications*, Blacksburg, Etats-Unis, 2001.

[PAT 09] PATEL A., SMITH W.A.P., « Shape-from-Shading Driven 3D Morphable Models for Illumination Insensitive Face Recognition », *British Machine Vision Conference*, Londres, Royaume Uni, 2009.

[PEN 89] PENTLAND A.P., « Local Shading Analysis », dans HORN B.K.P. (dir.), *Shape from Shading*, p. 443-487, MIT Press, Cambridge, Etats-Unis, 1989.

[PER 02] PEREZ P., HUE C., VERMAAK J., GANGNET M., « Color-Based Probabilistic Tracking », *European Conference on Computer Vision - Copenhagen*, p. 661-675, 2002.

[POL 04] POLLEFEYS M., VAN GOOL L., VERGAUWEN M., VERBIEST F., CORNELIS K., TOPS J., KOCH R., « Visual Modeling with a Hand-Held Camera », *International Journal of Computer Vision*, vol. 59, p. 207-232, 2004.

[ROM 02] ROMDHANI S., BLANZ V., VETTER T., « Face Identification by Fitting a 3D Morphable Model using Linear Shape and Texture Error Functions », *European Conference on Computer Vision - Copenhagen*, p. 3-19, 2002.

[ROM 03] ROMDHANI S., VETTER T., « Efficient, Robust and Accurate Fitting of a 3D Morphable Model », *International Conference on Computer Vision - Nice*, p. 59-66, 2003.

[ROM 05] ROMDHANI S., VETTER T., « Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior », *IEEE Conference on Computer Vision and Pattern Recognition - Beijing*, p. 986-993, 2005.

[ROS 08] ROSS D.A., LIM J., LIN R.-S., YANG M.-H., « Incremental Learning for Robust Visual Tracking », *International Journal of Computer Vision*, vol. 77, p. 125-141, 2008.

[RYD 87] RYDFALK M., *CANDIDE : A Parameterized face*, Rapport n°LiTH-ISY-I-0866, Linköping University, Suède, 1987.

[SAI 07] SAITO H., ITO Y., MOCHIMARU M., « Face Shape Reconstruction from Image Sequence Taken with Monocular Camera using Shape Database », *International Conference on Image Analysis and Processing - Modena*, p. 165-170, 2007.

[SAT 00] SATOH S., « Comparative Evaluation of Face Sequence Matching for Content-Based Video Access », *IEEE International Conference on Automatic Face and Gesture Recognition - Grenoble*, p. 163-168, 2000.

[SEI 06] SEITZ S. M., CURLESS B., DIEBEL J., SCHARSTEIN D., SZELISKI R., « A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms », *IEEE Conference on Computer Vision and Pattern Recognition - New York*, p. 519-528, 2006.



[STA 07] STALLKAMP J., EKENEL H.K., STIEFELHAGEN R., « Video-based Face Recognition on Real-World Data », *International Conference on Computer Vision - Rio de Janeiro*, p. 1-8, 2007.

[TUR 91] TURK M.A., PENTLAND A.P., « Face Recognition using Eigenfaces », *IEEE Conference on Computer Vision and Pattern Recognition - Maui*, p. 586-591, 1991.

[VET 97] VETTER T., « Recognizing Faces from a New Viewpoint », *IEEE International Conference on Acoustics, Speech, and Signal Processing - München*, p. 143-146, 1997.

[VIL 10] VILLEGAS M., PAREDES R., « Fusion of Qualities for Frame Selection in Video Face Verification », *International Conference on Pattern Recognition - Istanbul*, p. 1302-1305, 2010.

[VIO 04] VIOLA P., JONES M.J., «Robust Real-Time Face Detection », *International Journal of Computer Vision*, vol. 57, p. 137-154, 2004.

[WIS 97] WISKOTT L., FELLOUS J.-M., KRÜGER N., VON DER MALSBURG C., « Face Recognition by Elastic Bunch Graph Matching », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, n°7, p. 775-779, 1997.

