Évaluation et comparaison de méthodes de reconstruction 3D fine de visages à partir de plusieurs images non calibrées

Hassan Lhallabi, Géraldine Morin, Simone Gasparini, Sylvie Chambon, Xavier Naturel, Jérôme Guénard

▶ **To cite this version:**

Hassan Lhallabi, Géraldine Morin, Simone Gasparini, Sylvie Chambon, Xavier Naturel, et al.. Évaluation et comparaison de méthodes de reconstruction 3D fine de visages à partir de plusieurs images non calibrées. 19èmes journées francophones des jeunes chercheurs en vision par ordinateur (ORASIS 2023), Equipe "Signal et Image" du Laboratoire d'Informatique et Systèmes (LIS), UMR7020, May 2023, Carqueiranne, France. pp.1-8. hal-04219395

HAL Id: hal-04219395
https://hal.science/hal-04219395

Submitted on 27 Sep 2023

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

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

Évaluation et comparaison de méthodes de reconstruction 3D fine de visages à partir de plusieurs images non calibrées

Hassan Lhallabi[1,2] Sylvie Chambon[1] Géraldine Morin[1]
Simone Gasparini[1] Xavier Naturel[2] Jérôme Guenard[2]

[1] IRIT, Université de Toulouse, Toulouse INP, Toulouse, France
[2] Fittingbox, Toulouse, France

hassan.lhallabi@irit.fr

Résumé

Dans cet article, nous effectuons un état de l'art des méthodes qui reconstruisent le visage à partir de plusieurs images non calibrées, par optimisation, réseaux de neurones ou une combinaison des deux. Nous souhaitons reconstruire la géométrie du visage. La texture et l'expression n'étant pas nécessaires, nous analysons ces méthodes aux objectifs variés dans ce cadre. La reconstruction d'un visage peut prendre en compte un a priori, nous en présentons les approches existantes. La comparaison entre les méthodes est difficile puisqu'elles ne sont pas testées sur les mêmes données et/ou selon la même méthodologie. Nous étudions les jeux de données existants, ainsi que leur pertinence pour notre cas d'utilisation : une acquisition des images par smartphone. Nous évaluons 3 *méthodes de l'état de l'art disponibles publiquement sur les deux jeux de données les plus appropriés.*

Mots Clef

Reconstruction 3D, Apprentissage Profond.

Abstract

In this paper we review the state of the art of methods for the reconstruction of the face from several uncalibrated images. In particular, we limit the scope to the reconstruction of the geometry of the face without taking into account the texture and different facial expressions. We focus on classic reconstruction methods based on optimisation, as well as on more recent methods based on machine learning and those that mix the two approaches. After reviewing the main approaches, we selected the most promising ones for which an implementation is publicly available and we compared the performances on a state-of-the-art public dataset as well as an internal dataset.

Keywords

3D reconstruction, Deep Learning.

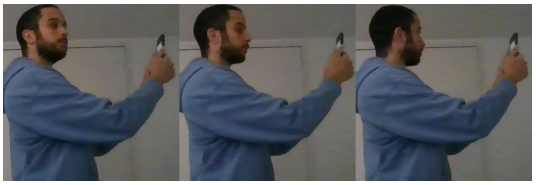

FIGURE 1 – Exemple de protocole possible, la personne peut aussi effectuer le mouvement avec son *smartphone* et garder le visage fixe.

1 Introduction

Contexte – Nous nous plaçons dans le contexte de la reconstruction 3D fine à partir de multiples points de vue. Plus précisément, dans nos travaux de recherche, nous visons un contexte d'application qui est celui de Fittingbox, une entreprise fournissant un logiciel d'essayage virtuel dans le milieu de la vente de lunettes Dans le cadre des ses activités, nous souhaitons reconstruire le visage à partir d'images. L'objectif de la reconstruction est d'effectuer des mesures (en garantissant une précision sous-millimétrique) sur le visage afin d'ajuster les lunettes essayées au visage de la personne.

Contraintes d'acquisition – Il est important de souligner que les objectifs d'une reconstruction 3D diffèrent d'une méthode à une autre. Certaines applications nécessitent d'ajouter une texture sur le visage [13], d'autres de transférer une expression d'un visage à l'autre [14] ou d'animer un avatar personnel [36]. Dans notre contexte, nous souhaitons retrouver la géométrie du visage la plus fidèle possible, avec ou sans texture.
Ainsi, en terme d'acquisition, nous devons respecter les contraintes suivantes :

1. Le visage est reconstruit avec une expression neutre.
2. L'acquisition (illumination, positions caméras, ...) n'est pas calibrée
3. L'acquisition est réalisée avec un smartphone par l'utilisatrice ou l'utilisateur, un exemple d'acquisition

possible est présenté en figure 1 .

4. Il est important pour nous que l'acquisition soit simple à réaliser pour la majorité des utilisatrices et utilisateurs. Un protocole sera donc fourni, textuellement et visuellement.

