Personalized Expression Synthesis Using a Hybrid Geometric Machine Learning Method and Mental Representation Analysis of Joyful Expression

Sarra Zaied

DOCTORAT BRETAGNE LOIRE / MATHSTIC

CentraleSupélec

THÈSE DE DOCTORAT DE

CENTRALESUPELEC
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par
« Sarra ZAIED »

« Personalized Expression Synthesis Using a Hybrid Geometric Machine Learning Method and Mental Representation Analysis of Joyful Expression»

Thèse présentée et soutenue à «Rennes», le « 14/12/2020 »
Unité de recherche : IETR
Thèse N° : 2020CSUP0008

Rapporteurs avant soutenance :

Mohamed DAOUDI	Professeur, Université de Lille, Lille
Lionel PREVOST	Professeur, Université Pierre et Marie Curie, Paris

Composition du Jury :

Rapporteurs :	Mohamed DAOUDI	Professeur, Université de Lille, Lille
	Lionel PREVOST	Professeur, Université Pierre et Marie Curie, Paris
Examinateurs :	Saida bouakaz	Professeur, INSA Lyon, Lyon
	Olivier ALATA	Professeur, Université de Jean Monnet, Saint-Etienne
Dir. de thèse :	Pierre-Yves RICHARD	professeur, CentraleSupélec, Rennes
Co-dir. de thèse :	Catherine SOLADIE	Maître de conférence, CentraleSupélec, Rennes

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Catherine Soladié for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank my thesis director : Pr. Pierre-Yves Richard, for his encouragement, insightful comments, and his supervision.

My sincere thanks also goes to Pr. Jean-Julien Aucouturier, for offering me the opportunity for working on diverse exciting and interested projects.

I thank my fellow labmates in FAST Team : Jinting Li, Sen Yen, Siwei Wang and Duong Nam Duong, for the stimulating discussions and for their encouragement. Also I thank my friend in CentraleSupélec Rennes :Karine Bernard.

Last but not the least, I would like to thank my family : my parents, my brothers, sisters and my fiance for supporting me spiritually throughout my life. They all keep me going and this book would not have been possible without them.

Résumé Français

Chapitre 1 : Introduction

La théorie de la rétroaction émotionnelle affirme que nos expériences émotionnelles sont sous l'influence rétroactive de nos propres expressions. Ce sujet est ouvert en psychologie depuis les écrits de William James au 19e siècle. Que le fait de sourire ou de froncer les sourcils puisse avoir un effet automatique dans l'expérience émotionnelle d'une personne ouvre d'importantes pistes thérapeutiques pour les troubles psychiatriques de l'émotion. Cette idée mène à plusieurs travaux de recherche comme les travaux présentés dans cette thèse. Le cadre principal de la thèse est un projet, ANR REFLETS, qui est basée sur la théorie de rétroaction émotionnelle. Ce projet vise à proposer une nouvelle technologie de transformation audio-visuelle permettant de canaliser le potentiel du mécanisme psychologique de la rétroaction émotionnelle faciale et vocale pour une application clinique dans le domaine du stress post-traumatique (PTSD). Pour ce faire, ce projet combine des forces de recherche significatives dans les domaines du :

- Traitement d'image (mené par l'équipe FAST : notre partie dans le projet).
- Traitement du signal audio (mené par Institut de Recherche et Coordination Acoustique/Musique (STMS | IRCAM).
- La psychologie cognitive (est menée par une équipe financée par une subvention European Research Council (ERC)).
- La psychologie clinique pour facteurs humains et PTSD (dirigé par L'institut de recherche biomédicale des armées (IRBA)).
- La conception et le développement d'un miroir commandé par ordinateur dans lequel le programme sera intégré (mené par l'un des principaux acteurs de l'industrie du luxe)

La technologie centrale du projet est un miroir commandé par ordinateur, dans lequel l'observateur se voit et s'entend de manière progressivement plus positive : sans qu'il le sache, son visage réfléchi (capturé par une caméra) est transformé de façon algorithmique pour apparaître plus souriant, et son discours (capturé par un microphone) est ré-synthétisé pour

être plus heureux et détendu. A l'aide du miroir, l'hypothèse émise est que les observateurs arriveront à croire que le reflet facial et vocal transformé et émis de ce dispositif leur est propre, ce qui va provoquer la réaction congruente de leurs muscles faciaux axés sur cette émotion. Ainsi, ils finiront par aligner leurs sentiments avec la transformation. Cliniquement, proposons d'utiliser ce mécanisme pour remédier au traumatisme émotionnel chez les patients atteints de PTSD. Donc notre mission dans ce projet est de proposer un système pour engendrer transformations visuelles. Notre système est capable de transformer de manière photo-réaliste et en temps réel un visage détecté en un visage plus souriant. La particularité est que nous synthétisons une expression de joie spécifique à la personne, tout en gardant l'identité de la personne et l'identité de son émotion.

A travers ce présent rapport, nous proposons d'abord une méthode hybride de synthèse des expressions faciales. Notre méthode est capable de manipuler l'expression d'une image d'entrée et de contrôler de manière paramétrique l'intensité du « sourire » vue sur le visage, tout en conservant d'autres caractéristiques telles que l'identité de l'émotion de chaque personne. Une partie de notre outil de synthèse est utilisé ensuite pour étudier la perception du joie chez les personnes. Cette seconde partie s'agit d'un travail collaboratif avec L'IRCAM pour investiger la perception à l'aide des représentations mentales. Une représentation mentale est l'image mentale qu'un individu se fait d'une situation, d'un concept ou une expression. Nous déterminons les représentations mentales du joie de chaque personne sur son visage à lui ainsi que sur les visages des autres, afin d'étudier la façon de perception de joie et sa relation avec la production de cette expression. Une analyse approfondie est faite par l'IRCAM pour comparer les représentations mentales à la manière qu'à la personne pour exprimer son expression de joie.

Les études existantes sur la rétroaction faciale montrent que l'induction d'un sourire ou d'un froncement de sourcils conduit à des changements congruents dans les réactions émotionnelles des participants [114, 138, 137, 80, 115]. Cependant, bien qu'elles soutiennent la notion générale selon laquelle l'expression émotionnelle influence l'expérience, ces expériences laissent beaucoup de challenges à relever.

Premièrement, un défi majeur avec les études précédentes de rétroaction émotionnelle [114, 138, 137] est qu'il a été jusqu'à présent impossible de contrôler l'intensité de l'expression émotionnelle des participants, et plus généralement de lier les propriétés de leur expression à un effet de rétroaction potentiel. Les recherches ne peuvent pas affirmer le fonctionnement du feedback émotionnel à cause ou malgré la prise de conscience de la situation expérimentale. Par exemple dans [114], il est généralement très difficile pour les participants d'ignorer qu'ils tiennent un stylo dans leur bouche. Le but de notre système

est de créer des transformations visuelles indiscernables des expressions personnelles. Les utilisateurs ne seront pas conscients des déformations. Pour cette raison, nous personnalisons l'expression faciale synthétisée. Par conséquent, notre système adapte le sourire synthétisé à la manière de sourire de chaque personne pour garder la crédibilité. Notre système permet, pour la première fois, d'explorer des paramètres pour rendre un sourire auto-contagieux ce qui permet d'étudier les réponses émotionnelles.

Deuxièmement, dans la majorité des études de rétroaction émotionnelle, les participants devaient faire certaines actions. Dans [138, 137], les participants doivent garder une expression faciale neutre et ils ne doivent pas parler ni changer la position de leur tête pendant l'expérience. L'utilisateur devait se positionner devant le miroir ou la caméra et rester à une distance précise pour que la caméra détecte son visage et effectue les déformations. L'originalité est que notre système s'adapte à la position du visage de l'utilisateur et le déforme pour apparaître plus souriant en fonction de son sourire sans aucune contrainte. De plus, nous pouvons manipuler l'intensité du sourire des expressions synthétisées en temps réel.

Troisièmement, les méthodes de synthèse existantes se concentrent soit sur la forme, soit sur la texture, mais rarement sur les deux. En effet, les méthodes géométriques [5, 138, 137] fournissent une déformation pertinente du visage mais elles manquent de détails dans les images générées. De plus les expressions générées ne sont pas spécifiques à la personne. Les modèles génératifs [86, 21, 131, 124, 86, 48],eux, réussissent à ajouter des détails de texture (rides et dents) mais les sourires générés ne sont pas non plus ceux de la personne. Pour surmonter cela, nous proposons une méthode hybride géométrique-apprentissage machine qui combine les avantages de ces méthodes pour générer une expression de joie personnalisée en forme et en texture. Ainsi, notre système est capable de modifier tous les aspects du visage en particulier la forme (par exemple la pente du sourire, la courbe de la bouche) et la texture (par exemple les rides, les dents).

Enfin, plusieurs recherches récentes [51, 95, 13] se sont intéressées à l'étude des représentations mentales, c'est-à-dire des représentations que l'on fait, par la pensée, d'une projection sensorielle, d'un concept ou d'une situation [73, 77]. De telles représentations sont construitesen se basant sur les sensations et la mémoire. Dans ces études, les représentations mentales sont construites néanmoins qu'elle sont déterminées sur des avatars et non pas sur des sujets réels. Pour ce faire, nous utilisons notre système pour déterminer les représentations mentales personnelles du sourire d'affiliation. Ces représentations définissent l'image construite par l'esprit qu'un individu a d'un sourire ou d'une expression positive. Contrairement aux études récentes [51, 95], nous utilisons des expressions réelles générées pour déterminer et analyser

la représentation mentale ce qui permet d'étudier l'attitude de la perception de joie de chaque personne que se soit sur son visage ou sur les visages des autres.

Nous avons quatre contributions dans cette thèse :

Apprendre un modèle géométrique spécifique à la personne : la contribution majeure est que notre système synthétise une expression de joie spécifique à la personne et permet de préserver la forme morphologique et l'identité de l'émotion en reproduisant la manière spécifique de sourire de chaque sujet. Pour ce faire, nous apprenons un modèle spécifique à la personne en utilisant son visage neutre et son sourire.

Manipulation de l'intensité synthétisée : Une autre contribution est que notre approche peut synthétiser des expressions de sourire avec différentes intensités à partir d'une seule image. Pour ce faire, nous proposons deux solutions. La première est que nous synthétisons ces expressions avec des intensités différentes en fonction de l'amplitude du sourire apprise (l'apex). Nous utilisons un coefficient de déformation, l'augmentation de ce coefficient augmente l'intensité du sourire et vice versa. La seconde solution est que nous générons une expression de sourire du neutre à l'apex en modélisant les caractéristiques dynamiques temporelles du sourire. L'originalité de notre modèle est que nous préservons l'identité émotionnelle de chaque personne en apprenant un modèle dynamique personnel.

Une méthode hybride géométrique-apprentissage machine : Une troisième contribution est que nous agissons à la fois sur la forme et la texture. Nous proposons une méthode hybride combinant des outils géométrique et d'apprentissage machine. La partie géométrique vise à préserver la forme. La partie apprentissage machine, UC-GAN, offre une texture faciale réaliste et permet d'affiner naturellement les détails de texture globale des images synthétisées comme les rides et l'apparence des dents. Cependant, notre UC-GAN manque de cohérence dans l'ajout de la texture des dents car les dents ajoutées ne sont pas celles de la personne. Comme nous souhaitons personnaliser au maximum l'expression synthétisée, pour surmonter cette limitation, nous proposons d'utiliser l'une des deux méthodes suivantes : soit modifier l'image par l'équation de Poisson (Poisson editing method), soit en utilisant un auto-encodeur pour personnaliser la région des dents.

Analyse de la représentation mentale : Une dernière contribution concerne la détermination de représentations mentales (RM) de l'expression de joie sur des sujets réels. Nous menons ce travail en collaboration avec l'IRCAM. L'originalité est que nous proposons un nouvel outil pour construire des RM d'expression de joie sur de vrais sujets. Cet outil est utilisé pour déformer des visages réels et déterminer les RMs du participant. Le but de ce travail collaboratif est d'étudier l'attitude de perception d'un groupe de participants. Nous

générons leurs RMs sur leurs propres visages et leurs RMs sur les visages des autres. Ces RMs sont analysés par l'IRCAM pour étudier trois hypothèses posées. Ils étudient le lien entre les RMs de l'expression de joie de chaque participant sur son visage et sur le visage des autres. De plus, ils recherchent le lien entre les RMs de l'observateur et leur vraie souriant. Finalement, 3 questionnaires sont utilisée pou étudier les traits de leur personnalité de chaque participant.

Le document est organisé comme suit : le chapitre 2 donne un aperçu sur la synthèse des expressions, en particulier pour l'expression de joie. Le chapitre 3 décrit notre système qui synthétise des expressions personnalisées à l'aide d'une méthode hybride géométrique-apprentissage machine. Le chapitre 4 présente les résultats expérimentaux de notre méthode. Les résultats expérimentaux sont analysés en fonction de nos contributions et comparés aux méthodes récentes. Le chapitre 5 détaille le travail collaboratif avec l'IRCAM qui vise à déterminer et analyser les représentations mentales de l'expression de joie. Le chapitre 6 résume nos contributions à cette thèse et présente les perspectives des futures travaux.

Chapitre 2 : État de l'art

Comme expliqué dans le chapitre précédant, nous nous intéressons à la synthèse d'expressions de joie pour agir positivement sur l'état émotionnel. Ce chapitre effectue un aperçu des expressions faciales, plus précisément sur l'expression de la joie (sourire). Tout d'abord, nous explorons comment les mouvements faciaux influencent l'état émotionnel. Deuxièmement, nous discutons comment la synthèse d'expression joie/tristesse peut évoquer l'émotion. Troisièmement, un revue des méthodes de synthèse automatique d'expressions est présentée.

Les mouvements faciaux influencent l'état émotionnel :

La théorie de la rétroaction émotionnelle déclare que nos expériences émotionnelles sont sous l'influence rétroactive de nos propres expressions [52, 82]. Dans les études de rétroaction émotionnelle [11, 19], les expressions faciales ont été définies comme un moyen de communication affective pour les humains. L'hypothèse de rétroaction faciale a deux effets complémentaires : faire une expression faciale devrait augmenter l'intensité de l'expérience émotionnelle, et inhiber les expressions faciales devrait la diminuer [19]. Strack et al. [114] ont mené des études nécessitant l'action des participants pour analyser l'influence des mouvements des muscles faciaux sur l'humeur humaine dans différentes situations. Ils ont demandé aux participants d'évaluer le côté drôle des dessins animés en utilisant un

stylo qu'ils tenaient dans leurs bouches. Conformément à l'hypothèse de rétroaction faciale, lorsque les participants tenaient le stylo avec leurs dents ce qui produit une expression de sourire, ils jugeaient les dessins animés plus drôles que lorsqu'ils tenaient le stylo avec leurs lèvres. Ces résultats ont montré que les mécanismes inhibiteurs et facilitateurs ont contribué aux réponses affectives observées. Basée sur cette théorie, plusieurs recherches ont été menées pour étudier le lien entre le stress et les expressions faciales [79, 63, 70]. Tara et al. [63] a étudié si la manipulation des expressions faciales pouvait influencer les réponses cardiovasculaires et émotionnelles au stress. Les participants ont été invités à accomplir deux tâches stressantes différentes en tenant des baguettes dans leur bouche afin de produire un sourire Duchenne, un sourire standard ou une expression neutre. La sensibilité a été manipulée en demandant explicitement à la moitié des participants de sourire et en donnant à l'autre moitié des instructions de non-sourire. Les résultats ont révélé que tous les participants souriants, qu'ils soient conscients ou non d'un sourire, avaient des fréquences cardiaques plus faibles pendant la récupération du stress que le groupe neutre. Ces résultats montrent qu'il existe des avantages physiologiques et psychologiques à maintenir des expressions faciales positives pendant le stress.

La synthèse d'expression joie/tristesse peut évoquer l'émotion :

Plusieurs études montrent qu'un sourire conduit à des changements congruents dans les réactions émotionnelles du participant [114, 138, 137, 121, 115, 80]. Une personne se voyant souriante et radieuse appréciera naturellement ce moment, tandis que si son visage lui paraît triste, elle en sera affectée négativement. En conséquence, notre perception de nous-mêmes joue un rôle clé dans notre humeur. Un aspect important est que les émotions sont contagieuses. La contagion émotionnelle n'est pas seulement transmise de personne à personne, mais elle peut être réalisée par la personne elle-même via un miroir par exemple, en voyant son visage souriant. Dans le cadre de stimuler les émotions, un prototype a été développé [138, 137] permettant de déformer le visage d'un utilisateur avec un sourire ou une expression triste en temps réel. À la fin de l'expérience, les participants donnent leurs retour (ce qu'ils ont ressenti et remarqué pendant l'expérience pour voir s'ils étaient conscients ou pas de la déformation). Les résultats indiquent qu'un changement dans les expressions faciales influence les émotions des individus.

Les méthodes de synthèse d'expressions :

Les applications citées ci-dessus sont basées sur des méthodes de synthèse d'expression. Au cours des deux dernières décennies, plusieurs chercheurs ont travaillé sur la synthèse

d'expressions faciales en utilisant différentes méthodes. Ces travaux peuvent être divisés en deux catégories. La première catégorie recourt principalement aux techniques géométriques. Ces méthodes déforment directement les points détectés du visage pour générer le visage expressif. La deuxième catégorie vise à créer des modèles d'apprentissage automatique tels que des modèles génératifs pour synthétiser des images d'expression faciale.

Les méthodes géométriques :
Les techniques d'interpolation font partie des premières approches géométriques utilisées pour la synthèse des expressions faciales [66]. Cette technique de morphing est capable de déterminer un mouvement fluide entre deux images à des positions extrêmes, sur un intervalle de temps. Néanmoins, les résultats sont médiocres car les deux images d'entrée contribuent au résultat final [126]. Pour surmonter certaines des limitations et restrictions des interpolations, les chercheurs se sont tournés vers la déformation (Warping). Cette technique réalise une transformation fluide en incorporant des maillages 2D pour maintenir la cohérence tout au long du processus de déformation. Elle effectue la transformation entre deux images ou modèles cibles basés sur la déformation des maillages 2D [138, 87, 40, 39, 134, 80, 122, 23]. Les résultats des méthodes de déformation dépendent d'abord de l'efficacité de l'extraction des détails d'expression à partir du sujet source et deuxièmement du transfert de ces détails au sujet cible. La plupart des travaux déplacent les points caractéristiques de l'expression cible «relativement» en fonction des mouvements de ces points sur le visage de la personne source. Le sujet source est généralement une autre personne ou une expression moyenne sur une base de données. Cependant, chaque personne a sa propre façon d'exprimer son expression. Par conséquent, toutes les expressions générées peuvent être réalistes mais ne sont pas réelles.
Les méthodes d'apprentissage machine :
Récemment, des techniques d'apprentissage profond génératif ont été explorées comme la machine Boltzmann [93], Auto-Encodeur [136, 144, 135] et les réseaux antagonistes génératifs (GAN) [131, 124, 86, 48, 21]. Les réseaux antagonistes génératifs (GAN) [76, 112, 72, 86] ont récemment obtenu des résultats impressionnants pour la synthèse d'expression. Ces modèles sont capables de générer des expressions photo-réalistes et d'ajouter des détails de texture aux images. Néanmoins, les expressions synthétisées ne sont pas celles de la personne. Les expressions générées dans ces recherches proviennent de modèles génériques qui sont appris sur plusieurs personnes. Les modèles génératifs apprennent différentes manières de sourire (base de données d'apprentissage) mais, sans aucune connaissance préalable d'un nouveau sujet dans la base de données de test, le GAN ne peut pas deviner la façon propre de

sourire de ce sujet.

Comme les expressions sont très personnelles et que chaque personne sourit différemment, notre défi est de synthétiser des expressions joyeuses qui sont spécifiques à chaque personne. Pour ce faire, nous personnalisons d'abord la forme à l'aide d'une technique géométrique puis nous affinons la texture globale du sourire généré à l'aide d'une méthode d'apprentissage automatique. Enfin, nous proposons deux solutions pour affiner les détails d'expression locale de l'expression générée. Les détails de notre méthode hybride sont donnés dans le chapitre suivant.

Chapitre 3 : Synthèse d'expression de joie personnalisée à l'aide d'une méthode hybride géométrique-apprentissage machine

Notre objectif dans cette thèse est de proposer un système qui permette de synthétiser l'expression de joie. Comme les expressions sont personnelles, nous sommes donc amenés à personnaliser l'expression synthétisée pour préserver l'identité émotionnelle de chaque personne et garder la crédibilité. Pour ce faire, nous proposons dans ce chapitre une méthode hybride pour synthétiser une expression de joie spécifique à la personne. Notre système apprend un modèle spécifique en utilisant une image neutre et une autre souriante de la personne et génère une expression de joie avec une intensité manipulée. Nous affinons la texture globale de l'expression en utilisant un modèle génératif, "Generative Adversarial Network" GAN en nous basant sur des connaissances préalables sur la manière de sourire de chaque personne. Les textures locales telles que les dents sont ensuite personnalisées comme illustrée dans la Fig. 1. Les détails de notre méthode d'apprentissage hybride sont expliqués dans ce chapitre.

Méthode géométrique

Comme les expressions sont spécifique à chaque personne, notre première contribution est de personnaliser l'expression générée en **forme**. Notre système synthétise l'expression de joie spécifique à la personne et préserve l'identité de l'émotion en reproduisant la manière spécifique de sourire de chaque sujet à l'aide d'une méthode géométrique. Pour ce faire, nous utilisons les connaissances préalables sur la manière de sourire du sujet. Dans cette section, nous présentons les différentes étapes pour générer une expression de forme personnalisée comme illustré dans la Fig. 2.

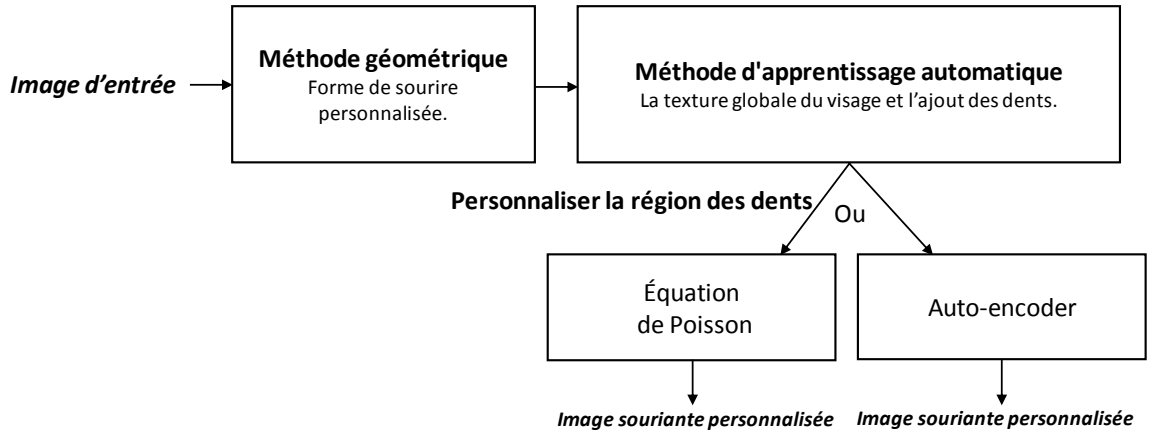


FIGURE 1 Pour générer des expressions de joie spécifiques à la personne, nous proposons un système composé de 3 parties. La première partie vise à personnaliser la forme de l'expression (partie géométrique). La deuxième partie consiste à affiner la texture globale à l'aide d'un modèle génératif (GAN). La troisième partie permet de personnaliser la région des dents en utilisant soit un encodeur automatique, soit une méthode d'édition par l'équation de Poisson.

Apprentissage du modèle spécifique à la personne

La première étape de notre méthode consiste à apprendre un modèle paramétrique spécifique à la personne en utilisant une image neutre X_n et une image souriante X_s de cette personne. Pour la détection des points caractéristiques, nous utilisons Kurma Tracker de Dynamixyz [24] qui détermine avec précision les coordonnées de 149 points caractéristiques dans chacune des images.

Un sourire est exprimé avec la montée des coins de la bouche et des joues, ainsi que la levée des paupières inférieures [28]. Nous avons sélectionné les 10 points caractéristiques correspondant aux coins de la bouche et aux points inférieurs des yeux pour apprendre et appliquer les déformations. Nous effectuons une triangulation de Delaunay sur le visage neutre pour avoir un maillage du visage. Chaque point de contrôle du sourire X_s^i est situé à l'intérieur d'un triangle du maillage du visage neutre ($X_n^{u_i}$, $X_n^{v_i}$, $X_n^{w_i}$) comme le montre la Fig. 3 pour le point 64 du visage souriant. Nous calculons les coordonnées barycentriques (α_i, β_i, γ_i) pour chacun des 10 points du visage souriant X_s^i. Ces coordonnées sont les paramètres de notre modèle spécifique à la personne. Le modèle personnalisé est donc composé de 10 vecteurs à 6 composantes. Ces composantes sont les 3 index de sommets de triangles (u_i, v_i, w_i) et les 3 coordonnées barycentriques (α_i, β_i, γ_i) associées à chacun des 3 sommets. Le

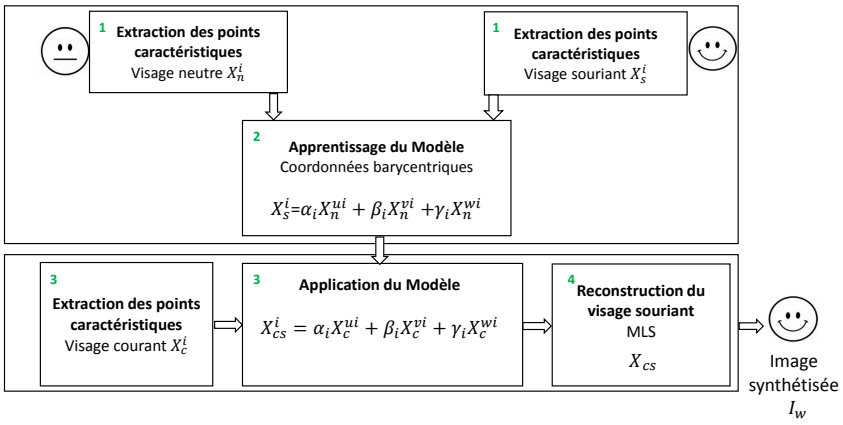


FIGURE 2 La première partie de méthode est composé de 4 étapes. Dans la première étape, nous déterminons les points caractéristiques d'un visage neutre X_n et d'un visage souriant X_s du sujet. Dans la deuxième étape et en nous basant sur les coordonnées barycentriques, nous apprenons un modèle spécifique à la personne à partir des points détectés des deux visages. Ce modèle apprend la déformation à appliquer pour rendre le visage plus positive. Dans la troisième étape, nous détectons le visage courant de la personne X_{cs} et nous calculons les nouvelles positions des points de sourire X_{cs}^i en fonction du modèle appris. Enfin, nous utilisons une méthode de déformation 2D MLS [105] pour déformer toute le visage courant X_c en utilisant X_{cs}^i.

calcul de X_s^i est formulée comme suit :

$$X_s^i = \alpha_i X_n^{u_i} + \beta_i X_n^{v_i} + \gamma_i X_n^{w_i} \tag{1}$$

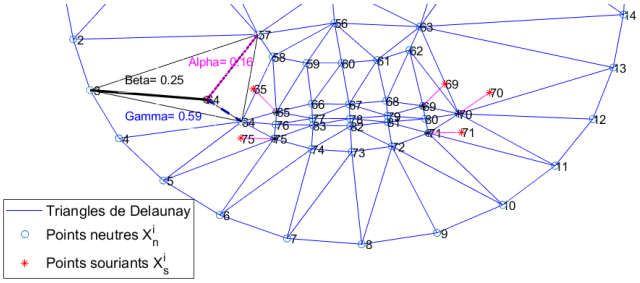


FIGURE 3 Les positions des points souriants X_s^i par rapport aux triangles du visage neutre X_n^i. En souriant, chaque point de contrôle du sourire est situé dans un triangle du visage neutre (X_s^{64} est dans le triangle des sommets X_n^{57}, X_n^3, X_n^{64}).

Génération de l'expression de joie

Une fois que le modèle spécifique est appris, nous pouvons modifier chaque nouvelle image X_c du sujet. Nous détectons les points du visage courant X_c^i. Ayant les différents coefficients α_i, β_i et γ_i de chacun des 10 points du visage souriant, et connaissant les coordonnées des triangles (u_i, v_i, w_i) dans lesquels se trouvent ces points, nous déterminons les positions des 10 nouveaux points souriants X_{cs}^i en utilisant l'équation (2). L'utilisation d'une déformation relative par rapport aux points du visage détecté permet de s'affranchir du problème d'orientation du visage.

$$X_{cs}^i = \alpha_i X_c^{u_i} + \beta_i X_c^{v_i} + \gamma_i X_c^{w_i} \tag{2}$$

Pour construire l'image souriante, nous appliquons la méthode Moving Least Squares (MLS) [105, 5] pour transformer le visage détecté X_c à partir des nouveaux points calculés. Compte tenu du temps nécessaire pour effectuer les déformations et les appliquer, nous faisons un compromis temps/esthétique comme dans [5] : nous appliquons l'algorithme du MLS sur des grilles autour de chaque œil et de la bouche, et non pas sur chaque pixel de l'image.

Manipulation de l'intensité de l'expression

La deuxième contribution de notre système est la synthèse des expressions de joie avec différentes intensités à partir d'une seule image. Pour ce faire, nous proposons deux solutions : La première est que nous synthétisons ces expressions avec des intensités différentes en

fonction de l'amplitude du sourire appris (Apex). Nous utilisons un coefficient de déformation, l'augmentation de ce coefficient augmente l'intensité de l'expression et vice versa. Nous manipulons l'intensité en utilisant l'équation suivante :

$$X_{cs}^{i,d} = X_c^i + (X_{cs}^{i,1} - X_c^i) \times d \tag{3}$$

Où d est le coefficient de déformation. L'augmentation de ce coefficient augmente l'intensité de sourire et vice versa.

— Si $d = 0$, le résultat est un visage inchangé X_{cs}^0=X_c.
— Si $d = 1$, le résultat est une expression de joie X_{cs}^1 correspondant à l'intensité de celui qui a été appris X_s.

L'avantage de cette solution est que nous synthétisons une expression de sourire manipulée basée sur deux images : une expression neutre et une expression Apex. Cependant, manipuler l'intensité basée sur l'amplitude ne respecte pas l'évolution de l'expression dans le temps. Par exemple, synthétiser une expression avec un $d = 0,5$ ne génère pas l'image du milieu d'une expression. Pour aborder ce cas, nous proposons une deuxième manière de manipuler l'intensité du sourire. Nous proposons de générer une expression de sourire de l'onset à l'apex en modélisant la dynamique temporelle du sourire. Dans ce cas, nous avons besoin d'une vidéo souriante de la personne pour apprendre cette dynamique.

L'originalité de cette solution est qu'elle respecte les propriétés temporelles du sourire de la personne. Nous commençons par apprendre la déformation entre les images d'une première vidéo sourire du sujet (de l'onset à l'apex). Nous utilisons notre partie géométrique basée sur les coordonnées barycentriques pour déterminer un modèle par image. Nous avons appliqué quelques changements sur le système proposé précédemment comme illustré dans la Fig. 4. Au lieu d'apprendre une déformation entre 2 expressions (neutre et apex) et de manipuler l'amplitude du sourire avec le coefficient d, nous proposons d'apprendre la déformation entre l'image neutre X_{n1} et chacune des images suivantes X_{s1}^f d'une première vidéo souriante de la personne telle que formulée dans l'équation suivante :

$$X_{s1}^{i,f} = \alpha_{i,f} X_{n1}^{u_i} + \beta_{i,f} X_{n1}^{v_i} + \gamma_{i,f} X_{n1}^{w_i} \tag{4}$$

où f représente l'index de l'image. Selon les résultats trouvés dans [3, 106], la durée Onset-Apex est d'environ 0,5 s. Par exemple, si les vidéos sont enregistrées à 50 fps, f est compris entre 0 et 25.

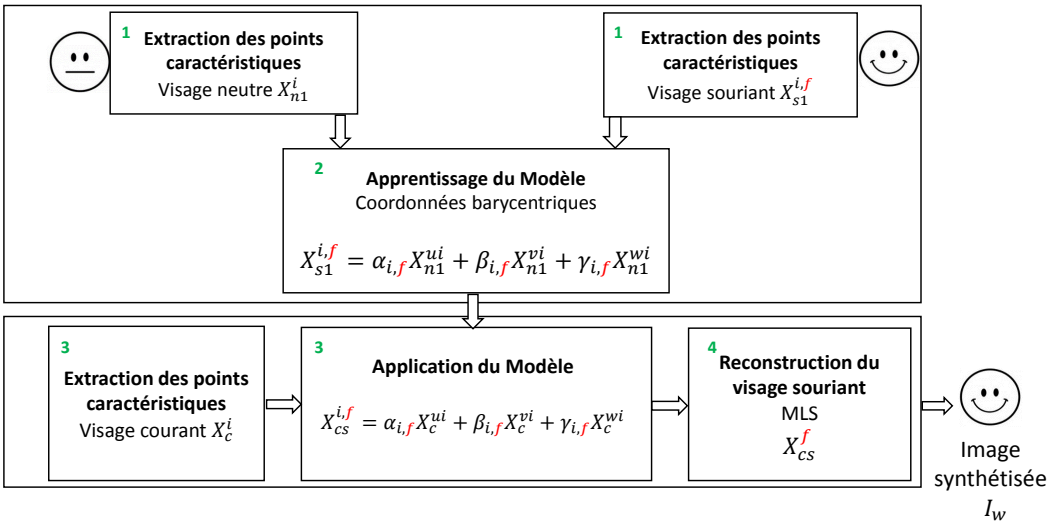


FIGURE 4 Pour apprendre un modèle dynamique on commence d'abord par déterminer les points caractéristiques du visage neutre X_{n1} et toutes les images souriantes X_{s1}^{f} de la vidéo souriante 1. Ensuite, nous apprenons un modèle spécifique à la personne par image à partir des points détectés de chacun des deux visages (neutre et l'autre image d'index f). En fin, le modèle appris est utilisé avec une nouvelle image X_c du sujet pour générer des nouvelles images souriantes X_{cs}^{f}.

Méthode d'apprentissage automatique : raffinement de texture via GAN

La méthode géométrique précédente permet de construire un modèle spécifique par personne, qui est utilisé comme une connaissance préalable pour générer des expressions de joie. L'expression générée est spécifique à la personne. Cependant, comme nous nous concentrons uniquement sur la forme, certains détails de texture manquent, tels que les dents et les rides. Nous proposons notre solution d'apprentissage automatique pour surmonter cette limitation.

Pour affiner les détails de texture sur les images synthétisées (ajouter des rides, des fossettes et des dents), nous utilisons un U-Net Conditioned Generative Adversarial Network. L'UC-GAN [50] a obtenu des résultats impressionnants grâce aux connexions utilisées dans le générateur [112, 124, 86, 123]. Nous utilisons un UC-GAN d'une part pour s'assurer que préserve l'identité de la personne, d'autre part, pour affiner les textures des images synthétisées d'une façon pertinente. Nous améliorons le traditionnel UC-GAN en combinant une connaissance préalable qui contient des informations de la manière de sourire de chaque personne. Notre première originalité est que nous utilisons les images synthétisées (déformées) I_{wj} générées par la méthode géométrique et les vraies expressions souriantes I_{sj} pour entraîner notre UC-GAN comme montre Fig. 5 où j est l'index de l'image. L'entrée de notre UC-GAN est le visage déformée I_{wj} qui ne posséde pas de détails globaux tels que les dents et les rides mais qui est personnalisé en forme. Ces détails sont affinés à l'aide de l'UC-GAN proposé.

Notre deuxième originalité est l'utilisation d'un vecteur L_{sj} ($[O_j, V_j]$) qui guide le GAN à ajouter les détails de l'expression tels que l'ouverture de la bouche. L_{sj} est composé du coefficient de l'ouverture des lèvres O_j, concaténé avec un vecteur V_j qui indique le niveau de l'intensité de l'expression à générer. Le coefficient O_j est le rapport entre la distance entre les yeux et l'ouverture de la bouche.

Personnalisation de la régions des dents

Nos résultats à ce niveau montrent que notre système génère des expressions proches de expressions réelles grâce à la méthode hybride. Cependant, nous avons remarqué qu'il existe deux limitations dans notre système. Le premier est que le GAN n'ajoute pas les dents de la personne réelle et le second est que le GAN ajoute des dents différentes à chaque image de l'onset à l'apex. Nous proposons donc 2 solutions pour surmonter ces limitations et renforcer

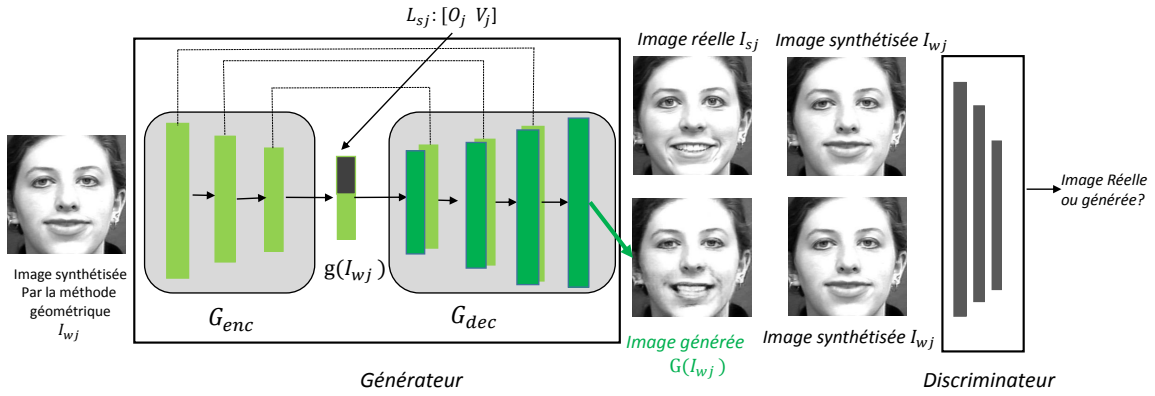


FIGURE 5 Le GAN est utilisé pour affiner les détails sur les images géométriques synthétisées I_{wj}. L'image I_{wj} générée par la méthode géométrique, est transmise au générateur. La représentation de l'image $g(I_{wj})$ est concaténée avec un vecteur L_{sj} qui est composé du coefficient de l'ouverture des lèvres O_j dans l'image réelle et d'un vecteurV_j qui caractérise l'intensité cible. Le décodeur génère l'image de joie avec plus de détails. Le discriminateur D prend deux couple d'images ; l'image synthétisée I_{wj} avec la réelle I_{sj} et la l'image synthétisée I_{wj} avec celle générée $G(I_{wj})$ pour déterminer si cette dernière est une image réelle ou générée.

notre système.