[WOO 89] WOODHAM R. J., « Photometric Method for Determining Surface Orientation from Multiple Images », dans HORN B.K.P. (dir.), *Shape from Shading*, p. 513-531, MIT Press, Cambridge, Etats-Unis,1989.

[WU 11] WU C., WILBURN B., MATSUSHITA Y., THEOBALT C., « High-Quality Shape from Multi-View Stereo and Shading under General Illumination », *IEEE Conference on Computer vision and Pattern Recognition – Colorado Springs*, p. 969-976, 2011.

[XIA 04] XIAO J., BAKER S., MATTHEWS I., KANADE T., « Real-time Combined 2D+3D Active Appearance Models », *IEEE Conference on Computer Vision and Pattern Recognition - Washington*, p. 535-542, 2004.

[YAN 06] YANG T., LI S.Z., PAN Q., LI J., ZHAO C., « Reliable and Fast Tracking of Faces under Varying Pose », *IEEE International Conference on Automatic Face and Gesture Recognition - Southhapton*, p. 421-428, 2006.

[YAN 98] HSUAN YANG M., AHUJA N., « Detecting Human Faces in Color Images », *IEEE International Conference on Image Processing - Chicago*, p. 127-130, 1998.

[ZHA 00] ZHANG Z., « A Flexible New Technique for Camera Calibration », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, p. 1330-1334, 2000.

[ZHA 04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S.M., « Spacetime Faces : High Resolution Capture for Modeling and Animation », *ACM Transactions on Graphics (SIGGRAPH) – Los Angeles*, vol. 23, n°3, p. 548-558, 2004.

[ZHA 10] ZHANG C., ZHANG Z., *A Survey of Recent Advances in Face Detection*, Rapport n°MSR-TR-2010-66, Microsoft Research, 2010.

[ZHA 99] ZHANG R., TSAI P.-S., CRYER J.E., SHAH M., « Shape from Shading : A Survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, n°8, p. 690-706, 1999.

[ZHU 04] ZHU Z., JI Q., « 3D Face Pose Tracking From an Uncalibrated Monocular Camera », *International Conference on Pattern Recognition - Cambridge*, p. 400-403, 2004.

[ZOL 11] ZOLLHÖFER M., MARTINEK M., GREINER G., STAMMINGER M., SÜSSMUTH J., « Automatic Reconstruction of Personalized Avatars from 3D Face Scans », *Computer Animation and Virtual Worlds*, vol. 22, p. 195-202, 2011.

Chapitre 5

# Reconnaissance faciale 3D

## 5.1. Introduction

La reconnaissance faciale 3D permet de palier certains problèmes liés à la pose et aux conditions d'éclairage. En effet, l'information 3D, une fois obtenue par les capteurs appropriés, est invariante aux changements des conditions d'éclairage et de pose. Néanmoins, la déformation faciale causée par les expressions a constitué un des défis auquel les chercheurs et les industriels tentent de répondre. La reconnaissance faciale 3D nécessite donc l'acquisition 3D de visage. Les capteurs 3D commerciaux mais aussi les solutions proposées par la communauté de recherche ont des limitations. Nous pouvons citer : la portée des capteurs qui est de 1 à 2 mètres ; les conditions d'éclairage contraintes ; la précision ; et finalement la durée de l'acquisition.

Il existe à ce jour deux principaux paradigmes de reconnaissance faciale utilisant la modalité 3D :

– la reconnaissance symétrique où les données dans la galerie et de test sont de même nature, plus précisément 3D ou 3D + texture ;

– la reconnaissance asymétrique qui utilise des données de galerie et de test hétérogènes. La galerie est alors constituée de données 3D ou 3D texturées alors que les données de test sont uniquement des images de texture ou *vice-versa*. L'avantage de ce dernier paradigme est que l'utilisation de l'information 3D limitée. On parle aussi de la reconnaissance assistée par le 3D.

---

Chapitre rédigé par Mohsen ARDABILIAN, Przemyslaw SZEPTYCKI, Di HUANG et Liming CHEN.