Méthodes de reconstruction 3D fine du visage – Nous pouvons présenter les différents types d'acquisition d'images pour reconstruire le modèle 3D d'un visage existants, en considérant le nombre d'images utilisées ainsi que le degré de contrôle sur l'acquisition. Les approches que nous considérons comme les plus difficiles s'appuient sur l'utilisation d'une seule image dans un environnement quelconque [37, 6, 38]. Des approches reconstruisent le visage à chaque instant d'une vidéo monoculaire [13, 14, 36] où l'expression du visage varie. Dans un cadre très contrôlé (deux caméras calibrées ou plus), la stéréovision permet de retrouver avec précision la structure 3D du visage comme par exemple dans [33] où la reconstruction 3D du visage est réalisée avec une précision de l'ordre du dixième de millimètre. Enfin, d'autres approches s'intéressent à reconstruire le visage avec un nombre réduit d'images (3 ou moins) [5, 23, 32]. Pour conclure, les approches les plus proches de notre contexte applicatif, c'est-à-dire qui s'appuient sur plus d'une dizaine d'images avec un protocole d'acquisition déterminé et avec une expression fixe et neutre sont celles de [2, 21].

Plan de la présentation – Nous avons indiqué que la reconstruction 3D du visage à partir d'images a d'abord été faite par des méthodes d'optimisation classique [6] mais, depuis quelques années, les méthodes par réseaux de neurones se multiplient. Ces deux approches posent des hypothèses différentes :

- L'optimisation lors de l'inférence essaie de retrouver la surface qui explique le mieux les données étant donné l'a priori.
- L'utilisation d'un réseau de neurones assume que l'apprentissage effectué permet de prédire avec suffisamment de précision, en fonction de l'objectif de la reconstruction souhaité, la surface dans tous les cas que l'on rencontrera.

Le temps d'inférence dépend de la taille du réseau et du problème d'optimisation mais en règle générale le réseau de neurones requiert un apprentissage préalable qui permet un temps d'inférence plus court. Certaines méthodes atteignent même le temps réel [37, 31]. Des méthodes hybrides incorporent de l'optimisation dans une approche par réseau de neurones. L'objectif est de retrouver la surface avec une meilleure précision tout en utilisant la capacité de représentation non-linéaire des réseaux, au prix d'un temps d'inférence plus élevé.
Ainsi, dans la suite de ce papier, dans un premier temps, nous nous intéressons à la modélisation du visage. En effet, il existe des approches de reconstruction qui prennent en compte cet a priori. Nous présentons les modèles paramétriques existants avant d'aborder les approches qui les utilisent puis celles qui affinent la surface issue d'un modèle donné ou qui reconstruisent une surface générique. Dans un second temps, nous étudierons les méthodes de reconstruction de la surface d'un visage à partir d'images. Nous aborderons séparément les méthodes par optimisation, et par apprentissage, puis celles qui considèrent ces deux aspects, en particulier en précisant leurs avantages et leurs inconvénients. Nous terminerons ce travail par une comparaison de plusieurs méthodes disponibles publiquement et applicables à notre cas d'utilisation afin de déterminer le type d'approche que nous souhaitons développer par la suite.

2 Modèles de visages

L'a priori sur le visage prend historiquement la forme d'un modèle paramétrique appelé 3DMM, *3D Morphable Model* [10]. Ces modèles peuvent être linéaires, cf. § 2.1, non-linéaires, cf. § 2.1, ou représentés implicitement par des réseaux de neurones, *cf*. § 2.2. Les deux premières méthodes sont explicites en opposition aux méthodes par réseaux de neurones qui apprennent implicitement un espace latent. Les 3DMM permettent de réduire l'espace de recherche et de forcer la solution à être plausible : toute combinaison de paramètres dans la distribution du modèle donnera une surface représentant un visage. Certaines approches résolvent et/ou prédisent les paramètres de tels modèles puis affinent la surface obtenue en utilisant un autre modèle ou en corrigeant les positions des sommets du maillage. Dans la suite, toutes les approches présentées sont des approches faisant intervenir une base de données d'apprentissage. Nous précisons les attributs exploités dans ces bases de données et nous donnerons des détails sur celles-ci uniquement dans le § 4.1.

2.1 Modèle moyen du visage

Blanz et Vetter [6] ont introduit cette approche. À partir d'une base de visage 3D, les étapes pour construire un tel modèle sont les suivantes :

- Aligner tous les modèles 3D ou nuages de points 3D. Il est nécessaire d'estimer les paramètres de la transformation permettant ce recalage. Pour cela, un visage de référence, dit *template* est déformé pour être recalé sur chaque visage 3D de la base. Blanz et Vetter [6], qui supposent une transformation non-rigide, adaptent un algorithme de flot optique alors que l'approche présentée dans [3] s'appuie sur un algorithme d'ICP, *Iterative Closest Point*, non-rigide calculant la transformation optimale à chaque itération. Cette approche a permis de construire les modèles LSFM, *Large Scale Facial Model* [7] et HIFI3D++, *High-Fidelity* [8]. Par opposition aux méthodes purement géométriques En supplément de la géométrie, les auteurs de [20], s'appuient sur la minimisation d'une erreur photo-métrique entre une photographie du visage et l'image de rendu du maillage de référence.

- Calculer le visage moyen noté $\bar{S}$, à partir de ces visages recalés.

