Reconstruction d'un clone de visage 3D à partir de patchs de forme

Jérôme MANCEAU, Catherine SOLADIÉ, Renaud SÉGUIER

CENTRALESUPELEC/IETR UMR6164, équipe Fast (Facial Analysis, Synthesis and Tracking)
Avenue de la Boulaie, Cesson-Sévigné, France
jerome.manceau@supelec.fr, catherine.soladie@centralesupelec.fr
renaud.seguier@centralesupelec.fr

Résumé – Un clone de visage 3D sémantique peut être utilisé comme prétraitement dans des applications comme l'analyse des émotions. Toutefois, ces clones doivent avoir la forme du visage bien modélisée tout en gardant la spécificité des individus. Dans notre technique, nous utilisons un capteur RVB-Z pour obtenir la spécificité des individus et un modèle déformable de visage 3D pour marquer la forme du visage. Nous gardons les parties appropriées de données de profondeur appelés Patch. Cette sélection est effectuée en utilisant une erreur de distance et la direction des vecteurs normaux de chaque point. Selon l'emplacement, nous fusionnons soit les données des capteurs soit les données obtenues avec le modèle déformable. Nous comparons notre méthode avec un processus de fitting classique. Les tests qualitatifs montrent que nos résultats sont plus précis qu'une méthode de fitting classique et les tests quantitatifs montrent que notre clone possède à la fois les spécificités de la personne et la forme du visage bien modélisée.

Abstract – Semantic 3D face clone can be used as pretreatment in applications such as emotion analysis. However, such clones should have well-modeled facial shape while keeping specificity of individuals. In our technique, we use a RGB-D sensor to get the specificity of individuals and a 3D Morphable Face Model to mark facial shape. We keep the suitable parts of depth data called Patch. This selection is performed using an error of distance and the direction of the normal vectors. Depending on the location, we merge either sensor data or 3D Morphable Face Model data. We compare our method with a classical fitting process. The qualitative tests show that our results are more accurate than a conventional fitting method and quantitative tests show that our clone has both the specificities of the person and the shape of the face well-modeled.

1 Introduction

Dans le domaine de la vision par ordinateur, les clones de visage sont souvent utilisés dans des applications tels que les jeux sérieux et les interactions hommes machines [1]. Ce type de fonctionnalité doit pouvoir être utilisée par n'importe quelle personne à son domicile et être entièrement automatique. Les caméras RVB-Z à faible résolution ont l'avantage d'avoir un faible coût. C'est pourquoi elles ont été récemment utilisées dans le domaine du clonage de visage. R.A. Newcombe et al. [2] présentent une reconstruction 3-D de scènes ou d'objets à l'aide d'une caméra RVB-Z basse résolution. Y. Cui et al. [3] et Q. Sun et al. [4] proposent des méthodes pour reconstruire un visage 3D à partir de données RVB-Z capturées avec une caméra Kinect. Ces méthodes ne fournissent pas de clones 3D sémantique. Elles ne peuvent donc pas être directement utilisés comme prétraitement dans des applications nécessitant la connaissance de la correspondance des points du maillage avec le visage. Pour palier à ce problème, de nombreuses méthodes utilisent un modèle déformable de visage. Ils permettent de reconstruire un visage 3D, d'éliminer le bruit et d'augmenter la résolution des données. M. Zollhöfer et al. [5] présentent un algorithme pour le clonage 3D à partir de données RVB-Z obtenues avec une caméra Kinect. C. Cao et al. [6] ont créé une base de données de visages 3D de 150 personnes. Ils ont utilisé l'algorithme Kinect Fusion [2] et un modèle déformable de visage pour reconstruire le visage 3D de chaque personne. M. Zollhöfer et al. [7] présentent une méthode itérative pour cloner le visage 3D d'une personne. Ces techniques dépendent fortement de la qualité des visages du modèle. En effet, les spécificités des individus ne peuvent être trouvées que si elles appartiennent à la base de données. Il est donc essentiel d'utiliser une base de données composée de visages diversifiés. Le maillage 3D sémantique résultant obtenu à partir de ces procédés peut être utilisé dans diverses applications comme un prétraitement pour la détection du regard. Par exemple, K.A. Funes Mora et J. Odobez [1] utilisent des clones 3D pour estimer la pose de la tête et la direction du regard d'une personne.