L'édition de l'image par l'équation de Poisson : Afin de synthétiser une texture de bouche spécifique à la personne, nous proposons d'utiliser la technique d'édition par équation de Poisson pour modifier les dents générés par ceux de la personne. L'idée est d'utiliser une autre vidéo souriante de cette personne pour ajouter ses propres dents aux images générées $G(I_{W\,j})$ par notre système hybride. Un masque M_j est généré autour de la région des dents de $G(I_{W\,j})$. La région des dents à ajouter $S_{I_{sk}}$ est déterminée à partir d'une image réelle I_{sk} d'une autre vidéo souriante de cette personne. Pour chaque région de dents à modifier $T_{G(I_{W\,j})}$, nous déterminons automatiquement sa région correspondante dans la deuxième vidéo en comparant les distances d'ouverture de la bouche dans les deux images. Pour ce faire nous calculons les distances d_j entre les deux points des coins de la bouche de l'image généré $G(I_{W\,j})$ et recherchons dans la deuxième vidéo l'image I_{sk} qui a la distance la plus proche d_k à d_j et l'utilise comme source.

Auto-encodeur (AE) : Une autre solution pour personnalisée la région des dents est d'exploiter les avantages de l'AE pour encoder les dents de la personne ce qui permet de guider le GAN à ajouter des informations de texture spécifiques à la personne. Lors de la

phase d'apprentissage, l'AE prend comme entrée la bouche de la personne m_j pour apprendre l'extraction des informations significatives et génère un code C_j. Lors de la phase de test, l'AE génère un code personnalisé C_j, qui contient des caractéristiques des dents. Ce code C_j est concaténé avec le vecteur L_{sj}. Avec cette solution, nous nous attendons à ce que la combinaison aide le GAN à générer une texture plus personnelle. L'avantage d'utiliser un AE est qu'il permet d'affiner la texture quelle que soit la position ou l'orientation du visage de la personne.

Chapitre 4 : Résultats expérimentaux

Nous présentons les résultats expérimentaux de notre méthodes dans ce chapitre. Tout d'abord nous présentons deux études statistiques des sourires réels qui prouvent que chaque personne a sa propre manière de sourire. Afin de démontrer la pertinence du système proposé, le reste du chapitre est basé sur nos principales contributions présentées précédemment. Nous examinons la pertinence de ces contributions une après une.

Notre contribution principale est que notre système synthétise une expression de joie spécifique à la personne et conduit à préserver la forme morphologique et l'identité de l'émotion en reproduisant la manière spécifique de sourire de chaque sujet. Nous comparons nos résultats avec les méthodes récentes de l'état de l'art [5, 124]. Les résultats qualitatifs et quantitatifs montrent que notre méthode synthétise des expressions proches des expressions réelle de chaque sujet.

La deuxième contribution est que nous synthétisons l'expression avec une intensité différente en utilisant une seule image neutre de la personne. L'originalité est que nous proposons de manipuler l'intensité en fonction de l'amplitude du sourire apprise ou en fonction de la dynamique temporelle du sourire de chaque personne. Les résultats comparant les deux manières de manipulation d'intensité sont ainsi présentés dans ce chapitre.

Une autre contribution principale est évaluée. Nous agissons à la fois sur la forme et la texture en introduisant notre méthode hybride combinant des outils d'apprentissage géométrique et machine pour affiner la texture globale. Les résultats quantitatifs et qualitatifs sont présentés dans ce chapitre et montrent que le GAN raffine la texture des expressions synthétisées avec la méthode géométrique en ajoutant les détails globaux tel que les rides et l'apparition des dents.

La quatrième contribution concerne la personnalisation des dents en utilisant la méthode d'édition par équation de Poisson ou en utilisant un auto-encodeur. Ces deux outils sont testés pour affiner la texture locale de la région des dents. Les résultats montrent que les deux méthodes contribuent efficacement à affiner la texture local des dents.

Chapitre 5 : Détermination et analyse des représentations mentales de la joie

Ce chapitre représente notre travail collaboratif avec l'IRCAM (Institut de recherche et de coordination Acoustique / Musique). Notre collaboration vise à étudier la manière dont chaque personne perçoit l'expression de la joie sur son propre visage et sur le visage des autres. Par conséquent, nous nous sommes dirigés vers la détermination et l'analyse des représentations mentales (RM) de l'expression de joie de sujets réels (percevoir leur visage et celui des autres). La RM porte sur la façon dont nous stockons les informations dans la mémoire et comment nous les représentons dans notre esprit ou les manipulons à travers nos processus de raisonnement.

La synthèse des expressions aléatoires

Tout d'abord, nous proposons un outil qui consiste à générer des expressions faciales aléatoires à partir d'un visage neutre réel. Pour cette étape, nous avons utilisé notre méthode de déformation proposée dans les chapitres précédents. Nous déformons les visages de 2 acteurs et les visages de 24 participants pour générer 1400 expressions aléatoires pour chacun d'entre eux.

Le test de perception et determinations de RMs

Deuxièmement, nous (avec l'IRCAM) avons mené une expérience psycho-physique, dans laquelle les participant doivent catégoriser les expressions faciales aléatoires générées. Au cours du test nous affichons les images par paire, et le participant doit choisir l'image la plus souriante dans chaque paire. Chaque participant est invité à passer ce test sur son visage et sur les visages des 2 acteurs. À partir de leurs réponses, nous reconstruisons les 3 représentations mentales de chaque participant (**Self** et 2 **Actros** MRs), en utilisant la procédure de "Reverse correlation" [2, 97, 78].

L'analyse des RMs

Une fois les représentations mentales sont déterminées, l'IRCAM analyse trois hypothèses liées à la perception d'expressions de joie. Les résultats de cette analyse sont présentés dans l'annexe A. Ils ont étudié le lien entre les MRs de l'expression de joie sur le visage du participant (**Self** MR) et sur le visage des autres (2 **Actors** MRs). De plus, ils ont recherché le lien entre les MRs de l'observateur et leur façon de sourire réelle (**Production**), ainsi que les

traits de leur personnalité. Les résultats liés à la première hypothèse prouvent qu'il existe une corrélation entre la perception du visage de soi et celle des autres visages. Ensuite, la seconde hypothèse s'est avérée non valable car il n'y a pas de lien entre les MRs du participant et la manière dont il produit son expression de joie réelle. Sur la base de nos résultats, nous pouvons dire que la manière de percevoir n'est pas liée à la façon dont nous produisons réellement nos propres expressions. En analysant la personnalité des participants, nous avons validé la troisième hypothèse, qui justifie le lien entre les représentations mentales et la personnalité du participant.

Chapitre 6 : Conclusion

Cette thèse fait partie d'un projet ANR nommé "REFLETS". Notre mission est de proposer une solution permettant de générer du miroir en produisant un pseudo-sourire grâce au traitement d'image. Dans cette thèse, nous avons étudié la nature de l'expression de joie en se basant sur deux axes.

Le premier est la production d'une expression joyeuse. Nous avons prouvé que chaque personne a sa propre façon de produire son expression de joie. Par conséquent, nous avons proposé un système qui peut déformer le visage de l'utilisateur en un visage plus souriant. Notre méthode hybride géométrique-apprentissage machine combine les avantages de ces techniques pour générer une expression de joie personnalisée en forme et en texture. Nous avons comparé nos résultats avec deux méthodes récentes et nos résultats montrent que nous générons des sourires proche de l'expression réelle de la personne.

Le deuxième axe élaboré dans cette thèse est la perception de la joie. Nous avons conduit un travail collaboratif avec l'IRCAM qui vise à étudier la façon dont chaque personne perçoit l'expression de joie sur son visage et sur le visage des autres en déterminant leurs représentations mentales. Nous avons proposé un système qui vise à déformer des visages réels pour générer des expressions aléatoires. Ces expressions sont utilisées pour procéder un test de perception et déterminer les représentations mentales sur le visage du sujet et sur le visage des autres. Les représentations mentales déterminées sont analysées par l'IRCAM. Les résultats prouvent que la manière dont chaque personne produit son sourire n'affecte pas sa perception. Il apparaît donc que les sujets ne perçoivent pas comme ils produisent. En outre, ils perçoivent la joie avec le même raisonnement soit pour le visage de soi, soit pour

les autres visages.

Concernant les perspectives de nos travaux, nous abordons les points suivants :

— **Amélioration de notre système** : Comme précédemment cité, nos expressions synthétisées se sont avérées être les plus proches de la vérité terrain. Cependant, il existe encore des différences avec ces derniers. De plus, le GAN influence la résolution des images générées. Une solution possible pour améliorer la résolution est d'augmenter les données dans la partie GAN est nécessaire pour améliorer les performances du système.

— **Système temps-réel** : A ce moment, seule la méthode géométrique est implémentée pour déformer le visage détecté pour apparaître plus joyeux en temps réel. Après avoir amélioré nos performances GAN, nous pouvons également mettre à jour notre méthode hybride pour générer les expressions en temps réel.

— **Résultats psychologiques sur les patients PTSD** : À ce stade, notre première version est encore en phase de test et nous n'avons pas encore les résultats des tests psychologiques sur les PTSD à l'hôpital de Percy. Ces résultats contribueront à améliorer notre système et à l'adapter aux applications réelles.

— **Utiliser notre outil de représentation mentale dans d'autres contextes** : L'originalité de notre outil est qu'il est capable de déterminer la représentation mentale de toutes les expressions (joie, peur, dégoût, ect ..). Avec cet outil, il serait également possible de déterminer les représentations mentales d'émotions complexes comme le stress et de définir également une configuration faciale de ces expressions.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

My thesis takes place at CentraleSupelec Rennes in the Facial Analysis Synthesis Tracking (FAST) team which is part of the Institute of Electronics and Telecommunications (IETR) of Rennes. The thesis is a part of ANR REFLETS project. This chapter first introduces the general context and the reason for proposing REFLETS project in section 1. Then the research scientific hypotheses, the objectives of REFLETS project and details of our mission are presented in section 2. Thirdly, we present the challenges and our contributions in this thesis. Finally, we give the organization of the following chapters of this manuscript.

1.1 Background and general context

Facial expressions are incredibly linked to emotions. They give us information about the mental state of the person we meet. We understand that persons are happy if signs of enjoyment in their face and their voice appeared [75, 25]. In social contexts, emotions influence what we think of other persons, and how we behave in front of them [82]. They also influence our moods [114, 121] because emotions are contagious [45]. Research on emotions has produced several discoveries that have led to important real-world applications [138, 137, 80]. That's why many researchers have been working on synthesizing facial expressions in different contexts such as multimedia applications, Human-Computer Interaction (HCI) and video conferencing [16, 83, 84, 144, 101, 43]. A wide range of applications in affective interactions, such as facial animation [117] and facial editing [122, 23] have been facilitated by the advances of synthesizing facial expressions. Moreover, expressions become an important aspect of embodied agents designed for human-computer interactions [85, 33].

Recently, tools for facial expression synthesis have been used to manipulate and to evoke emotions [138, 115, 5]. Keita et al. [115] have constructed a video system, called FaceShare,

which can deform the user's face with a smile in response to his partner's smile in order to act on his emotional state. So the pseudo-smile makes the mimicker closer to his partner, for his face is deformed according to the expressions of the latter. It also improves the flow of the conversation for both the mimicker and the mimickee, who sees the mimicker's deformed face. The systems proposed in [138, 115, 5] are based on several psychiatric research [63, 81] which have shown that the movement of the facial muscles influences the emotional experiences. This theory is named "peripheral emotional feedback" [52, 82]. They proved also that smiling can relax the body, and it can also lower the heart rate and blood pressure. It has a positive effect on a person's mood causing a reduction in stress and a brief feeling of relaxation and happiness [63]. Moreover, the smile makes us appear more attractive to others. It lifts our mood, as well as the moods of those around us because emotions are contagious [45]. That has led to a huge potential for clinical improvement in psychiatric disorders, especially for Post-traumatic stress disorder (PTSD) patients.

PTSD is a psychiatric disorder that can occur when someone experiences a traumatic event. One may feel sadness, fear or anger ; and one may feel detached or estranged from other people [35]. A range of psychological treatments is currently used for PTSD sufferers, such as eye movement desensitization and reprocessing (EMDR) [110, 109], effective–behavior therapy [127, 92] and cognitive therapy [94, 9]. Based on the peripheral emotional feedback theory and emotional contagion, significant research forces are combined and proposed a project "REFLETS" that aims at building a novel health system to positively act on the emotional experience of Post-traumatic stress disorder (PTSD) patients by changing in real-time their reflected faces and voices. We participated in the kick-off of REFLETS project on 02/10/2017, where we met all the partners who presented their different tasks in several fields such as computer science, cognitive psychology, and clinical psychology. More details about the project are given in the following section.

1.2 REFLETS project

The REFLETS project aims at improving the care methodology of PTSD patients via a technological device acting on their ability to perceive and to regulate their own emotions. This section presents the project. Subsection 1.2.1 raises the scientific hypothesis. In Subsection 1.2.2, we present the partners of the project and the research fields. The final expected product of the project is defined in subsection 1.2.3. At the end of this section, we detail our mission in the project.

1.2.1 Scientific hypotheses

Emotional feedback hypothesis (EFH) has been an ongoing subject of debate in psychology since William James [52, 82]. This theory states that our emotional experiences are under the retroactive influence of our own expressions. Thus, smiling can have an automatic effect on one's emotional experience and has enormous potential for clinical remediation in psychiatric disorders [42]. In [114], researchers investigated the hypothesis that people's facial movements influence their emotional state. Their results showed that both inhibitory and facilitatory facial mechanisms have acted to the observed affective responses.

The REFLETS project aims at exploiting the advantages of this theory, by building novel health technology able to channel the psychological mechanism of facial and vocal emotional feedback for clinical application to post-traumatic stress disorders (PTSD) as well as well-being applications in the general population. To do so, REFLETS combines significant research forces in several fields.

1.2.2 Partners and research domains

As mentioned above, the goal of REFLETS project is to develop an application which can act positively on patients' moods. For that reason, the project reunites different research domains such as :

— Computer Science : An Audio-Visual software is proposed by two teams : FAST (Facial Analysis, Synthesis, and Tracking of UMR6164 IETR CNRS/CentraleSupelec) and PDS (Perception and Sound Design team of UMR9912 STMS (CNRS/IRCAM)). Our FAST team works in the field of image processing and its main activity concerns the analysis of emotions (depression, PTSD, Stress) in the context of virtual and augmented reality. Therefore, the visual transformation software of REFLETS project is carried out by FAST team through the system proposed in my thesis. We work also in collaboration with Dynamixyz, a software company located in Rennes. This company provides us with an efficient software [24] for detecting facial landmarks in real-time from multiple cameras behind the mirror, in order to support our visual smile transformation software. The audio software is developed by PSD team. This team conducts research on sound perception, sound design, and cognition by combining knowledge in signal processing/synthesis, psychology, and cognitive sciences.

— Cognitive Psychology : HumanEvo team is in charge of carrying out affective social studies.

— Clinical Psychology : Human factors and post-traumatic stress disorder studies are handled by the Armed Forces Biomedical Research Institute (IBRA).
— Hardware technology : The development of an augmented mirror-like is conducted by one of the main players in the luxury industry (Chanel). This hardware is used to analyze and reflect the transformed face and voice of the observer.

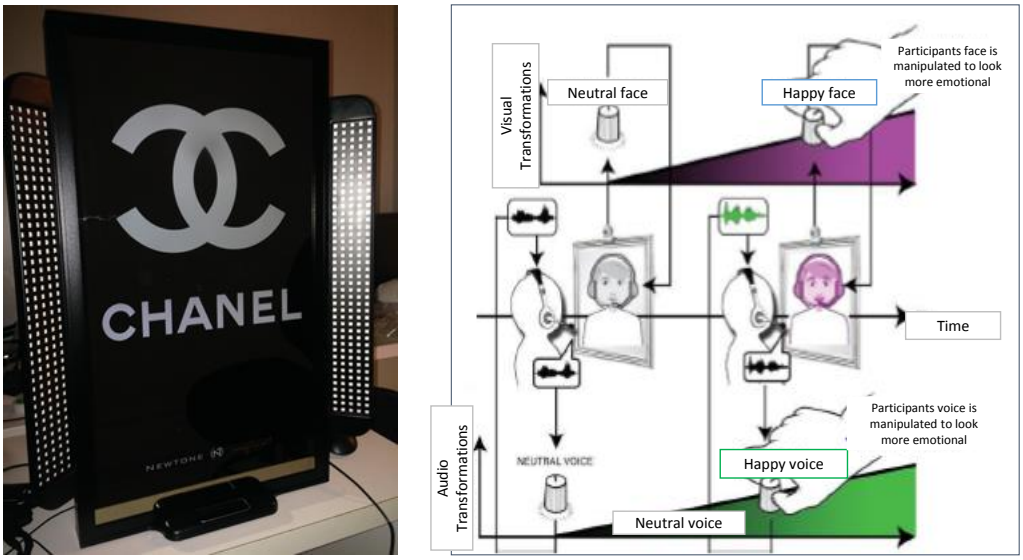


FIGURE 1.1 Prototype of computer-augmented mirror (left) and its proposed functionality (right) : the observers' reflected face (captured by a set of cameras) is algorithmically and photo-realistically transformed to appear more smiling, and their speech (captured by a microphone) is re-synthesized in real-time to seem more happy and relaxed [91].

1.2.3 Objective and final product

REFLETS project has a major objective, which is the development of realistic and multi-modal synthesis system to evoke user's positive emotions. This software can modify in real-time the audio-visual reflection of the user, to make his face appear more smiling by increasing the curvature of his mouth and change his voice sounds to seem brighter by modifying their formant frequencies in a way that simulates smiled speech. To this aim, a novel audio-visual deformation framework is proposed, based on learning a person-specific smile model for each user [140, 141] and on modifying the spectral envelope of the speech signal [6, 5].

The system is moving towards implementation in a computer-augmented mirror device, in which observers both see and hear themselves in a more positive way. Their reflected face captured by a set of cameras is algorithmically transformed to appear with a personal smile. Moreover, their speech captured by a microphone is re-synthesized to seem more happy and relaxed. Using the mirror and based on the theory of peripheral feedback, we expect that observers will come to believe the emotional tone of their transformed facial and vocal reflection as their own, and align their feelings with the transformation [5, 6, 140]. Clinically, we propose to use this mechanism to cure emotional numbing in Post-Traumatic Stress Disorder (PTSD) patients.

1.2.4 Our Mission

Among the objectives of the project is the development of audio-visual synthesis software. Our mission is to propose a solution which can generate mirroring by producing a pseudo-smile through image processing. In this thesis, we study the nature of joyful expression basing on two axis. The first axis is the production of joy, we prove that each person has his own way to produce his joyful expression. Therefore, we propose a system which can deform the user's face into a personal smiling face. This axis is treated in **chapter 2, 3 and 4**. The second axis is the perception of joy. Our smile synthesis system is also used in a collaboration work in neuroscience. We propose a new system which aims at determining the mental representation of the subject on a given face. This system allows us to study how each person perceives the expression of joy on their face and on the faces of others. Moreover, we analyzed whether this perception is influenced by the way that have the person to produce his joyful expression and by his personality traits. This axis is handled in **chapter 5**. We present our contributions in the following section.

1.3 Challenges and Our Contributions

Perceived a joyful expression or produce it are two acts that play an important social role : they could activate the chemistry of our body to feel better in a few minutes. Therefore, to act on the emotional state of an observer, we headed-on proposing a system for evoking positive emotion by synthesizing a joyful expression. Moreover, we studding the perception of joyful expression on real subjects to determine their mental representations on their faces and the faces of others.

1.3.1 Challenges

The relation between the expression and experience of emotions is confirmed in the field of psychology [52, 104]. Existing studies on facial feedback show that induction of a smile or a frown leads to congruent changes in the participants' emotional reactions [114, 138, 137, 80, 115]. However, although they support the general notion that emotional expression influences experience, these experiments leave a lot of mechanistic details largely unresolved.

First, one major challenge with previous emotional feedback studies [114, 138, 137] is that it has been so far impossible to control the intensity of the emotional expression of participants, and more generally link the properties of their expression with a potential feedback effect. Researches can not affirm how emotional feedback works because, or despite, the awareness of the experimental situation. For example in [114], it is typically very difficult for the participants to ignore that they are currently holding a pen in their mouth. The goal of our system is to create visual transformations that are indistinguishable from genuine expressions. The users will not be aware of the deformations. For this reason, we personalize the synthesized facial expression. Therefore, our system adapts the synthesized smile to the way of smiling of each person to keep the credibility. Our system allows, for the first time, to explore parameters to make a smile self-contagious and allows to investigate emotional feedback responses.

Second, in the majority of emotional feedback studies, participants were required to do some actions. In [138, 137], the participants must keep the neutral facial expression and they should not speak or change their head position during the experience. The user had to position in front of the mirror or the camera and remain at a precise distance so that the camera detects the face and makes the deformations. The originality is that our system adapts the deformation to the position of the user's face and deforms it to appear more smiley according to their smile without any user's action. Moreover, we can manipulate the smile intensity of the synthesized expressions in real-time.

Third, the existing synthesis methods either focus on shape or on texture, but rarely on both. Indeed, the geometric methods [5, 138, 137] provide a relevant shape deformation but they lack local details in the generated images. The generative models [86, 21, 131, 124, 86, 48] succeed in adding texture details (wrinkles and teeth) but the generated smiles are not those of the person. To overcome this, we propose a hybrid geometric-machine learning method that combines the benefits of the geometric and machine learning methods to generate a joyful expression personalized on shape and texture. Thus it is able to modify all the aspects

of the face especially the shape (e.g smile slope, mouth curve) and texture (e.g wrinkles, teeth).

Finally, various recent researches [51, 95, 13] were interested in studying mental representations, meaning representations that one makes, by thought, of a sensory projection, of a concept or a situation [73, 77]. Such representations are constructed by the confluence of sensations and memory. Actually, in these studies, random AUs and avatars are used by the experimental protocol to determine the mental representations of the 6 basis expressions. These representations are not determined on real subjects. To do so, we propose a system to determine personal mental representations of affiliation smile. These representations define the mind-constructed image that an individual has about a smile or a positive expression. Unlike recent studies [51, 95], we use generated real expressions to determine and analyze the mental representation of each person.

1.3.2 Contributions

In this thesis, we propose first a hybrid geometric-machine learning approach to synthesize photo-realistic and personalized joyful expressions while keeping the identity of the emotion. A part of our approach is used then to determine and analyse mental representations of joy expressions. We have four contributions thesis :

Learn person-specific model : One contribution is that our system synthesizes in real-time person-specific joyful expression and leads to preserving the morphology shape and the identity of the emotion by reproducing the specific way of smiling of each subject. To do so, we use previous knowledge of the way of smiling of the subject. We learn a person-specific model for each subject using their neutral face and their smiling face (apex).

Manipulated intensity : Another contribution is that our approach can synthesize smile expressions with different intensities from a single image. To do so we propose two solutions. The first one is that we synthesize these expressions with different intensities based on the smile amplitude of the learned expression (apex). We use a deformation coefficient, such that increasing this coefficient increases the smile intensity and vice versa. The second one is that we generate a smile expression from onset to apex by modeling the temporal dynamic characteristics of the smile. The originality of our model is that we preserve the emotion identity on shape of each person by learning a personal-dynamic model.

A hybrid geometric-machine learning method : A third contribution is that we act on both shape and texture. The originality is introducing a hybrid method combining geometric and machine learning tools. As the frames synthesized with our geometric model missed some details like wrinkles and teeth, we use an GAN to refine the texture details in the

generated images. The geometric part aims at preserving the identity shape and optimizes the distortion made on the image. The GAN offers a realistic facial texture and allows to naturally refine the global-texture details of the synthesized images such as wrinkles and appearance of teeth. However, our GAN has a lack of consistency in adding the teeth texture because the added teeth are not those of the person. As we would like to personalize the synthesized expression as much as possible, to overcome this limitation, we propose to use one of two methods : either Poisson editing method or an Auto-encoder to personalize the teeth region.

Mental representation analysis : A last contribution concerns the generation of self and other-person mental representations (MRs) of joy expressions on real subjects. We conduct this collaborative work with IRCAM. The originality is that we propose a novel tool to build MRs of joyful expression on real participants. This tool is used for deforming real faces and determine the participant's MRs. The aim of this collaborative work is studying the perception attitude of a group of participants. We generate their MRs on their own faces and their MRs on the faces of others. These MRs are analyzed by the IRCAM to investigate three posed hypotheses. They investigate the link between the MRs of joyful expression on the participant's face and on the faces of others. Moreover, they search the link between the observer's MRs and their smiling manner, as well as their personality's traits.

1.4 Thesis Organization

The document is organized as follows :

— **Chapter2** conducts an overview of expressions synthesis, especially for the joyful expression. First, we introduce the smile and it's morphology and neurology control. We explore how the facial movements influence human emotional state. Second, the existing systems for expression synthesis for evoking emotion are discussed. Thirdly, a review on the state of the art synthesis expressions methods is presented. Finally, we give our perspective on this research domain and the contributions of our method.

— **Chapter3** details our system that can synthesize personalized expression using a hybrid Geometric-Machine Learning method. One originality of our method is that we use previous knowledge of the smiling way of each person to personalize the synthesized expression. The second originality is that we use a hybrid geometric-machine learning method to synthesize the personalized joyful expression. Our framework is composed of 3 parts : a geometric part to build a person-specific model, then the model is used to generate personalized expressions with different intensities.

And eventually, we use a GAN to refine the textures of the generated expressions. The third part is the teeth region personalization. First sections of this chapter present these 3 sub-steps of our hybrid method. Finally, we conclude the chapter.

— **Chapter4** presents the experimental results of our method. The experimental results are analyzed regarding our contributions and compared with recent state of the art methods. This chapter is organized as follows : first, we introduce the databases used to conduct the tests. Secondly, we justify with a statistical study why we need to personalize the synthesized joyful expression. This study shows that the smile is specific and occurs differently for each person and justify why we proposed an algorithm able to learn a person-specific model for each person and transform a captured face to appear more joyful according to his own expression. Thirdly, we present qualitative and quantitative results on shape and texture with our hybrid geometric-machine learning method. They show that the joyful expression synthesized with our method is closer to the ground truth than the expression generated with two generic methods. Finally, a brief conclusion is given at the end of this chapter.

— **Chapter5** present a collaborative work with IRCAM which aims at investigating the way that has each person to perceive the expression of joy on her face and on the faces of the others. Therefore, we propose a tool to build mental representations (MRs) of joyful expression on real participants. We give first a review of mental representation research. Second, we detail the visual deformation system used to preparing the data for a perception experiment. Third, our experimental protocol for determining participants' MRs is detailed. Then we present the determined MRs. These MRs are analyzed by the IRCAM, we give in the last section a brief details about the found results. These results are detailed in Appendix A.

— **Chapter6** summarizes our contributions in this thesis and presents the perspectives of future work.

Chapitre 2

State of the art on smile expression synthesis on real faces

As presented in chapter 1, we are interested in synthesizing joy facial expressions to positively act on the observer's emotional state. This chapter conducts a review on facial expressions, focusing on the smile. First, we explore how the smile influences human emotional state. Secondly, the Smile/Sad expression synthesis for evoking emotion is discussed. Thirdly, a review on expression synthesis methods is presented. Finally, we conclude the chapter with our perspectives on this research domain and the contributions of our system.

2.1 The Smile expression

There are several muscles in the face, most of which are controlled by the seventh cranial nerve (The facial nerve) [99]. In this section, we present at first the neurological and muscle control of smile, then the different types of smiles and the influence of facial expressions on the human emotional state. Finally, we define the mirror effect and the emotions contagion, on which we built our system.

2.1.1 Neurological and muscle control

The facial nerve [99] divides into three bundles and then into several sub-branches to respectively innervate the upper, middle and lower face muscles as illustrated in Fig. 2.1 ; that allows the face to twist and contort into a variety of expressions. Each expression can be described with a set of action units using the Facial Action Coding System (FACS).

Indeed, **FACS** is a system that decomposes facial movements into units of actions [32]. It dissects the expressions observed and classifies them into actions made from one or more facial muscles. These actions are called AU (Action Unit). According to FACS, there are around 30 main AUs, whose combinations make it possible to describe around 10,000 facial expressions [26]. From a set of AUs, it is possible to reconstruct a coded facial expression and assign a meaning to it. The intensities of AUs are annotated by appending letters A–E (for minimal-maximal intensity) to the action unit number (e.g. AU1A is the weakest trace of AU1 and AU1E is the maximum intensity possible for the individual person). There are other letters used, such as "R" which represents an action that occurs on the right side of the face and "L" for actions which occur on the left side.

Emotional facial expressions have been the subject of numerous studies related to FACS coding [71, 17, 44]. Seven emotional facial expressions are considered universal and go beyond ethnic, cultural and sexual considerations [30]. The first who talks about the universality of expressions and emotions is Charles Darwin [18]. He confirmed that smile is present in all cultures, in different situations, such as salutation and seduction. Many years later, experiments were carried out by Paul Ekman [25] with a judgment study, to determine whether the expressions (including the smile) are universal or cultural : people of various nationalities were photographed, expressing joy, sadness, anger, disgust, surprise, and fear. People of different ages and cultures have, for the most part, recognized each of them well, out of empathy. These results show the universality of the emotions. The universality hypothesis has been studied also by E. Jack et al. [51]. They showed that facial expressions of emotion are culture specific, refuting the notion that human emotions are universally represented by the same set of distinct facial expressions. For each facial expression, there are therefore prototypes of these emotional facial expressions. The table 2.1 presents the prototypes of the 7 basic emotions.

In the smile expression [28] there are 2 potential muscles are activated. The first are the zygomaticus major and minor muscles which control the corners of mouth (AU12). The second muscle is the obicularis occuli (AU6) and it encircles the eye socket as shown in Fig 2.1. The movement of these muscles causes an excitation of the anterior part of the hypothalamus, a gland located at the base of the brain. Like a wave, it transmits a nervous impulse to the limbic system, the seat of the emotions. Therefore, the brain releases molecules called neuropeptides, which work toward fighting off stress [107, 67]. Other neurotransmitters like dopamine, serotonin (an antidepressant) and endorphins are also released when a smile happens. Then, the muscles relax and the facial reactions of contentment appear.

TABLE 2.1 The 7 basic emotions with their considering action units combinations [28].

Emotion	FACs AUs codes
Happiness	6 +12
Sadness	6+15 1+4+15 1+4+11+15B
Anger	4+5+7+23 4+5+7+17 with 25 4+5+7+17+24 4+5+7+23+25 with 26
Fear	1+2+4+5+20 with 26 or 27 1+2+4+5+25 with 26 or 27
Surprise	1+2+5B with 26 1+2+5B with 27
Disgust	9+15+16
Contempt	R12A+R14A

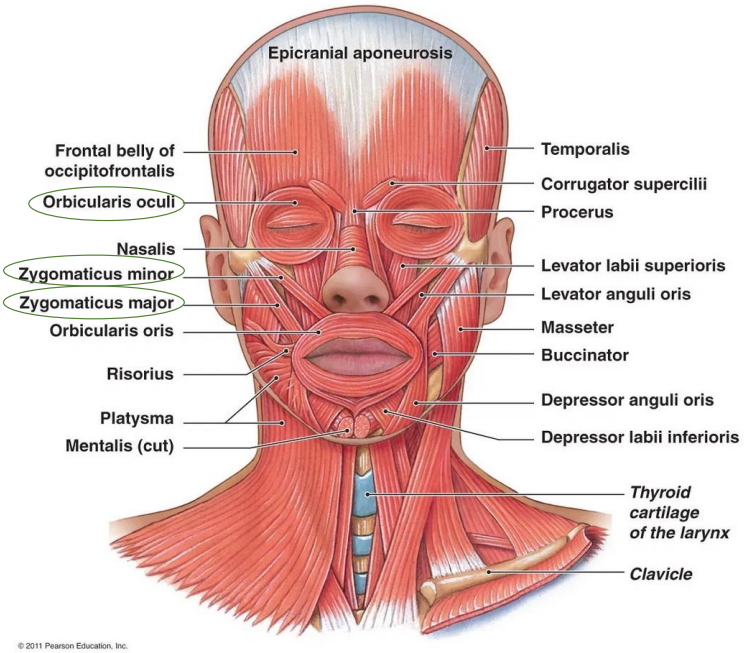


FIGURE 2.1 The upper, middle and lower face muscles. In green ; the activated muscles during a smile : the major and minor zygomaticus muscles and the orbicularis oculi muscle.

2.1.2 Different types of smiles

Ekman and Friesen have shown in their study [29] that the enjoyment smile (Duchenne's smile) has distinct markers that differentiate it from other types of smiles. For example, the presence of orbicularis oculi action in conjunction with the zygomatic major is essential in the Duchenne's smile with a symmetrical action of the zygomatic major on both sides of the face. In [36, 29], they have shown that the enjoyment smile seems to be associated with positive emotion, whereas other types of smiles are not.

There are many types of smiles which are adapted to diverse social situations to express a wide range of emotions. Many classifications are possible, according to the intensity, the situations of appearance, the interior state of the individual. This has been proved by several studies [29, 3, 12, 36, 100]. For example, Duchenne's smile appears more often when watching enjoyable movies [27] and is associated with activity in the frontal and left anterior temporal lobes, which are areas related to positive affect. In that way, this smile can positively affect the emotional state of the person and the moods of those around. Moreover, [100] showed that reward smiles are symmetrical and accompanied by eyebrow raising, affiliative smiles involve lip pressing, and dominance smiles are asymmetrical and contain nose wrinkling and upper-lip raising as shown in Fig. 2.2.

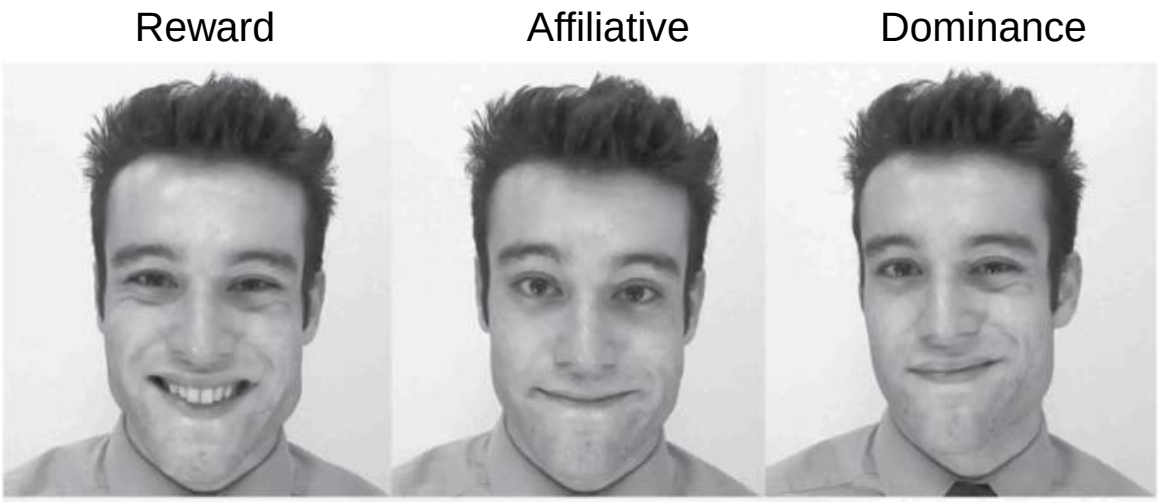


FIGURE 2.2 Image frames showing the difference between reward, affiliative, and dominance smiles. Reward smiles are symmetrical and accompanied by eyebrow raising, affiliative smiles involve lip pressing, and dominance smiles are asymmetrical and contain nose wrinkling and upper-lip raising [100].

In [3], they investigated also the difference between 3 types of smiles. In their studies, the observers judged more than one hundred smiles as perceived amused, embarrassed/nervous,

or polite. Their results showed that perceived smile meanings are related to a specific variation in smile morphological and dynamic characteristics. They proved that amused smile includes usually AU6 (check raiser) with an open mouth, a larger amplitude, a larger maximum onset and offset velocity, and a longer duration in comparison with polite perceived smile. However, the amused smile includes lower maximum offset velocity, and smaller forward head pitch in comparison with embarrassed/nervous perceived smile.

2.1.3 Influence of facial movements on the emotional state

The two psychologists William James (1842-1910) and Carl Georg Lange (1834-1900) were the first that have defined the James-Lange theory [68, 53, 54]. For them, emotion reflect is a response to physiological changes. Lange said "We feel sad because we cry, angry because we hit someone and scared because we tremble." Therefore, the emotional experience would come from physiological feedback of our movements and comportment. That reasoning postulates that the production of certain facial expressions provokes particular emotions because of the physiological link to the reflex reactions. As said also James “We don’t laugh because we’re happy, we’re happy because we laugh.”. In the same context, Ekman has shown [31, 25] that facial expressions have autonomous consequences, such as changes in heart rate, skin temperature, skin conductance, and blood volume. These involuntary reactions are witnesses to the emotion felt. The positive effect is amplified if the subject imagines something pleasant. In another side, Walter Cannon (1927) [14] disagreed with the James-Lange theory of emotion. First, he suggested, emotions could be experienced even when the body does not reveal a physiological response. In other cases, he noted, physiological responses to different emotions can be extremely similar. For example, your heart might race because you have been exercising, not because you are afraid. He stated that the experience of emotion does not depend on interpreting the body’s physiological reactions. Instead, he believed that the emotion and the physical response occur simultaneously and that one is not dependent on the other. Another well-known physiological theory is the Schachter-Singer (1962) [104] theory of emotion. This theory states that the physiological arousal occurs first, and then the individual must identify the reason for this arousal to experience and recognize it as an emotion.