Analyse en composantes principales. A partir de ces visages recalés dans la même topologie, ainsi que du visage moyen, Blanz *et al*. introduisent une approche par Analyse en Composantes principales (ACP) pour calculer une nouvelle base de dimension réduite sur laquelle représenter le visage. Cette approche a ensuite été largement reprise dans la littérature [7, 20, 8]. La méthode consiste à calculer les vecteurs propres de la matrice de covariances des écarts à la moyenne afin d'obtenir le modèle :

$$S = \bar{S} + \sum_{i=1}^{m} \alpha_i s_i, \tag{1}$$

avec s_i les vecteurs propres et α_i les coefficients associés. On obtient ainsi une représentation dans une base de plus faible dimension en sélectionnant les m premiers vecteurs propres qui sont ceux qui expliquent le plus de variance.
Ainsi, cette ACP permet de passer d'une morphologie s'appuyant sur une représentation en dizaines de milliers de dimensions à une représentation en centaines de dimensions, plus précisément 200 pour [6], 300 pour [20] et 526 pour [8]). Cette méthode modélise très bien les variations générales présente dans la base des visages mais les résultats obtenus dans les publications citées indiquent qu'il est difficile de modéliser avec précision certains détails. Il est également possible d'obtenir des modèles séparés pour modéliser les expressions et/ou la texture du visage.

Processus Gaussien. Il est également possible de modéliser le visage à partir d'un processus gaussien [15] en s'appuyant sur ces deux éléments :

- ν, fonction de $\mathbb{R}^3$ dans $\mathbb{R}^3$ modélise à chaque point de l'espace la déformation moyenne (elle est définie comme la fonction nulle puisqu'on modélise la déformation à partir d'un visage moyen).
- K, fonction de $\mathbb{R}^{3*3}$ dans $\mathbb{R}^{3*3}$ modélise pour chaque paire de points dans l'espace, la covariance entre chaque couple de coordonnées.

Le processus gaussien est utilisé d'une manière analogue à l'ACP pour modéliser les variations au visage moyen à partir d'un jeu de données. De plus, Gerig *et al*. [15] combinent plusieurs fonctions K pour apporter des connaissances a priori autres que la variation à la moyenne dans un jeu de données. Ils implémentent entre autres des fonctions symétriques et B-spline en s'appuyant sur plusieurs échelles et localités différentes pour donner plus d'expressivité à leur modèle paramétrique.

2.2 Réseaux de neurones

Il existe une famille d'approches qui s'appuient sur les espaces latents extraits des réseaux de neurones pour encoder la géométrie du visage. Plus précisément en utilisant implicitement un espace latent de plus faible dimension modélisant le visage, comme par exemple 256 pour [37] et 85 pour [1]. Dans ce type d'architecture le réseau de neurones est divisé en :

- Un encodeur qui transforme les coordonnées 3D d'un visage en un vecteur de plus petite dimension. Abrevaya *et al*. utilisent un champ de hauteur [1] pour ensuite le traiter par convolutions 2D. Zhou *et al*. traitent directement les coordonnées 3D ainsi que les couleurs associées par des convolutions adaptées à un maillage.
- Un décodeur responsable de reconstruire l'entrée du réseau à partir de ce vecteur.

En faisant un tel apprentissage sur une base suffisamment grande de modèles 3D numérisés (21k [37], 5.2k [1]), le réseau apprend un espace latent représentant l'ensemble des visages. Zhou *et al*. [37] effectuent en parallèle un second apprentissage auto-supervisé en reconstruisant le visage à partir d'une image, *cf*. § 3.2, ce qui leur permet d'utiliser une grande base (260k images).
Les opérations d'un réseau étant non-linéaires, l'espace obtenu l'est aussi, contrairement aux modèles obtenus par ACP par exemple. On peut ensuite utiliser le décodeur pour passer d'un vecteur de l'espace latent à la représentation 3D. Ce second réseau peut ensuite être utilisé dans le même contexte que les modèles explicites pour effectuer la transformation entre l'espace de petite dimension et la représentation 3D.
D'autres approches conditionnent un réseau en entrée par un vecteur analogue aux paramètres d'un 3DMM [19, 23]. Les vecteurs sont optimisés avec les poids du réseau lors de l'apprentissage. Ces réseaux font partie de la catégorie hybride que nous présentons au § 3.2, ils allient apprentissage et optimisation à l'inférence.
Ces modèles paramétriques peuvent être limités par le manque de variabilité des données utilisées pour les construire. De plus, comme on réduit le nombre de dimensions de l'espace de recherche, modéliser certains détails n'est pas possible, même avec une très grande base. Il est nécessaire d'affiner ce modèle si l'on veut atteindre une plus grande précision.
La structure du visage permet de prendre en compte un a priori fort sur l'objet à reconstruire. Des modèles explicites ou implicites existent pour le prendre en compte mais ces approches peuvent ne pas être suffisantes pour atteindre la précision souhaitée. Les modèles sont pour la plupart suffisamment génériques pour être utilisés dans les mêmes méthodes de reconstruction, que nous étudions dans la section suivante.

3 Reconstruction du visage

À présent, nous abordons les trois grandes familles de reconstruction du visage que nous souhaitons distinguer : les approches classiques par optimisation, les méthodes récentes par réseaux de neurones et enfin celles qui permettent de combiner les deux types de techniques. Nous