La première contribution de cet article est l'utilisation de patchs de profondeur (ensemble soigneusement choisi de points) pour reconstruire le visage 3D. Lorsque nous utilisons un modèle déformable de visage, certaines des spécificités morphologiques de l'individu que nous voulons cloner peuvent disparaître. En effet, l'ensemble d'apprentissage du modèle déformable ne contient pas toutes les formes et les détails possibles du visage inconnu. C'est pourquoi, nous identifions les parties (patchs) de chaque maillage 3D qui sont pertinentes. Nous utilisons à la fois une distance d'erreur et la direction des vecteurs normaux de chaque point du maillage pour calculer les patchs. La deuxième contribution est la façon dont nous fusionnons les

La recherche a été conduit avec le support de Miles (FUI projet) et de ARED (Région Bretagne).

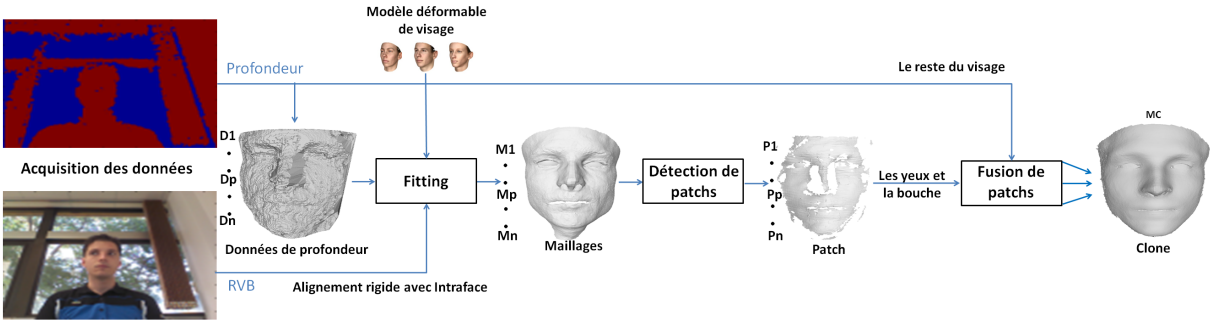

FIGURE 1 – Notre système de détection et de fusion des patchs de formes.

patchs en exploitant à la fois les données obtenues avec le modèle déformable de visage et celles obtenues avec le capteur de profondeur. Les données du modèle déformable sont utilisées aux niveaux des yeux et de la bouche. En effet, les données des capteurs ne sont pas précises et sont très bruitées sur ces parties du visage. Nous utilisons les données de la caméra RVB-Z sur le reste du visage. Cette méthode de fusion augmente le réalisme et la précision du clone 3D.

Cet article est organisé comme suit. La section 2 décrit notre algorithme. La partie 3 donne nos résultats. La section 4 conclut le papier.

2 Méthode de détection et de fusion de patchs

Pour chaque personne, nous capturons les données RVB-Z de **n** vues du visage. Ces données sont bruitées et de basse résolution (voir processus global dans la figure 1). Nous ajustons un modèle déformable de visage M(§2.1) sur chaque trame de profondeur D_p (p = 1 à n). Après cette étape de fitting, nous obtenons **n** maillages sémantique M_p correspondant aux **n** trames D_p. Ensuite nous détectons **n** patchs de profondeur P_n (§2.2) correspondant aux données de différents maillages M_p. Enfin, nous fusionnons (§2.3) les différents patchs que nous avons détectés pour générer un clone 3D complet M_C.

2.1 Fitting avec un modèle déformable de visage 3D

Dans l'étape de fitting, nous effectuons d'abord un prétraitement avec un filtre bilatéral [8] sur chaque trame D_p pour éliminer une partie du bruit. Puis, à chaque itération, nous calculons les matrices de rotation et de translation qui alignent chaque trame D_p avec le maillage M (transformation rigide). Nous utilisons l'algorithme Iterative Closest Points [9] pour trouver ces deux matrices. Pour réduire le nombre d'itérations de l'algorithme ICP, nous initialisons l'angle de rotation en utilisant la pose calculée par Intraface [10]. Il permet de trouver l'angle de rotation à partir des données de couleur 2D. Ensuite, nous déformons le maillage M, de sorte qu'il prenne la forme de la trame D_p (transformation non-rigide). Nous utilisons l'algorithme d'optimisation de Gauss-Newton pour trouver les paramètres du modèle déformable qui minimise l'erreur de distance entre la trame D_p et le maillage M afin de générer autant de modèle déformés M_p que de trames en entrée.