The theory of Emotional Feedback Hypothesis (EFH) states that our emotional experiences are under the retroactive influence of our own expressions [52, 82]. In emotional feedback studies [11, 19] facial expressions have defined as a means of affective communication for humans. Facial Feedback Hypothesis (FFH) has two complementary effects : posing a facial expression should increase the intensity of the emotional experience, and inhibiting

facial expressions should decrease it [19]. Strack et al. [114] has conducted studies requiring participants' action to analyze the influence of facial muscle movements on the human mood in different situations. They instructed participants to rate the funniness of cartoons using a pen that they held in their mouths as shown in Fig. 2.3. In line with the facial feedback hypothesis (FFH), when participants held the pen with their teeth, they rated the cartoons as funnier than when they held the pen with their lips. These results have shown that both the inhibitor and facilitator mechanisms have contributed to the observed affective responses.

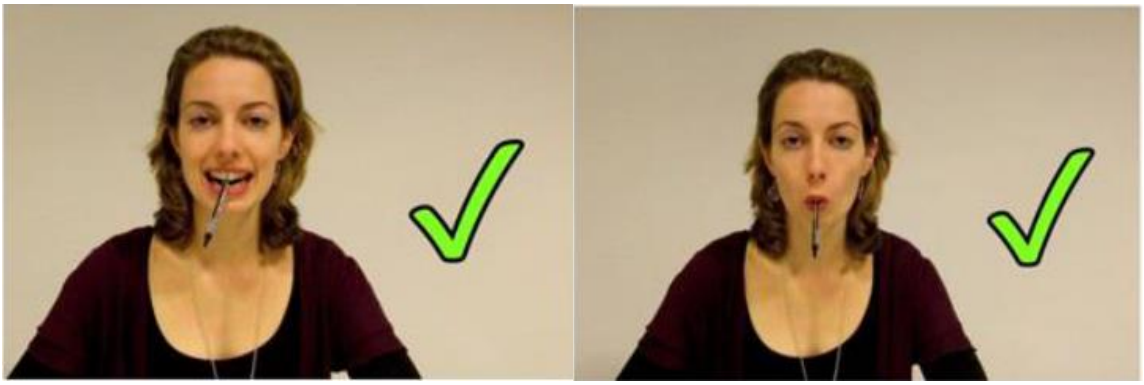


FIGURE 2.3 A common experimental manipulation used in emotional feedback experiments : participants are instructed to either hold a pen in their teeth (left ; simulating a smile) or between their closed lips (right ; simulating a frown) [91]. They found [114] that people holding the pen with the teeth (smile position, simulation of the muscles responsible for putting a smile) annotated the comics more funny and reported more intense humorous responses when the cartoons were presented compared to the other group. This demonstrate that facial expressions could influence our emotional experience and our emotions

As part of studying facial expression reactions and its influence on emotional experience, Davis et al. [19, 81] tested this facial feedback hypothesis (FFH) by comparing the impact on self-reported emotional experience after injections of BOTOX which paralyze facial expression muscles and injections of Restylane HA which is a cosmetic filler that does not affect facial muscles. They assessed the emotional response to positive and negative video clips before and after treatment. They predicted that if facial feedback can influence the emotional experience, people who received BOTOX injections would show a decrease in self-reported emotional experience compared to any changes demonstrated by people who received HA injections. The comparisons between the results of the groups showed that the participants in BOTOX condition had a significant overall decrease of the emotional

experience. Botox reduced facial feedback by reducing the accuracy of the perception of facial emotions of observers [81] and it affects the ability to pick up on others' subtle facial cues and to empathize with their emotions. Eventually, their experiment attests the decrease in responses to positive clips in the BOTOX group and an unexpected increase in responses to negative clips in the Restylane group. These data suggest that feedback from facial expressions can influence the emotional experience in certain circumstances.

Based on the effects of facial expressions on the emotional state, several researches have been conducted to study the link between stress and facial expressions [79, 63, 70]. Tara et al. [63] studied whether manipulating facial expressions would influence cardiovascular and emotional responses to stress. Participants were asked to accomplish two different stressful tasks by holding chopsticks in their mouths in order to produce a Duchenne smile, a standard smile or a neutral expression. Sensitivity was manipulated by explicitly asking half of the participants to smile and giving the other half no smile instruction. The results revealed that all of the smiling participants, regardless of whether they were aware of a smile, had lower heart rates during stress recovery than the neutral group with a slight benefit for Duchenne's smile group. Moreover, participants in smiling groups who were explicitly asked to smile, reported less decrease in positive affect during a stressful task than the neutral group. These results show that there are physiological and psychological benefits of maintaining positive facial expressions during stress.

Based on those researches, Laughter Yoga is born. The indian physician Madan Kataria published a book called "Laugh for no reason" [59], in which he explained the Laughter Yoga benefits. He argues that our body cannot differentiate between the two : fake laughter and real laughter, so the two expressions promote the same benefits to our body. Moreover, Woodbury [129] explains that the laughter in a positive social setting where everyone else is also laughing (with Duchenne smile) could produces physiological and biochemical changes that include the stimulation of oxytocin and promoting well-being. Keltner et.al [61] declared that laughing is an effort that after a while turns into a Duchenne laugh and therefore becomes spontaneous and even contagious.

2.1.4 Mirror effect and emotional contagion

Another main aspect of emotion is contagion. The imitation is pervasive and automatic in humans. During a conversation [46, 56, 115], humans unconsciously tend to imitate the facial expressions, posture, body language, and tone and rhythm of speech of the other person. The mirror neuron system activity (MNS) [49] is described as a pre-reflective, automatic mechanism of mirroring which consists of the activation of the same area of neurons when

observing an action and when doing it. Wicker et.al [125] have shown that the same neurons activate when experiencing disgust and when perceiving it in the face of another person. According to Hatfield [47], the emotions recognition of others could be fulfilled by the facial feedback signals generated when we automatically imitate the expressions displayed on their faces. Moreover, these muscular reactions can lead to feeling the same emotion by activating the same neurons as if the emotion were experienced personally.
With the effect of contagion of emotions, the facial muscles involved in the smile can be put in action and these muscular movements can then send the signal to the brain of the felt emotion. According to [37], the smile expresses pleasure and fun. It causes a reduction in stress and a brief feeling of relaxing and happiness [63]. Our natural instinct [128] to reproduce each other's facial expressions would help us be more empathetic and feel the same emotions as the person we are imitating. So when a person in front of us smiles, we tend to smile too. Therefore, positive emotions like enthusiasm and joy (even negative ones) pass easily from person to person, often without their being aware of it, and the emotional contagion occurs within milliseconds. So, when we see someone smiling, our mirror neurons for smiling fire up [47], and create a sensation in our own mind of the feeling associated with smiling.
The emotional contagion is not only transmitted from person to person, but it can be realized by the person herself via a mirror for example by seeing her smiling face. To understand this phenomenon, a prototype has been developed [138, 137] allowing to deform the face of a user with a smile or sad expression in real-time as demonstrated in Fig. 2.4 (A). The participant was asked to click on targets that appeared around her own face, which was shown on the display, with a mouse. Each participant performed three sets of mouse-click task sessions and evaluation sessions. During each mouse-click task session, participants were presented with one of three facial expression changes (Sad face, neutral face and smiley face). At the end of the experiment, participants should explain what they felt and noticed during the experiment to see if they were aware of the deformation. Their results indicate that a change in facial expressions influences the emotions of individuals. Most of the participants were unaware of the study goal.

The next section presents how researchers have used facial feedback and the emotional contagion theories in real applications.

A B

FIGURE 2.4 (A) : The facial expressions deformations Smile face and Sad face .(B) : Experiment protocol for the facial deformation in real-time [138]. This study proves that user's emotional state is influenced by the feedback of deformed facial expressions

2.2 Health and well-being applications of smile synthesis on emotion analysis and manipulation

All the cited studies on facial feedback have shown that the hidden induction of a smile leads to congruent changes in the participant's emotional reactions [114]. A person seeing himself smiling and radiant will naturally appreciate this moment while if his face appears dismal and sad, he will be negatively affected. As a result, our perception of ourselves plays a key role in our mood. In this section, we present how this theory changes user feelings and improves their daily lives basing on health and well-being applications.

2.2.1 Health applications

Series of work has studied clinical cases in which facial expressions were impaired and how their modified feedback affected the mood of the patients. [120] tested a group of people with varying degrees of facial paralysis and found that those who had specific impairment in the ability to smile rated highest on depression scores. These patients experience reduced physiological feedback associated with smiling as well as the social consequences of the inability to communicate positive emotion. Therefore, the impairment in the ability to smile elevates the risk for depression. Depressed individuals can be recognized by their depressed facial expression, in which the facial muscles create a distressed or sad appearance. With

a similar rationale, [34] treated 10 people who had symptoms of depression with Botox injections in the forehead ; thus artificially paralysing muscles responsible for frowning. They have used the botulinum toxin A to treat the glabellar frown lines on faces of these depressed patients. Their results show that muscles and skin movement in the face positively contribute to the mood and emotion of 9 patients of 10 and found this improved their depression.

Rickard et al. [96] have used various forms of auditory and visual feedback in the context of emotional memory disorders analysis. They have shown that relaxing music played after exposure to negative stimuli can reduce the memory consolidation of unwanted material. Their findings provide some promise for the use of music as a potential treatment for decreasing the negative impact of emotionally laden information, including those associated with some forms of PTSD.

2.2.2 Well-being applications

The creativity is the ability to find original solutions. It is attributed to cognitive processes, the social environment and personality. To improve creativity, Nakazato [80] proposed a system that can modify in real-time the face of interlocutors, during video conferences as shown in Fig. 2.5. The evaluation of the creativity task have focused on the number of responses to an open problem during video conferences with a modified face with a smile expression and without. This application made the participants appear more smiling, more friendly and aimed at providing quality work in optimal conditions of satisfaction.

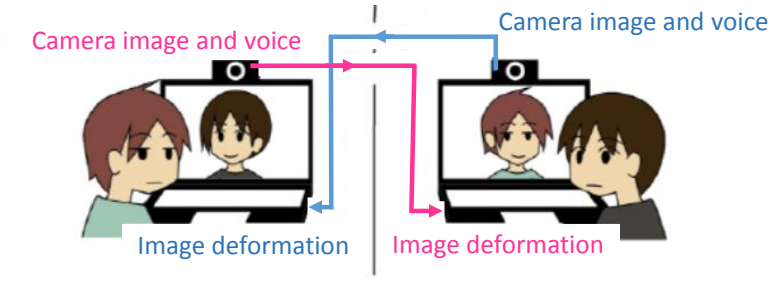


FIGURE 2.5 Deforming the interlocutors faces with a smile expression during a brainstorming task in video conference [80]. The results proved the changes in the number of brainstorming responses to validate how the system positively affected the level of creativity.

In [115] Suzuki et.al conducted an experiment in which pairs of participants had brief conversations via FaceShare. That system can deform in real-time a user's facial expression into a smile, congruently with that of their partner's smile. The results proved that mirroring

using the pseudo-smile improves the flow of the conversation for both the mimicker, whose face is deformed according to the expressions of his partner and the mimickee, who perceives the mimicker's deformed face.

Another example is the system proposed in [138]. They used their framework of face deformation with smile/sad faces to manipulate the preferences of the users. They tested whether the preferences of an individual can be manipulated through a face deformation. While deforming their faces, these users had to choose their favorite scarf. 8 different scarves were used and participants had to choose their favorite from different scarf selections. The results showed that the majority of participants like the scarf with which they saw themselves artificially smiling.

With all these empirical demonstrations of emotional feedback mechanisms [114, 121, 115, 80], there are still a lot of possible applications of these mechanisms in health or well-being contexts such as clinical applications for PTSD patients.

2.3 Facial expressions synthesis on real faces

The applications cited above are based on expression synthesizing methods. During the last two decades, several researchers have been working on facial expressions generation on real faces. These works can be divided into two categories. The first category mainly resorts to computer graphic techniques. These methods directly deform detected landmarks of the face to generate the expressive face. The second category aims at building machine learning frameworks such as generative models to synthesize facial expression images. In this section, we give more details on those 2 categories.

2.3.1 Geometric methods

Facial expression synthesis is an act of generating new face shapes from a given face. This process should preserve the person identity and the distinct facial characteristics of the initial face. The traditional methods are based on geometry controlled image deformation. These methods use the feature difference between the source and target expression images to generate feature positions for a new face. In this sub-section, we present an overview of some of these geometric methods.

2.3.1.1 Interpolation techniques

The interpolation techniques are among the first geometric approaches used for facial expression synthesis [66]. This morphing technique is able to determine smooth motion between two key-frames at extreme positions, over a time interval. Interpolations are fast and easily generate primitive facial expressions. However, their ability to create a wide range of realistic facial configurations is severely restricted [89, 4]. Moreover, the results are poor because the two input images are used to generate the output facial expressions [126] as figured in Fig. 2.6. The first frames in the sequence are similar to the source image and the last frames are similar to target image. Middle image of the sequence is the average of the input and output frames.

FIGURE 2.6 Expression generation using interpolation [126]. Interpolation can only generate expressions in-between the given expressions but cannot be used to generate expressions for a new face.

2.3.1.2 Warping techniques

To overcome some of the limitations and restrictions of interpolations, researchers moved towards warping. Warping achieves a fluid transformation by incorporating 2D meshes to maintain geometric alignment throughout the deformation process. The warping performs the transformation between two target images or models based on different algorithms such as mesh warping [87, 40], field morphing [8] or multilevel free-form deformations [108, 57].

The mesh 2D warping algorithm is the most used technique for face deformation [122, 23] and facial expressions synthesis [87, 40, 39, 134]. It produces realistic facial expressions using shape and texture models based on the motion between corresponding points in the two facial images. The 2D warping can be used for creating one target person's expression

face from her neutral face. Many algorithms can be used to automatically determine the motion or the deformations between two expressions such as the active appearance model AAM. An AAM is the 2D shape represented by a triangulated 2D mesh with a set of vertices. Therefore the shape vectors consists of the 2D coordinate of the vertices that make up the mesh which is used for the warping process. In [87], to generate an extreme (amplified) expression, they extracted at first a motion vector of predefined feature points (between a neutral and an expressive face). Then, they deformed neutral facial expression into the extreme facial expressions by multiplying the motion vectors of feature points by the wanted amplifying vector and added it to the Active Appearance Model (AAM). Finally, the extreme facial expression image is recovered by the piecewise affine warping method [126]. The Active Shape/Appearance Models (ASM/AAM) have achieved good results in facial feature extraction [39, 40, 87]. Lee et al. [69] used these models with a thin plate spline (TPS) warping method [98] to synthesis different facial expressions. The ASM, can globally match the object shapes very well, but with lower local accuracy, i.e.the performance deteriorates under local shape variations. Xu et al. [133, 134] used a Bayesian Shape Model (BSM) for extracting facial features. Then, facial warping is performed using piecewise affine algorithm based on the mean face model. Unfortunately, the proposed system ignored the personal characteristics of each face which led to synthesizing expressions which are not personal and lack texture details as illustrated in Fig. 2.7.

2.3.1.3 Warping in YUV domain

The warping result could be affected by the variation of illumination and noise during extracting facial details such as facial wrinkles from source or target image in spatial domain. To solve this problem, [43, 118] used the YUV color space to manipulate facial expressions. For example in [43] to synthesize the target expression they converted the Y channels of the two expressions into the frequency domain by discrete cosine transformation. Then, the expression details including the micro-expressions of the source subject are extracted by analyzing their frequency distribution. Finally, the expression details are assembled to the target Y channel by the Poisson editing method. The transferred result is combined with the U and V channels of the distorted target expression to synthesize the final expression. The proposed method is robust to the variations of source subject, target subject and expression category. However, the result expression looks not so much like the real expression (smile) of the person. People often have very different local features on their faces and their specific characteristics are still absent in the expression results. For example, consider a source person whose mouth is small in the neutral face and whose mouth becomes much larger in the

FIGURE 2.7 First row : the original expressions. Second row : the warped face images. The synthesized expressions using the mean face model are photo-realistic but are not those of the person and lack texture details [134].

smile expression face. If they apply this smile expression to a target person whose mouth is already large in the neutral face, the warping method here leads to an unnatural large mouth. According to our work, the limitation of this approach is that the techniques are far from permitting either synthesizing real expression or real-time deformations.

2.3.1.4 Real-time warping techniques

Researchers in [138, 137, 80] proposed real-time frameworks to manipulate and synthesize real facial expressions. They used predefined models to generate sad and smile expressions. Their models are based on translation. They use FaceTraker library [103] to detect the face in real-time. They deform the detected face to create the smiling one by raising the corners of the mouth and cheek and lifting the lower eyelids. For the sad face, they create it by lowering the corner of the mouth and lifting the inside of the upper eyelid of the eyebrow. The deformations are obtained by a predefined translation of the landmarks and using Moving Least Squares MLS [105]. The main limitation in this researches is that the face changes look similar for all subjects. Moreover in some works [138, 137] the subject should not change his head position during the experience. This constraint is improved by the framework of Arias et al. [5]. They proposed a system which provides audiovisual real-time

transformations. The proposed system adapts the position of the user's face and deforms it to appear more positive. Yet, the deformations are still based on the same model for all the subjects. However, each person has his own way to make expressions. Recently, researchers [117, 116, 115, 15] have been working on synthesizing real-time facial expressions using 3D models. In [115], the smile expression is generated in real-time by means of 3D deformation using the detected face landmarks (the contours, eyes, nose, and mouth). They deformed a user's face into a smile by lifting the lower eyelids, both ends of the mouth, and the cheeks [28]. They applied a 3D deformation method [145], which changes the positions of the mesh vertexes of 3D objects by manipulating a set of point handles as illustrated in Fig. 2.8. Finally, the moved mesh vertexes are projected onto the original 2D image, and the texture is re-pasted. The technique is robust to changes in facial rotation and distance from the camera because it uses depth information. However, the deformations in these researches are still based on translation and the generated expressions are not personalized.

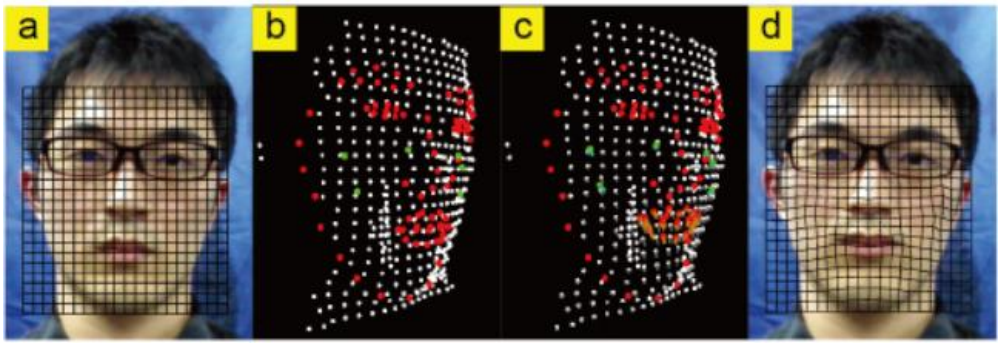


FIGURE 2.8 a : The Face is divided into a mesh. b : The detected landmarks are defined as the point handles. c : The point handles are moved according to predetermined parameters, and the mesh vertexes are then moved accordingly. d : The moved mesh vertexes are projected onto the original 2D image, and the texture is refined [115].

The results of warping methods depend first on the efficiency of the expression details extraction from the source subject and second on the blending of these details to the target subject. Most of the existing geometric warping algorithms are performed where the feature positions in the target expression are moved "relatively" according to the movements of their corresponding feature positions in the source person's face. The source subject usually is another person or a mean model expression on a database. However, each person has her own way to make her expression. Therefore, all the generated expressions may be realistic but are not real.

2.3.2 Machine learning methods

Geometric methods demonstrate promising results, even of high-resolution. However, the face warping phase depends heavily on locations of detected face landmarks. More importantly, the identity characteristics in the given face cannot be well protected due to global or local texture warping and rendering. Recently, generative deep learning techniques have been explored in this topic using Variational Auto-Encoder [136, 144, 135] and generative adversarial networks (GAN) [131, 124, 86, 48, 21].

In order to preserve as much detail as possible and ensure the generation of real expressions, Goodfellow et.al proposed a new generative model [41] to synthesize plausible facial expressions as figured in Fig 2.9 . The main idea is to train paired generator and discriminator networks at the same time, where the goal of the discriminator is to classify between "real" images and generated "fake" images, and the generator aims at fooling the discriminator so that the generated images are indistinguishable from real images. Once trained, the generator is used to synthesize expressive images driven by a compact vector of noise. As illustrated in Fig 2.9 the generated images lack fine details and tend to be blurry or of low-resolution. It is a hard task for a simple GAN to take fine-grain control of the synthesized images, e.g. widen the smile or narrow the eyes. To handle the resolution limitations, Yeh et al. [136] proposed a variational auto-encoder to learn a flow map between two different expressions (source and target faces). Then they applied the flow map learned to transfer the expression to a new subject's face. Although the generated face image has a high resolution, nonetheless the synthesized faces are poorly detailed mainly because identity cues are more or less lost. With a focus on preserving the person identity [144] proposed the conditional difference adversarial autoencoder (CDAAE) with a long-range feedforward connection from the encoder to the decoder. The network allows to decouple identity and expression information. Though the method preserved the person identity information, the generated facial images are of low resolution (64×64) and some expression details (teeth and wrinkles) and hair information are lost as illustrated in Fig. 2.10.

With an unconditioned GAN [41], there is no control of the data being generated. To this aim, researchers in [76, 112, 72] proposed Conditional generative adversarial networks (CGAN) to generate photo-realistic facial expressions while providing continuous adjustment for the target expression. By conditioning the model on additional input information, it is possible to direct the data generation process. The condition information should be based on a class label or on some part of data to reinforce the generation process [76]. For example,

FIGURE 2.9 Samples generated using the model trained with Toronto Face Database [41]. The GAN is able to generate expressive images from a noise input vector however the generated expressions lack fine details and tend to be blurry or of low-resolution.

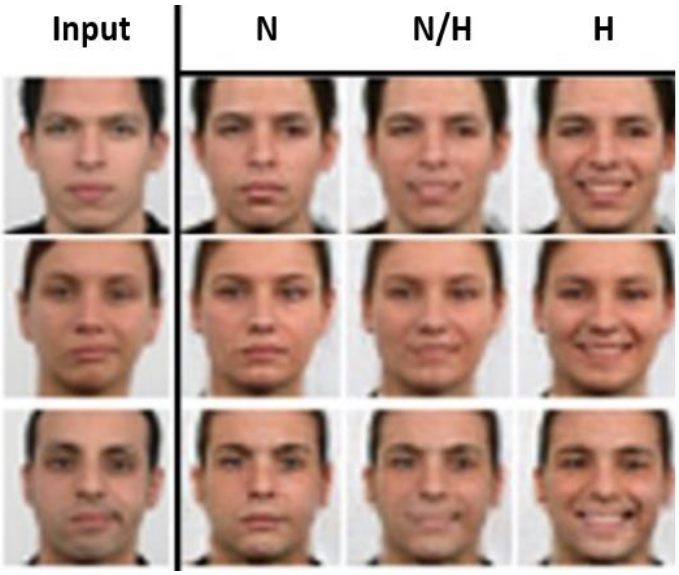


FIGURE 2.10 Synthesis results of N :neutral and H :happiness classes and their interpolation using the CDAAE network [144]. the proposed system preserves the person identity but the generated expressions are of low resolution and some expression details such teeth and hair are lost.

Olszewski et al. [86] proposed a method to enable end-to-end synthesis of high-resolution facial textures conditioned on the target identity and source facial expressions. Ding et al. [22] proposed an Expression Generative Adversarial Network (ExprGAN) for photo-realistic facial expression editing to control both the type of the expression and its intensity simultaneously. However, the generated expressions tend to be a little blurry, do not have much expression details and miss the original person's identity. To preserve the face shape and person identity Song et al. [112] proposed a G2-GAN to synthesize photo-realistic facial expressions using the face geometry information in active appearance models (AAM) to control the expression synthesis process. The GAN is conditioned by the target expression geometry to control the expression synthesis process. However, the generated expressions tend to be blurry and lack of details as illustrated in Fig. 2.11

FIGURE 2.11 Results of Oulu-CASIA database for facial expression synthesis [112].The proposed G2-GAN synthesize photo-realistic facial expressions. However the system need fiducial points the original expression, which is difficult to operate in practical usage. Moreover, the generated expressions tend to be blurry and lack of details.

To address this issue, Lu et al. [72] have proposed a novel Couple-Agent Face Parsing based Generative Adversarial Network (CAFP-GAN) for intensity-controllable facial expression synthesis. The network requires neutral facial images and labels for various expressions

synthesis. The controllable condition in this work is a face parsing map of the target expression. Such a condition can provide a robust representation of every pixel of facial parts to efficiently manipulate expressions. However, the synthesized expressions are not those of the person, the emotion identity cues are lost as figured in Fig. 2.12

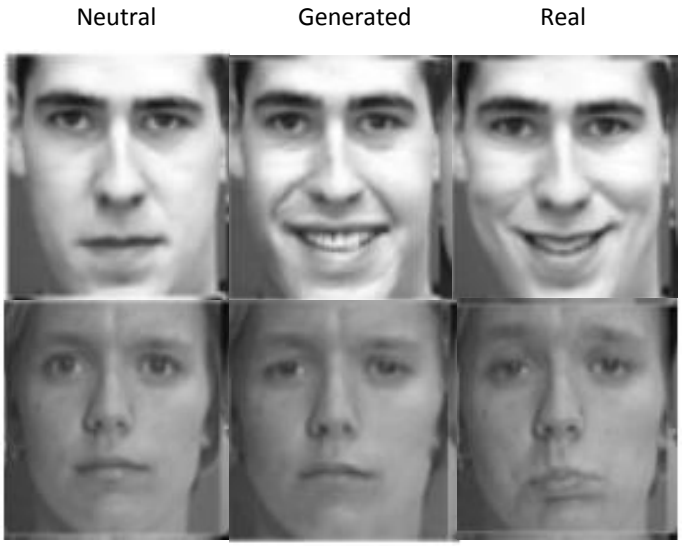


FIGURE 2.12 Results of CK+ database for facial expression synthesis. First column the input neutral expressions. Second column the generated expressions using CAFP-GAN and the third column is the real expressions [72]. The proposed model could efficiently manipulate expressions. However, the synthesized expressions are not those of the person, the emotion identity cues are lost.

The proposed models based on GANs achieve an impressive texture refinement and allow to synthesize different photo-realistic facial expressions. However, the deformations are based on a model that is learned on several subjects. So that the synthesized expressions are not the subject's own. Note That our goal setting is different from the one tackled by most image generation methods found in the literature. These methods attempt to manipulate facial expression, but the full complexity of natural human expressions and identity-specific characteristics are hard to model only by generative networks. These networks do not systematically model the intricate correlations between the expressions and facial movements for each person.

2.4 Conclusion

Nowadays, various geometric and machine learning methods are employed to synthesize facial expressions. The geometric techniques offer high-performance shape deformation but lead to images lacking texture details such as wrinkles and teeth. Systems proposed in [5, 138, 80] provide real-time visual transformations. Yet, the deformations are still based on the same model for all the subjects. However, each person has his own way to make expressions. The main limitations in these researches is that the face changes look similar for all subjects and the synthesized expressions lack texture details.

On the other hand, the machine learning methods (e.g. Generative Adversarial Network GAN) have recently obtained impressive results for expression synthesis. These models are able to generate photo-realistic expressions and add texture details to the images. Nevertheless, the synthesized expressions are not those of the person. The expressions generated in these researches come from generic models which are learned on several people. The generative models learn different ways of smiling (learning database) but without any prior knowledge of a new subject in the test database, the GAN cannot guess the own way of smiling of that subject. So, these models do not take into account the fact that people make their expressions in different ways.

As the expressions are highly personal and each person smiles differently, our challenge is to synthesize joyful expressions that are specific to each person. To do so, we personalize at first the shape using a geometric strategy and then we refine the global texture of the generated smile using a machine learning method (GAN). To preserve the identity of the emotion and to generate person-specific expressions, we use a previous knowledge of the way of smiling of the subject to train our proposed GAN. Finally, we propose two solutions to refine local expression details of the generated expression. Details of our hybrid method are given in the following chapter.

TABLE 2.2 Summary of the advantages and limitations of the state-of-art methods.

	Method	Advantages	Limitations
Geometric	Interpolation techniques	Fast and easily generate primitive facial expressions.	Restricted ability to create a wide range of realistic facial expressions.
	Warping techniques	Fluid transformations. Realistic facial expressions.	Based on the same model. Ignored the personal characteristics of each face which led to synthesizing expressions which are not personal and lack texture details.
	Warping techniques in YUV	Robust to the variations of source subject,target subject and expression category.	Local features of the target faces and their specific characteristics are still absent in the expression results.
Machine learning	Variational Auto-Encoder	Easy to sample a latent space for expression generation or interpolation.	Generated images are often blurry. The generated expressions are not those of the person.
	Generative adversarial networks	Generate photo-realistic expressions and add texture details to the images.	The synthesized expressions are not those of the person. The models are learned on several people which ignore the specific characteristics of each subject.

Chapitre 3

Synthesis of personalized joyful expression using our Hybrid Geometric-Machine Learning method

As introduced in chapter 2, the geometric methods used to synthesize smile expressions apply the same deformations on the face of all the subjects, thus preserving their own identity but not their specific way of smiling. Now there is no question that each person has her own way to make an expression. Similarly generative models have recently obtained impressive results for image synthesis applications and generate natural expressions but not personalized ones. Our goal in this thesis is to propose a system for evoking positive emotion by synthesizing facial expression. As the expression are personal, so we are led to personalize the synthesis expression to preserve the emotion identity of each person and to keep the credibility. To do so, we propose a hybrid geometric-machine learning method for synthesizing person-specific joyful expression. Our system learns a person-specific model using a neutral and an expressive frame and generate joyful expression with manipulated intensity. We refine global-texture using a generative model, "Generative Adversarial Network" GAN basing on prior knowledge about the way of smiling of each person. The local-texture such as teeth are then personalized.

Our framework is illustrated in Fig. 3.1. We first propose a geometric method to personalize the shape of the synthesized expression (section 3.1). As the synthesized expression lacks details such as wrinkles and teeth, we use a GAN (section 3.2) to refine the global-texture details. Yet, the synthesized teeth are not those of the person. To address this limitation, we propose refining such a local-texture details in the mouth region by adding the teeth of the

person. To this aim, we test and compare two methods : the poison editing method and an auto-encoder (section 3.3). The details of our hybrid geometric-machine learning method are explained in this chapter.

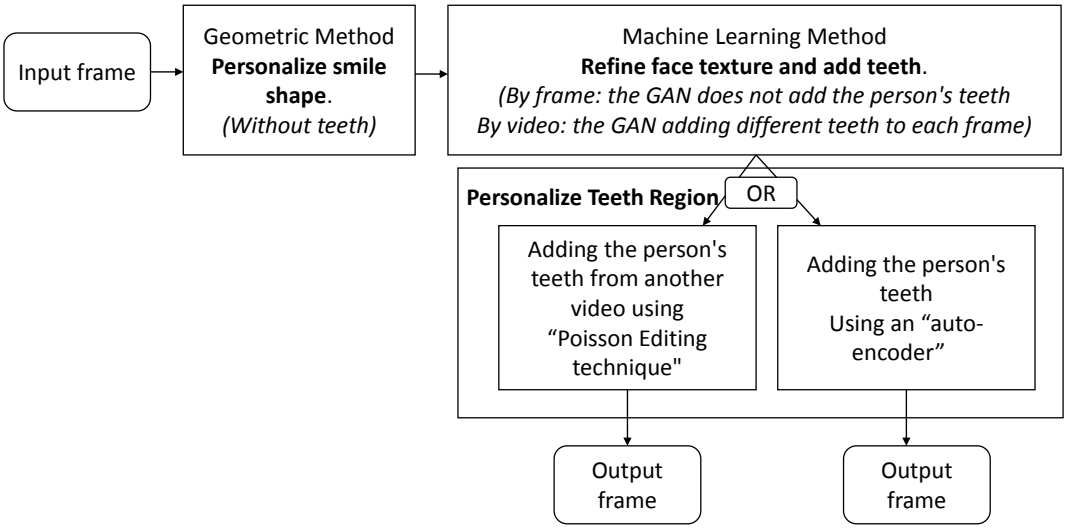


FIGURE 3.1 To generate expressions close to the real expression of each person, we propose a framework composed of 3 parts. The first part aims at personalizing the shape expression (geometric part). The second part consists on global texture refinement using a generative model (GAN). The third parts allows to personalize the refined texture using either an Auto-encoder or a Poisson editing method.

3.1 Geometric method : Shape personalization of the smile expression

As the expressions are personal and each person smiles differently, our first contribution is personalizing the generated expression **shape**. Our system synthesizes person-specific joyful expression and preserves the identity of the emotion by reproducing the specific way of smiling of each subject using a geometric method. To do so, we use previous knowledge of the way of smiling of the subject. In this section we present the different steps to generate personalized shape expression as illustrated in Fig. 3.2. Our system is composed of 4 steps. In the first step, we track one neutral face X_n and one smiling face X_s of the subject to extract the features which are the landmarks positions (subsection 3.1.1). After the alignment of the detected landmarks, we learn a person-specific model from the detected landmarks of the

two faces based on the barycentric coordinates (subsection 3.1.2). Once the model is learned, at run-time, we detect the current face X_c of the person to determine the position landmarks X_c^i. We calculate the new smile landmarks positions X_{cs}^i based on the learned model. Finally, We employed a 2D warping method MLS [105] to deform the whole current face of that person X_c basing on the determined deformed shape X_{cs}^i (subsection 3.1.4).

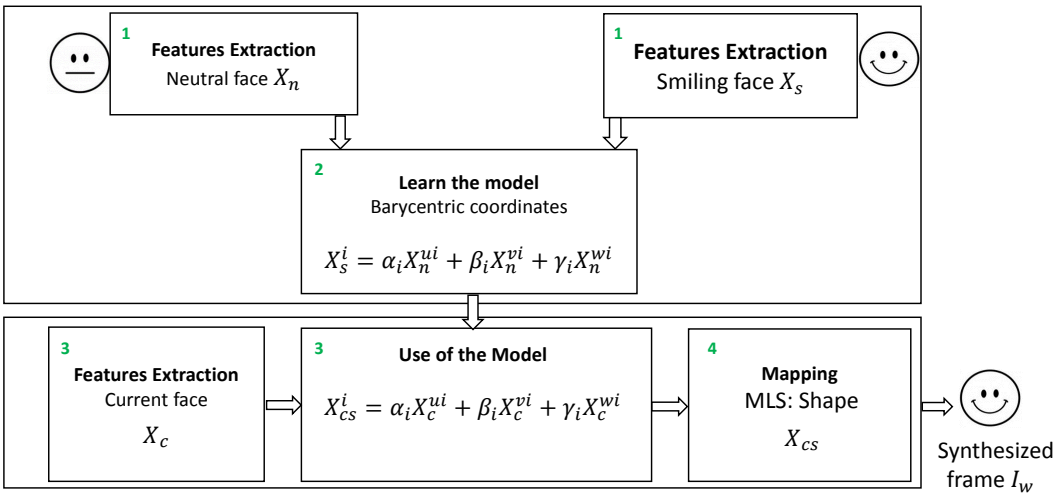


FIGURE 3.2 Our framework is composed of 4 steps. In the first step, we track one neutral X_n and one smiling X_s face of the subject to extract the features which are the landmarks positions. In the second step and basing on the barycentric coordinates, we learn a person-specific model from the detected landmarks of the two faces. This model learns the deformation to apply on a face to make it more positive. In the third step, we detect the current face of the person X_{cs} and we calculate the new smile landmarks positions X_{cs}^i based on the learned model. Finally, we employ a 2D warping method MLS [105] to deform the whole current face X_c from the deformed shape X_{cs}^i.

3.1.1 Pre-processing : Tracking and alignment

To deform a detected face and make it more positive, we use the facial detected landmarks related to the joyful expression such as the eyes and mouth landmarks. These landmarks are moved to determine the smiling shape. In this section we present the steps (tracking and alignment) to do before moving these points.

3.1.1.1 Facial landmarks detection

The first pre-processing step consists in detecting features on the human face for each frame of a smiling video. We use Kurma Tracker of Dynamixyz [24]. This tracker detects the face and determines precisely the coordinates of facial landmarks as well as the orientation, scale, and position of the face. This tool has high performances, in particular its computing time and its ability to detect a face in several states of illumination, and in different poses. Through this API, we can detect 149 landmarks that mark different parts of the face as figured in Fig. 3.3 :

- 0 to 16 : landmarks of the contour of the face.
- 17 to 40 : landmarks for the contour of the eyebrows (12 landmarks surrounding each eyebrow).
- 41 to 52 : landmarks for the eye area (6 landmarks for each eye).
- 53 to 63 : landmarks that mark the nose.
- 64 to 75 : landmarks for the outline of the mouth.
- 76 to 83 : landmarks for the [inner] mouth contour.
- 84 to 148 : the remained landmarks mark different area of cheeks.

In addition to the landmarks identified by the Tracker, it can automatically determine and recognize the orientation of the user's head facing the camera by calculating 3 types of angles according to the reference (x, y, z) as illustrated in Fig. 3.4 :

- Angle which determines the inclination of the head along the x-axis.
- Angle which determines the elevation of the head along the y-axis.
- Angle which determines the look (left/right) of the head along the z-axis.

3.1.1.2 Alignment step

The second pre-processing step consists in the alignment of the neutral and smile face of each person (could be the smile Apex or another intensity of smile). In order to perform the most precise calculation and to reduce measurement errors, we use the homography [65] to align the 2 faces. As depicted in Fig. 3.5, the determination of the homography matrix is based on the blue landmarks of the contour, the nose, and the eyebrows (0 to 40 and 53 to 63) because the position of those points does not change much between the two expressions.