1. Un champ de hauteur est une image encodant des déplacements à partir d'une surface référence

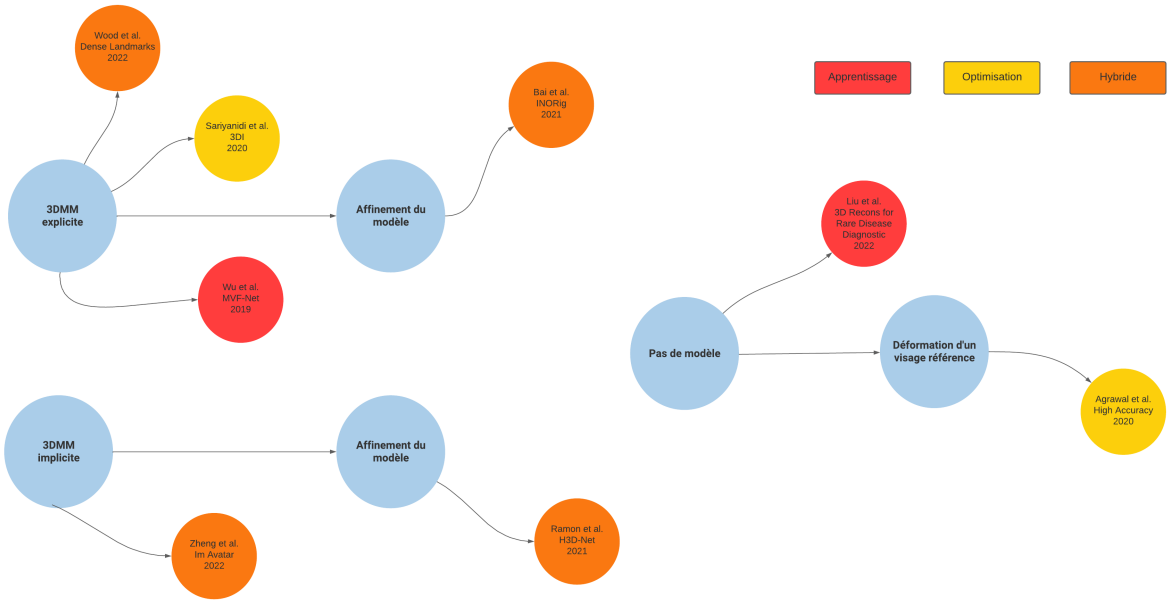

FIGURE 2 – Schéma récapitulatif des trois familles de méthodes de reconstruction multi-vues abordées dans cet article (par apprentissage, par optimisation ou hybride qui combine les deux types d'approches). En bleu, nous indiquons si la méthode s'appuie sur un modèle ou non et l'utilisation qui peut en être faite (explicite/implicite/affinement).

présentons un schéma récapitulatif des méthodes multi-vues abordées en figure 2

3.1 Méthodes par optimisation

Optimisation par rendu. Avec la publication du premier 3DMM, Vetter *et al*. [6] introduisent une méthode pour reconstruire le visage à partir d'une seule image. Elle consiste à minimiser la distance entre l'image source et une image rendue étant donnée les paramètres du 3DMM, de caméra et d'illumination. Les auteurs utilisent une descente de gradient stochastique. Dès lors, de nombreuses méthodes optimisent les paramètres de ce type de modèle. Sariyanidi *et al*. [24] introduisent une méthode d'optimisation sous contraintes appelée 3DI (*3D estimation via Inequality constraints*). D'une part pour rester dans l'espace des paramètres du 3DMM probables, d'autre part pour que les projections de certains sommets restent proches de points détectés sur les images. Ils adoptent une nouvelle manière de comparer les images sources et rendues : la corrélation de gradients [27]. Ils ne pénalisent pas une différence L_1 ou L_2 entre les images mais utilisent une fonction robuste : la similarité cosinus entre les gradients normalisés. Cela permet également de réduire l'impact des zones peu texturées qui contiennent une information photométrique faible.

Stéréovision à multiples vues. Agrawal *et al*. [2] introduisent une chaîne algorithmique afin de reconstruire le visage à partir d'une vidéo acquise sur smartphone : ils utilisent plusieurs algorithmes de l'état de l'art pour trouver les paramètres caméras avant de reconstruire la 3D. Plus précisément, cette approche fait intervenir des techniques d'ajustement de faisceaux, *Bundle adjustment*, en séparant l'ajustement géométrique [16] de l'ajustement photométrique [11]. Puis, elle s'appuie sur 68 points clés ainsi que la technique *PatchMatch* [12] pour prédire une carte de profondeur par vue et en déduire un nuage de points 3D. Pour terminer, les contraintes liées à l'a priori du visage sont exploitées (disposition des yeux, nez, oreilles), ce qui permet de modéliser une déformation non-rigide. Pour chaque point du visage référence que l'on déforme, le point cible est la médiane des points suffisamment proches de la normale, cette modélisation permet ainsi d'être plus robuste au bruit que d'autres méthodes comme la reconstruction de surface utilisant Poisson [17].

3.2 Méthodes par apprentissage

Rendu différentiable. Le cas d'application le plus direct est la structure encodeur (régression des paramètres d'un 3DMM et de rendu) décodeur (rendu différentiable) pour faire de l'apprentissage auto-supervisé en comparant les images d'entrées avec la sortie du rendu [26, 32]. Cela permet d'utiliser (surtout dans le cas d'une seule image) de grands jeux de données contenant des images plutôt que d'avoir à réunir une vérité terrain 3D pour chaque visage. Dans leur méthode MVF-Net, Wu *et al*. [32] généralisent cette approche au cas à 3 vues : une de face, les autres légèrement à droite et à gauche. Les auteurs pénalisent la consistance photométrique entre les vues en comparant la reprojection de la texture échantillonnée de la vue centrale dans les zones visibles des vues de côté :

$$I_{A\to B}[\mathbf{u}] = I_A[Pr(Pr^{-1}(\mathbf{u}, \mathcal{X}, \mathcal{P}_A), \mathcal{P}_B)] \qquad (2)$$

pour un pixel **u** donné, la reprojection de la vue A vers la vue B est la projection suivant les paramètres caméras $\mathcal{P}_\mathcal{B}$, de la retroprojection du pixel **u** suivant les paramètres caméras $\mathcal{P}_\mathcal{A}$ sur le modèle issu des paramètres 3DMM $\mathcal{X}$. Il est constaté que cela conduit à un mauvais alignement, notamment à cause des zones peu texturées du visage : les intensités des pixels sont proches les unes des autres. Une pénalisation est ajoutée sur le flot optique entre $I_{A\to B}$ et I_B.