2.2 Détection des patchs

Nous voulons seulement garder les parties des maillages qui sont adéquates et précises. Par exemple, pour une trame de profondeur de profil droit nous voulons garder le patch du maillage qui correspond au bon profil (voir figure 2). Nous appelons "patch" tous les points isolés de chaque maillage que nous voulons garder. Nous utilisons deux critères pour la détection des patchs.

Distance d'erreur : L'utilisation d'une erreur de distance permet d'éliminer le bruit du capteur et les erreurs de fitting. Ce critère est basé sur la distance d'erreur point à point entre chaque trame D_p et le maillage M_p correspondant. Cette erreur de distance doit être inférieure à un certain seuil (1 mm). Sur la vue de face de la figure 2 (a4), il y a un trou au niveau du nez car le capteur ne donne pas cette information. Le maillage M_p, calculé par le fitting, ne possède pas un tel trou. Les données de maillage M_p ne reflètent donc pas l'identité du sujet pour cette partie du visage. L'erreur de distance supprime ces informations.

Vecteur normal : La caméra RVB-Z capture plus précisément les zones où l'axe optique est perpendiculaire à surface de l'objet. Pour cette raison, nous ne gardons que les points des patchs qui ont un vecteur normal parallèle à l'axe optique de la caméra. Dans la figure 2 (a5), nous constatons que ce critère de sélection élimine les points que la caméra ne saisit pas correctement. Par exemple, nous notons que pour une trame de profondeur de face, les points situés sur les côtés du nez sont mal capturées par la caméra et ne sont pas précis.

Dans notre processus, nous utilisons les deux critères détaillés ci-dessus. Pour qu'un point soit conservé, il doit avoir un vecteur normal parallèle à l'axe optique de la caméra et la distance entre le maillage M_p et la trame D_p doit être inférieure à un seuil. Nous tolérons un angle de plus ou moins 20 degrés pour les vecteurs normaux. Cela nous permet de détecter les zones qui ne sont pas capturées par la caméra et d'éliminer le bruit et les erreurs de fitting. La figure 2 (a6), 2 (b6) montre que seuls les points appropriés sont conservés.

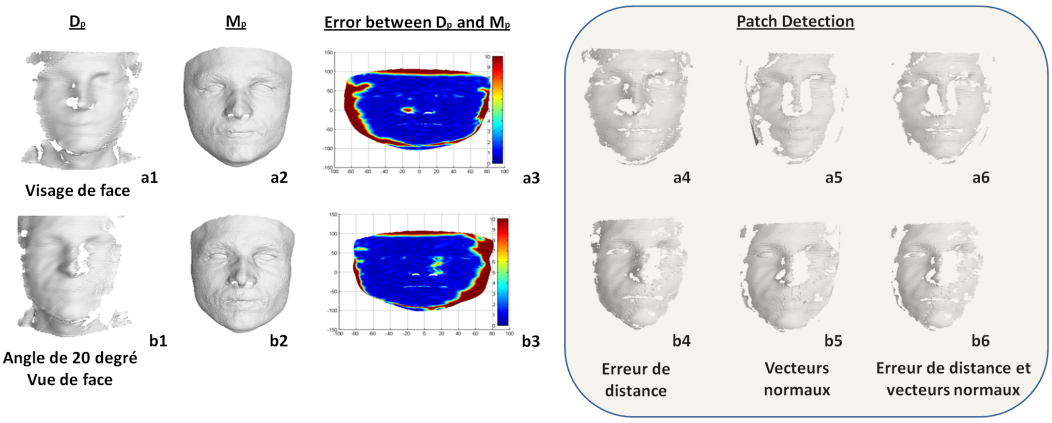

FIGURE 2 – Exemples de détection des patchs : $a1$: une trame de face. b_1 : une trame de profil droit (le visage est re-synthétisé de face). Nous utilisons un fitting pour calculer deux maillages sémantique (a_2 et b_2). L'erreur entre la trame et le maillage (a_3 et b_3) donne les patchs qui sont adéquats. Nous sélectionnons les parties du maillage (patch) où l'erreur est petite (a_4 et b_4) et où les vecteurs normaux sont pertinents (a_5 et b_5). Dans notre processus, nous utilisons ces deux critères (a_6 et b_6).