3.1.2 Learning person-specific model shape

The Duchenne smile [28] involves the rise of the mouth corners and cheek, and also the lift of the lower eyelids. In our work, we want that the observers see themselves in a

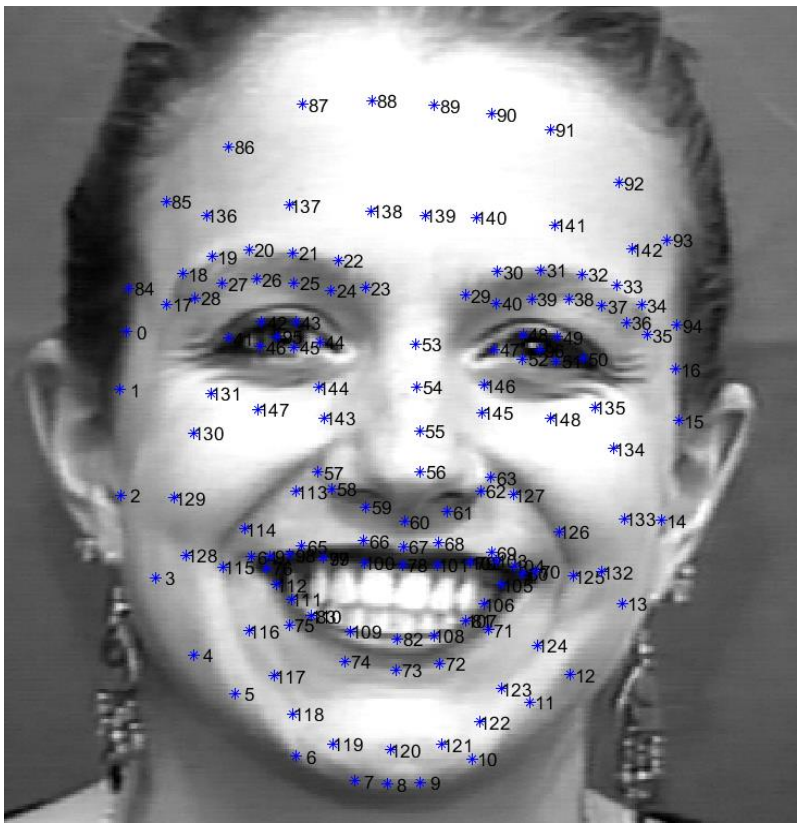


FIGURE 3.3 Illustration of the points identified by the Kurma Tracker [24].

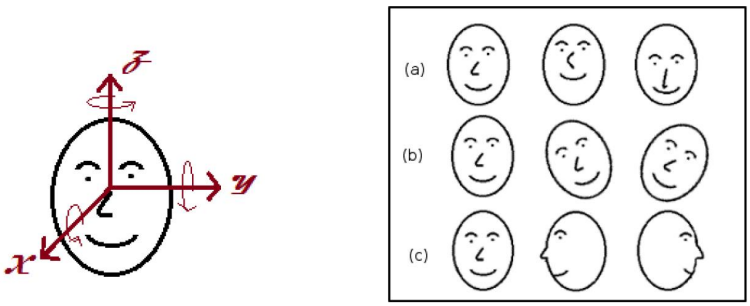


FIGURE 3.4 Head orientation axes : (a) Elevation of the head. (b) Inclination of the head. (c) look of the head.

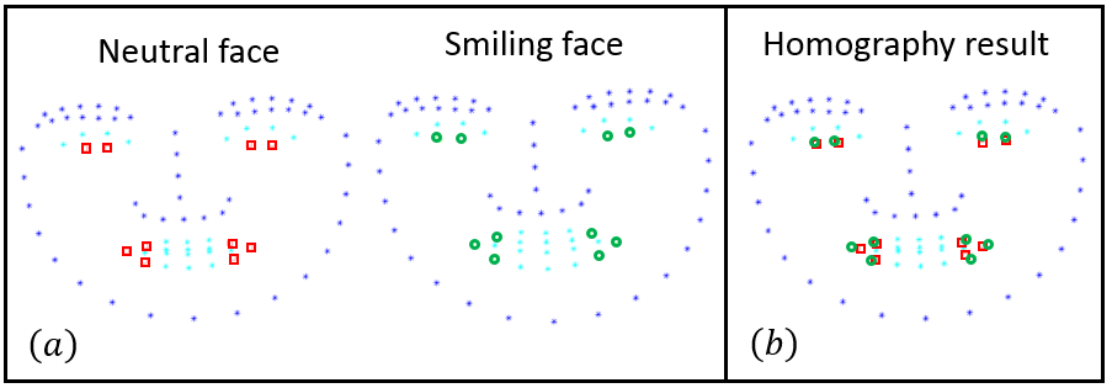


FIGURE 3.5 Pre-treatment of the learning phase. (a) : The neutral and the smiling detected landmarks. (b) : The alignment result.

gradually more positive way ; without their knowing, their reflected faces are algorithmically transformed so the generated smiles must not be perceived consciously. We selected the 10 landmarks corresponding to the corners lips and eyelids to learn the deformation between neutral X_n and smiling X_s faces as illustrated in Fig.3.6. The left mouth corner landmarks have the indexes 64, 65, 75 and the right mouth corner landmarks have the indexes 69, 70, 71. The eyes indexes are (45, 46) and (51, 52) for the left eye and right eye respectively. This choice is based on results of a previous perception test conducted by our team which aims at comparing the impact of adding points for the deformation.

The coordinates of the neutral landmarks are used to perform a Delaunay triangulation as shown in Fig. 3.7. After the alignment, each smiling landmark X_s^i is located inside a neutral face triangle ($X_n^{u_i}$, $X_n^{v_i}$, $X_n^{w_i}$) as shown in Fig. 3.7. We compute the barycentric coordinates of each of these smiling landmarks X_s^i. Fig.3.8 gives an example of the barycentric coordinates determination for the landmark X_s^{64}. This landmark is situated in the triangle X_n^{114}, X_n^{128} and, Xn_n^{115}. Then the barycentric coordinates (α_{64}, β_{64}, γ_{64}) are given by :

$$X_s^{64} = \alpha_{64} X_n^{114} + \beta_{64} X_n^{128} + \gamma_{64} X_n^{115} \tag{3.1}$$

These coordinates are the parameters of our person-specific model. Our person-specific model is thus composed of 10 vectors with 6 components. These components are the 3 vertices indexes (u_i, v_i, w_i) of the triangle (114, 128 and 115 in our example) and the 3 barycentric coordinates (α_i, β_i, γ_i) associated to each of the 3 vertices. Therefore X_s^i is

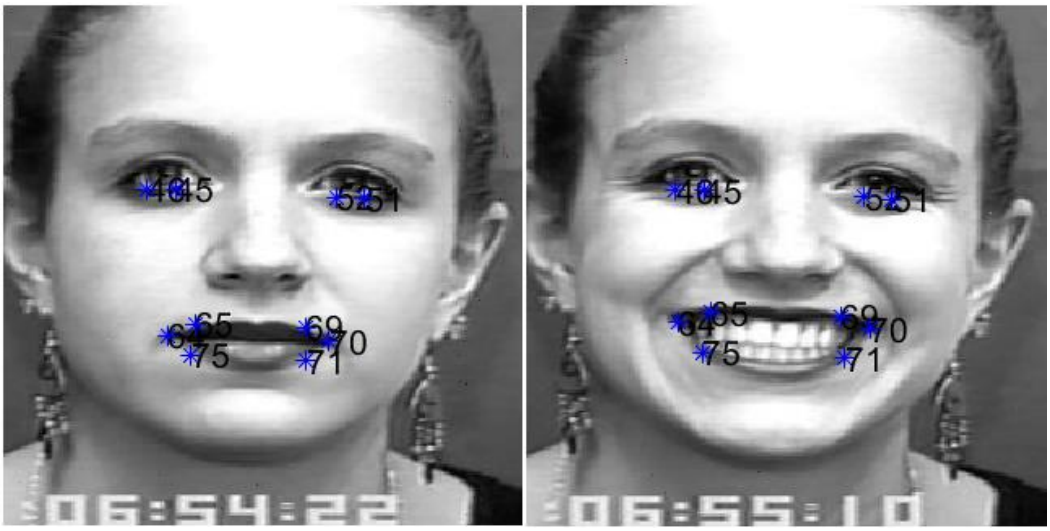


FIGURE 3.6 The selected the 10 landmarks corresponding to the corners lips and eyelids to learn the deformation between neutral X_n and smiling X_s faces. These landmarks are used to learn the deformation between the two expressions.

formulated as below :

$$X_s^i = \alpha_i X_n^{u_i} + \beta_i X_n^{v_i} + \gamma_i X_n^{w_i} \tag{3.2}$$

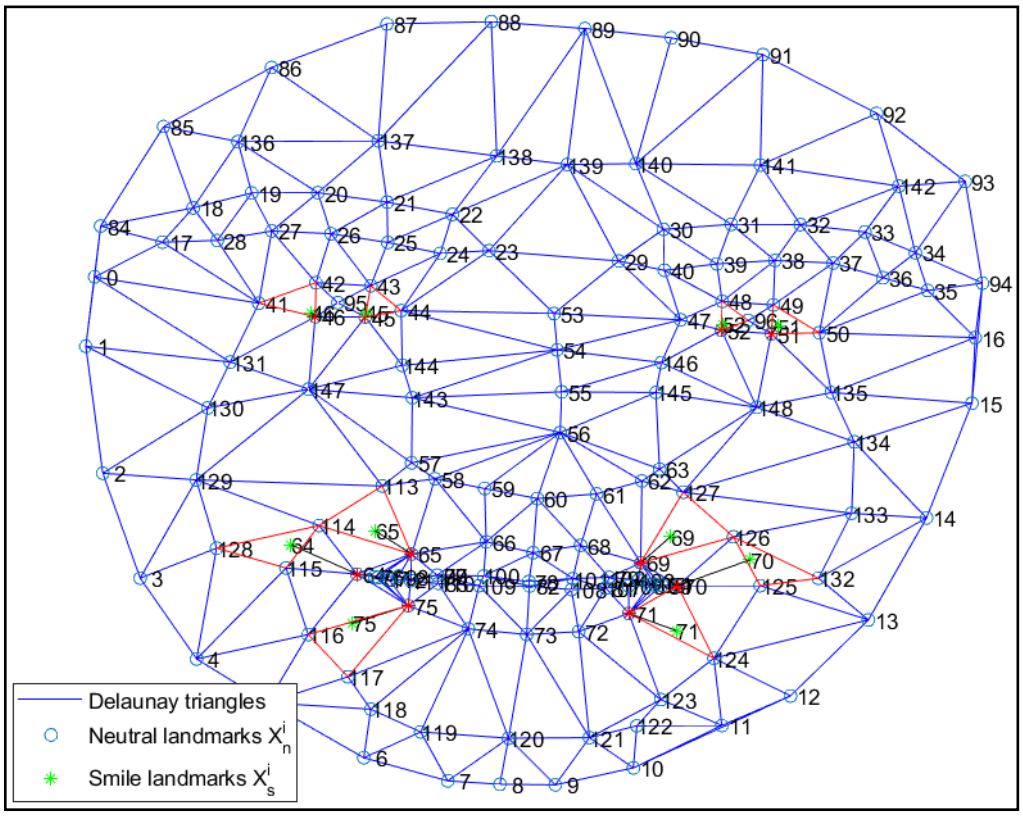


FIGURE 3.7 Positions of the smiling landmarks relative to the neutral landmarks. Each landmark of the smiling face is inside a triangle of the neutral face.

The parameters of the model are learned from a neutral face and a smiling face of the same person. This justifies that our method builds a specific model for each person to re-synthesize her own joyful expression, whereas, with Arias et al. method [5], the parametric model is always the same for all the users.

3.1.3 Test : Shape generation

The Run-time test starts with the detection of the current face X_c. We use the detected landmarks X_c^i of that face and the learned model parameters to generate the joyful expression learned, no matter the detected expression. Having the different coefficients α_i, β_i and γ_i of each smiling landmark, and knowing the coordinates of the triangles in which this smiling

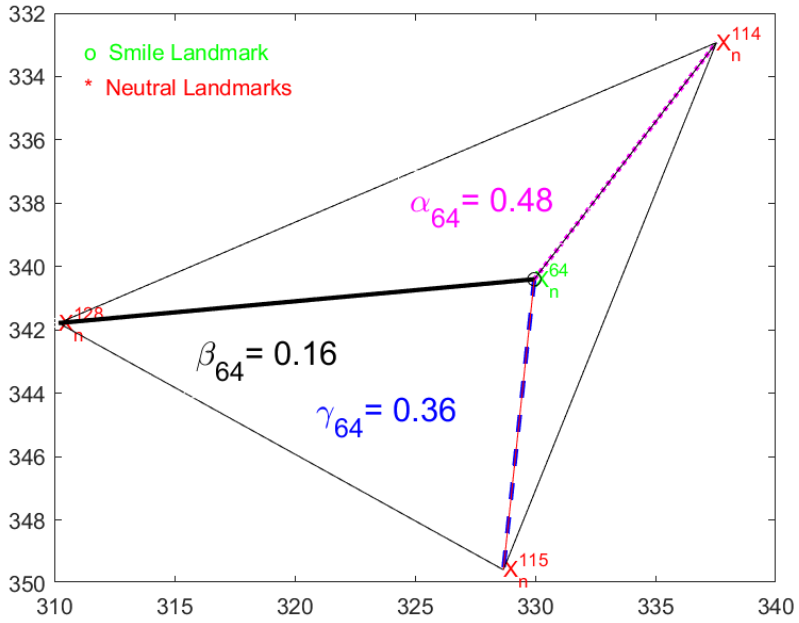


FIGURE 3.8 Example of barycentric coordinates determination for the landmark X_s^{64}. In the learning phase, we use X_s^{64} and the 3 vertices X_n^{114}, X_n^{128} and X_n^{115} to compute the model parameters α_{64}, β_{64} et γ_{64}.

landmark is situated, we determine the positions of the 10 new smiling landmarks $X_c s^i$ using the Equation 3.2 in which the index n is replaced by the index c.

$$X_{cs}^i = \alpha_i X_c^{u_i} + \beta_i X_c^{v_i} + \gamma_i X_c^{w_i} \tag{3.3}$$

This equation aims at determining the 10 new smiling positions of the detected landmarks, which allows to deform the detected face on a more positive one.

3.1.4 Test : 2D deformation

At this step the target smile landmarks X_{cs}^i are determined. In this subsection we present how we generate the warped texture from a detected face X_c. To generate the smiling expression we use a 2D warping method named Moving Least Squares method MLS [105, 5]. As shown in Fig.3.9, this method needs the detected face X_c, the landmarks X_c^i of that face and, the determined smile landmarks X_{cs}^i to generate the deformation. The MLS is very efficient for image deformation and optimizes the distortion made on the image in real-time.

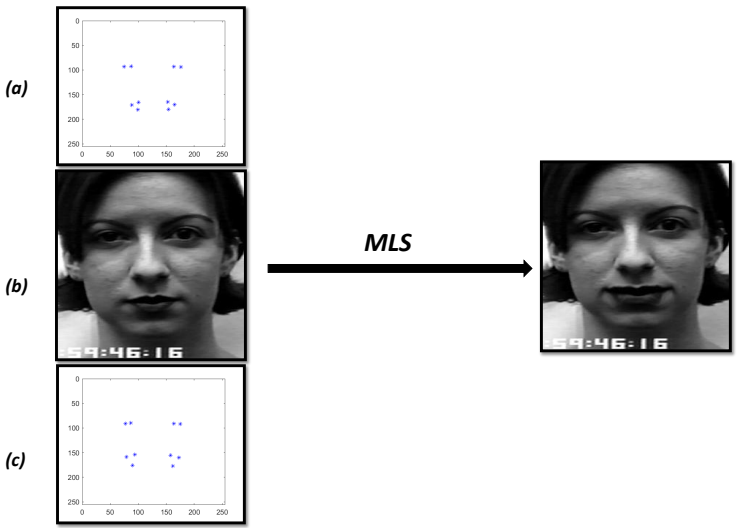


FIGURE 3.9 The MLS inputs are : (a) the detected landmarks X_c^i. (b) the detected face and (c) the target smiling landmarks determined by the model X_{cs}^i.

As performed in [5] we apply the MLS on a neighborhood of pixels around the eyes and mouth to generate the most as possible a photo-realistic expression. Moreover, given the time needed to perform the deformations and apply them, we make a time/esthetic's compromise as in [5] ; we apply the algorithm on grids around each eye and around the mouth, not on

each pixel of the image. First we draw a grid in these areas ; a fine mesh around the eyes and a fine mesh around the mouth and another coarser which surrounds it as illustrated in Fig.3.10. Then, we perform the MLS method on these meshes to warp the input texture and generate the expressive frame.

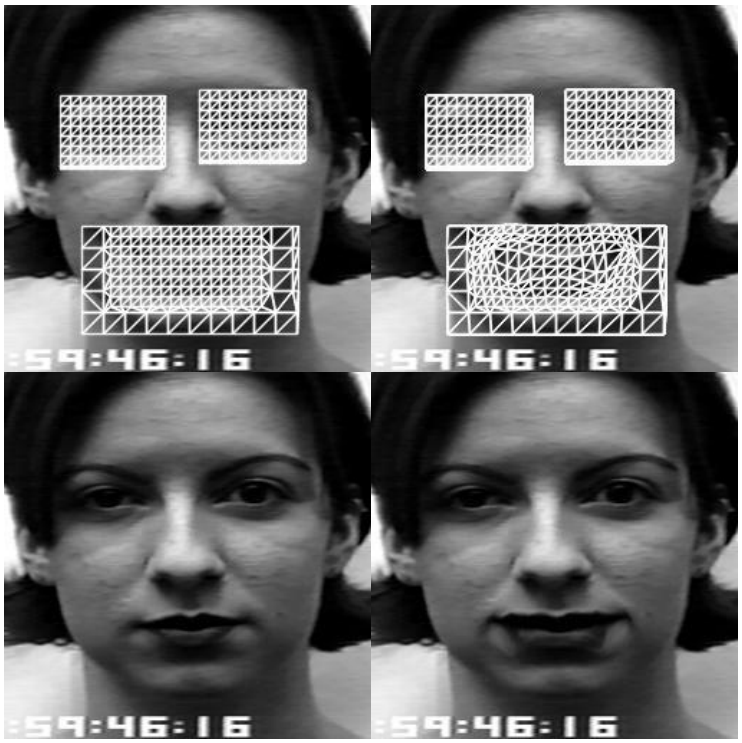


FIGURE 3.10 Example of photo-realistic deformation. First column : The neutral expression. The second column : Warping the meshes of the neutral face to generate a smile expression.

3.1.5 Smile Intensity manipulation

The second contribution of our system is synthesizing smile expressions with different intensities from a single image. To do so, we propose two solutions. The first one is that we synthesize these expressions with different intensities based on the amplitude of the learned smile (Apex frame). We use a deformation coefficient, such that increasing this coefficient increases the smile intensity and vice versa. The second one is that we generate a smile expression from onset to apex by modeling the temporal dynamic of the smile. In this case, we need a smiling video of the person. More details are given in the following subsections.

3.1.5.1 Intensity manipulation based on amplitude

One of our method originalities is that we can generate a person-specific expression with different intensities according to the amplitude of the smile (Apex) as presented in Fig. 3.11.

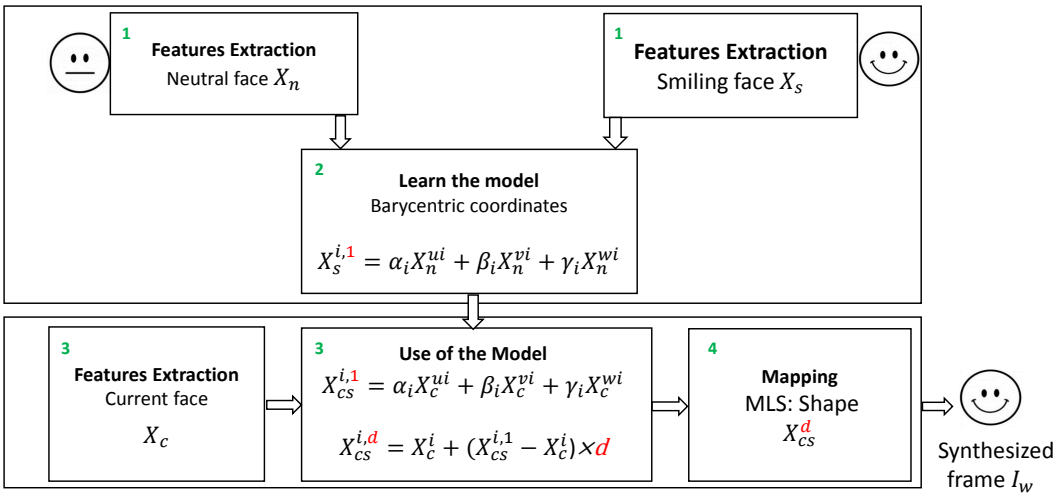


FIGURE 3.11 This step consists on tracking a current face with Kurma tracker [24] to determine the position of its landmarks X_c^i. Using the 10 detected landmarks related to the smile and the learned model (the different coefficients α_i, β_i and γ_i of each of the 10 points and the coordinates of their corresponding triangles (u_i, v_i, w_i)), we determine the positions of the new 10 landmarks $X_{cs}^{i,d}$ to deform the current face to a smile face X_{cs}^d with a coefficient d to amplify or decrease the smile intensity.

We manipulate the intensity using the following equation :

$$X_{cs}^{i,d} = X_c^i + (X_{cs}^{i,1} - X_c^i) \times d \tag{3.4}$$

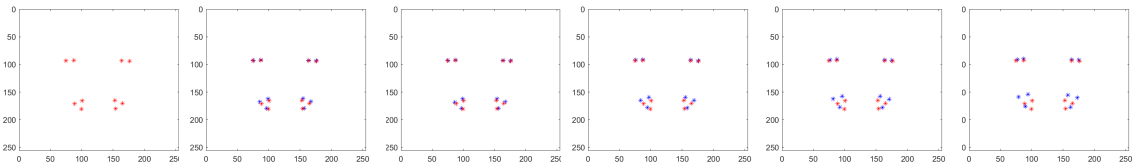
Where d is the intensity coefficient. Increasing this coefficient increases the smile intensity and vice versa. $X_{cs}^{i,1}$ is the positions of the learned landmarks in the smiling frame.

— If $d = 0$, the result is an unchanged face X_{cs}^0=X_c.

— If $d = 1$, the result is an expression of joy X_{cs}^1 that matches the intensity of the one that was learned X_s (Apex).

Fig. 3.12 shows an example of the generated landmarks with different intensities for one subject.

The advantage of this solution is that we synthesize a manipulated smile expression basing on two frames : one neutral and one Apex expression. However, manipulating intensity based



$d = 0$ $d = 0.2$ $d = 0.4$ $d = 0.6$ $d = 0.8$ $d = 1$

FIGURE 3.12 The 10 smile landmarks generated with coefficient variation from 0 to 1 with step of 0.2. The red landmarks are the neutral landmarks. The blue landmarks are the landmarks calculated using the coefficient d.

on the amplitude couldn't respect the evolution of the expression in time, that is to say that synthesizing an expression with a $d = 0.5$ could not synthesize the middle frame intensity of a smile expression. To address this case we propose in the following subsection a second way to manipulate the smile intensity.

3.1.5.2 Intensity manipulation based on temporal dynamic characteristics

The muscles have an elasticity which can be modulated according to an activation function. This function of facial muscles is non-linear and forms a specific dynamics. In this subsection, we detail an alternative method to manipulate the expression intensity to generate real expression over time. Therefore, we propose to manipulate the smile intensity according to the temporal dynamics from neutral to apex.

The temporal dynamics of smile is learned using a first video to generate another personalized video of that person. The originality of this solution is that it respects the smile temporal properties of the person. The first step of our method is learning the deformation between frames of a first smile video of the subject (from Onset to Apex). We use our geometric part based on the barycentric coordinates to determine a per-frame model. We applied some changes on the previously proposed system as illustrated in Fig. 3.13. Instead of learning a deformation between 2 frames and manipulating the amplitude of smile with the coefficient d , we propose to learn the deformation between neutral frame X_{n1} and each of the following frames X_{s1}^{f} of a first smiling video of the person as formulated in the following equation :

$$X_{s1}^{i,f} = \alpha_{i,f} X_{n1}^{u_i} + \beta_{i,f} X_{n1}^{v_i} + \gamma_{i,f} X_{n1}^{w_i} \tag{3.5}$$

where f represents the frame index. According to the results found in [3, 106] the Onset-Apex duration is about 0.5 s. For example, if the videos are recorded at 50 fps, so f is between 0 and 25.

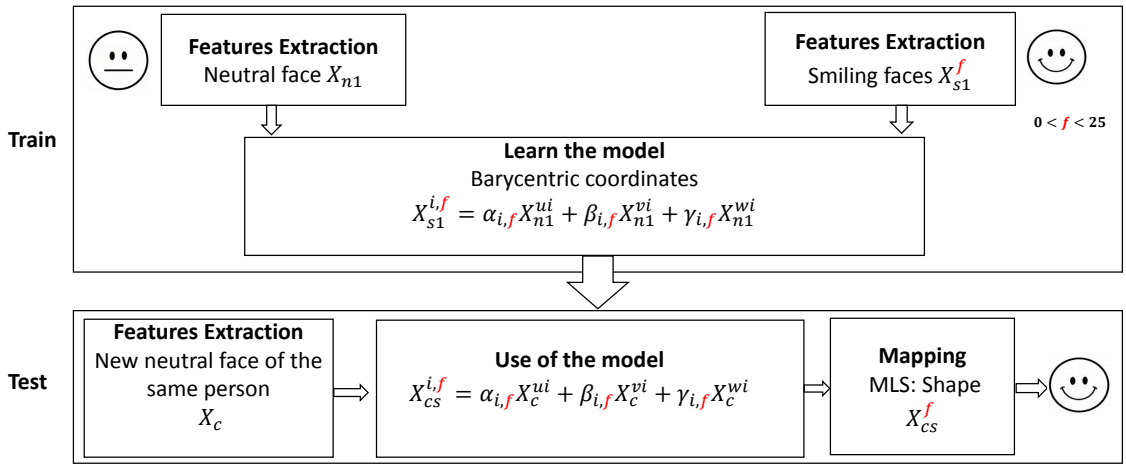


FIGURE 3.13 Our framework is composed of 2 steps. In the first steps, we track the neutral X_{n1} and all the smiling X_{s1}^{f} frames from the smiling video 1 to extract the landmarks. Then, we learn a per-frame person-specific model from the detected landmarks of each two faces (neutral and other frame of index f). In the second step, the learned models are used with a new neutral frame X_c of the subject to generate the new smiling frames X_{cs}^{f}.

The different learned model parameters of each smile landmark $X_{s1}^{i,f}$ are used with a new neutral frame of that person X_c to determine the positions of the new smiling landmarks $X_{cs}^{i,f}$ as mentioned in the following equation :

$$X_{cs}^{i,f} = \alpha_{i,f} X_c^{u_i} + \beta_{i,f} X_c^{v_i} + \gamma_{i,f} X_c^{w_i} \tag{3.6}$$

Fig. 3.14 shows an example of the generated landmarks according to the smile dynamic of the subject in her own smiling video.

The advantage of this solution is that we synthesize a smile expression from Onset to Apex with one single image, while respecting the identity of the emotion and smile temporal properties of the person. One disadvantage may be is that we need to have a whole smiling video of the person, unlike the model with amplitude manipulation which needs just one smiling frame (Apex frame).

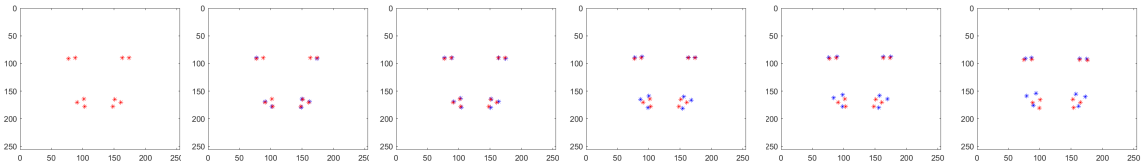


FIGURE 3.14 The landmarks generated according to the dynamic smile of the subject. The red landmarks are the neutral landmarks. The blue landmarks are the landmarks calculated using the frames of the smiling video.

3.1.6 Conclusion on the geometric method

In this section, we proposed an algorithm able to learn a person-specific shape model for each person and transform a captured face in real-time to appear more joyful according to his own expression. Moreover, using our geometric method, we are able to generate smile expression with different intensities while respecting his way of smiling in shape and dynamics (temporal proprieties). In this way, for each subject we built a geometric model that is used as a previous knowledge to generate personal joyful expressions. The generated joyful expression is personal in shape and photo-realistic. However, as we only focus on shape, it misses some details in texture such as teeth and wrinkles. We present our solution to overcome this limitation in the next section.

3.2 Machine Learning method : Texture refinement via GAN

The Generative Adversarial Networks (GAN) have achieved impressive results in generation topics. This methodology was successfully employed in several researches for facial expressions synthesis [124, 86, 123]. Therefore, to infer deformations from a source texture onto a target texture, we choose to use a generative adversarial network. In this section, we give an overview of the generative adversarial networks, then we present the specific GAN used in our method. Finally, we conclude the section with the advantages and the cons of this method.

3.2.1 Conditional Generative adversarial network

Generative adversarial networks [41] are composed of two networks that are simultaneously trained via an adversarial process. The so-called generator G captures the data distribution and the so-called discriminator D estimates the probability that the generated data came from the training data rather than G. To improve the traditional GAN results, the conditional GANs (CGAN) [50] have been proposed. As illustrated in Fig. 3.15 they take as input both a noise vector z, along with additional input y in order to produce the output x. In a CGAN both the generator and discriminator are conditioned on some extra information y. The y condition can be any kind of auxiliary information, such as class labels or data from other modalities. In the generator the prior input noise z, and y are combined in joint hidden representation. The generator synthesizes an image $\tilde{x} = G(z,y)$ to fool D and D tries to distinguish the real images x and the fake one $\tilde{x}$. Therefore, CGANs aim to force the generated images to be indistinguishable from natural images. The objective function of the G and D is written as follows[76] :

$$\min_G \max_D E_{x,y}[logD(x|y)] + E_{y,z}[log(1 - D(G(z|y)))] \tag{3.7}$$

3.2.2 U-Net Conditioned Generative Adversarial Networks (UC-GAN)

To refine the texture details on the synthesized frames (add wrinkles, dimples, and teeth), we use a U-Net Conditioned Generative Adversarial Network (UC-GAN). The UC-GAN [50] achieved impressive results thanks to its skip connections. It was employed in several researches on facial expressions [112, 124, 86, 123]. These skip connections between G_{enc} and D_{enc} aims at increasing the resolution of the output. As shown in Fig.3.16 the encoder features are transmitted along these connections with the conditioned information to the decoder, which can help ensure the low-level information to be shared between the input and output. The conditioned U-Net GAN connects each layer k and layer n-k, where n is the total number of layers. Each skip connection simply concatenates all channels at layer k with those at layer n-k.

3.2.2.1 Specific input

In our framework, we use UC-GAN for 2 reasons. First, to ensure that the output preserves the identity of the input face, secondly, to efficiently refine textures of the synthesized images. We improve the traditional UC-GAN by combining a prior knowledge information of the

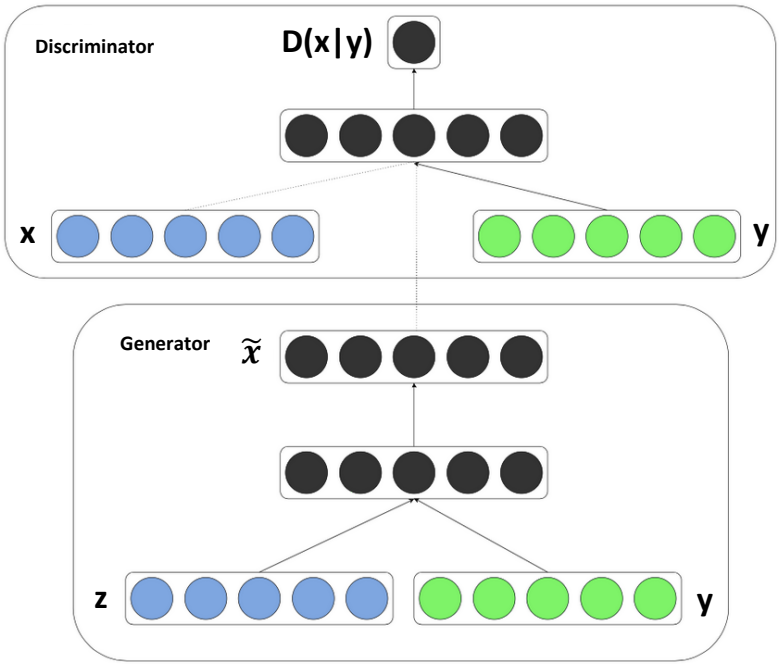


FIGURE 3.15 The architecture of CGAN [76]

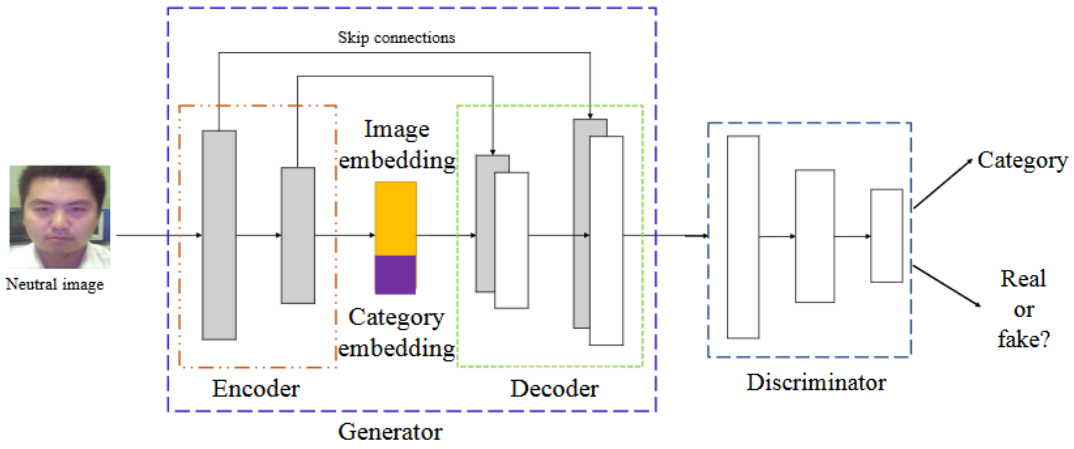


FIGURE 3.16 Illustration of the proposed UC-GAN model in [124].

way of smiling of each person. Our first originality is that we use the synthesized (warped) images I_{wj} generated with the MLS (our geometric method) to train our UC-GAN as Fig. 3.17 shows where j is the frame index. The input of our UC-GAN is the deformed face I_{wj} which lacks global details such as teeth and wrinkles. These details are refined using the proposed UC-GAN.

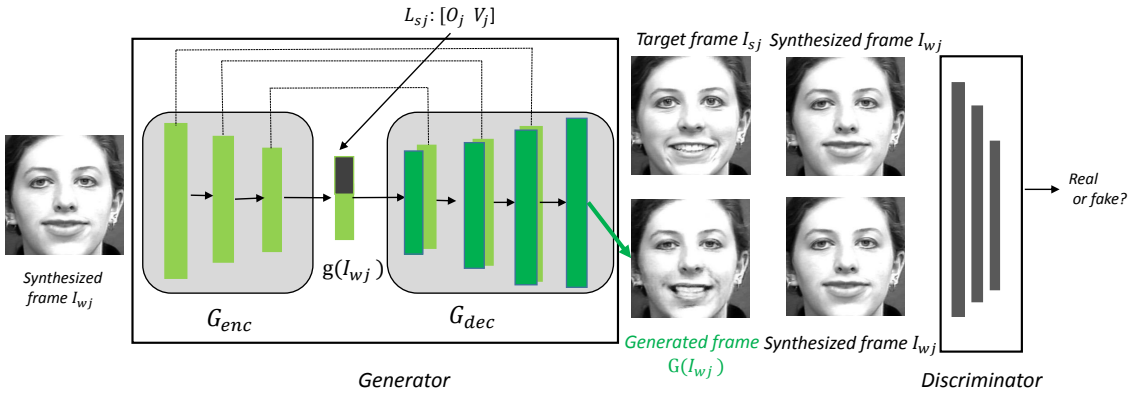


FIGURE 3.17 The GAN is used to refine details on the synthesized geometric images I_{wj}. The synthesized images I_{wj} generated by the geometric step, are fed to the Generator. The encoding representation of the image $g(I_{wj})$ is concatenated with a label vector L_{sj} which is composed of the coefficient of the lips opening O_j and one hot vector V_j to characterize the intensity. The decoder generates the expressive frame with more details. The discriminator D takes two couple of images; the synthesized frame I_{wj} with the real frame I_{sj} and the synthesized frame I_{wj} with the generated frame $G(I_{wj})$ to determine if the latter is a real or a fake expression.

3.2.2.2 Specific label vector

Our second originality is that the encoding representation of the image $g(I_{wj})$ is concatenated with a label vector L_{sj} which is composed of the coefficient of the lips opening O_j and one hot vector V_j to characterize the intensity.. The label L_{sj} ([O_j,V_j]) helps the GAN to correctly add the expression details such as the opening of the mouth, which indicates that it should add teeth. The coefficient O_j is calculated using the eyes and the mouth open distances as shown in Fig. 3.18.

$$O_j = \frac{Dist_1}{Dist_2} \tag{3.8}$$

The V_j is a one hot vector of j values which characterizes the intensity level. The intensity levels depends on the fps of the smiling videos. The Onset-Apex duration is 0.5 s [3, 106].

For example if the videos are recorded at 30 fps so the Onset-Apex corresponds to the first 15 frames of a smiling video.