Stéréovision à multiples vues. De nombreux réseaux [29] tentent de retrouver la géométrie 3D d'objets à partir de plusieurs images via de la stéréovision. Liu *et al*. [21] appliquent un de ces réseaux à la reconstruction du visage. En faisant l'hypothèse d'acquisition d'images en tournant autour d'un axe vertical avec un smartphone, les paramètres caméras sont estimés grâce à la technique de [25]. Ensuite, le nuage de points 3D est estimé par *PatchmatchNet* [28] et enfin la surface du visage est estimée par une reconstruction de Poisson [17].
Le réseau *PatchmatchNet* prend en entrée un ensemble d'images de l'objet à reconstruire et prédit une carte de profondeur d'une image référence en utilisant les autres images appelées sources. Le but de ce réseau est d'effectuer plusieurs prédictions de profondeur par pixel puis de calculer l'espérance de la profondeur étant donné un calcul de score de similarité entre les différentes vues. Les images sont d'abord traitées par un réseau par convolutions et les similarités sont calculées sur les sorties de ce réseau, appelées *feature maps*. Ainsi, le réseau n'apprend pas à déterminer comment faire correspondre un pixel à un autre mais il considère explicitement le calcul du correspondant via les paramètres caméras estimés.
Ce réseau est entraîné de manière supervisée via la vérité terrain des cartes de profondeur. Le jeu de données utilisé ne comporte pas de visage, on perd l'assurance de retrouver un visage vraisemblable qu'on a en utilisant un modèle.

3.3 Méthodes hybrides

Optimisation intégrée au réseau. *Riggable 3D Face Reconstruction via In-Network Optimization*, INORig [5, 4] incorpore une étape d'optimisation dans un réseau de neurones. Cette méthode alterne itérativement prédiction de *features map* pour chaque image, transformation d'un espace paramétrique vers un maillage du visage puis optimisation des paramètres (position caméra, illumination, expression, identité). L'objectif de l'optimisation est la consistance entre les multiples vues (dans les *features maps*) aux projections des sommets du maillage retrouvé.

Représentation implicite. Au lieu de traiter les images pour prédire des paramètres et/ou des déplacements sur une surface, des réseaux de neurones sont construits pour modéliser implicitement la surface d'un objet [22, 35]. La principale idée est que l'entrée du réseau est une position dans l'espace 3D le long d'un rayon issu d'un pixel. La sortie est directement une couleur RGB, en utilisant le canal σ qui représente l'opacité, comme dans l'approche *Neural Radiance Fields of view*, NeRF [22], ou en séparant la géométrie et la texture dans *Implicit Differentiable Renderer*, IDR [35]. Le principe de ces réseaux est de faire un surapprentissage de la scène lors de l'inférence, ce qui revient à optimiser les poids du réseaux pour minimiser la distance entre les intensités des pixels et les couleurs prédites. Des méthodes pour incorporer un a priori sur le visage dans ce type de réseau sont développées [23, 36]. Ces réseaux sont conditionnés par un vecteur z, responsable de modifier la géométrie du visage représenté. Une première phase d'apprentissage sur une base de visages 3D optimise les poids du réseau ainsi que les vecteurs donnés en entrée : le réseau apprend l'a priori. Lors de l'inférence, *High-fidelity 3D head reconstruction*, H3D-Net [23], optimise la géométrie seulement avec le vecteur z avant de relâcher les poids du réseau pour affiner le résultat.

Prédiction de points clés puis optimisation. Un champ de recherche s'intéresse au problème du *face alignment* : prédiction de points clés sur le visage par réseau de neurones (généralement 68, comme précédemment). Il est possible de faire correspondre à ces points les sommets d'un 3DMM puis d'optimiser les paramètres du modèle ainsi que de la caméra pour minimiser la distance entre la projection des sommets et la position des points. Ces 68 points n'apportent pas assez d'information pour reconstruire avec précision la géométrie du visage. Wood *et al*. effectuent une prédiction de points clés denses (703 points) [31], ils utilisent un jeu de données synthétique se rapprochant le plus possible du réel [30] afin d'avoir une vérité terrain d'images ainsi que les points clés associés. Leur réseau prédit la position des points mais aussi une mesure de confiance associée. Les distances entre ces positions et la projection des sommets du modèle pondérées par les mesures de confiance sont minimisées pour retrouver la géométrie du visage.