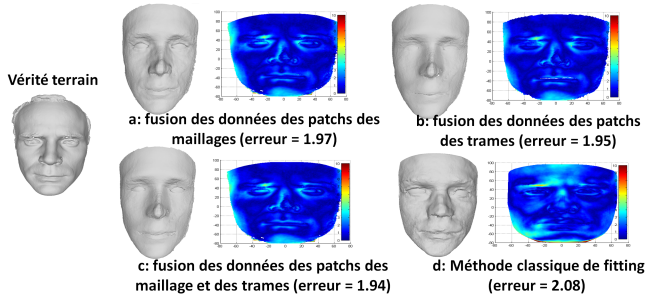

FIGURE 3 – Comparaison quantitative avec la réalité de terrain.

2.3 Fusion de patchs

Comme tous les maillages sont sémantique, nous connaissons la position exacte de chaque patch sur le visage (yeux, front ...). Nous reconstruisons le clone en fusionnant ces patchs. Cette fusion est effectuée sur différents types de données.

Données des maillages (M_1 à M_n, cf.figure 1) : nous utilisons les données des patchs des différents maillages M_p. Ce type de données fournit un clone réaliste de la personne. Mais certaines spécificités peuvent ne pas apparaître dans le clone, si elles n'ont pas été apprises lors de la construction du modèle déformable de visage 3D (cf.figure 3 (a)).

Données des trames (D_1 à D_n, cf.figure 1) : nous utilisons les données des trames D_p pour reconstruire le clone sémantique M_C. Pour chaque point des patchs P_p, nous connaissons le point de la trame correspondant D_p le plus proche. Nous remplaçons les valeurs des points des patchs par les valeurs des points des trames les plus proches. Avec ce type de données, nous obtenons un clone sémantique de haute résolution qui contient de nombreuses spécificités de l'individu. Mais les données acquises par le capteur RVB-Z ne sont pas précises et sont très bruitées au niveau des yeux et de la bouche (cf figure 3 (b)).

Dans notre processus, nous utilisons les deux types de données décrites ci-dessus (cf figure 3 (c)). Plus précisément, nous utilisons les points des maillages M_p pour les yeux et la bouche et les points des trames D_p pour le reste du visage. L'utilisation des données des trames fournit un clone avec de nombreuses spécificités de la personne et l'utilisation des données des maillages donne un clone réaliste au niveau des yeux et de la bouche. Pour chaque point du clone, il peut y avoir plusieurs patchs qui se chevauchent (comme le front du visage dans la figure 2). C'est pourquoi, nous faisons une fusion de chaque point de ces patchs. Nous effectuons la moyenne pondérée robuste sur les points qui se chevauchent. Ainsi, nous ne prenons pas en compte les valeurs aberrantes dans le calcul de la moyenne. La pondération est calculée à partir de la distance d'erreur point à point calculée au paragraphe 2.2. Les paires de points avec une courte distance sont les plus importants. C'est pourquoi, la pondération est inversement proportionnelle à la distance entre les points appariés. Nous éliminons les points qui sont loin de la valeur médiane avec un seuil (2mm).

3 Résultats expérimentaux

Nous comparons qualitativement notre méthode au processus classique de fitting [5, 6]. Les méthodes de fitting classiques réalisent d'abord une fusion des trames et ensuite un fitting. Ici, la méthode de fitting auquel nous nous comparons utilise l'ICP décrite au paragraphe 2.1, la fusion des trames étant réalisée par Kinect Fusion. Nous comparons également quantitativement notre méthode avec une réalité de terrain acquise avec un capteur haute résolution.

Protocole expérimental : Nous utilisons une caméra Kinect version 1 qui est équipée d'un capteur de couleur et d'un capteur de profondeur. Pour le fitting, nous utilisons le modèle déformable de visage Basel Face Model (BFM) [11]. Le sujet effectue un mouvement de rotation de la tête. Il doit faire une expression neutre lors de l'acquisition des données. Notre base de données de test se compose de six sujets (voir figure 4).

Résultats qualitatifs : La figure 3 montre que la méthode

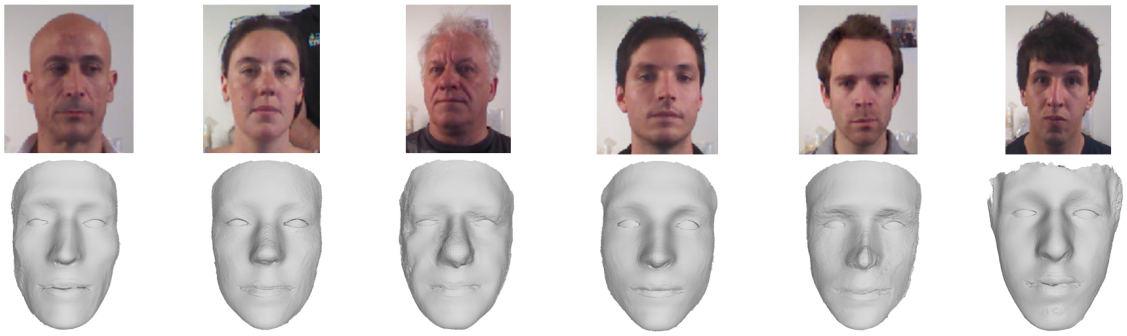

FIGURE 4 – Résultats de notre méthode sur 6 personnes.