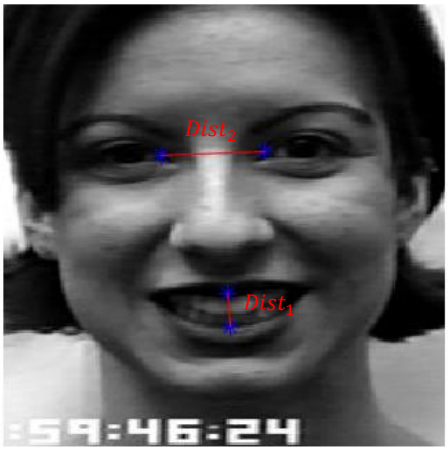


FIGURE 3.18 The two Distances $Dist_1$ (mouth) and $Dist_2$ (eye) are used to calculate the coefficient of the lips opening O_j which represents the first value in the label vector.

In the training phase, our UC-GAN take as input the synthesized (warped) images I_{wj} to be encoded $g(I_{wj})$ as illustrated in Fig. 3.17. This coded is concatenated with a label vector L_{sj} (Sub-subsection 3.2.2.2) to add specific information. All of them are fed to the G_{dec} to generate the smiling image $G(I_{wj})$. Then, that generated image is passed thought the discriminator to be compared to the real image I_{sj}. The discriminator tries to maximize the difference between $G(I_{wj})$ and I_{sj} which helps the generator at each iteration to perform well the generation of the smiling frame $G(I_{wj})$.

Once the GAN is trained, we test it using a new warped face (I_{wj}) with their corresponding label vector to generate the target expression $G(I_{wj})$. Then the expression is evaluated to prove the efficiency of our proposed framework (see chapter 4).

3.2.2.3 Loss functions

Generators and discriminators are trained alternatively towards adversarial goals. The adversarial losses for generator and discriminator [124] are written as follows :

$$L_G = L_{sce}(D(I_{wj}, G(I_{wj}, L_{sj})), 1) \tag{3.9}$$

$$L_D = L_{sce}(D(I_{wj}, I_{sj}), 1) + L_{sce}(D(I_{wj}, G(I_{wj}, L_{sj})), 0) \tag{3.10}$$

Where the L_{sce} is the sigmoid cross-entropy.

The generator intended not only to fool the discriminator but also to synthesize images similar to the target ground truth I_{sj} as far as possible. To this aim, we use a pixel loss L_p [50] to achieve small distance between the generated image and the ground truth one.

$$L_p = ||G(I_{wj}, L_{sj}) - I_{sj}||_1 \tag{3.11}$$

Furthermore, to capture the structural information of the images and achieve more realistic joyful expressions, we adopt the feature matching loss term F_m of [132, 102]. This function forces the synthetic image I_{wj} and the real smile image I_{sj} to share the same features. Thus, with the help of the feature matching term, the reconstructed results will have more realistic features.

$$F_m^l = ||\phi^l(G(I_{wj})) - \phi^l(I_{sj})||^2 \tag{3.12}$$

Where $\phi^l(G(I_wj))$ represent the features of the input I_{wj} at the l-th layer of the discriminator. Therefore, as [132] we modify the loss functions of D and G with F_m^3 and F_m^2 respectively. Note that the layer l for updating L_D and L_G can be different. Like in [132], we use the second convolutional layer of D to update L_G which maintains the main structure features of the input. In this way, we enforce the $G(I_{wj})$ to be similar to I_{sj} on pixel and structure. Then, we use the third layer to update L_D which aims at investigating if the real frame I_{sj} is close to the generated frame $G(I_{wj})$. This step helps to visually sharp results with higher image quality.

$$L_G^* = L_{sce}(D(I_{wj}, G(I_{wj}, L_{sj})), 1) + \lambda_1 F_m^2 \tag{3.13}$$

$$L_D^* = L_{sce}(D(I_{wj}, I_{sj}), 1) + L_{sce}(D(I_n, G(I_{wj}, L_{sj})), 0) + \lambda_2 F_m^3 \tag{3.14}$$

The final objective loss function is defined as :

$$F_{loss} = L_D^* + L_G^* + \lambda L_p \tag{3.15}$$

Where λ, λ_1, λ_2 are the hyper parameters that balances the loss functions.

Our results at this level (see chapter 4) show that our system generates expressions close to the ground truth thanks to the personalization of the smile shape by the geometric part and the texture refinement using the GAN. However, we noticed that there are two limitations in our system. The first is that the GAN does not add the real person's teeth and the second is that the GAN adds different teeth to each frame from onset to apex. In the next section, we present the proposed methods to overcome these limitations and strengthen our system.

3.3 Teeth refinement

Our hybrid (Geometric-machine learning) framework can synthesize from an input (closed-mouth) a smile expression (open-mouth) and infer the inner mouth region. However, due to the lack of training data for the mouth interior, the generated texture tends to be of rather low resolution. Moreover, the GAN has a lack of consistency in adding the teeth texture because the added teeth are not those of the person. As we would like to personalize the synthesized expression as much as possible to overcome this limitation, we propose to use one of two methods : either Poisson editing method or Auto-encoder as shown in Fig. 3.1. We detailed these solutions in the following subsections.

3.3.1 Poisson editing technique

The Poisson editing (PE) method proposed in [88] achieves good results in the cloning field. It allows to change the appearance of image regions such as texture, illumination, and colors [1, 130]. Perez et al. [88] method is based on a guided interpolation. They define a vector field v that is used as a guidance field to solve interpolation problem and find the solution of Poisson equations to get intensity of the unknown pixels in the new blended region in the final composite image. This method provides good results and seamlessly clones the source region on the target image.

In order to synthesize a person-specific mouth texture, we propose to use Poisson editing technique to blend the person's teeth to her generated expression. The idea is to use another smiling video of that person to add her own teeth. The technique is illustrated in Fig. 3.19, we use the generated image $G(I_{Wj})$ by our hybrid system (Geo-GAN) as the target image and a mask around M_j the teeth region of this image. We want to blend a source teeth-region $S_{I_{sk}}$ on the target image generated $G(I_{Wj})$. The source teeth region $S_{I_{sk}}$ is determined from a real frame I_{sk} in another smiling video of that person. For each target teeth region $T_{G(I_{Wj})}$,

we automatically determine its corresponding region in the second video by comparing the opening mouth distances in the two frames. To do so we calculated the distances d_j between the two landmarks of the corners of the mouth of the generated frame $G(I_{Wj})$ and search in the second video the frame I_{sk} which has the closest distance d_k to d_j and use it as a source.

$$\begin{cases} d_k = d(I_{sk}^{64}, I_{sk}^{70}) \\ d_j = d(G(I_{Wj})^{64}, G(I_{Wj})^{70}) \\ \text{if } d_k \approx d_j, T_{G(I_{Wj})} = S_{I_{sk}} \end{cases} \tag{3.16}$$

The advantage of this geometric method is that it blends the mouth region with high coherence between the generated frames and the mouth region. However, Poisson Editing is limited to blend the mouth regions only on frontal faces on both videos. This technique works only if the source face and the target face are in the same position or orientation. With oriented faces or faces in complex positions, we propose as an alternative of PE, to employ a machine learning method which is detailed in the next subsection.

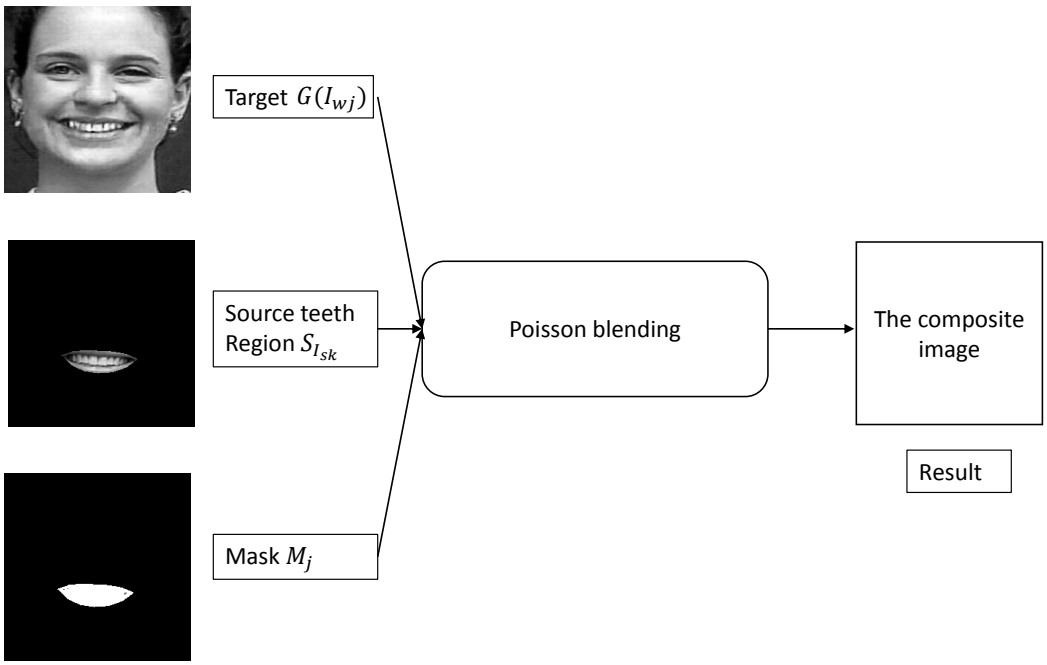


FIGURE 3.19 Poisson editing method used to blend the teeth region in the generated images by our Geo-ML system. The target image is the generated image $G(I_{Wj})$ with teeth that are not those of the person. The mask M_j is the region around the teeth region of the generated image $G(I_{Wj})$. The source teeth region $S_{I_{sk}}$ is determined from a frame in the second smiling video of that person.

3.3.2 Auto-encoder method

An auto-encoder (AE) aims at learning efficient data representation in an unsupervised manner by minimizing the discrepancy between the original data and its reconstruction [64]. The AE generates from input data a reduced encoding representation as close as possible to its original input. Typically a neural network (AE) is used for dimensional reduction, more precisely, the representation retains meaningful information about the input [142]. Therefore, we propose to exploit AE advantages in our framework to guide the GAN to add specific texture information from the subject to the generated frame as shown in Fig. 3.20. At the training phase, the AE takes as input the mouth of the person m_j to extract the meaningful information and generates a code C_j.

At the test phase, the AE generate a personalized code C_j, that contains vital amount information of the input mouth m_j. This code C_j is concatenated with the latent representation $g(I_{wj})$ with the label vector L_{sj} which is composed of the coefficient of the lips opening O_j and one hot vector V_j to characterize the intensity. With this solution, we expect that the combination helps the GAN to generate more personal-specific texture. The advantage of using an AE is that it can refine the face texture whatever the position or the orientation of the person face.

3.4 Conclusion

In this chapter, we proposed a hybrid (Geometric-machine learning) approach aiming at learning a person-specific model for each person and transforming a captured face to appear more joyful. Our method generates for each subject a photo-realistic joyful expression according to her own expression. The geometric part aims at personalizing the smile shape and the machine learning part allows to refine the texture of the synthesized images using a generative network. To ameliorate this refinement, we propose to use Auto-encoder or Poisson Editing method in addition to the GAN. The following chapter will present the experimental results and the comparison with the state of the arts methods.

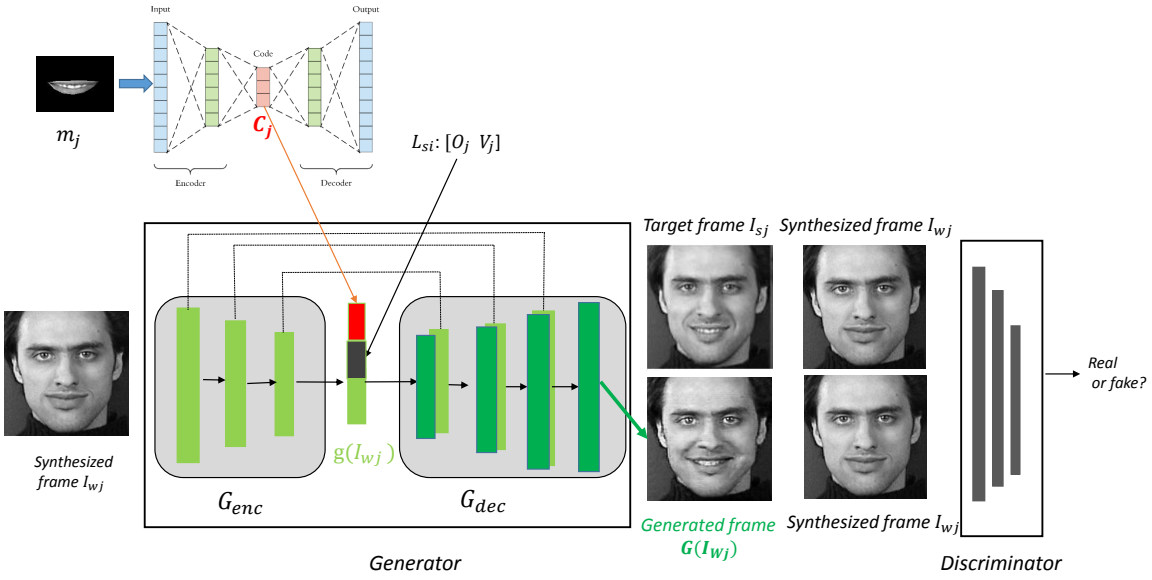


FIGURE 3.20 The code C_j generated with the AE is concatenated with the latent representation $g(I_{wj})$ and the label vector L_{sj} which is composed of the coefficient of the lips opening O_j and one hot vector V_j to characterize the intensity. This code guides the GAN to add specific texture information to the generated frame $G(I_{wj})$.

Chapitre 4

Experimental results of joyful expression synthesis

In this thesis, we propose a hybrid geometric-machine learning approach to synthesize photo-realistic and personalized joyful expressions while keeping the personal identity of the emotion in shape and in texture. This chapter aims at proving the efficiency of our hybrid system. The proposed method is compared with two state of the art methods : a geometric method [5] and a machine learning method [124].

The databases used in the experiments are introduced at first in section 4.1. Then in section 4.2, we present two statistical studies that prove that each person has their own manner to smile. In order to demonstrate the relevance of the proposed system, the rest of the chapter is based on our main contributions presented in Fig. 4.1. We examine the relevance of these contributions one after one in section 4.3. Our main contribution is that our system synthesizes in real-time person-specific joyful expression and leads to preserving the morphology shape (Section 4.3.1) and the identity of the emotion by reproducing the specific way of smiling of each subject. We compare our results with the state of the art methods [5, 124] in section 4.3.1.1. The second contribution is that we synthesize expression with different intensity using a single neutral image of the person. The originality is that we propose to manipulate the intensity based on the amplitude of the smile or based on the temporal-dynamic of the smile which respects the properties of smiling of each person. The results comparing the two manners are given in section 4.3.1.2. Another main contribution is evaluated. We act on both shape and texture by introducing a hybrid method combining geometric and machine learning tools to refine global texture. Quantitative and qualitative results are presented in section 4.3.2. Furthermore, we contribute on the teeth personalization by using an AE or Poisson editing method to refine local texture of mouth region. The results

are given in section 4.3.3. Finally, we conclude the chapter in section 4.4 by indicating the performances evolution of the method.

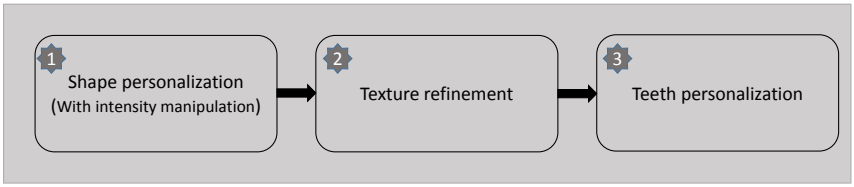


FIGURE 4.1 Organization of the section 4.3 on our contributions.

4.1 Databases

Since our goal is to synthesize person-specific smile expression with manipulated intensity on real subjects, real smile videos are needed for testing. We use for our experiments different databases in order to evaluate our process and compare our method with the state of the art methods [5, 124]. We choose to use databases that contain a large number of videos for the training of our system to ensure relevant learning. In addition, we search for databases that have two videos for each subject. The first video is used for the training and the second one for the test. We describe the used 4 databases in the following subsections.

4.1.1 CK Database

This database [58] includes 486 sequences from 97 subjects. Each sequence begins with a neutral expression and proceeds to a peak expression as illustrated in Fig. 4.2. The videos are recorded with 30 fps and a resolution of 640 $\times$ 490. We select the 88 subjects who have smile sequences. Since almost all the sequences are gray-scale, we use gray scale images in our experiments.

4.1.2 MMI Database

The database MMI [119] consists of over 2900 videos and high-resolution images of 75 subjects. It is fully annotated for the presence of AUs in videos. 56 smile videos of 28 subjects are used in our experiment which are annotated with AU6 and AU12 (corresponding to smile). Each subject has 2 smiling videos as shown in Fig. 4.3 so, we have the opportunity

FIGURE 4.2 Three examples of recorded sequences of CK database. Each sequence begins with a neutral expression and proceeds to a peak expression. We just present 4frames from each sequence for each subject.

to learn a person-specific model on one video (using a neutral and a smiling face) and test it on another one. Those tests are referred to MMI* in the experiments (Subsection 4.3).

4.1.3 Oulu-CASIA Database

The database contains 80 subjects [143] with the 6 basic expressions for each subject. The videos are captured with VIS camera with strong illumination and are at 30 fps. The resolution is 320 $\times$ 240. All the subjects of this database have a smiling video. We use these smile sequences to carry out our experiments.

4.1.4 UVA-NEMO Database

UVA-NEMO Smile Database [20] is a large-scale smile database which has 1240 smile videos (597 spontaneous and 643 posed) from 400 subjects. Videos are in RGB color and recorded with a resolution of 1920×1080 pixels at a rate of 50 frames per second under controlled illumination conditions. Each subject has between 1 and 4 spontaneous/deliberate smile videos. For further illumination and color normalization, a color chart is present on the background of the videos. Each video starts and ends with neutral or near-neutral expressions.

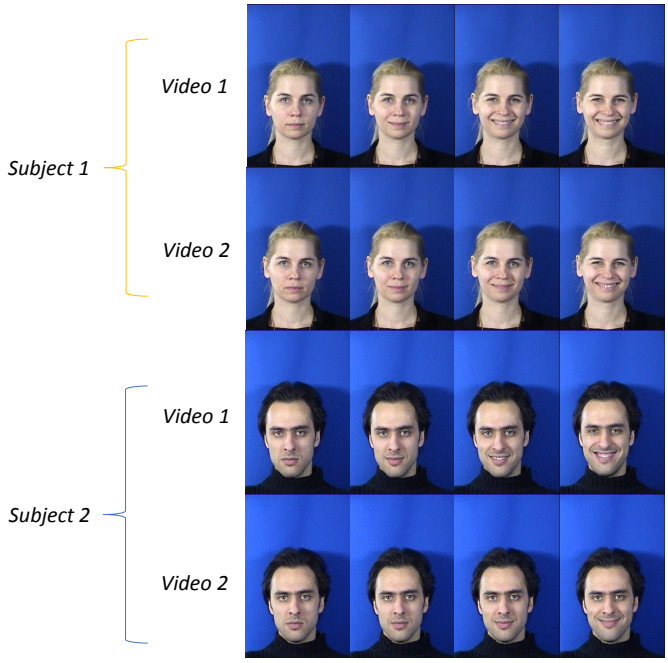


FIGURE 4.3 Four smile videos from MMI databases. Each subject has 2 smiling videos. We present just 4 frames from each sequence for 2 subjects. Each sequence begins with a neutral expression and proceeds to a peak expression then to a neutral expression again.

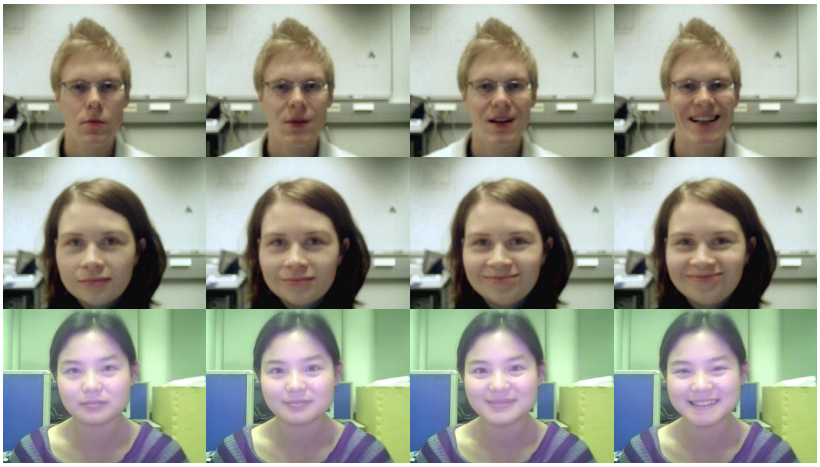


FIGURE 4.4 Smiling images of 3 subjects from sequences of Oulu-Casia database. Each sequence begins with a neutral expression and proceeds to a peak expression.

FIGURE 4.5 Four subjects from videos of UVA-NEMO database. Each subject has between 1 and 4 spontaneous/deliberate smile videos.

To achieve our experiments we need databases with smile videos or a smile frames also a databases which has more than one video smile per subject. That is why we chosen the previous described database which correspond well to our needs for the training and the test of our system. The table 4.1 describe the different characteristics of the used databases.

TABLE 4.1 Characteristics summary of the used databases.

Characteristics	Databases			
	CK	MMI	Oulu-CASIA	UVA-NEMO
Modality	Frames/Videos	Videos	Frames/Videos	Videos
Number of Subjects used in the experiments	88	28	80	400
Number of smile videos per subject	1	2	1	1-4
Expressions	Posed	Posed	Posed	Posed and Spontaneous
Action Units	No	Yes	No	No
FPS	30	30	30	50
Video duration	0.51s	2.5s	0.7s	3s
Resolution	640×490	369×576	320×240	1920×1080

4.2 Statistic analysis : How differently people smile ?

The cited databases are used at first to justify the two hypotheses that we have taken to build our model. The first one is that we considered that the smile trajectory (i.e. the displacement of the mouth corner) is straight, so that we modeled the variation of intensity by a linear equation for each landmark using the d in Equation (3.4). The second hypothesis is that the smile is specific to each person and we need a prior knowledge of how each subject made their own smile to synthesize a personalized expression. Therefore, we learn specific model parameters from their own neutral and smiling expressions. To this aim, we investigate the trajectory of smiles to explain why we propose such a system. We carry out this studies with UVA-Nemo and MMI* databases where each subject has 2 smiling videos. More details are given in the following subsections.

4.2.1 Trajectories analysis : Straight trajectory

We started the analysis by cropping and tracking all the frames for each subject of the databases using the Kurma Tracker [24]. We align the faces of each subject according to the first frame using Procrustes alignment. Fig. 4.6 shows the trajectories of the landmark of the left corner of the mouth for 3 subjects. A linear regression $Y = a_{GT}X + b_{GT}$ is used to estimate the landmarks trajectories and to determine the ground truth (GT) slopes a_{GT} for each video sequence. We noticed that the trajectories are approximately straight [113] but with rather different slopes between subjects. We remark also that the slopes smiles of the same subject is quite similar in his two smiling videos. We found the same results for the other landmarks of the mouth. Consequently, we deduced that the smile is straight and occurs in a different way for each subject. These results justify the use of a coefficient d for our person-specific method to variate the intensity of the expression based on amplitude (see chapter 3, Equation(3.4)). In the following subsection, we investigate the similarity between the smiles of the same subject by studying the inter and intra-variability of smiles between subjects.

4.2.2 Trajectory analysis : Inter and Intra-variability

A complementary study of the smile trajectory (UVA-Nemo database) allows us to analyze the smile trajectories within subjects. In this database we found several videos for each subject that help to investigate the smile trajectories for each subject in their videos and compare it to the smiles of other subjects.

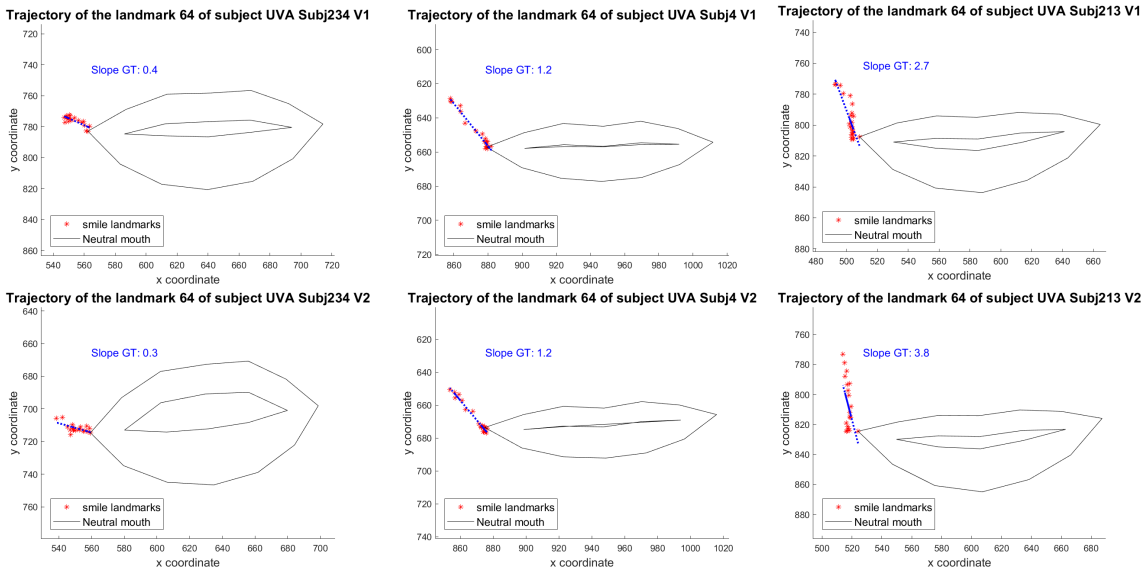


FIGURE 4.6 Examples of 3 subjects smile trajectory. Each column represents the trajectories of the landmarks during smile for one subject in two videos. The subject 234 have a flat smile with a slighter slope than subject 213 who has rather a high slope explained by a growing smile

To analyze this relation between the way of smiling of each subject and that of others, we conduct a statistical study that investigates the inter and intra-variability of the smile trajectories subjects. The study consists on determining the correlation coefficient between the smile trajectories. We calculate the correlations between the smile trajectories of the same subject (intra-variability) and the correlations between the trajectories of all the subject (inter-variability). The results are shown in Fig. 4.7. Such a study proves that the intra-variability correlation is bigger than the inter-variability correlation. Therefore, we found a strong link between videos of each person and this shows also that the smile is personal and occurs in different way for each person.

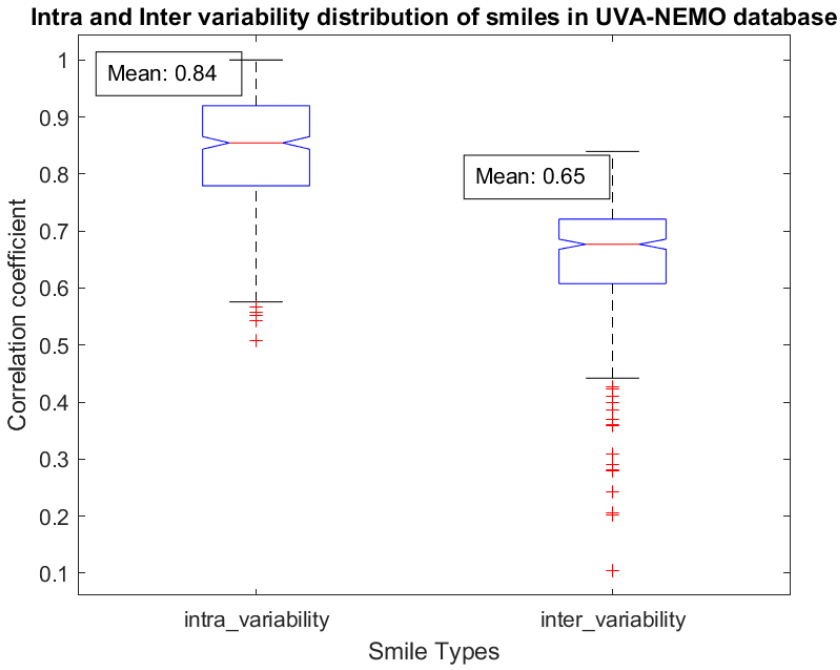


FIGURE 4.7 The Inter and Intra-Variability of smile trajectories within subjects. The intra-variability correlation is more important than the inter-variability correlation.

Based on the previous results on the smile trajectories (Smile is straight, personal and occurs with different way for each subject) we headed towards proposing a system that personalizes the synthesized expression. First, to preserve the identity of the emotion and second to keep the credibility that can help on acting on the emotional state of the PTSD patient in our use case. The smile study allows us to prove the high correlation between the

smile expressions of the same person. The originality of our proposed system is that we learn a person-specific model. The proposed framework generates joyful expressions while respecting the person's real smile. The following section is consecrated to the quantitative and qualitative results of our proposed method.

4.3 Efficient personalized smile : shape and texture results

To validate our approach, we test our person-specific method along with two generic ones : a geometric method and a machine learning method [5, 124] to re-synthesize a smiling expression for each subject on the 4 databases and compare their results. Arias et al.[5] proposed a system of smile/sad face generation based on 2D warping method. They tracks first the morphological features of the face, such as the eyes and lip corners and stretches its position using a predefined mean model to synthesize the expressive face. These deformations are adaptive to the position of the user (more precisely, to camera-user distance and head pose). However, these deformations are still based on a mean model yet each person is smiling differently. Their model is described by the following equation :

$$X_i^s = X_i^n \pm (Q_r \times \Delta_{xy}) \times s \times \alpha_v \tag{4.1}$$

Where Δ_{xy} are the learned mean model parameters and α_v is the intensity of the smile distortion (they use α_v =1.5 to generate a smile expression). To adapt to face-camera distance and head pose, s is computed as the distance between the two eyes multiplied by cosine of the angle yaw, and Q_r is the rotation matrix corresponding to the roll.

Wang et al. [124] build a generative model to synthesize facial expressions from neutral frames. The GAN allows to synthesize different photo-realistic facial expressions and generate natural and reasonable face expressions (e.g. different smiles) with global details. However, The deformations are based on a model which is learned on several subjects. As a result the synthesized expressions are not the subject's own. Yet, each person has her own way to make expressions. Then, these methods [5, 124] generate non-personalized expressions.

This section aims at proving the efficiency of our whole system. The section is organized based on our contributions as illustrated in Fig. 4.1. First, the qualitative and quantitative results of the geometric part on shape personalization are presented in subsection 4.3.1 and subsection 4.3.2. In Subsection 4.3.3 and 4.3.4 we detailed the qualitative and quantitative results on texture after adding the machine learning part which allows to refine the texture of the previously synthesized images. Finally, we prove the effectiveness of using an AE or

PE method in addition to the GAN to refine local mouth texture and personalizing the teeth region (Subsection 4.3.5 and 4.3.6).

4.3.1 Results on shape personalization

The geometric part of our system (see Fig. 3.1) consists on personalizing the shape of the synthesized expression by learning a person-specific model. With that learned model we generate the learned smile expression (Apex). We assign for this expression an intensity of 1 (d=1). Furthermore, we generate different smile intensity by manipulating the coefficient d (based on the smile amplitude) or using the temporal dynamics of the subject's smile (based on the temporal variation of a smile). In this subsection, we prove first the efficiency of our method in synthesizing the learned smile (apex frame). Then we present the qualitative and quantitative results on shape personalization with the two ways of the intensity manipulation.

4.3.1.1 Efficient smile shape synthesis

To prove the efficiency of our method, we compare our results with two state of the art methods [5, 124]. In this subsection we give the qualitative and quantitative results of synthesizing the learned smile shape. Our geometric part aims at learning a person-specific model for synthesizing joyful expression using only two frames (Neutral and Apex). To synthesize the learned Apex expression we use a coefficient intensity d=1 (See chapter3 section 3.1.2). We conduct this experiment on MMI* and UVA-NEMO databases as they contain more than one smile video per subject. Therefore, we learn a person-specific model using one smiling video and test it on another smiling video of that subject.

4.3.1.1.1 Qualitative results

Fig. 4.8 illustrates the results obtained with our method and two state of the art methods. We observe that with the geometric method of Arias et al. [5], the lips corners are systematically raised (steep slope) for the 2 subjects. We notice that subject 1825 has an asymmetric smile while the smile generated with [5] is symmetric. Indeed, this method applies the same deformation whatever the way of smiling of the subject. Visual fidelity shows that the smile generated with our method seems to be closer to the real one.

The results with the GAN proposed by Wang et al. [124] show that the smiles generated are different for the 2 subjects but are not those of these subjects. Thus the GAN generates realistic but not real smiles as shown on Fig. 4.8 (third column). On the contrary, our method

synthesizes personalized smile using the person-specific model of the person as input with a new detected face.

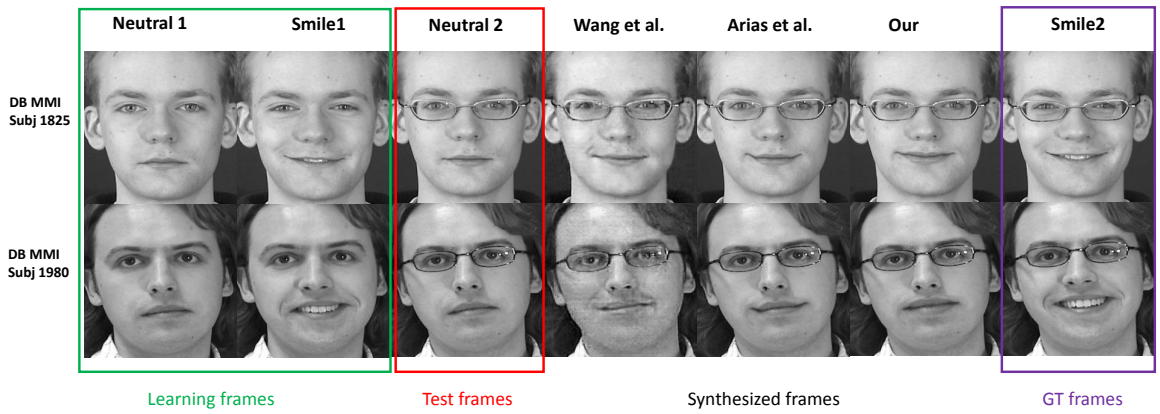


FIGURE 4.8 Examples results with different methods on two subjects of the database. We learned the smile deformation using two frames of the subject (Neutral1 and smile1). Then we test the model on a new neutral frame of the subject (Neutral2). We noticed that the frames synthesized with Wang et.al [124] are realistic but are not those of the subject. For the synthesized frames with Arias et al. [5], we remark that the corner of the lips is systematically raised for the 2 subjects which justifies that the method applies the same deformation whatever the way of smiling of the subject. However, Our results synthesize 2 different type of smiles for the two subjects. The first subject smile seems to be asymmetric and the second subject smile is symmetric and flat.

4.3.1.1.2 Quantitative results

We conduct our experiments on MMI* and UVA-NEMO databases as they contain more than one smile video per subject. To evaluate the results, we proposed to analyze the slopes of the generated trajectories with the 3 methods. As shown in Fig. 4.9, we calculated the slopes of the trajectories of the generated smiles and the trajectories of the learnt expression of the subject 1825 in Fig.4.8. We noticed that in the synthesized frames with Arias et al. [5], the corner of the lips is systematically raised which is justified with a steep slope. Moreover, the smile synthesized with Wang et al. [124] seems to be flat in Fig.4.8 and it's proved with a low slope. However, the slope of the smile generated of our method is the closest slope to the real smile ones. This result shows that our method generate a conform smile (mouth shape) to the person's one.

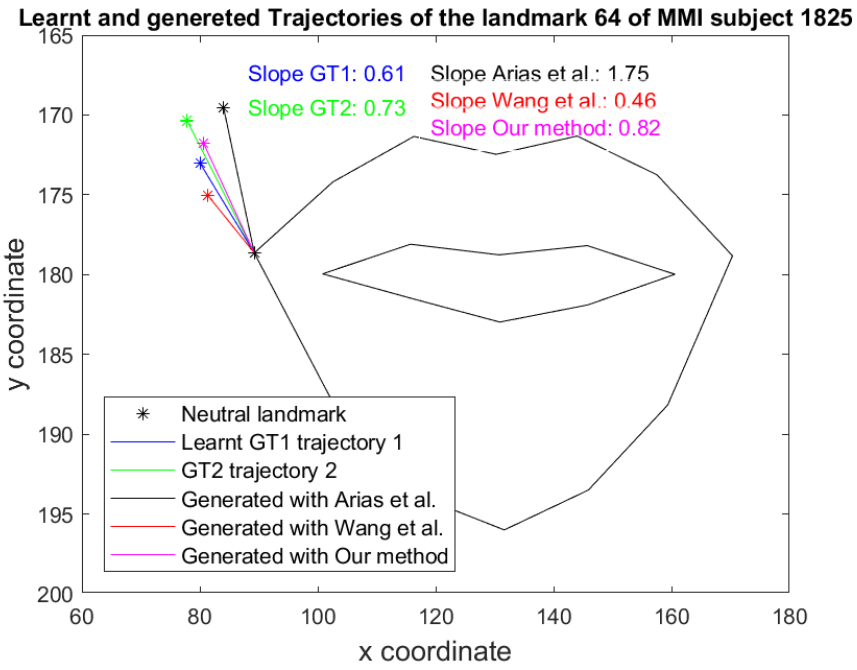


FIGURE 4.9 The two real smile trajectories and the synthesized ones with the different methods. We notice that the slope of the trajectory generated with our method is the closer to the real ones.

To go further evaluation we use the determined slopes to calculate the angles between the trajectories. The angle is determined by :

$$\theta = \tan^{-1} \left| \frac{s_{GT} - s}{1 + s_{GT} s} \right| \tag{4.2}$$

Where s_{GT} is the slope of the real smile expression (of the second smile) and s is the slope of the generated expression. Table 4.2 gives the results of the calculated angles with the 3 methods on the landmark of the left mouth corner (64). The angles represent the error between the generated trajectory and the target one. As shown our method gives the smaller mean. That means that we generate smiles which seems to be closer to the ground truth. We give the results of the rest of the mouth landmarks on the Appendix B (section B.1).