4 Évaluation et comparaison

4.1 Jeux de données

Le premier modèle [6] est construit à partir de 200 visages. Des modèles basés sur plus de données sont ensuite élaborés : 10000 visages 3D sont acquis pour construire LSFM [7], 3800 pour FLAME [20], 2000 pour HIFI3D++ [8]. Les jeux de données sont équilibrés entre hommes et femmes. Cependant, les jeux de données ne sont pas tous représentatifs de tous les âges et ethnies. Booth *et al*. [7] présentent une distribution sur tous les âges, même si les personnes entre 5 et 30 ans sont surreprésentés. Chai *et al*. [8] équilibrent la proportion de personnes originaires d'Europe et d'Amérique du nord avec celles des personnes originaires d'Asie ; les autres ethnies restent sous-représentées.
Afin d'évaluer une méthode de reconstruction, il est nécessaire de réunir des acquisition 3D de visages et pour chacun, un ensemble de prises de vues de la personne

Nom	individus	vues	illumination	acquisition
H3DS [23]	23	60	non constante	caméra
3DFAW-video [18]	26	vidéo 6 s	constante (smartphone) et non constante (caméra)	smartphone et caméra
Facescape [34]	359	60	constante	caméra
interne	15	20-50	non constante	smartphone

TABLE 1 – Jeux de données sélectionnés

Nom	3DFAW-vidéo					Interne				
	moyenne	médiane	écart-type	min	max	moyenne	médiane	écart-type	min	max
MVFNet [32]	1.53	1.37	0.48	0.91	2.61	**1.84**	1.62	1.05	**1.05**	5.61
INORig [4]	1.53	1.29	0.66	0.81	3.77	1.87	1.63	1.00	1.13	5.35
PatchmatchNet [21]	**1.21**	**1.21**	**0.32**	**0.68**	**1.87**	**1.84**	**1.55**	**0.86**	1.15	**4.67**

TABLE 2 – Erreurs points à surface (vérité terrain vers prédiction) des méthodes MVFNet, INORig, PatchmatchNet sur les deux jeux de données sélectionnés, unité en millimètres.

correspondante. L'acquisition 3D peut-être la même pour chaque approche. Cependant, chacune a des besoins différents pour les images. Pour comparer toutes les méthodes ensemble, il faut utiliser un jeu de données permettant d'évaluer la plus restrictive des méthodes. Ainsi, on sélectionne les jeux de données contenant des dizaines de vues pour chaque personne (ce qui est nécessaire pour la méthode générique [21]).
Il convient également de prendre en compte l'illumination de la scène : certaines méthodes supposent que les images sont prises simultanément sous différents angles de vue. En cas de mouvement du visage à la place de la caméra, cela peut entraîner des changements d'illumination sur le visage, contrairement à un mouvement de la caméra tout en gardant le visage fixe.Les jeux de données correspondant à nos critères sont répertoriés dans le tableau 1. Nous rappelons que dans l'introduction nous avons fait l'hypothèse d'une acquisition de multiples images avec un smartphone. Afin d'évaluer les méthodes sur les deux types d'acquisitions envisagés (illumination constante ou non), nous sélectionnons notre jeu de données interne (acquisition effectuée suivant le protocole présenté en figure 1) ainsi que la partie smartphone de 3DFAW-vidéo [18] qui est disponible publiquement.

4.2 Évaluation

Nous souhaitons dans ce papier déterminer quel est le type d'approche le plus intéressant pour notre cas d'utilisation et nos objectifs. Dans ce cadre une évaluation des méthodes est nécessaire puisqu'il est difficile de comparer les résultats de chaque publication entre elles : elles ne sont pas testées sur les mêmes jeux de données et/ou selon la même méthodologie. Il serait idéal de comparer les meilleures méthodes de chaque paradigme (optimisation, réseau, hybride avec ou sans a priori), cependant elles ne sont pas toutes disponibles publiquement. Nous évaluons donc, les méthodes multi-vues de chaque paradigme dont le code est disponible publiquement :

- PatchMatchNet [28, 21], méthode par apprentissage sans a priori;
- MVF-Net [32], méthode par apprentissage avec a priori;
- INORig [4], méthode hybride avec a priori.

Nous avons identifié les méthodes suivantes qui ne sont pas disponibles mais présentent des résultats intéressants :

- *High Accuracy Face Geometry Capture using a Smartphone Video* [2], méthode générique par optimisation qui déforme ensuite un visage référence sur un nuage de points. Les auteurs présentent une bonne performance sur leur propre jeu de données avec une médiane d'erreur de $0.95\,\mathrm{mm}$.
- H3D-Net [23], méthode hybride représentant implicitement le visage. Il serait pertinent d'évaluer leur méthode sur un plus grands nombre d'images par personnes. Ils se sont limités à 3 sur 3DFAW-vidéo [18].
- 3DI [24], méthode par optimisation avec a priori, dont l'implémentation est disponible mais ne permet pas la reconstruction à partir de plusieurs images. Il serait intéressant d'évaluer leur méthode sur un plus grand nombre d'images.

Méthodologie. L'alignement entre le visage 3D prédit et la vérité terrain est une étape importante de l'évaluation. Elle consiste généralement [5, 23] en un alignement entre des points clés 3D annotés puis une étape d'ICP. Chai *et al*. [8] remettent en cause cette méthodologie. Ils identifient deux défauts :

- Une erreur locale peut avoir un impact sur l'alignement global. Par exemple un nez trop long peut faire avancer tout le visage, ce qui augmente les erreurs mesurées sur les autres zones (yeux, joues, ...) même si elles sont bien reconstruites.
- On aligne puis mesure généralement l'erreur dans un sens uniquement. Cette asymétrie peut amener des erreurs de correspondances lors du calcul des erreurs.