de fitting classique fournit un clone sémantique où les traits du visage sont prononcés. Mais le clone obtenu avec la méthode de fitting classique possède moins de spécificités de l'individu. En effet, les modèles déformable de visage ne contiennent pas toutes les formes possibles et détails du visage du sujet et leurs bases de données d'apprentissage sont limitées. Dans notre procédé les traits du visage sont bien modélisés tout en ayant plus de spécificités de la personne que le clone 3D créé par la méthode de fitting classique. Par exemple, nous pouvons voir que les yeux de notre clone 3D sont plus réalistes.

Résultats quantitatifs : La comparaison quantitative illustre la précision de nos résultats (cf.figure 3). Nous apparions chaque point du clone que nous voulons comparer avec le point le plus proche de la vérité terrain, et nous calculons l'erreur de distance globale entre les paires de points. Nous n'avons pas calculé l'erreur au niveau des yeux parce que la vérité terrain n'est pas correcte sur cette zone du visage due au type de capteur. La figure 3 montre que l'erreur est plus petite avec notre méthode en particulier au niveau du front et de la zone du menton. Nous observons que l'erreur globale est plus petite avec notre méthode (Erreur : 1,94) qu'avec le processus classique de fitting (Erreur = 2,08) [6]. A noter qu'il n'y a aucune différence quantitative significative entre les clones obtenus avec les 3 types de fusion (1.94, 1.95, 1.97). Néanmoins, les résultats qualitatifs sont meilleurs avec les données des trames et des maillages car ils conservent les zones les plus adéquates des deux types de données.

4 CONCLUSION

Nous avons proposé un système qui fournit un clone 3D sémantique à partir d'un capteur à faible coût. Notre méthode permet de retrouver plus facilement les spécificités du visage d'un individu et produit un clone plus réaliste. Dans nos travaux futurs, nous voulons travailler sur la reconstruction de la texture 3D.

Références

[1] K A. Funes Mora and J Odobez. *Gaze Estimation From Multimodal Kinect Data*. IEEE Conference in Computer Vision and Pattern Recognition, Workshop on Gesture Recognition, 2012.

[2] R A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A J. Andrew, P. Kohli, J. ShottonS. Hodges and A. Fitzgibbon. *KinectFusion : Real-time Dense Surface Mapping and Tracking*. Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality.

[3] Y. Cui, S. Schuon, S. Thrun, D. Stricker and C. Theobalt. *Algorithms for 3D Shape Scanning with a Depth Camera*. IEEE Trans. Pattern Anal. Mach. Intel, 2009.

[4] Q. Sun, Y. Tang, P. Hu and J. Peng. *Kinect-based automatic 3D high-resolution face modeling*. Image Analysis and Signal Processing (IASP), 2012 International Conference.

[5] M. Zollhofer, M. Martinek, G. Greiner, M. Stamminger and J. SuBmuth. *Automatic Reconstruction of Personalized Avatars from 3D Face Scans*. Comput. Animat. Virtual Worlds, 2011.

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong and K. Zhou. *FaceWarehouse : A 3D Facial Expression Database for Visual Computing*. IEEE Transactions on Visualization and Computer Graphics, 2014.

[7] M. Zollhofer, J. Thies, M. Colaianni, M. Stamminger and G. Greiner. *Interactive model-based reconstruction of the human head using an RGB-D sensor*. Computer Animation and Virtual Worlds, 2014.

[8] S. Fleishman, I. Drori and D. Cohen-Or. *Bilateral Mesh Denoising*. ACM Trans. Graph, 2003.

[9] K. Low. *Linear least-squares optimization for point-toplane ICP surface registration*. Chapel Hill, University of North Carolina, 2004.

[10] X. Xuehan and F. De la Torre. *Supervised Descent Method and its Applications to Face Alignment*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[11] P. Paysan, R. Knothe, B. Amberg, S. Romdhani and T. Vetter. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*. AVSS, 2009.