TABLE 4.2 Mean and standard deviation of angles calculated with the 3 methods on the 2 databases MMI* and UVA-NEMO for the landmark of the left corner of the mouth for all the subjects.

Method	MMI*		UVA-NEMO	
	mean	SD	mean	SD
Real smile1 and smile2 (baseline)	10.91	12.53	15.01	13.86
Arias et al. [5]	12.00	13.70	15.65	13.93
Wang et al. [124]	15.15	13.34	13.40	15.26
Our geometric method	**7.50**	**9.68**	**12.54**	**14.91**

4.3.1.2 Intensity manipulation

As explained before (section 4.3.1), as our method generates the learned Apex expression, it can also generate different intensities. The first way to manipulate the expression intensity is based on the variation of the coefficient d. This variation aims at synthesizing intermediate expression (d<1) between the neutral and the learned Apex or an amplified smile expression (d>1). As the correlation is high between the smile videos of the same subject, we propose a second way to manipulate the intensity (temporal-dynamic way) which consists of learning the temporal dynamic of the smile from one video and test it on another neutral frame. The test frame is extracted from a second video of that subject. We give the qualitative and quantitative results of the two ways in the following subsections.

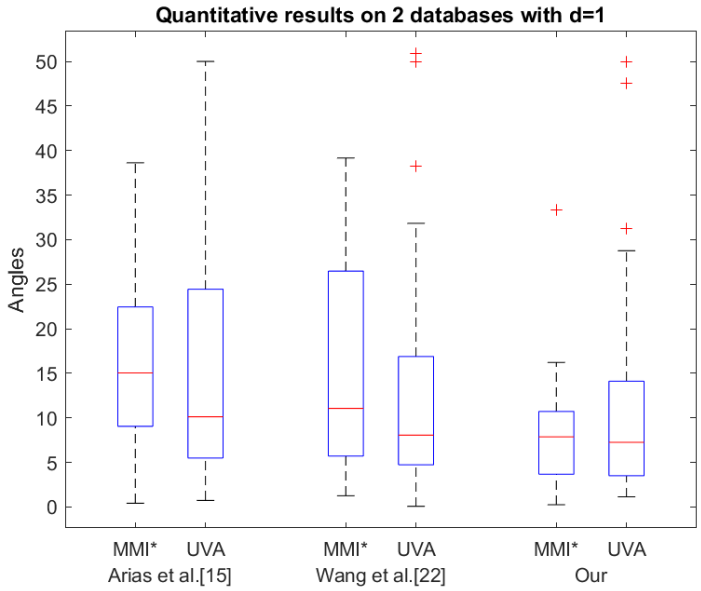


FIGURE 4.10 Representation of the angles calculated with the 3 methods on the 2 databases MMI* and UVA-NEMO for the landmark of the left corner of the mouth for all the subjects.

4.3.1.2.1 Qualitative results

We use the databases described previously to carry out these experiments. The databases MMI* and UVA-Nemo have two smile videos per subject, so we have the opportunity to learn a model on one video and test in on the second one. As illustrated in Fig. 4.11 the two proposed ways perform well the intensity manipulation. The videos $V1_{GT1295}$ and $V2_{GT1296}$ are two smiling videos of the subject of MMI* database.

To manipulate the expression of that subject based on their smile amplitude (using the d coefficient), we learn the person-specific model using the neutral and apex frame of the video $V1_{GT1295}$. We test that model on the neutral frame of the second video $V2_{GT1296}$ to generate $\widehat{V}2_{L1296}$ (Fig. 4.11, third row). The dynamic manner aims at generating a new smiling video $\widehat{V}2_{D1296}$ (Fig. 4.11, fourth row) based on the temporal dynamic model learnt (from neutral to apex) on the video $V1_{GT1295}$ and tested on the neutral frame of the second video $V2_{GT1296}$. As shown, the two ways of intensity manipulation allow different smile levels to be generated (neutral to apex). The quantitative results are given in the next subsection.

4.3.1.2.2 Quantitative results

As presented in the previous section, we personalize the shape to synthesize person-specific joyful expression with different intensities. We proposed two solutions to manipulate the

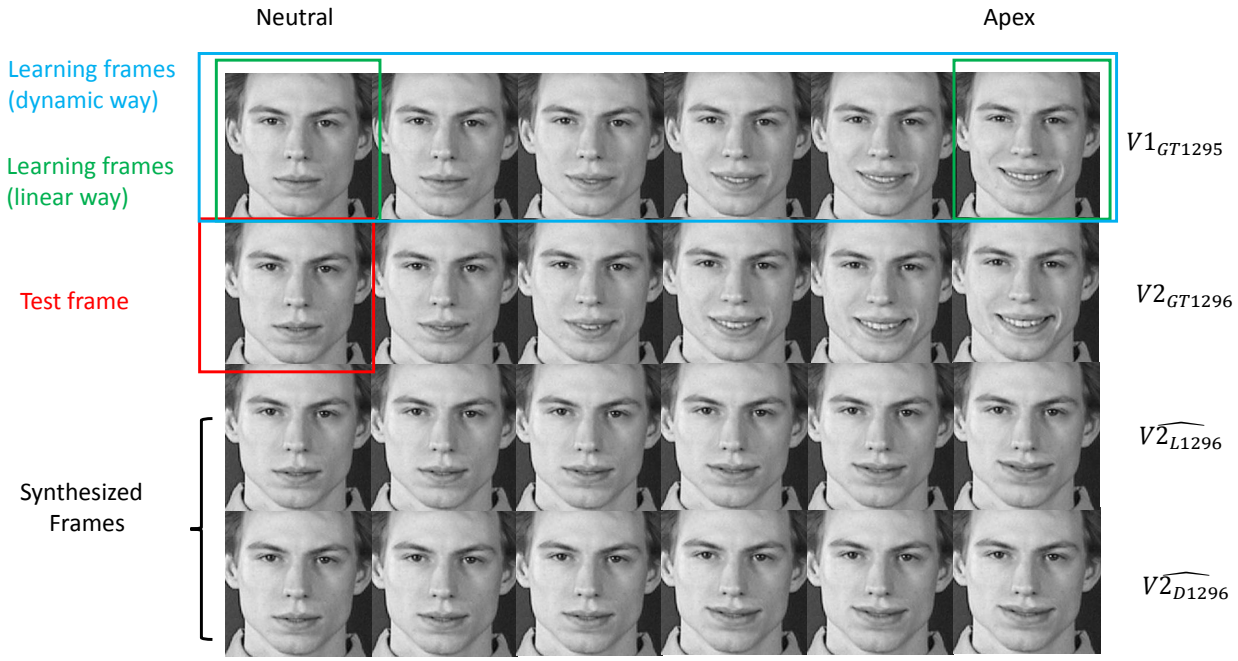


FIGURE 4.11 Examples of generated frames with different intensities. First row : $V1_{GT1295}$ first smiling video of the person. Second row : $V2_{GT1296}$ second smiling video. Third row : $\widehat{V}2_{L1296}$ the frames generated with the amplitude variation of the intensity learned from $V1_{GT1295}$. Fouth row : $\widehat{V}2_{D1296}$ the generated frames with the dynamic learned on $V1_{GT1295}$ and tested on the first frame of $V2_{GT1296}$.

generated smile intensities. In this subsection, we detailed the quantitative results achieved with these two solutions.

— **Intensity manipulation based on the smile amplitude**
The first originality of our system is that we can generate different intensities from a model learned using only 2 frames (neutral and apex) by the variation the coefficient d. To evaluate if the generated smile is that of the person or not, an evaluation metric is proposed. We choose to calculate the angles between the ground truth trajectory and the trajectories generated with the three methods (Ours, [5] and [124]) as shown in Fig. 4.12. To determine these angles, we use the linear regression $Y = a_{GT}X + b_{GT}$ of the ground truth trajectory. We assume that $Y = aX + b$ is the defined trajectory of the landmarks generated with one of the three methods. Then we use these defined trajectories to calculate the angles.
The angle is determined by :

$$\theta = \tan^{-1} \left| \frac{a_{GT} - a}{1 + a_{GT} a} \right| \tag{4.3}$$

Having the angles of each landmark trajectory, we calculate the mean and the standard deviation (SD) of these angles on the 4 databases. Our experiments experiments on CK, Oulu-CASIA and MMI databases aim at evaluating the efficiency of our method on generating the intermediate frames. The experiments on MMI* in which each subject has 2 smile videos intended to prove the relevance of our system in generating a person-specific smile expression on a new frame of the subject.

The angles mean represents the error between the ground truth (GT) trajectory and a generated trajectory so it is expected to be as small as possible. Table 4.3 shows the statistical results for the landmark located on the left corner of the mouth. We choose to represent the results of that landmark because the mouth corners are the most representative regions of a smile form. The results of the rest of the landmarks are given in Appendix B (section B.2.1). The results show that our method has the smaller mean and generates trajectories which are closer to the ground truth than the generic methods [5] and [124] (mean closer to 0). In addition, we can consider that our method is more stable than the other two methods because of the low value of SD. We observe that the results with MMI* are less good with our method than the results with MMI because the learnt model (first video) is tested on another video of

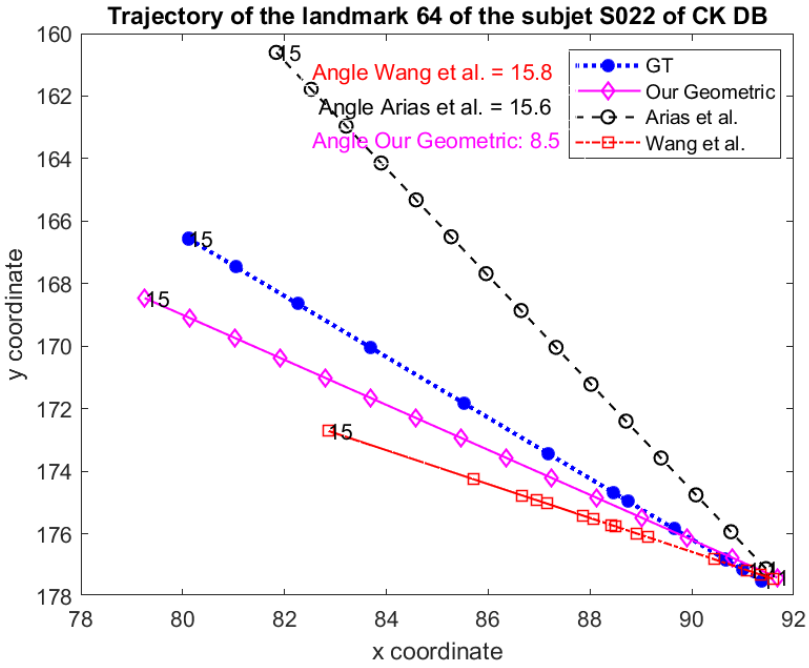


FIGURE 4.12 Example of the evaluation metric used for the evaluation of the smile expressions generated with the 3 methods. The point displacements are presented with a linear regression to calculate the angles between the trajectories of one method and the ground truth of the landmark located at the left corner of the mouth. We can notice that the angle between the GT trajectory and our method trajectory is smaller than the angle calculated with the geometric method [5] and with the GAN method [124].

the subject. In other words, we compare the generated smile with another video of the subject not with the learnt one. But we still get the closest results to the GT.

TABLE 4.3 Mean and standard deviation of angles calculated with the 3 methods on the 3 databases for the landmark of the left corner of the mouth for all the subjects.

Method	CK		Oulu-CASIA		MMI		MMI*	
	mean	SD	mean	SD	mean	SD	mean	SD
Arias et al. [5]	12.10	9.73	17.81	15.94	12.00	13.70	12.00	13.70
Wang et al. [124]	16.16	12.24	19.06	18.21	15.26	12.66	15.26	12.66
Our	**6.65**	**7.56**	**5.58**	**5.37**	**4.38**	**5.27**	**7.50**	**9.68**

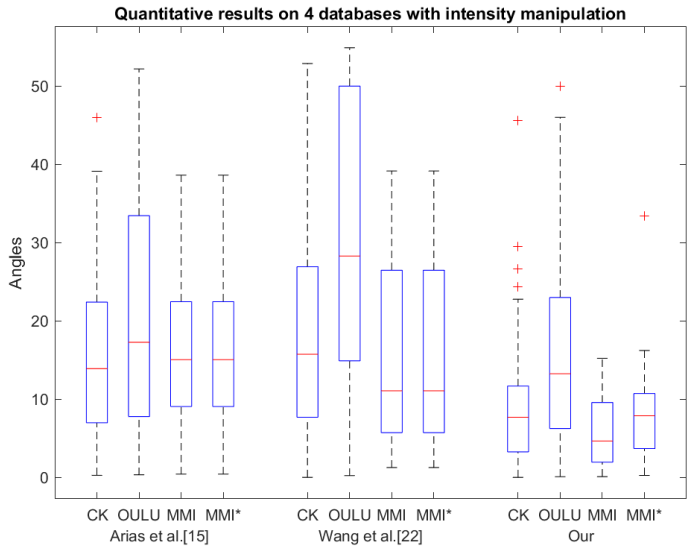


FIGURE 4.13 Representation of the angles calculated with the 3 methods on the 4 databases CK, OULU, MMI* and MMI* for the landmark of the left corner of the mouth for all the subjects.

— **Intensity manipulation based on the temporal dynamic of the smile**
We conduct these experiments to evaluate the performance of the dynamic model and compare it to the previously proposed way of smile manipulating (based on smile amplitude). We learn a dynamic person-specific model on one smiling video (from Onset to Apex) and test it on another one. Those tests are performed on MMI* and UVA-NEMO because in these 2 databases each subject has 2 smiles videos.
At first, we learn the dynamic model on video $V1_{GT}$ and test it on the first frame of another video $V2_{GT}$ of the person to generate $\widehat{V2_D}$. Moreover, we generate $\widehat{V2_L}$ using

the model learned on the neutral and apex frames of the video $V1_{GT}$. We generate also the $\widehat{V2_M}$ using the mean model on all the database. The mean model is the mean smile displacement of all the subjects on a database.
Once the videos are generated ($\widehat{V2_D}$, $\widehat{V2_L}$, $\widehat{V2_M}$), we determine the dynamic of each of the model landmarks. Each landmark dynamic is determined by the Euclidean distance between the position of this landmark in the neutral frame and the following frames of a video.
To evaluate the performance of the dynamic intensity manipulation, we propose to calculate the MSE and the correlation between dynamic landmarks in the real smiling video $V2_{GT}$ of the person and the generated ones $\widehat{V2_D}$, $\widehat{V2_L}$ and $\widehat{V2_M}$ using the dynamic person-specific model, the model based on amplitude and the mean model respectively.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(D^i_{V2_{GT\,j}} - D^i_{\widehat{V2_{D}}_j})^2 \tag{4.4}$$

Where $D^i_{V2_{GT\,j}}$ is the Euclidean distance between the position of the landmark i in the neutral frame and its position in the frame j of the video $V2_{GT}$.
Fig. 4.14 illustrates an example of the generated and real smile dynamics of 2 subjects. If we compare $V2_{GT}$ and $\widehat{V2_M}$ (Fig. 4.14 (a)), we notice that the intensity of the expression (Apex) is not taken into account by the average model, which influences the MSE values. While the model based on amplitude $\widehat{V2_L}$ takes into account the target intensity (Fig. 4.14 (a,b)), this is why the error MSE is quite small. Moreover, Fig. 4.14 (a) shows that the target intensity is achieved by our model based on temporal dynamic $\widehat{V2_D}$. Fig. 4.14 (b) shows that neither the average nor the model based on amplitude considers the elasticity of the movement (i.e how much time (frames) is needed to pass from neutral to the apex for each subject). However, the variability in Neutral-Apex time between subjects is taken into account by the dynamic model but not by the amplitude or average models. That explains the highest correlation and the smaller error achieved with our dynamic way of intensity manipulation.
The evaluation results of the left corner landmark (landmark 64) dynamic are presented in table 4.4. The results show that the videos $\widehat{V2_D}$ are more correlated to $V2_{GT}$ than $\widehat{V2_L}$ and $\widehat{V2_M}$. We notice that the dynamic way of intensity manipulation conserve the correlation between the real smile videos of each subject. Therefore, our method makes it possible to generate the closest expression to the ground truth with

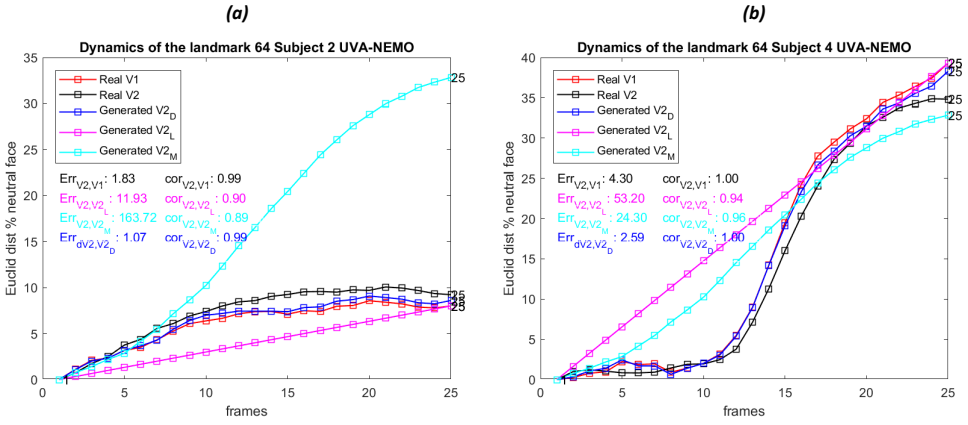


FIGURE 4.14 Example of the generated and real smile dynamics of 2 subjects. We notice that the dynamic model (blue) respects the Neutral-Apex time for each subject unlike the model based on amplitude (magenta) $\widehat{V2_L}$ or the average model (cyan) $\widehat{V2_M}$. The intensity of the expression in (a) (Apex) is taken into account by the dynamic model $\widehat{V2_D}$ but not with $\widehat{V2_L}$ and $\widehat{V2_M}$ models. Moreover, the target intensity is usually achieved by the dynamic model $\widehat{V2_D}$.

different intensities (cf. Fig. 4.14). The results of the rest of the landmarks are given in Appendix B (section B.2.2).

TABLE 4.4 Quantitative results on shape : Mean error and correlation results on MMI* and UVA-NEMO databases. The results show that the dynamic model generate the closest dynamic to the real smile of each subject.

Method	MMI*		UVA-Nemo	
	MSE	correlation	MSE	correlation
$(V2_{GT}, \widehat{V2_M})$	14.67	0.82	83.75	0.96
$(V2_{GT}, \widehat{V2_L})$	12.94	0.82	38.38	0.95
$(V2_{GT}, \widehat{V2_D})$	**12.04**	**0.85**	**30.62**	**0.98**

4.3.2 Qualitative results on texture

As described previously, the geometric part of our system provides an efficient personalized shape deformation to synthesize a smile expression. However, we need to add details to make it real. To refine global texture details of the synthesized expression (person-specific

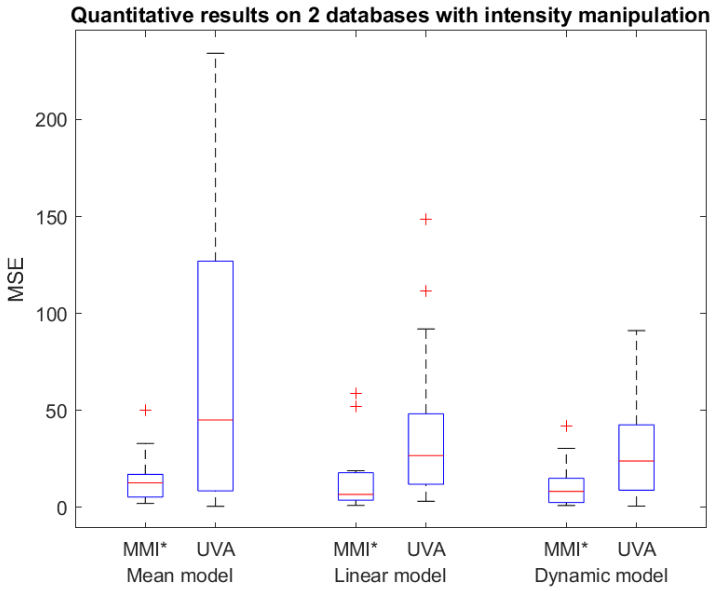


FIGURE 4.15 Representation of the results of the MSE of 2 databases MMI* and UVA-NEMO.

shape), we proposed to use a GAN. The qualitative and quantitative results are presented in the following subsections.

4.3.2.1 GAN training

The originality is that the input of the GAN is the frames already deformed in shape (using the geometric part). All the synthesized images are normalized, aligned and cropped to 256×256 size to train the GAN. In the training phase, we perform random flipping of the input images to encourage the generalization of the network. We use leave-one-out cross-validation to train the GAN. We use all the synthesized smile images of the subjects of the 4 databases for the training and the frames of one remaining subject for the test. This method led to train one model per subject. The generated frames with the two ways of intensity manipulating (based on the smile amplitude and based on the temporal dynamic) could be used for the refinement.

For the GAN, we adopt the architecture from [50]. The generator G is a U-Net auto-encoder which takes as input the synthesized image as shown in Fig. 4.16. The G_{enc} contains 8 convolutional layers. The first one is a simple convolutional layer with a 5×5 kernel and stride 2. The others layers are composed of Leaky ReLU as activation function, convolution

with 5×5 stride 2 and a batch normalization. G_{dec} is composed of 8 deconvolutional layers with 5×5 stride 2 and Leaky ReLU. To share the information between the input and the output features, the decoder layers have skip connections with their corresponding layers of G_{enc}. Adaptive Moment Estimation optimizer (ADAM) is used to train our model with β=0.5, 0.0002 as learning rate, λ_1=1, λ_2=0.1 and λ = 100 (See equation 3.15, Chapter 3).

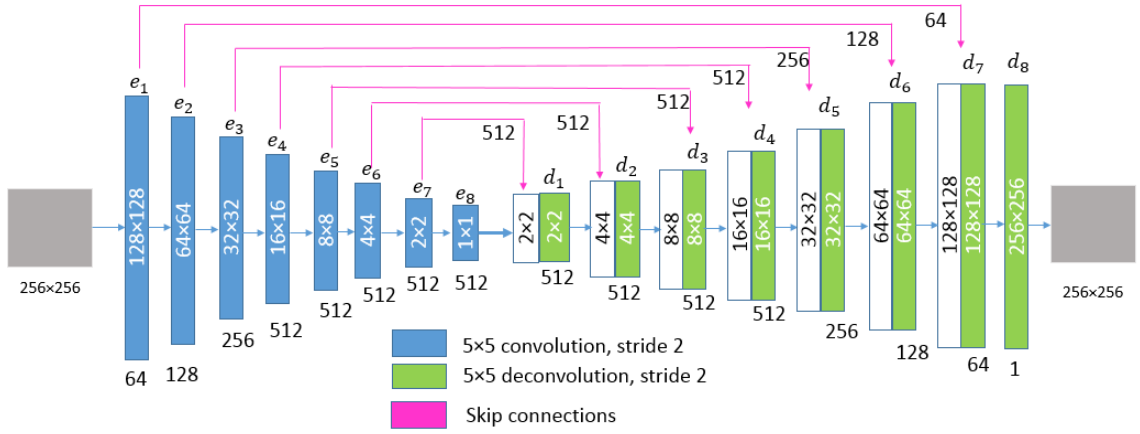


FIGURE 4.16 The architecture of the proposed GAN. The G_{enc} contains 8 convolutional layers. The first one is a simple convolutional layer with a 5×5 kernel and stride 2. The others layers are composed of Leaky ReLU as activation function, convolution with 5×5 stride 2 and a batch normalization. G_{dec} is composed of 8 deconvolutional layers with 5×5 stride 2 and Leaky ReLU. To share the information between the input and the output features, the decoder layers have skip connections with their corresponding layers of G_{enc}.

4.3.2.2 Texture refinement of the frames synthesized with amplitude variation

In this subsection, we compare the results of refinement on the frames synthesized with the amplitude variation and the results of state of the art methods [5, 124]. This way of manipulation aims at synthesizing for each subject a smile sequence. To add global details to these synthesized expressions, they must go through the GAN to synthesize more refined frames. We give qualitative and quantitative results in these frames in the following paragraphs.

4.3.2.2.1 Qualitative results

In this section, we present the results of refinement on the frames. The frame synthesized by the geometric method is fed to the trained GAN to add the missing details. Fig. 4.17

illustrates the results [1] obtained for 2 intensities for 4 subjects. The Qualitative results confirm that our method gives closer results to the GT smiles than [5] and [124].

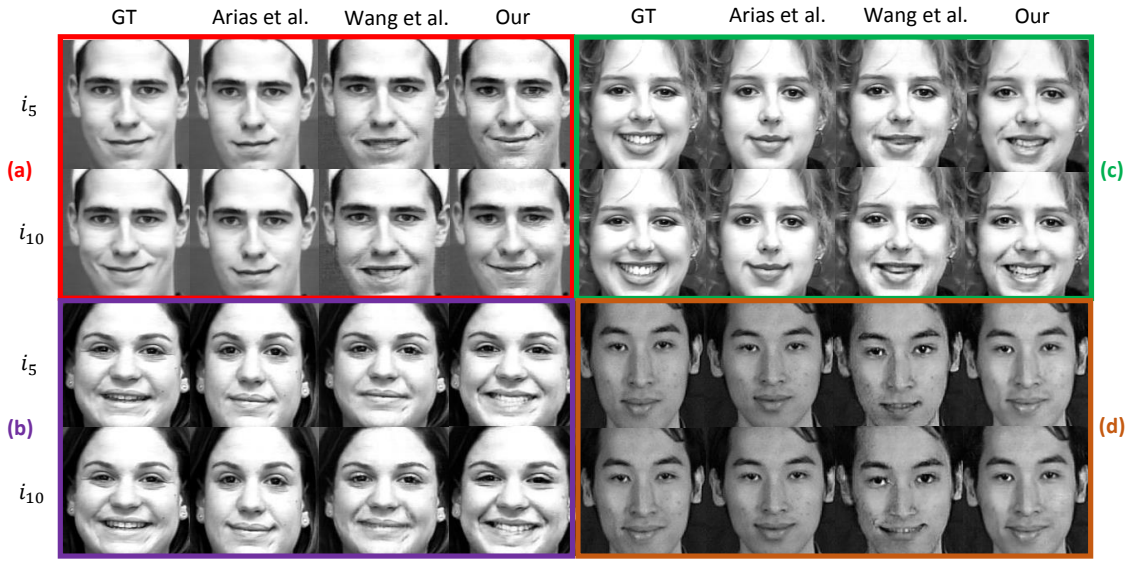


FIGURE 4.17 The Ground Truth smile (GT) and the result frames of 4 subjects from the 2 databases (MMI and CK) with two intensities for each subject.

Concerning the shape, we observe that with the geometric method of Arias et al. [5], the corner of the lips is systematically raised (steep slope) for all the subjects. We notice that the subjects (a) and (b) have a flat smile (low slope) in their real smiles, while the smiles generated with [5] is growing and not realistic. The GAN in Wang et .al [124] generates different smiles but we can perceive that are not those of the subjects. For the subject(c), the real smile is opened but the GAN generates a tight smile.

Concerning the texture, we notice that with Arias et al. all the generated smiles are without teeth whatever the texture shape of the smile. For the subjects (a) and (d), the real smile is without teeth but the GAN of [124] generates the teeth for these subjects. On the contrary, for the subject (b) the real smile has occurred with teeth but the GAN generates it without. Then, the GAN generates realistic joyful expression but not those of the person. Unlike our method which respects the smile shape and global details of the subject, so if the subject smiles with his teeth we generate an open mouth smile, if not the synthesized mouth will be closed. Quantitative results are presented in the following subsection.

1. More results are available on https://drive.google.com/file/d/1hY15DNrrYNjxnf9JmVFJhWUYhteTYnQa/view?usp=sharing

4.3.2.2.2 Quantitative results

The metrics we use to check the performances of our method are the same used to evaluate our geometric part. We determine the angles between the ground truth smile trajectory and the trajectories of the generated smile as illustrated in Fig. 4.18. Table 4.5 shows the statistical results for the landmark located on the left corner of the mouth. The results of the rest of the landmarks are given in Appendix B (section B.2.1). The results show that our method generates trajectories which are the closer to the ground truth trajectories (a mean closer to 0). In addition, we can consider that our method is more stable than the other two methods because of the low value of SD. We notice that the GAN influences the resolution of the images, that is why the results of the hybrid method are less good than those of our geometric method, but the generated expression remains closer to the GT than that generated with the other two methods [5] and [124]. As a baseline, we investigate the impact of landmarks noise represented by a Gaussian distribution of mean 0 and variance 1 on the landmarks trajectories. The results shows that the angles are influenced of 3.18 on CK database.

TABLE 4.5 Mean and SD of angles calculated with the 3 methods on the 4 databases for the landmark of the left corner of the mouth for all the subjects.

Method	CK		Oulu-CASIA		MMI		MMI*	
	mean	SD	mean	SD	mean	SD	mean	SD
Arias et al. [5]	12.10	9.73	17.81	15.94	12.00	13.70	12.00	13.70
Wang et al. [124]	16.16	12.24	19.06	18.21	15.26	12.66	15.26	12.66
Our geometric method	**6.65**	**7.56**	**5.58**	**5.37**	**4.38**	**5.27**	**7.50**	**9.68**
Our hybrid method	7.65	8.25	6.85	7.26	5.83	7.01	9.20	10.78

4.3.2.3 Texture refinement of the frames synthesized according to the temporal dynamic of smile

In this subsection, we compare the results of refinement on the frames $\widehat{V2_{RL}}$ and $\widehat{V2_{RD}}$ which are generated based on the amplitude and the temporal-dynamic of the smile respectively. The refinement of the frames aims at adding the texture and generate more detailed expressions. Qualitative and quantitative results are presented in the following paragraphs.

4.3.2.3.1 Qualitative results

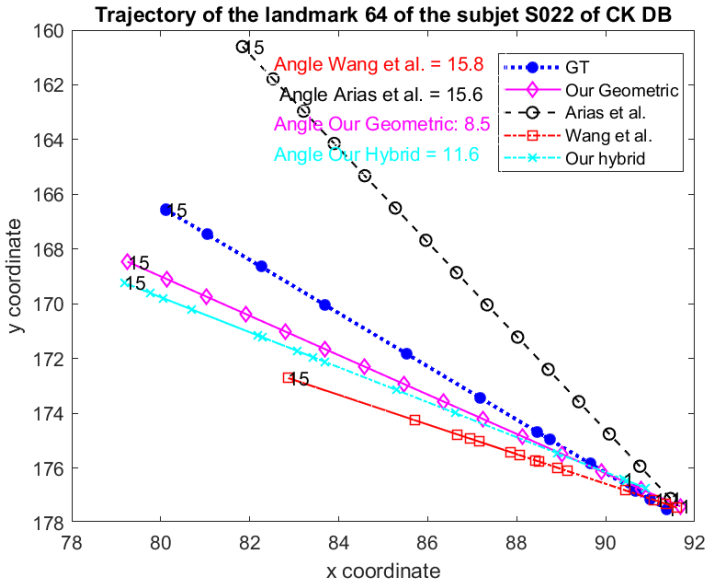


FIGURE 4.18 The same metric is used for the evaluation of the smile expressions generated with the hybrid method. We calculate the angles between the real smile trajectory and the generated ones. As shown the GAN influence the trajectory generated by the geometric method but the generated expression still the closer to the GT trajectory.

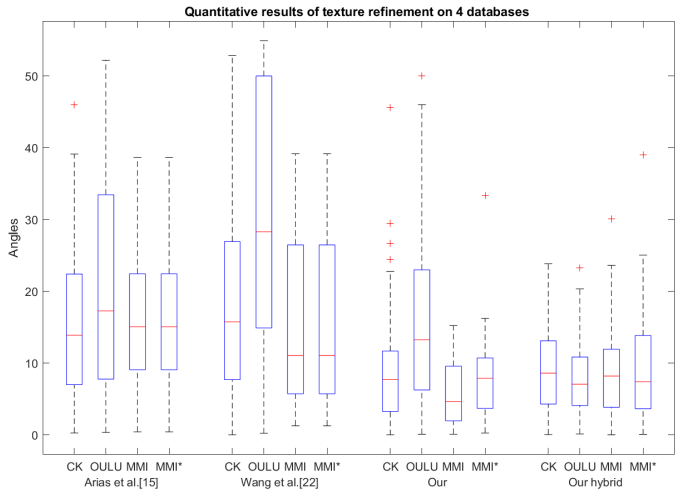


FIGURE 4.19 Representation of the angles of refinement results on the 4 databases with the different methods.

For the MMI* and UVA-NEMO databases, since each subject has 2 videos, we use one smile video for learning the person-specific model while the second video is used for the test. Fig. 4.20 illustrates the results of the refinement on the frames generated based on dynamic $\widehat{V}2_{RD1296}$ and based on amplitude $\widehat{V}2_{RL1296}$. Our UC-GAN is used to add the wrinkles and teeth to the smiling frames generated by the geometric part. The generated smiles are compared to $V2_{GT1296}$.

Comparing to results in Fig. 4.11 the proposed hybrid method influences the resolution of the images but shows it's efficiency to add such global texture details like appearance of teeth and global wrinkles. The visual fidelity shows that the refined frames ($\widehat{V}2_{RL1296}$, $\widehat{V}2_{RD1296}$) are close to the real smile. They maintain as much as possible the global texture of the real smile of the subjects. However, they appear as not very real because the GAN doesn't add the person's teeth.

4.3.2.3.2 Quantitative results

As presented in the previous subsections, at first we personalize the shape to synthesize person-specific joyful expression with different intensities (based on amplitude : $\widehat{V}2_L$ and based on temporal dynamic $\widehat{V}2_D$). Then, we refine the global texture on these synthesized videos using our UC-GAN to generated the refined videos ($\widehat{V}2_{RL}$ and $\widehat{V}2_{RD}$). In this sub-subsection we detailed the quantitative results achieved on the refined frames with two

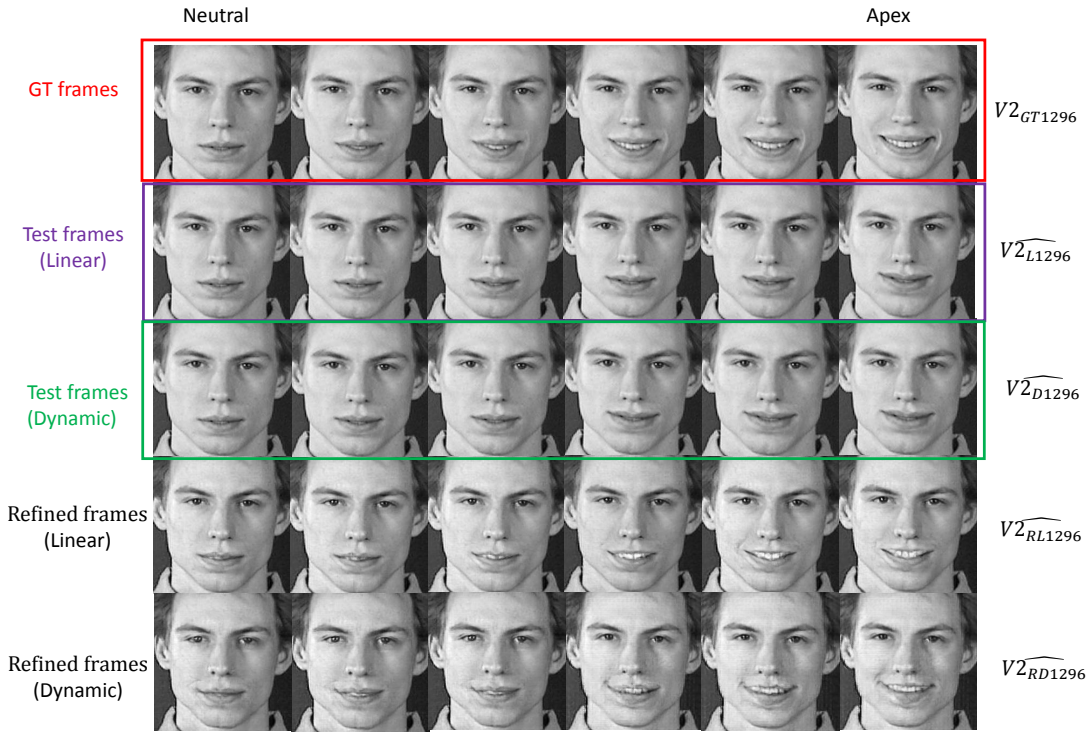


FIGURE 4.20 Examples of the refined frames. First row : $V2_{GT1296}$ the ground truth frames. Second row : $\widehat{V}2_{L1296}$ the frames generated with the amplitude variation of the intensity learned from $V1_{GT1295}$. Fouth row : $\widehat{V}2_{D1296}$ the generated frames with the dynamic model learned on $V1_{GT1295}$ and tested on the first frame of $V2_{GT1296}$. These frames are refined using our proposed GAN to generate more real expressions $\widehat{V}2_{RL1296}$ and $\widehat{V}2_{RD1296}$.

proposed manners of intensity manipulation. The results are shown in Table 4.6. We notice that the results of the landmark 64 (left mouth corner) are better than that in the Table 4.4 because we keep almost the same correlation and we approach of the real $V2_{GT}$. The results show that the refined videos of the dynamic model are the closest to $V2_{GT}$. The results of the rest of the landmarks are given in Appendix B (section B.2.2).

TABLE 4.6 Quantitative results on texture : Mean error and correlation results on MMI* and UVA-NEMO databases.