Les auteurs proposent donc une méthodologie bidirectionnelle et prenant en compte 4 régions du visage : fronts et

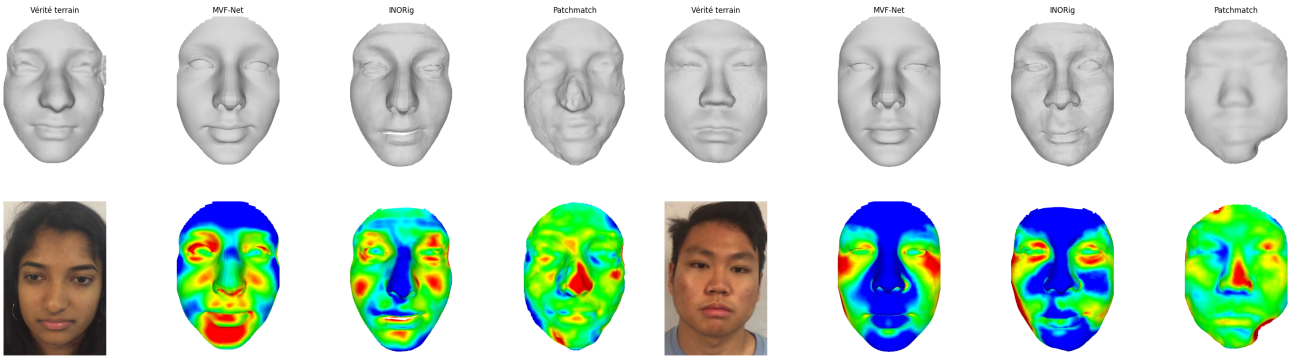

FIGURE 3 – Résultats sur deux exemples de 3DFAW-vidéo, l'erreur est signée, l'échelle va de -3mm (rouge) à 3mm (bleu)

yeux, joues et menton, nez, bouche. L'alignement pour une région R est fait grâce à un masque 3D des points appartenant à R, seuls ces points sont pris en compte. Une étape d'alignement non rigide déforme ensuite la vérité terrain sur la prédiction, les correspondances points à surface sont déterminées entre le visage prédit et la vérité terrain déformée. Ces correspondances sont utilisées pour calculer les distances entre le visage prédit et la vérité terrain non déformée. Les auteurs ont montré quantitativement l'apport de cette méthode d'évaluation par rapport aux méthodes traditionnelles. C'est la raison pour laquelle nous souhaitions utiliser cette nouvelle méthodologie, malheureusement elle est difficilement adaptable à des méthodes de reconstruction qui n'ont pas été pensées pour (par exemple elle nécessite une topologie constante pour chaque prédiction ce qui n'est pas le cas de la méthode [21]).
De ce fait, nous utilisons une méthode classique d'alignement rigide avec annotations de points clés puis ICP.

4.3 Protocole d'évaluation

Nous utilisons une méthode d'alignement similaire à celle de H3D-Net [23]. Tout d'abord, les prédictions sont alignées de manière rigide sur la vérité terrain en utilisant 7 points clés (coins des yeux, de la bouche et base du nez), puis par ICP, *Iterative Closest Point*. Ensuite, nous découpons les maillages en gardant uniquement les points situés à l'intérieur d'une sphère de rayon 9 centimètres centrée sur le bout du nez de la vérité terrain. Cette étape permet de ne conserver que la zone principale du visage et ainsi d'éliminer les erreurs dues à des zones non reconstruites par les méthodes (par exemple INORig ne reconstruit pas les oreilles et le haut du front). Finalement, une dernière étape d'ICP entre les maillages découpés permet d'améliorer l'alignement. Nous mesurons ensuite, pour chaque visage, l'erreur moyenne entre les points de surface de la vérité terrain et ceux de la prédiction.

Les résultats sont répertoriés dans le tableau 2.
Nous sélectionnons 3 images (une de face, les deux autres légèrement de profil) par personne pour MVF-Net [32]. Nous sélectionnons entre 15 et 20 images pour INORig [4] et PatchmatchNet [21]. Nous utilisons les paramètres par défaut de chaque méthode et effectuons la reconstruction de surface en utilisant l'implémentation de Meshlab [9]. Pour échantillonner chaque nuage de points obtenu avec un objectif de 100k points, nous calculons les normales en utilisant 100 voisins et 8 étapes de lissage, puis nous procédons à la reconstruction de surface en utilisant la méthode de Poisson avec un poids d'interpolation de 0.

4.4 Analyse des résultats obtenus

Les résultats de notre évaluation sont présentés dans le tableau 2. Nous constatons que les performances sont proches. Les différences sont de l'ordre de quelques dixièmes de millimètres et aucune des méthodes évaluées ne peut garantir une précision sous-millimétrique. La méthode présentant les meilleurs résultats est PatchmatchNet [21], c'est aussi celle n'utilisant pas de modèle et donc certaines parties du visages peuvent être très mal reconstruites (voir figure 3). Ces résultats sont encourageants et il serait intéressant d'utiliser la méthode de déformation d'un visage référence de [2] qui est plus robuste aux bruits et aux zones non reconstruites.
Les performances moins satisfaisantes de MVF-Net [32] et INORig [4] pourraient aussi s'expliquer par un domaine différent entre les jeux de données utilisés pour leur apprentissage et les images acquises par *smartphone*. On observe un plus grand écart ($0.6\,\mathrm{mm}$ en moyenne) entre les deux jeux de données. Les meilleurs résultats sur 3DFAW-vidéo peuvent indiquer le gain de performance à utiliser un protocole où l'illumination du visage reste constante.