Method	MMI*		UVA-Nemo	
	MSE	correlation	MSE	correlation
$(V2_{GT}, \widehat{V2_M})$	14.67	0.82	83.75	0.96
$(V2_{GT}, \widehat{V2_L})$	12.94	0.82	38.38	0.95
$(V2_{GT}, \widehat{V2_D})$	12.04	0.85	30.62	0.98
$(V2_{GT}, \widehat{V2_{RL}})$	11.66	0.84	34.41	0.96
$(V2_{GT}, \widehat{V2_{RD}})$	**10.76**	**0.85**	**29.25**	**0.98**

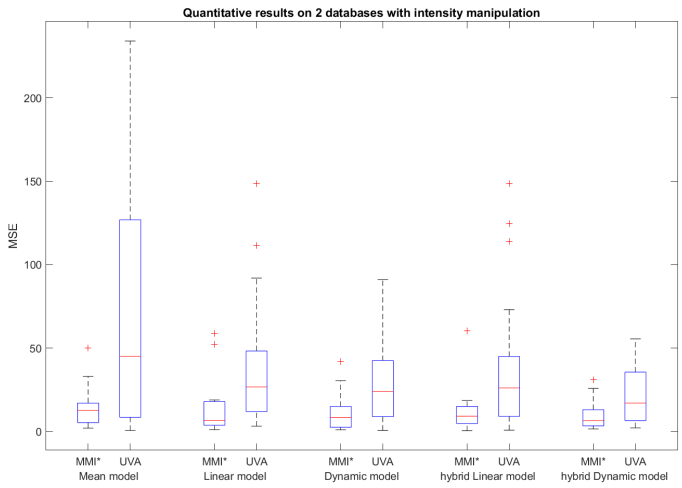


FIGURE 4.21 Representation of the MSE after the refinement step on the 2 databases with the different methods.

We notice that the GAN refines the global-face texture such as wrinkles and teeth appearance. However it cannot refine local details such as adding the personal teeth for each person. We proposed 2 solutions to overcome this limitation. The results are given in the next subsection.

We measured the latency of the overall algorithm. Our tests performed on an Intel processor Core i7 at 3.30 GHz with an NVIDIA geforce GTX 1070. The mean time to process a single frame is 65 ms. It is suitable for real time applications (15 fps) such as the mirror in our use case.

4.3.3 Teeth refinement

In the previous subsections we presented the efficiency of our hybrid system on generation person-specif joyful expression in shape and on adding global texture. However, we notified that the GAN cannot guess which type of teeth it should add for such a mouth or subject. To ameliorate our method, we proposed to use one of two methods that aim at enhancing the generated expression (Chapter 3, section 3.3). In this subsection we present the qualitative and quantitative results results of these two methods.

4.3.3.1 Qualitative results

The two proposed solutions to refine teeth (PE and AE ; Chapter 3, section 3.3) are used with the databases in which each subject has 2 smiling videos (MMI* and UVA-NEMO). The idea is that the geometric part (Geo) aims at personalizing the shape, the GAN contribute to add global texture and either the AE or PE are employed to add the teeth of the person to the generated frames. The results illustrated in Fig. 4.22 show that the resulting frames with Geo-GAN-PE seem to be much more realistic compared to the ones generated with Geo-GAN and the same for the frames generated with the Geo-GAN-AE. However, with PE we are limited to blend only frontal faces. Therefore, the real advantage of the GAN-AE is that it can learn different head positions to refine their corresponding texture.

4.3.3.2 Quantitative results

To evaluate the final refined expression, we proposed to calculate the 2D correlation between the real frames and the generated frames (Geo-GAN, Geo-GAN-AE and Geo-GAN-PE). Correlation coefficient R is widely used in statistical analysis, pattern recognition and images comparison [60]. For the quality assessment between the real frames and the generated frames, we calculate the MSE between the frames. Table 4.7 illustrated the results on MMI* and UVA-NEMO databases. We use these databases in this experiments because each subject has 2 videos smiles. The table proves that adding this step to refine local-mouth texture (teeth) succeeds in enhancing the generated frames and synthesizing more person-specific

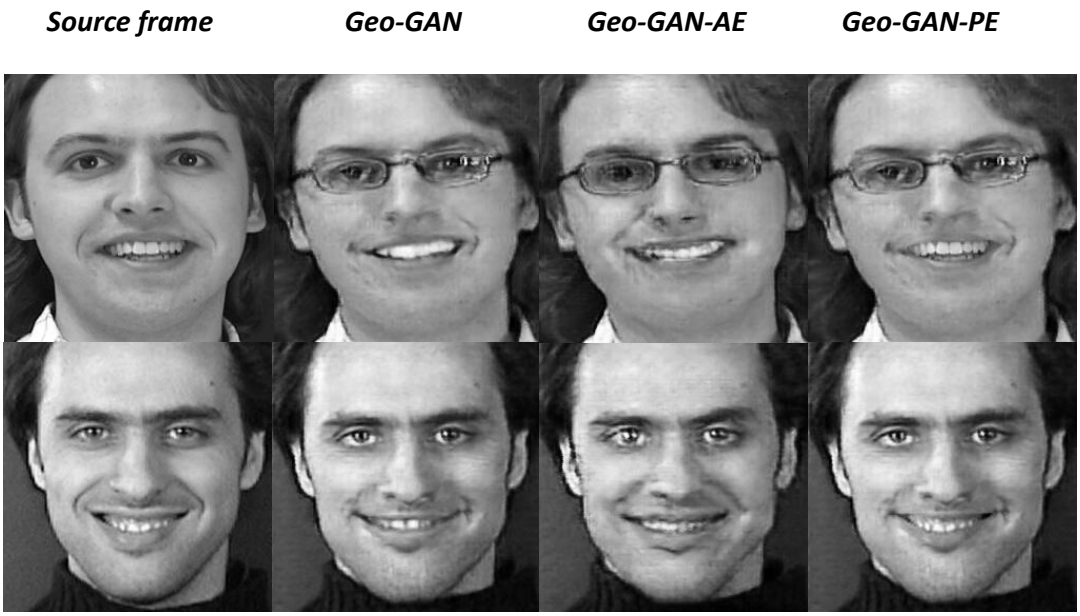


FIGURE 4.22 First columm : the source frames used from a second smiling video of the person to extract the mouth region. Second column : the generated frames with our system Geo-GAN. Third column : the generated frame using Geo-GAN-AE framework. Fourth column : the Blended frames Geo-GAN-PE result after refining the generated frames using the Poisson blending method . The frames generated with Geo-GAN-PE and Geo-GAN-AE seems to be much more realistic compared to the generated ones (Geo-GAN).

joyful expressions.

TABLE 4.7 Quantitative results : Mean error and 2D correlation results on MMI* and UVA-NEMO databases.

Method	MMI*		UVA-Nemo	
	MSE	R	MSE	R
Geo	45.21	0.91	22.74	0.93
Geo-GAN	31.38	0.93	13.58	0.96
Geo-GAN-AE	30.11	0.94	12.24	0.96
Geo-GAN-PE	**24.11**	**0.96**	**10.05**	**0.97**

We notice that using the GAN minimizes the MSE and enhances the correlation, which is justified by the appearance of the teeth and wrinkles on the generated images and makes the generated expression as a real one. The Geo-GAN-AE aims at adding a specific information about the person's mouth region which helps the GAN to add more specific information. The AE contributes to more refine the texture and specially the mouth, which is justified by the positive influence on the results of MSE and conserve the same correlation as Geo-GAN. Using the real teeth to blend the generated images (Geo-GAN-PE) improves the appearance and the resolution of the generated expression and makes it appear as a real expression (low MSE and high correlation).

4.4 Conclusion

In this chapter, we demonstrate that the proposed hybrid method is relevant for generating person-specific joyful expression. In addition, the global and local texture refinement makes the expression more realistic and closest to the real smile of the person. The evolution of the method performances is presented in table 4.8. Furthermore, our method synthesizes a smile expression while respecting the way of smiling of each subject unlike the two SOA methods (a geometric and a machine learning method).

In the previous chapters we studied a person-specific way of producing a joyful expression. The perception aspect is studied in the following chapter, which aims at investigating the way each person perceives the joyful expression on her own face and on the faces of the others.

TABLE 4.8 Evolution of the method performances with the proposed solutions.

Method	**Geometric (Geo)**	**Geo-GAN**	**Geo-GAN-AE**	**Geo-GAN-PE**
Person-specific shape	+	+	+	+
Smile texture	-	+	+	+
Wrinkles/ dimples	-	+	+	+
Specific teeth	-	-	+	+
Resolution	+	+/-	+/-	+

Chapitre 5

The perception of joyful expressions : Mental representation analysis

5.1 Introduction

After studying the person-specific way to express joy, we headed towards analyzing the perception of joyful expressions. Mirror-neuron-like systems [38] would fire when we observe an action (typically a facial expression of smile) performed by another person, mentally simulating that action for ourselves [45]. We expect that self-related production representations (Self facial movement) may subtend how we perceive others. But the question arises whether it is true that we perceive as we produce.

This chapter is a result of a collaborative work with IRCAM (Acoustic/Music research and coordination institute). Our collaboration aims at analyzing the cited hypothesis and investigating the way each person perceives the expression of joy on her own face and on the faces of the others. Therefore, we headed to determine and analyze the mental representations (MRs) of joyful expression of real subjects (perceiving their faces and that of others). The MR is about how we store information in memory and how we represent it in our mind's eye or manipulate it through our processes of reasoning. First, we propose a tool that consists on generating random faces expressions from a real neutral face. For this step, we used our deformation method proposed in the previous chapters. Second, we (along with IRCAM) conducted a psycho-physical experiment, in which several observers have to categorize the random facial expressions generated. From their responses, we reconstruct each individual observer's mental representations, using the "reverse-correlation" procedure [2, 97, 78]. That allows us to determine the mental representations that people have of a given face. Each observer is first invited to categorize the random expressions on their own face to determine

Self MR and next categorize 2 actors random expressions to determine 2 **Actors** MRs. Once the mental representations are determined, IRCAM analyzes three hypotheses related to the perception of joyful expressions as illustrated in Fig.5.1. The Results of this analysis are presented in Appendix A. They investigate the link between the MRs of joyful expression on the participant's face (**Self** MR) and on the faces of others (2 **Actors** MRs). Moreover, they search the link between the observer's MRs and their smiling manner (**Production**), as well as their personality's traits.

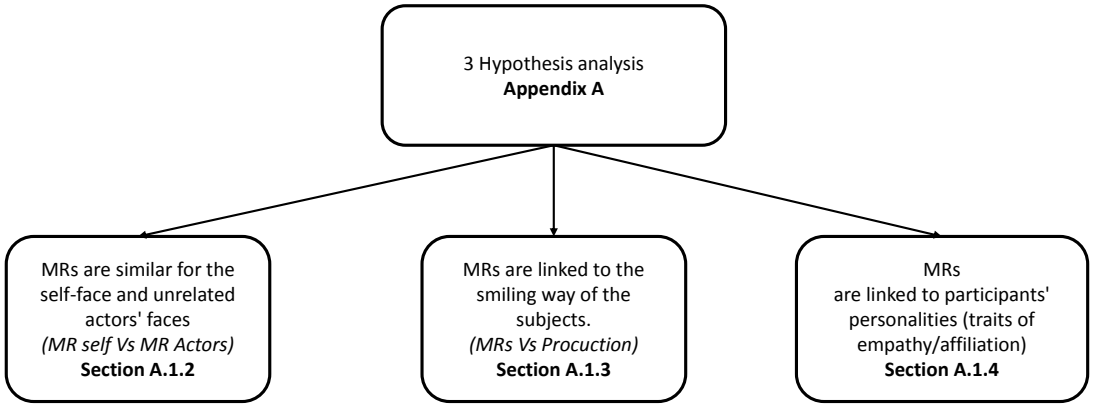


FIGURE 5.1 The 3 investigated hypothesis. The first one is that there is a significant association between the self-face and other-face representations. The second is that these representations are linked to the individual manner in which participants produce these expressions, either for the self-face or for unrelated actors. The third is that the traits of affiliation/empathy affect how people perceptually represent themselves and also affect how people represent others. The results of analyzing these hypotheses are given in appendix A.

My work in this chapter is proposing a tool to build MRs on real subjects. First, we conduct an overview of the research state of the art in mental representations (Section 5.2). Second, we introduce our proposed system for visual deformation in section 5.3. Then, in section 5.4, we describe the experimental procedure used to measure the different mental representations kernels. The kernel is defined as a landmarks vector deformation, which is computed for each participant, as the landmarks displacement from a neutral expression to a more joyful one (real smile or MRs). In section 5.5, we present and analyze the determined mental representations. As it is a collaborative work, statistical studies are conducted by IRCAM to analyze the similarity and the differences of the different determined kernels and investigate the three posed hypothesis. The results of these analysis are given in Appendix A.

5.2 Mental representation

The concept of mental representation (MR) [73] aims at explaining the nature of ideas, concepts, and other mental contents. The cognitive sciences [62, 111] consider that MRs are how we mentally represent information and define it as a direct mapping of what we see and hear about us in the world. For instance, the MR [10] is the sensory template that enables an observer's to categorize the visual stimuli corresponding to various facial expressions. It indicates the representation that the person has of a particular expression, either by an effort of memory or by imagination. For example, when someone tells someone else an event or a story, the mental image of that event is constructed in their mind and she imagines it.

MR has always seemed relevant in psychological research. Several researchers have already exploited this concept of perception and MR for their works. For instance, Jack et.al [51] used determined mental representations on avatars to study the universality of expressions. They showed that facial expressions of emotion are culture-specific, refuting the notion that human emotion is universally represented by the same set of six distinct facial expression signals. The same procedure of Jack et.al is used in [95] to determined mental representations of facial expression for subjects with prosopagnosia. The Prosopagnosia is a cognitive disorder of face perception in which the ability to recognize familiar faces, including self-face recognition, is impaired, while other aspects of visual processing and intellectual functioning remain intact. The study aims at analyzing whether the visual system relies on a unique representational system to code face features for identity and expression. The method of determining MR is very often identical. A software generates animations on avatar faces according to a set of Action Units (AU). It then generates a multitude of more complex face animations by temporally mixing these AUs. These animations are then presented to a subject who rates these faces according to the criteria proposed. To model the mental representation of each observer in [51, 95], they reverse correlated [2, 78] the generated random facial movements with the emotion response that these random facial movements elicited. The reverse correlation is a technique used at first in neurophysiology for studying how sensory neurons sum up stimuli that they receive at different times, to generate a response [97]. In [139] researchers used the same idea to categorize synthesized animations according to one of six possible facial expressions. They generated random animations and presented it to human observers for rating. The reverse correlation is also used in auditory to study temporal loudness weights [90] and to study intonation of interrogative vs declarative speech and rhythm of sung melodies [13]. Burred et .al [13] have developed a piece of software, CLEESE, to modify the intonation of the voice. From the recording of one word or more, they randomly generate thousands of other intonations having a different pitch

while keeping a realistic tone. Then, they conducted perception tests with these manipulated recordings and analyzed the results of the participant. In other words, they studied intonation of interrogative vs declarative speech to determine MRs of both of them. This interesting metric (Reverse correlation) and the protocol used by [51] provides directly understandable and credible results in various works. However, in these studies, they use random AUs and avatars for their experimental protocol to determine the mental representations rather than real subjects. Our work's originality is that we determine MRs on real subject's faces.

In our work, we headed to determine MR of the joyful expression on the own face of an observer (**Self**) and his MRs on other's faces (**Actors** MRs). The tool proposed for determining these mental representation is presented in Fig.5.2. Our process starts with a deformation block in which we randomly deform a neutral face using the visual deformation block. Then, we conduct a perception experiment, in which the participants have to categorize the generated random facial expressions in different trials. In each trial, we present 2 images and the participant chooses the most friendly expression. From their responses, we determine each participant's mental representation kernel [78] on that face. The kernel is defined as the landmarks displacement from a neutral expression to a more joyful one (mental representation). More details of our proposed framework are given in the next sections.

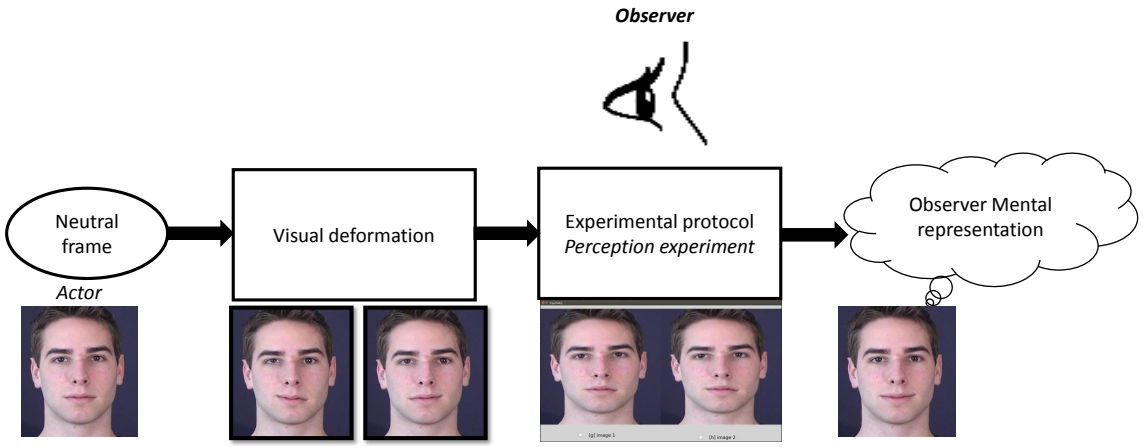


FIGURE 5.2 Overview of the process of determining Mental representation. Our framework starts with a deformation block in which we randomly deform a subject neutral face. Then, we conduct a perception test for a number of observers to reconstruct the mental representation of each of them.

5.3 Visual deformation

The first step for determining a mental representation that an observer has on a given face is generating randomly expressions from that face. Our framework aims at deforming a neutral face to synthesize random facial expressions as illustrated in Fig. 5.3. The originality of our method is that we determine the MRs on real faces. To this aim, we first detect the landmarks on a neutral face (subsection 5.3.1). Second, we randomly generate 1400 random positions for the the eyes, eyebrow, the forehead, the nose and mouth landmarks (subsection 5.3.2). Finally (subsection 5.3.3), we generate the 1400 random expressions using the warping method introduced previously (chapter 3 section 3.1.4).

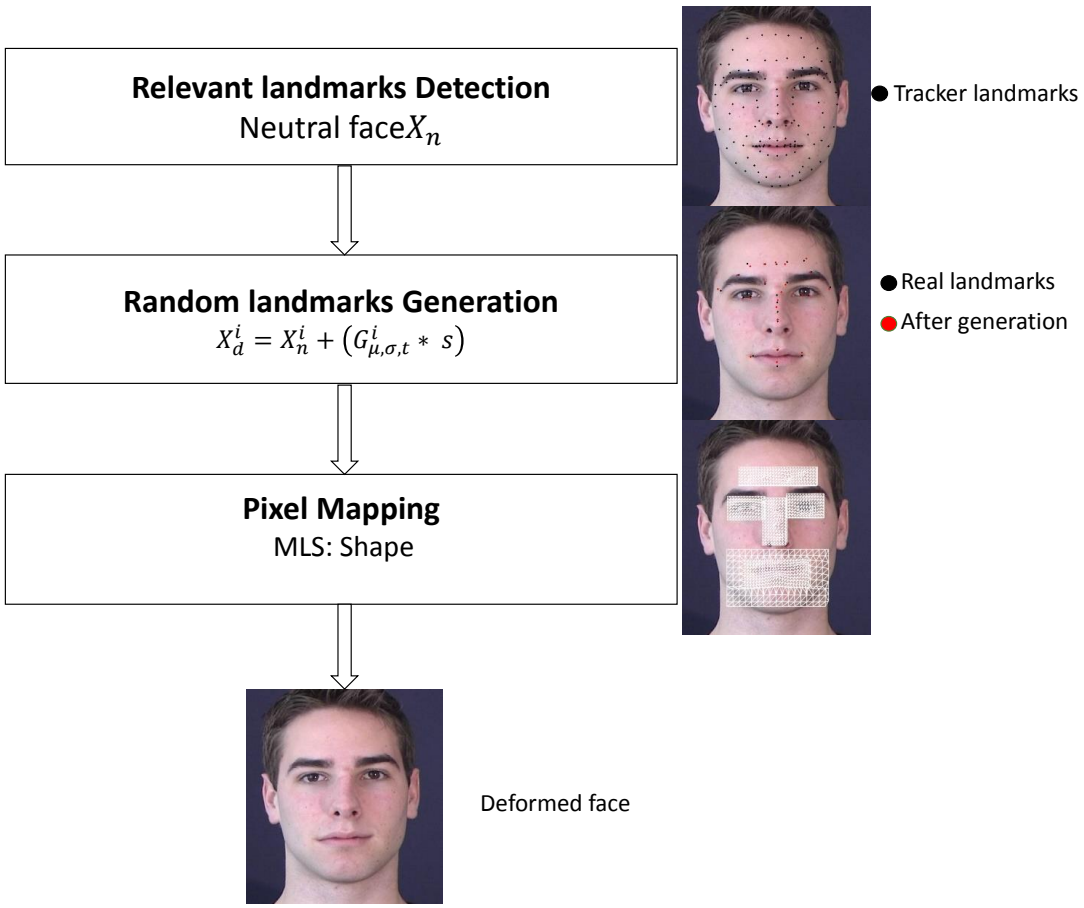


FIGURE 5.3 Overview of the visual deformation. In the first step, we track one neutral X_n face to extract the landmarks positions. Then, we randomly generate new positions for the selected landmarks around the eyes, eyebrow, the forehead, the nose, and mouth regions. The landmarks displacement is employed using a Gaussian distribution to avoid aberrant faces. In the last step, we employed a 2D warping method MLS [105] to deform the neutral face with the new random positions of the landmarks.

5.3.1 Relevant landmarks detection

Our algorithm starts with a tracking step. We used Kurma tracker of Dynamixyz [24] to detect the real face and determine precisely the coordinates of 149 landmarks. Selecting the subset of landmarks which are to be moved randomly is non trivial. This selection is necessary because the displacement of the totality of the landmarks causes the generation of aberrant faces because of the potentially opposed displacement of adjacent landmarks. So it risks to strongly affect the reliability of the final result. That's why we choose 23 landmarks of different face regions ; the landmarks of left eye (45, 46), left eyebrow (17, 20, 23), right eyebrow (29, 32, 35), right eye (51, 52), nose(53, 54, 55, 56), forehead (137, 138, 139, 140, 141), and mouth (64, 76, 70, 73) as shown in Fig.5.4 A.

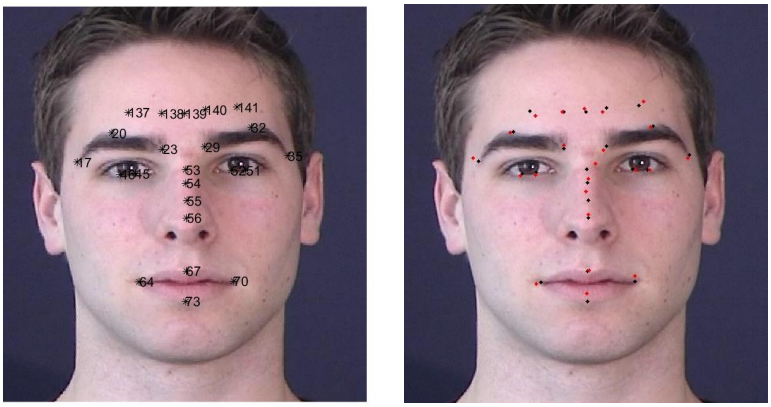


FIGURE 5.4 A : Positions of the selected landmarks detected using the Kurma tracker [24] (Black). To avoid the generation of aberrant faces, we choose to deform this set of landmarks. B : in Red are the random landmarks generated using the Gaussian distribution to generate a random expression.

5.3.2 Random landmarks generation

As shown in Fig.5.4 B, once the landmarks positions are determined we move these landmarks along two Gaussian distributions, one on the X coordinates and the other on the Y coordinates. In other words, the displacements of these landmarks are determined by sampling from a Gaussian distribution G. To avoid generating aberrant faces, while allowing the faces to be distinct from each other, we chose to truncate the Gaussian G for these

landmarks. G is centered on μ with a standard deviation σ and clipped at t as shown in Fig. 5.5. The standard deviation permits to statistically control the intensity of the deformation. We use another user-modifiable parameter s. This parameter acts as a multiplier on the size of the Gaussian and is the same coefficient for all the landmarks. The determination of the warped landmarks is formulated as below :

$$X_d^i = X_n^i + (G_{(\mu,\sigma,t)}^i \times s) \tag{5.1}$$

Where $s = 0.6$. The landmarks displacement is employed using a Gaussian distribution ($\mu = 0$, σ, clipped at $t = 2\sigma$). The choice of σ in this experiment was set empirically, based on the plausible face changes of the person and depending on the distance between the eyes d_{eyes}. We used the one-tenth of the eyes distance to configure σ.

$$\sigma = \frac{d_{eyes}}{10} \tag{5.2}$$

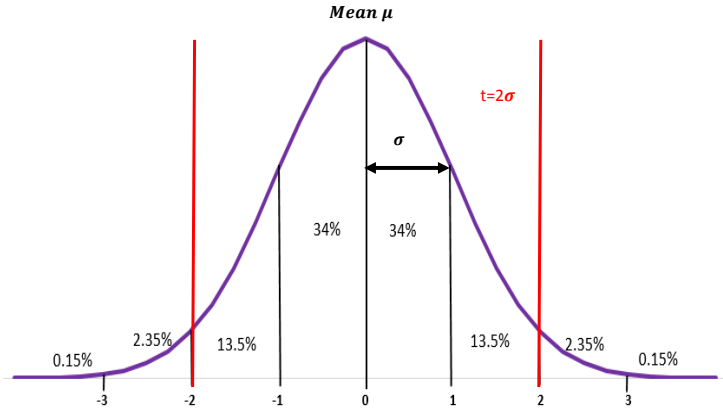


FIGURE 5.5 The Gaussian distribution used to generate the new random positions of the landmarks. G is centered on $\mu = 0$ with a standard deviation σ and clipped at $t = 2\sigma$ to avoid the generation of very weird expressions.

5.3.3 Pixel mapping

Once the random positions of the landmarks are determined, a pixel mapping is employed to deform the neutral face. To maintain the spatial coherence of the overall shape, we used the

Moving Least Squares method MLS [105, 5] to generate the deformation. The MLS is a very efficient 2D warping method for image deformation. As mentioned in chapter 3, we use MLS to deform a detected face and synthesize a more joyful expression. Given the time needed to perform the deformations of several regions, we make a time/esthetic's compromise as in [5]. We apply the algorithm on grids around each eye, the mouth, the nose, and the forehead regions, not on each pixel of the image. Fig. 5.6 illustrates an example.

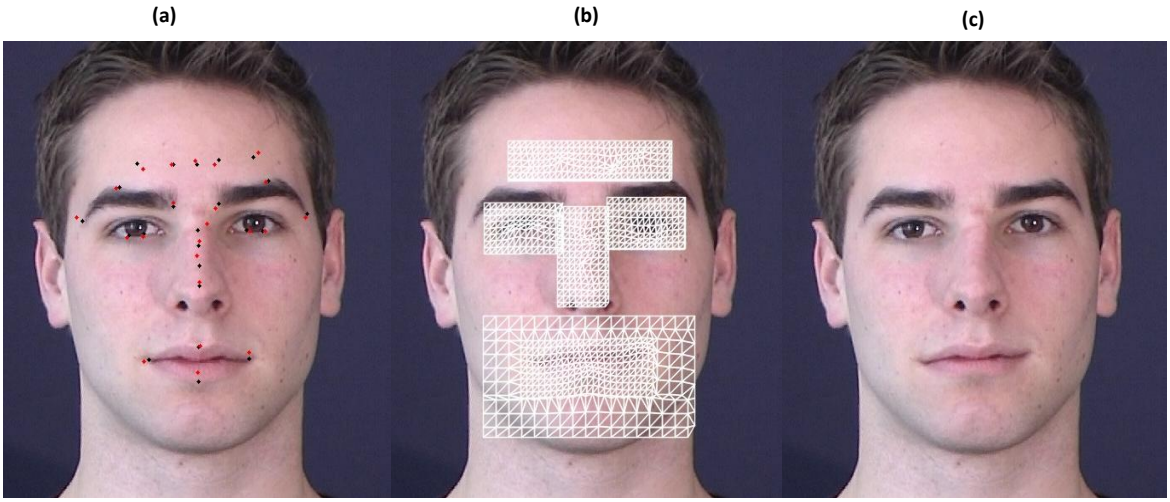


FIGURE 5.6 (a) : The neutral landmarks in black and the red landmarks are the random new positions. (b) : Grids around the interesting regions in which we apply the MLS algorithm. (c) : The deformed face result.

The described deformation method is used to prepare the deformed faces for a psychophysical experiment. In the following section, we give more details about this experiment.

5.4 Experimental protocol

The previous steps are performed to prepare the data for the perception test. We generate 1400 random expressions (700 pairs of expressions) for 2 actors (other subjects), and for each of the 24 participants involved in this experiment. Each of the 24 participants (12 male and 12 female, mean age = 22) was invited to rate the deformed images. We detail the tool used for determining a mental representation and the conducted experiments in the following subsections.

5.4.1 Mental representation tool

The tool used for determining a mental representation is presented in Fig.5.7. Each participant is invited to rate the deformed faces. After the registration of the participant, instructions were displayed to explain the unfolding of the perception test as shown in Fig. 5.8. 700 trials of randomly-manipulated faces were presented to the participant. As illustrated in Fig. 5.9, a trial is a pair of two randomly-modulated faces. For each pair, the participant is asked which of the two images presents the most "friendly" smile expression and he responds using the keyboard. During the experiment, the landmarks coordinates (23 landmarks) of image chosen as friendly F_+ or chosen as non friendly F_- are saved in each trial.

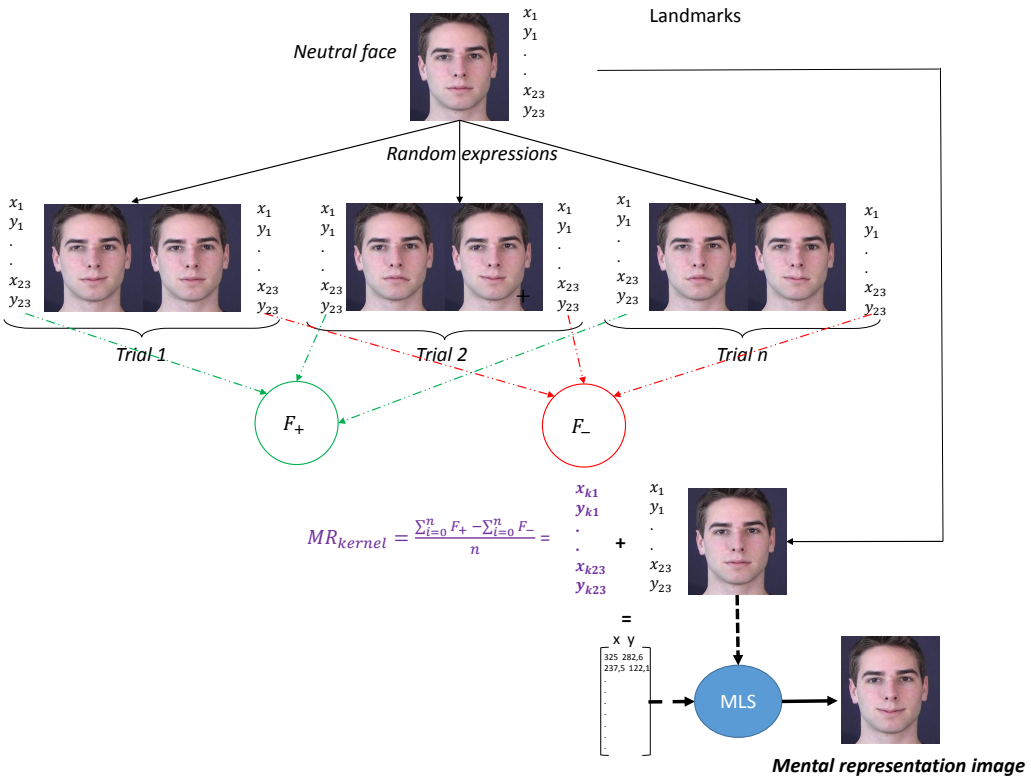


FIGURE 5.7 The conducted steps for determining the mental representation of a given face.

Once the perception task is performed, we determined the Mental representation kernel using the following equation :

$$MR_K = \frac{\sum_{i=0}^{n} F_+ - \sum_{i=0}^{n} F_-}{n} \tag{5.3}$$

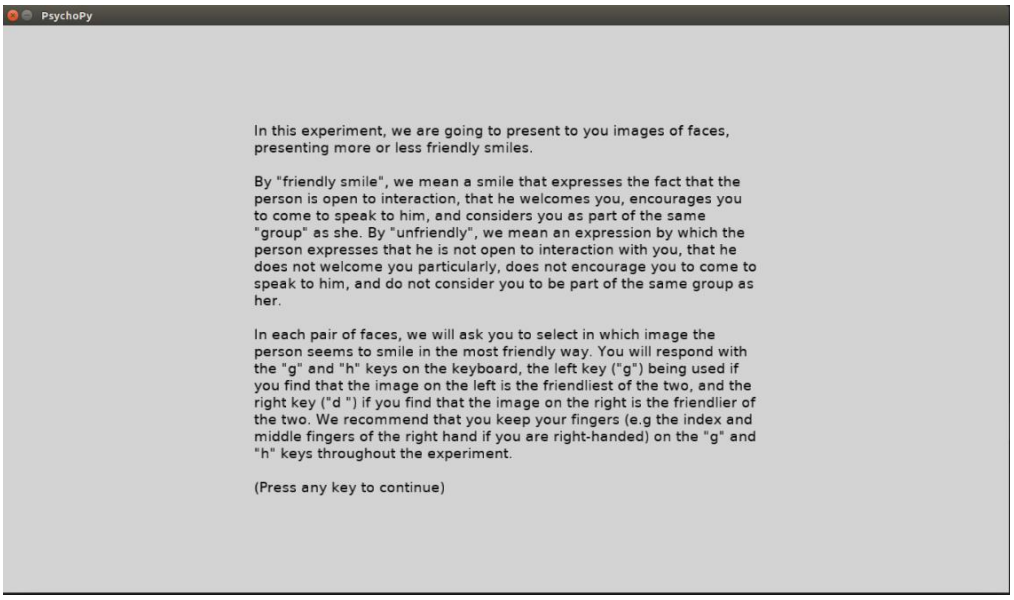


FIGURE 5.8 The instruction displayed for each participant to explain the perception test.

FIGURE 5.9 Trial example of Actor 2 displayed for the participant. He was asked to choose the image with the most friendly smile.

The kernel is defined as a 23-landmarks vector deformation. The kernel is computed as the mean of the deformed images classified as “friendly” F_+ minus the mean of the deformed images classified as “non-friendly” F_- [78]. Kernels were then normalized by dividing them by the absolute sum of their values.

As shown in Fig. 5.7, to generate the participant’s mental representation image on that face, we add the determined kernel MR_K to the neutral 23-landmarks of the rated face. The result of this addition gives the positions landmarks of the participant’s MR on the given face. To visualise the MR image (form), these landmarks are used to deform the neutral face using MLS warping method.

5.4.2 Experiments

Our experiments consist on determining the MRs of each participant on his own face (**Self**) and on the faces of other 2 actors (**Actors**). The experiments are carried out in two separate weeks to avoid that participants give the same responses for both self and actors stimuli (which would bias the comparison between the two sets of MRs).

First week : We record the real smiling videos of the participants. We asked them to produce a friendly (deliberate) smile expression. Then, each participant is invited to categorize 700 trials (1400 deformed faces) of the deformed faces of two actors and choose the most "friendly" smile in each trial. This part of the experiment aims at determining the smiling MRs on the faces of other persons (the **Actors** MRs). The recorded videos of the participants are used to determine the real smile kernel (**Production**) of each participant.

In the following week : The neutral frame of each participant’s video is used to generate 1400 random deformed faces. So, the participants came back to rate their own deformed faces. As before, participants were presented 700 pairs of stimuli (deformed images of their own face), and asked to choose the most "friendly" smile displayed on their faces. This part of the experiment aims at determining the participant’s MR on their own faces (the **Self** MR).

Once the perception tasks are performed, we determined the Mental representation kernels for the participants on their own faces (**Self**) and on the face of other actors (**Actors**) using the equation 5.3. We determine also the real way of smiling kernel of each participant which we named **Production**. The **Production** kernel is a 23-points vector deformation, which is computed for each participant, as the displacement between the neutral landmarks

positions and their positions in the frame that corresponds to the joyful expression (determined manually). Fig. 5.10 illustrates examples of the determined kernels.

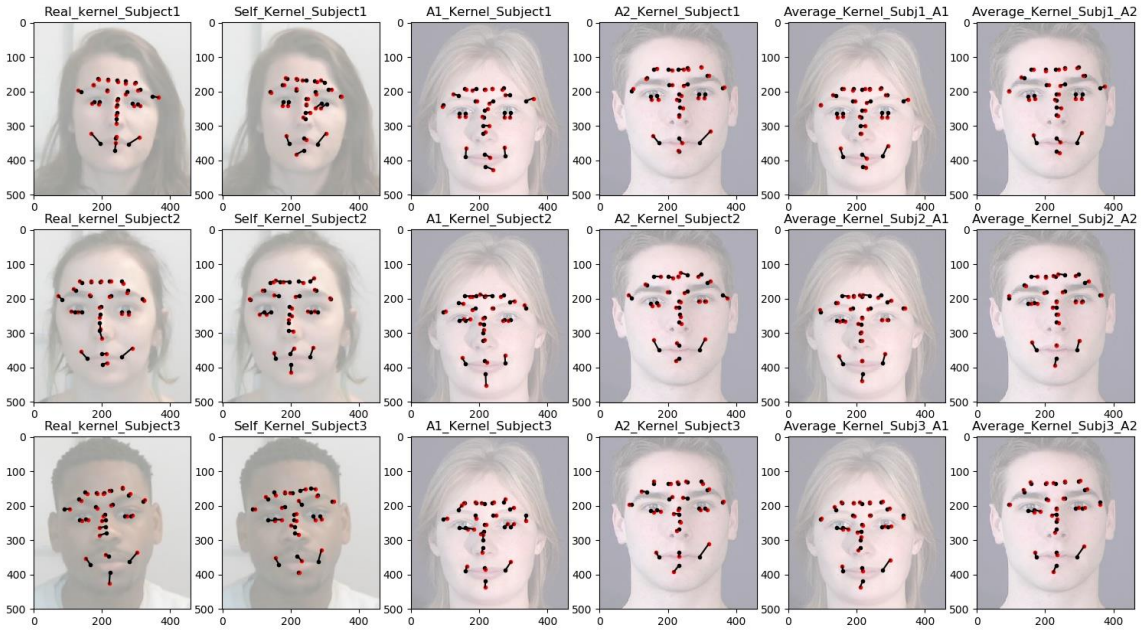


FIGURE 5.10 The determined kernels for 3 participants. In black are the neutral landmarks and in red the MR landmarks. First column : the**production** Kernels which represent the displacement of the landmarks when the participant produces her own real friendly expression. Second column : the self Kernel : the kernels resulted from the reverse correlation on their faces (participants). The third and fourth columns represent the kernels on the faces of 2 Actors (A1 and A2). The last two columns illustrate the average of Actors kernels applied on the faces of A1 and A2.