5 Conclusions et perspectives

Dans cet article, nous nous plaçons dans un contexte d'acquisition de visage par la personne utilisatrice, via son *smartphone*. Nous proposons un état de l'art des méthodes de reconstruction du visage s'appuyant sur plusieurs images non calibrées. Ces méthodes prennent en compte ou non un a priori sur le visage. Celui-ci prend la forme d'un modèle paramétrique, qui permet de réduire la dimension de l'espace de recherche tout en garantissant de reconstruire un visage vraisemblable. Ces modèles peuvent ne pas être suffisants pour atteindre une précision souhai-

tée et certaines méthodes, après une première reconstruction, affinent le visage obtenu. Les méthodes de reconstruction peuvent être classées en 3 catégories : les approches par optimisation, les techniques s'appuyant sur des réseaux de neurones et les méthodes hybrides. Nous évaluons 3 méthodes disponibles publiquement et identifions PatchmatchNet [21] comme celle donnant les meilleurs résultats, bien qu'ils restent proches des autres approches. Ces performances sont à mettre en perspective des performances d'autres méthodes dont le code n'est pas disponible, comme par exemple Agrawal *et al*. [2] qui obtiennent (sur un jeu de données différent) une médiane sous le millimètre. Enfin, nous nous sommes limités aux méthodes directement appliquées au visage. La reconstruction 3D est un champ de recherche très actif et il est possible que de nouvelles méthodes générales puissent être appliquées avec succès à la reconstruction du visage. De plus, il serait intéressant d'explorer les manières d'incorporer un a priori sur le visage dans ces méthodes générales, comme l'ont fait Ramon *et al*. [23].

Références

[1] V.-F. Abrevaya, S. Wuhrer, and E. Boyer. Multilinear autoencoder for 3D face model learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2018.

[2] S. Agrawal, A. Pahuja, and S. Lucey. High accuracy face geometry capture using a smartphone video. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[3] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[4] Z. Bai, Z. Cui, X. Liu, and P. Tan. Riggable 3d face reconstruction via in-network optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[5] Z. Bai, Z. Cui, J.-A. Rahim, X. Liu, and P. Tan. Deep facial non-rigid multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[6] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. *ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive techniques*, 2002.

[7] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Z. Chai, H. Zhang, J. Ren, D. Kang, Z. Xu, X. Zhe, C. Yuan, and L. Bao. REALY: Rethinking the Evaluation of 3D Face Reconstruction. In *European Conference on Computer Vision*, 2022.

[9] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*, 2008.

[10] B. Egger, W.-A.-P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter. 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.*, 39, 2020.

[11] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, 2017.

[12] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *IEEE/CVF International Conference on Computer Vision*, 2015.

[13] P. Garrido, L. Valgaert, C. Wu, and C. Theobalt. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. *ACM Transactions on Graphics*, 2013.

[14] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Transactions on Graphics*, 35, 2016.

[15] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *IEEE International Conference on Automatic Face & Gesture Recognition* , 2018.

[16] C. Ham, M.-F. Chang, S. Lucey, and S. Singh. Monocular depth from small motion video accelerated. In *International Conference on 3D Vision (3DV)*, 2017.

[17] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32, 2013.

[18] R. Krishnan Pillai, L. Attila Jeni, H. Yang, Z. Zhang, L. Yin, and J.-F. Cohn. The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *IEEE International Conference on Computer Vision Workshops*, 2019.

[19] M. Li, H. Haibin, Y. Zheng, M. Li, N. Sang, and C. Ma. Implicit Neural Deformation for Multi-View Face Reconstruction, 12 2021.

[20] T. Li, T. Bolkart, M. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 2017.

[21] Y. Liu, L. Li, S. An, P. Helmholz, R. Palmer, and G. Baynam. 3D Face Reconstruction with Mobile Phone Cameras for Rare Disease Diagnosis. In *AI 2022: Advances in Artificial Intelligence*, 2022.

[22] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng. *NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*. 2020.

[23] E. Ramon, G. Triginer, J. Escur, A. Pumarola, J. Garcia, X. Giro-i Nieto, and F. Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[24] E. Sariyanidi, C. J. Zampella, R. T. Schultz, and B. Tunc. Inequality-Constrained and Robust 3D Face Model Fitting. In *European Conference on Computer Vision*, 2020.

[25] J.-L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[26] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision Workshops*, 2017.

[27] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Robust and efficient parametric face alignment. In *IEEE/CVF International Conference on Computer Vision*, 2011.

[28] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[29] X. Wang, C. Wang, B. Liu, X. Zhou, L. Zhang, J. Zheng, and X. Bai. Multi-view stereo in the deep learning era: A comprehensive review. *Displays*, 70, 2021.

[30] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T.-J. Cashman, and J. Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[31] E. Wood, T. Baltrušaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljević, D. Wilde, S. Garbin, T. Sharp, I. Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, 2022.

[32] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K.-N. Ngan, and W. Liu. Mvf-net: Multi-view 3d face morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[33] Y. Xiao, H. Zhu, H. Yang, Z. Diao, X. Lu, and X. Cao. Detailed Facial Geometry Recovery from Multi-View Images by Learning an Implicit Function. *AAAI Conference on Artificial Intelligence*, 36, 2022.

[34] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[35] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020.

[36] Y. Zheng, V.-F. Abrevaya, M.-C. Bühler, X. Chen, M.-J. Black, and O. Hilliges. Im avatar: Implicit morphable head avatars from videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[37] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[38] W. Zielonka, T. Bolkart, and J. Thies. Towards Metrical Reconstruction of Human Faces. In *European Conference on Computer Vision*, 2022.