5.5 Kernels analysis

To better understand the kernels characteristics, we analyse in this section the statistic landmarks displacement. This study aims at determining which landmark is significantly displaced in each kernel, and which region is involved in a friendly smile and those not.

5.5.1 Actors kernels

We analyze in this section the landmark displacement on **Actors** kernels of the two actors (A1 and A2). To do so, we compute the dot product of each participant's using the equation

5.4. The dot product DP_Kp is the scalar product of the MR kernel MR_{Kp} of the participant p and the average kernel (of the 24 participants) on each actor MR_{Kav}. The dot product is based on the projection of participant kernel MR_{Kp} onto the average of the group kernels MR_{Kav} to determine the directional growth of that kernel to that average. This dot product is calculated for each landmark as follows :

$$DP^i_{Kp} = MR^i_{Kav} \cdot MR^i_{Kp} \tag{5.4}$$

Where DP^i_{Kp}=($X_{MR^i_{Kav}} \times X_{MR^i_{Kp}}$) + ($Y_{MR^i_{Kav}} \times Y_{MR^i_{Kp}}$). K is the Actor kernel index (A1 or A2), p is the participant index and i is the landmark index.

Fig. 5.11 and Fig. 5.12 illustrate the average landmarks displacement on A1 and A2 respectively. We noticed that the displacement on the mouth corners and the eyes is more important than in the nose and forehead regions for the 2 Actors. The dot product graphs justify the significant displacement of the eyes and mouth regions. We observe also the difference between the 2 mental representations on A1 and A2 mainly on the mouth regions. On the average of Actor 1 the corners mouth are highly raised unlike the Average of A2. However, the same regions are still involved in joyful expression mental representation (mouth and eyes).

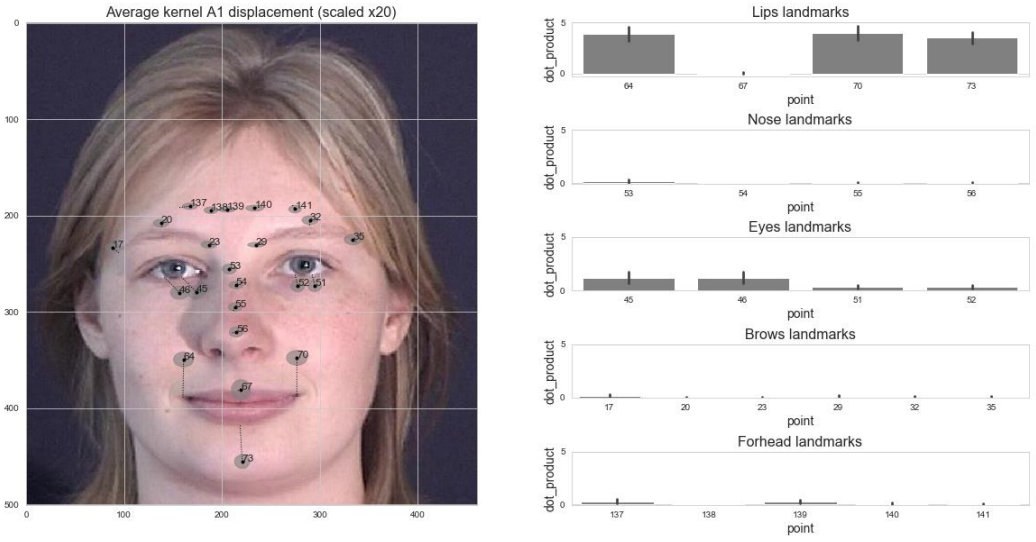


FIGURE 5.11 Left : the average displacement of the A1 kernels. Right : the mean dot products on the 23 deformed landmarks of A1 Kernels. The most displaced landmarks are the eye landmarks and the lips' landmarks.

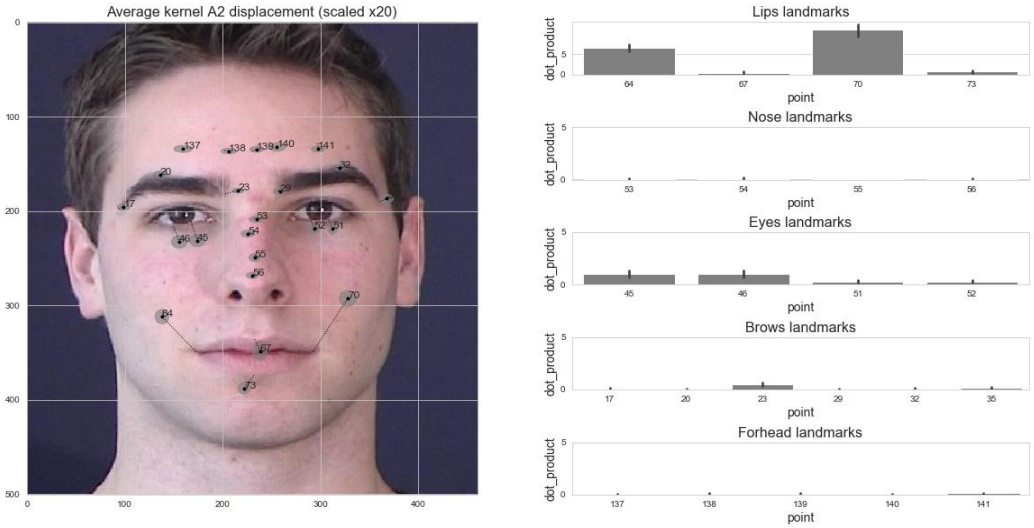


FIGURE 5.12 Left : the average displacement of the A2 kernels. Right : the mean dot products on the 23 deformed landmarks of A2 Kernels. The most displaced landmarks are the eye landmarks and the lips' landmarks.

To go further analyze, we investigate also the average MR of the 2 actors over the 24 participants. Fig. 5.13 illustrates the average of MR_{A1av} and MR_{A2av} displayed on the face of Actor 1. We notice that the mental representation of joyful expression includes the mouth and the eyes regions and not the nose and the forehead regions.

5.5.2 Self Kernels

The same previous analysis is done with the self kernels. As given in equation 5.5, we compute the dot product of each participant's **Self** kernel DP_{sp} using his **Self** kernel MR_{sp} and the average of the **Self** kernels over the 24 participants MR_{sav} . Fig. 5.14 illustrates the average of the **Self** kernels displayed on a participant face. The displayed kernel and the dot products results prove that the displacement on the mouth corners is more important than the other regions, where it is lacking. Comparing to the **Actors** MRs, we notice that the eyes displacement is quite small in **Self** MRs.

$$DP^i_{Sp} = MR^i_{Sav} \cdot MR^i_{Sp} \tag{5.5}$$

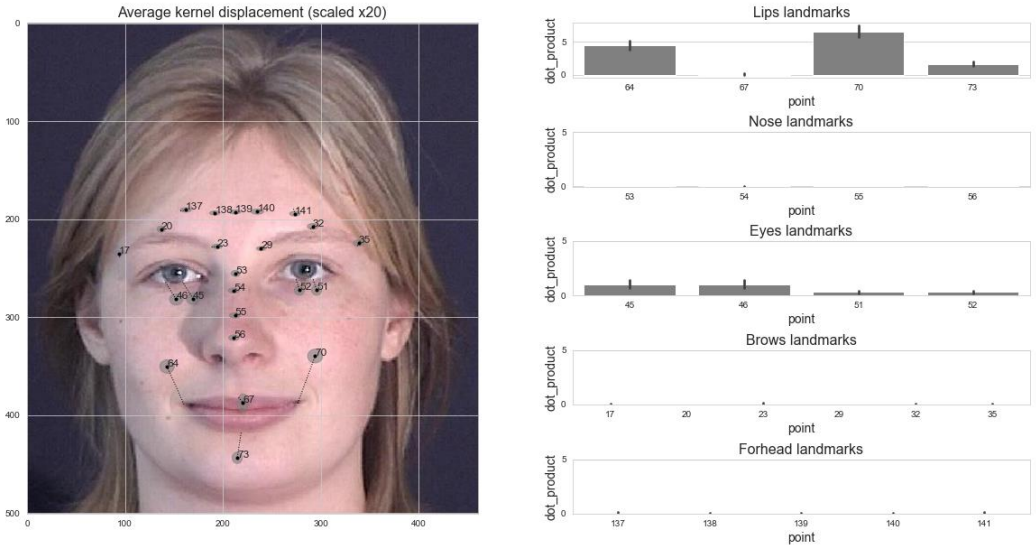


FIGURE 5.13 Left : The average displacement of the two actors displayed on A1. Right : the mean dot products on the 23 deformed landmarks of A1 and A2. The most displaced landmarks are the eye landmarks and the lips' landmarks.

Where Where DP^i_{Sp}=($X_{MR^i_{Sav}}$ $\times$X$_{MR^i_{Sp}}$) + ($Y_{MR^i_{Sav}}$ $\times$Y$_{MR^i_{Sp}}$). p is the participant index and i is the landmark index.

5.5.3 Production Kernels

As performed with **Actors** and **Self** kernels, To analyse the **Production** kernels we also compute the dot product using the following equation :

$$DP^i_{Pp} = MR^i_{Pav} \cdot MR^i_{Pp} \tag{5.6}$$

Where DP^i_{Pp}=($X_{MR^i_{Pav}}$ $\times$X$_{MR^i_{Pp}}$) + ($Y_{MR^i_{Pav}}$ $\times$Y$_{MR^i_{Pp}}$). p is the participant index, MR_{Pav} the average of the **Production** kernels and MR_{Pp} the participant's **Production** kernel.

Fig. 5.15 illustrates the average landmarks displacement of **Production** kernels. As before, we noticed that the displacement on the mouth corners and the eyes is more important than in the nose and forehead regions. We also observed a quite landmarks displacement in the brows regions.

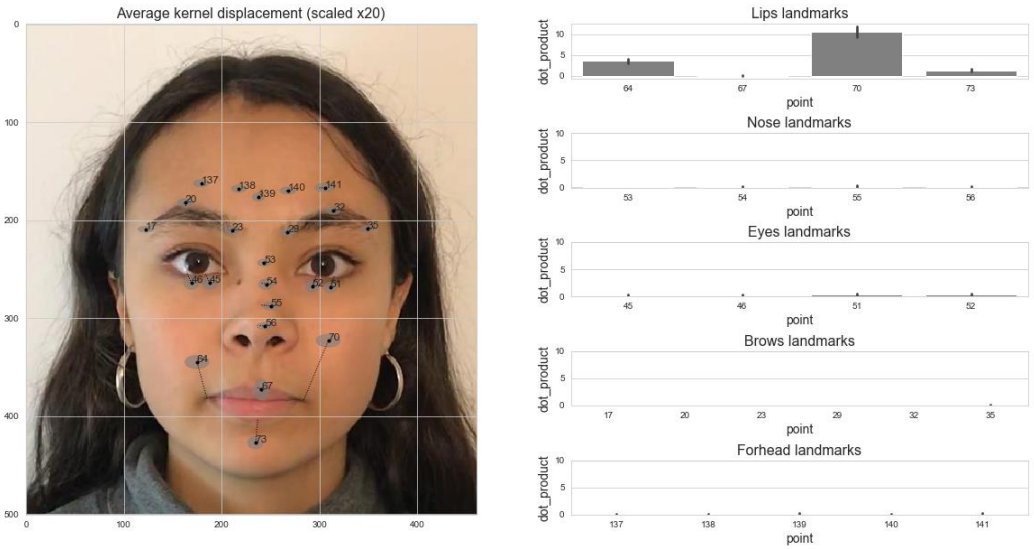


FIGURE 5.14 Left : the average displacement of the **Self** Kernels. Right : the dot product on the 23 deformed landmarks of Self Kernels. We observe that the most displaced landmarks in a **Self** mental representation of a joyful expression are the lips' landmarks. We observe also a quite displacement in the eyes regions.

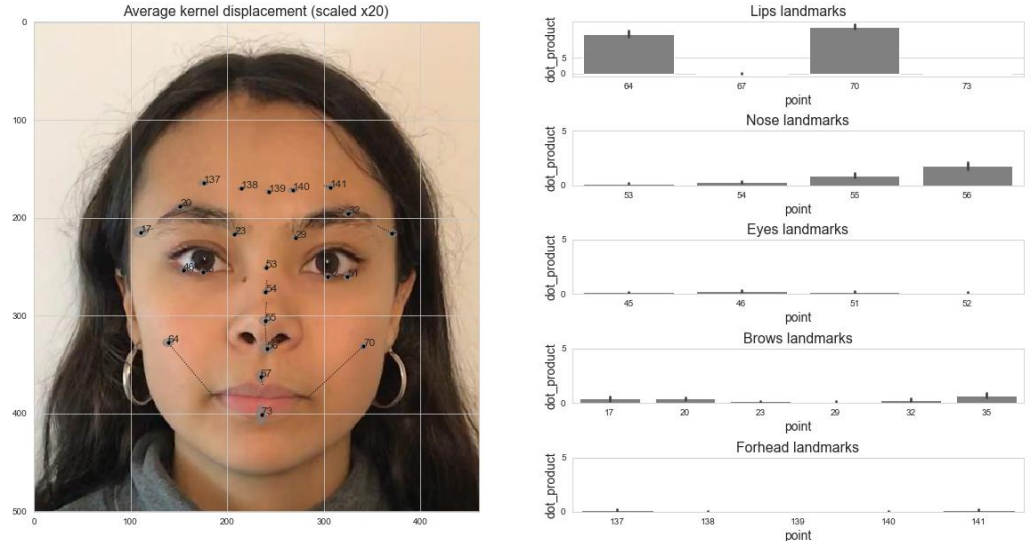


FIGURE 5.15 Left : the average displacement of the **Production** Kernels. Right : the dot product on the 23 deformed landmarks of **Production** Kernels. The most displaced landmarks in a real smile are the eye landmarks and the lips' landmarks. We notice also that there is quite displacement in the brows landmarks unlike in **Self** and **Actors** MR, where it is lacking

The determined kernels proved that the joyful expression involved the eyes, the mouth and maybe the brow regions. To analyse the differences and the similarities between these kernels as well as their relationship, more details are given in the next section.

5.6 Perception kernels interpretations

As mentioned in introduction of this chapter, my work intended to propose a tool to build MRs on real subjects. To this purpose we determined the 24 participant's **Self** and **Production** MR of joyful expression. We also determined the 2 **Actors** MRs for each participant. The analysis of the three posed hypothesis (see Fig. 5.1) have been made by IRCAM. The results are presented in Appendix A. First, in section A.1.1 they analyze the similarities and the differences of the determined kernels. Second, in section A.1.2 they investigate the link between the MRs of joyful expression on the participant's face (**Self** MR) and on the faces of other actors (they use the average of **Actors** MRs). Third, they search for the link between the observer's MRs and their smiling manner (**Production**) (Section A.1.3). Finally, 3 questionnaires (Section A.1.4) are used to explore the relation between participant's kernels and their personalities (traits of empathy/affiliation) (Section A.1.5 and Section A.1.5). In this section, we give an overview of the different found results.

5.6.1 Differences and similarities between kernels

Analyzing the average of the 3 sets of kernels determined in section A.1 1, as expected, we found that the eyes and the lips are mostly used to express a friendly smile. However, we also noticed differences between the determined kernels. The **Self** and **Actors** kernels of each participant seem to be more comparable, while **production** kernel differs. A strange mouth asymmetry of the **Self** and **Actors** kernels is proved. However, the **production** kernels are quite symmetric.

5.6.2 Is Self kernel more like Actors kernel ?

The link between **Self** and **Actors** kernels is tested in A.1.2. The results show that each participant implies the same reasoning which is specific to himself for the judgments of the smiling faces whatever it is for his face or for others. We found that a significant association

exists between the **Self** and **Actors** mental representations which could justify that the participants perceived others as they perceive themselves.

5.6.3 Is Production kernel more like Actors or Self kernel ?

The link also between the way of smiling of the person and their mental representation is investigated in section A.1.3. (**Production** vs **Actors** and **Production** vs **Self**). We found that **Self** and **Actors** representations are generally not correlated to how smiles are really produced. These results didn't support the idea that self-related production representations (Self facial movement) subtend how we perceive others. To go further in the analysis, the personality of each participant is analyzed to test if there is an influence of personality's traits on their perception manner.

5.6.4 Personality evaluation

3 questionnaires are used for personality traits analysis :

— The TAS20 (Toronto Alexithymia Scale) questionnaire shows that most of the participants are non-alexithymic and have scores between 32 and 50 (just 3 alexithymic participants). Alexithymic participant is a person who has an inability to recognize and verbalize their own emotions. The analysis proves that an important difference in lips stretching (Lip1) between participant's **Production** and **Self** kernels can be a sign that this subject is alexithymic.

— BESA questionnaire shows that all the participant are empathetic. The analysis shows that the more the subject is empathetic, the more important is the similarity between their **Self** and **actors** representations. That's why, our previous analysis (subsection 5.6.2) proves that the participants MRs (**Actors** and **Self**) are linked.

— IPIP-PIC questionnaire is used to judge how generally the person is affiliative/warm on the one hand, and dominance on the other. The results shows limited correlation between warmth and kernels. However, these results indicate that dominance is negatively correlated to lips stretching. This shows that dominance subjects didn't stretch a lot their lips in their real smile. That means that the real smile can reflect the subject's dominance.

The link between the different determined scores of the 3 questionnaires is tested also. We found no significant correlation between them. Except for the correlation between dominance and EOT (Externally-Oriented Thinking of the TAS20 questionnaire) which shows that if the

participant is dominant so probably he has a difficulty to focus his attention externally but this does not prove that he is alexithymic.

5.7 Conclusion

In this chapter, we proposed a tool to build MRs of joyful expression. This tool is used to determine the participant's MRs from deformations of real faces. The aim of this collaborative work was to study the perception attitude of a group of participants. We generated their MRs on their own faces and their MRs on the faces of others as well as their own manner of smiling. These MRs have been analyzed by IRCAM to investigate three hypotheses. The results related to the first hypothesis prove that there is a correlation between perception of self-face and others faces for each participant. Then, the second hypothesis proved not valid because there is no link between the participant's MRs and the manner they produce their real joyful expression. Based on our results, we can say that the perception manner is not related to how we really produce our own expressions. Analyzing the participants personalities, we validated the third hypothesis, namely the link between MRs and the participant's personality. The results prove that the alexithymia is related to the difference between the **Self** and **production** kernels. Then, empathy is related to the similarity between **Self** and **actors**.

Chapitre 6

Conclusion and perspectives

6.1 Conclusion

Smile expression is a universal symbol for happiness. We smile when we are happy or when we want to make a good impression. However, smiling is more than just a facial expression, it has psychological ramifications that can actually make us feel happier. For example, seeing someone who smiles for us, instantly we feel friendly. Moreover, smile is contagious so the joy that appears on the smiling faces transfers to ours, without any words or further communication needed. The smile lifts our mood, as well as the moods of those around us. That has led to a huge potential for clinical improvement in psychiatric disorders, especially for Post-traumatic stress disorder (PTSD) patients.

This thesis is a part of an ANR project named "REFLETS". Our mission was to propose a solution which can generate mirroring by producing a pseudo-smile through image processing. In this thesis, we studied the nature of joyful expression basing on two axes. The first one is the production of a joyful expression. We proved that each person has their own way to produce their joyful expression. Therefore, we proposed a system which can deform the user's face into a personal smiling face. The second axis is the perception of joy, which aims at investigating the way that has each person to perceive the joyful expression on their face and on the faces of the others by determining their mental representations of joy expression. The determined mental representations are analyzed by IRCAM.

To achieve our goal, we started by studying real smiles. Our results proved that the smile is personal and made in a different way for each person. Then we proposed a system which aims at generating joyful expressions that are indistinguishable from genuine expressions. Therefore, our system adapts the synthesized smile to the way of smiling of each person to keep the credibility. Our system allows, for the first time, to explore parameters to make a

smile self-contagious. Moreover, our system can manipulate the intensity of the synthesized expressions.

The existing synthesis methods either focus on shape or on texture, but rarely on both. Indeed, the geometric methods provide a relevant shape deformation but they lack local details in the generated expressions, which makes them far from the real person's smile. In contrast, the generative models succeed in adding texture details (wrinkles and teeth) but the generated smiles are not those of the person. To overcome this, we proposed a hybrid geometric-machine learning method that combines the benefits of the geometric and machine learning methods to generate a joyful expression personalized on shape and texture. Our method is able to modify all the aspects of the face, especially the shape (e.g smile slope, mouth curve) and texture (e.g wrinkles, teeth). We compared our results with two state-of-the-art methods and our results show that we generate the closest smile to the ground truth expression.

The second axis studied in this thesis is the perception of the joyful expression based on mental representations. Several researches were interested in studying mental representations. In these studies, random AUs and avatars are used by the experimental protocol to determine the mental representations of the 6 basis expressions. However, in our work we determine the mental representations of joyful expressions on real subject's faces. We proposed a system that aims at deforming real faces to generate random expressions. These expressions are used with a reverse correlation procedure to determine mental representations on the subject's face and on the faces of others. The determined mental representations are compared to the real way of smiling of each participant. Thus, we proved that the manner that has each person to produce their smile doesn't affect their perception. So it appears that subjects do not perceive as they produce. Besides, they perceive with the same lips configurations either for self-face or for the others faces.

Therefore, our contributions are highlighted as below :

— A learned specific-parametric model to personalize the shape of the synthesized smile. The person-specific model is learned using the detected landmarks of the two faces (neutral and smiling) based on the barycentric coordinates.

— Two different ways to manipulate the intensity of the synthesized expressions. First way is based on the smile amplitude to tuning the intensity using a coefficient ; next a dynamic way that consists on manipulating the intensity based on a temporal characteristics of the subject's smile.

— A hybrid geometric-machine learning method that aims at refining global details (e.g wrinkles, teeth). Global details like wrinkles are refined using a proposed GAN. The

originality is that we combine prior knowledge information about the way of smiling of each person. The GAN takes as input the synthesized (geometric part) expressions to be encoded. This code is concatenated with specific information about each subject to add the suitable texture.

— Two solutions for personalizing the teeth region in the generated expressions ; either using Poisson editing method or Auto-encoder.

— A tool developed to build mental representations (MRs) of joyful expression on real participants. This tool is used for deforming real faces and determine the participant's MRs. The aim of this collaborative work is to study the perception attitude of a group of participants. To this purpose we generate their MRs on their own faces and their MRs on the faces of others, as well as their manner of smiling to be analyzed.

6.2 Perspectives

Concerning the perspectives of our work, we discuss the followings points :

— **Ameliorate our system** : As previously cited, our synthesized expressions proved to be the closest to the ground truth. However, there are still different to the latter. Moreover, The GAN influences the resolution of the generated images. Going further to get the better results is usually possible. For example, the data augmentation in the GAN part is necessary to improve the system performances.

— **Real-time system** : At that time, only the geometric method is implemented to deform the detected face on more joyful expression in real-time. After enhancing our GAN performances, we can also update our hybrid method to turn in real-time.

— **Psychology results on PTSD patients** : At this point our first version is still in the testing phase and we do not yet have the results of the psychological tests on PTSD at Percy Hospital. These results will contribute on enhancing our system and adapt it to the real world applications.

— **Using our mental presentation tool in other contexts** : The originality of our tool is that it is able to determine the mental representation of any expressions (joy, fear, disgust, ect..). With this tool it would be possible also to determine mental representation of complex emotions like stress and define a facial configuration of these expressions.

Publications

— International Journal

— International Conference

ZAIED, Sarra, SOLADIE, Catherine, et RICHARD, Pierre-Yves. Person-Specific Joy Expression Synthesis with Geometric Method. In : 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019. p. 504-508.
ZAIED, Sarra, SOLADIE, Catherine, et RICHARD, Pierre-Yves. Personalized Expression Synthesis Using a Hybrid Geometric-Machine Learning Method. In : International Conference on Image Analysis and Processing. Springer, Cham, 2019. p. 24-34.

— National Conference

ZAIED, Sarra, SOLADIE, Catherine, et RICHARD, Pierre-Yves. Synthèse personalisée de l'expression de joie. In : Gretsi. 2019.

— Community activity

Oral presentation "Réunion du GDR ISIS : Journée Action, Visage, geste, action et comportement" : Synthèse d'expressions faciales personnalisées par méthode hybride géométrique-apprentissage machine. Novembre 2019

Bibliographie

[1] Afifi, M., Hussain, K. F., Ibrahim, H. M., and Omar, N. M. (2014). Video face replacement system using a modified poisson blending technique. In *2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 205–210. IEEE.

[2] Ahumada Jr, A. and Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, 49(6B) :1751–1756.

[3] Ambadar, Z., Cohn, J. F., and Reed, L. I. (2009). All smiles are not created equal : Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of nonverbal behavior*, 33(1) :17–34.

[4] Arai, K., Kurihara, T., and Anjyo, K.-i. (1996). Bilinear interpolation for facial expression and metamorphosis in real-time animation. *The Visual Computer*, 12(3) :105–116.

[5] Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., and Aucouturier, J.-J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*.

[6] Aucouturier, J.-J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., and Watanabe, K. (2016). Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 113(4) :948–953.

[7] Baron-Cohen, S. and Wheelwright, S. (2004). The empathy quotient : an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34(2) :163–175.

[8] Beier, T. and Neely, S. (1992). Feature-based image metamorphosis. *ACM SIGGRAPH computer graphics*, 26(2) :35–42.

[9] Blanchard, E. B. and Hickling, E. J. (2004). *After the crash : Psychological assessment and treatment of survivors of motor vehicle accidents*. American Psychological Association.

[10] Brinkman, L., Todorov, A., and Dotsch, R. (2017). Visualising mental representations : A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, 28(1) :333–361.

[11] Buck, R. (1980). Nonverbal behavior and the theory of emotion : the facial feedback hypothesis. *Journal of Personality and social Psychology*, 38(5) :811.

[12] Bujisic, M., Wu, L. L., Mattila, A., and Bilgihan, A. (2014). Not all smiles are created equal. *International Journal of Contemporary Hospitality Management*.

[13] Burred, J. J., Ponsot, E., Goupil, L., Liuni, M., and Aucouturier, J.-J. (2019). Cleese : An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PloS one*, 14(4).

[14] Cannon, W. B. (1927). The james-lange theory of emotions : A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4) :106–124.

[15] Cao, C., Weng, Y., Lin, S., and Zhou, K. (2013). 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4) :1–10.

[16] Choi, C. S., Aizawa, K., Harashima, H., and Takebe, T. (1994). Analysis and synthesis of facial image sequences in model-based image coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 4(3) :257–275.

[17] Cohn, J. F., Ambadar, Z., and Ekman, P. (2007). Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3) :203–221.

[18] Darwin, C. (1874). L'expression des emotions chez l'homme et les animaux.

[19] Davis, J. I., Senghas, A., Brandt, F., and Ochsner, K. N. (2010). The effects of botox injections on emotional experience. *Emotion*, 10(3) :433.

[20] Dibeklioglu, H. (2017). Visual transformation aided contrastive learning for video-based kinship verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2459–2468.

[21] Ding, H., Sricharan, K., and Chellappa, R. (2018a). Exprgan : Facial expression editing with controllable expression intensity. *AAAI*.

[22] Ding, H., Sricharan, K., and Chellappa, R. (2018b). Exprgan : Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[23] Ding, X., Hu, R., Han, Z., and Wang, Z. (2018c). A novel frontal facial synthesis algorithm based on individual residual face. In *International Conference on Multimedia Modeling*, pages 14–22. Springer.

[24] Dynamixyz (2017). Genfacetracker :person-independent real-time face tracker : http ://www.dynamixyz.com.

[25] Ekman, P. (1993). Facial expression and emotion. *American psychologist*, 48(4) :384.

[26] Ekman, P. (2002). Facial action coding system (facs). *A human face*.

[27] Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The duchenne smile : emotional expression and brain physiology : Ii. *Journal of personality and social psychology*, 58(2) :342.

[28] Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System : Investigatoris Guide*. Consulting Psychologists Press.

[29] Ekman, P. and Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of nonverbal behavior*, 6(4) :238–252.

[30] Ekman, P. and Keltner, D. (1997). Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication : Where nature meets culture*, pages 27–46.

[31] Ekman, P., Levenson, R., and Friesen, W. (1983). Automatic nervous system activity distinguishes between emotions. *Science*, 221(1) :210–1208.

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

[33] El Haddad, K., Cakmak, H., Dupont, S., and Dutoit, T. (2016). Laughter and smile processing for human-computer interactions. *Just talking-casual talk among humans and machines, Portoroz, Slovenia*, pages 23–28.

[34] Finzi, E. and Wasserman, E. (2006). Treatment of depression with botulinum toxin a : a case series. *Dermatologic Surgery*, 32(5) :645–650.

[35] for Mental Health (UK, N. C. C. et al. (2005). *Post-traumatic stress disorder : the management of PTSD in adults and children in primary and secondary care*. Gaskell.

[36] Frank, M. G. and Ekman, P. (1993). Not all smiles are created equal : The differences between enjoyment and nonenjoyment smiles. *Humor-International Journal of Humor Research*, 6(1) :9–26.

[37] Freitas-Magalhães, A. and Castro, É. (2009). The neuropsychophysiological construction of the human smile.

[38] Gallese, V., Gernsbacher, M. A., Heyes, C., Hickok, G., and Iacoboni, M. (2011). Mirror neuron forum. *Perspectives on psychological science*, 6(4) :369–407.

[39] Ghent, J. and McDonald, J. (2003). Generating a mapping function from one expression to another using a statistical model of facial shape. In *Proceedings of the Irish machine vision and image processing conference*. Citeseer.

[40] Ghent, J. and McDonald, J. (2005). Photo-realistic facial expression synthesis. *Image and Vision Computing*, 23(12) :1041–1050.

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

[42] Gross, J. J. and Jazaieri, H. (2014). Emotion, emotion regulation, and psychopathology : An affective science perspective. *Clinical Psychological Science*, 2(4) :387–401.

[43] Haiyang, L., Chunna, T., Bo, Y., and Wei, W. (2016). Fast dct based expression extraction and lifelike transferring for facial images. In *Signal Processing (ICSP), 2016 IEEE 13th International Conference on*, pages 1697–1700. IEEE.

[44] Hamm, J., Kohler, C. G., Gur, R. C., and Verma, R. (2011). Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2) :237–256.

[45] Hatfield, E., Cacioppo, J., and Rapson, R. (1992a). Emotion and social behavior : Review of personality and social psychology.

[46] Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1992b). Primitive emotional contagion. *Review of personality and social psychology*, 14 :151–177.

[47] Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science*, 2(3) :96–100.

[48] Huang, Y. and Khan, S. (2018). A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 437–445. ACM.

[49] Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60 :653–670.

[50] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., and Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19) :7241–7244.

[52] James, W. (1884). What is an emotion ? *Mind*, 9(34) :188–205.

[53] James, W. (2007). *The principles of psychology*, volume 1. Cosimo, Inc.

[54] James, W. and Burkhardt, F. H. (1983). The principles of psychology, the works of william james.

[55] Jolliffe, D. and Farrington, D. P. (2006). Development and validation of the basic empathy scale. *Journal of adolescence*, 29(4) :589–611.

[56] Juckel, G., Heinisch, C., Welpinghus, A., and Brüne, M. (2018). Understanding another person's emotions—an interdisciplinary research approach. *Frontiers in psychiatry*, 9 :414.

[57] Kalra, P., Mangili, A., Thalmann, N. M., and Thalmann, D. (1992). Simulation of facial muscle actions based on rational free form deformations. In *Computer Graphics Forum*, volume 11, pages 59–69. Wiley Online Library.

[58] Kanade, T., Cohn, J. F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE.

[59] Kataria, M. (2002). *Laugh for no reason*. Madhuri International.

[60] Kaur, A., Kaur, L., and Gupta, S. (2012). Image recognition using coefficient of correlation and structural similarity index in uncontrolled environment. *International Journal of Computer Applications*, 59(5).

[61] Keltner, D. and Bonanno, G. A. (1997). A study of laughter and dissociation : distinct correlates of laughter and smiling during bereavement. *Journal of personality and social psychology*, 73(4) :687.

[62] Knowlton, J. Q. (1966). On the definition of "picture". *AV Communication Review*, 14(2) :157–183.

[63] Kraft, T. L. and Pressman, S. D. (2012). Grin and bear it : The influence of manipulated facial expression on the stress response. *Psychological science*, 23(11) :1372–1378.

[64] Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2) :233–243.

[65] Kriegman, D. (2007). Homography estimation. *Lecture Computer Vision I, CSE A*, 252.

[66] Krinidis, S., Buciu, I., and Pitas, I. (2003). Facial expression analysis and synthesis : A survey. In *10th International Conference on Human-Computer Interaction (HCI 2003)*, pages 22–27. Citeseer.

[67] Lane, R. D. (2000). Neural correlates of conscious emotional experience. *Cognitive neuroscience of emotion*, pages 345–370.

[68] Lange, C. (2011). On periodical depressions and their pathogenesis. introduction and translation by johan schioldann. classic text no. 85. *History of Psychiatry*, 22(1) :108–130.

[69] Lee, C.-S. and Elgammal, A. (2006). Nonlinear shape and appearance models for facial expression analysis and synthesis. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 497–502. IEEE.

[70] Lerner, J. S., Dahl, R. E., Hariri, A. R., and Taylor, S. E. (2007). Facial expressions of emotion reveal neuroendocrine and cardiovascular stress responses. *Biological psychiatry*, 61(2) :253–260.

[71] Lien, J. J., Kanade, T., Cohn, J. F., and Li, C.-C. (1998). Automated facial expression recognition based on facs action units. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395. IEEE.

[72] Lu, Z., Hu, T., Song, L., Zhang, Z., and He, R. (2018). Conditional expression synthesis with face parsing transformation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1083–1091.

[73] Man, D. and Vision, A. (1982). A computational investigation into the human representation and processing of visual information.

[74] Markey, P. M. and Markey, C. N. (2009). A brief assessment of the interpersonal circumplex : The ipip-ipc. *Assessment*, 16(4) :352–361.

[75] Meltzoff, A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198(4312) :75–78.

[76] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv :1411.1784*.

[77] Morgan, A. (2014). Representations gone mental. *Synthese*, 191(2) :213–244.

[78] Murray, R. F. (2011). Classification images : A review. *Journal of vision*, 11(5) :2–2.

[79] Näätänen, P., Ryynänen, A., and Keltikangas-Järvinen, L. (1999). The influence of alexithymic characteristics on the self-perception and facial expression of a physiological stress state. *Psychotherapy and psychosomatics*, 68(5) :252–262.

[80] Nakazato, N., Yoshida, S., Sakurai, S., Narumi, T., Tanikawa, T., and Hirose, M. (2014). Smart face : enhancing creativity during video conferences using real-time facial deformation. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 75–83. ACM.

[81] Neal, D. T. and Chartrand, T. L. (2011). Embodied emotion perception : amplifying and dampening facial feedback modulates emotion perception accuracy. *Social Psychological and Personality Science*, 2(6) :673–678.

[82] Niedenthal, P. M., Mermillod, M., Maringer, M., and Hess, U. (2010). The simulation of smiles (sims) model : Embodied simulation and the meaning of facial expression. *Behavioral and brain sciences*, 33(6) :417–433.

[83] Noh, J.-y. and Neumann, U. (2001). Expression cloning. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 277–288.

[84] Nojavanasghari, B., Huang, Y., and Khan, S. (2018). Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv :1801.09092*.

[85] Ochs, M., Pelachaud, C., and Mckeown, G. (2017). A user perception–based approach to create smiling embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1) :1–33.

[86] Olszewski, K., Li, Z., Yang, C., Zhou, Y., Yu, R., Huang, Z., Xiang, S., Saito, S., Kohli, P., and Li, H. (2017). Realistic dynamic facial textures from a single image using gans. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5429–5438.

[87] Park, S., Shin, J., and Kim, D. (2008). Facial expression analysis with facial expression deformation. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.

[88] Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318.

[89] Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D. H. (1998). Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH*, pages 75–84.

[90] Ponsot, E., Arias, P., and Aucouturier, J.-J. (2018). Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America*, 143(1) :EL19–EL24.

[91] project REFLETS (2017). https ://www.ircam.fr/projects/pages/reflets/.

[92] Rachman, S. J. and Hodgson, R. J. (1980). *Obsessions and compulsions*. Prentice Hall.

[93] Reed, S., Sohn, K., Zhang, Y., and Lee, H. (2014). Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439.

[94] Resick, P. A. and Schnicke, M. (1993). *Cognitive processing therapy for rape victims : A treatment manual*, volume 4. Sage.

[95] Richoz, A.-R., Jack, R. E., Garrod, O. G., Schyns, P. G., and Caldara, R. (2015). Reconstructing dynamic mental models of facial expressions in prosopagnosia reveals distinct representations for identity and expression. *Cortex*, 65 :50–64.

[96] Rickard, N. S., Wong, W. W., and Velik, L. (2012). Relaxing music counters heightened consolidation of emotional memory. *Neurobiology of learning and memory*, 97(2) :220–228.

[97] Ringach, D. and Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, 28(2) :147–166.

[98] Rohr, K., Stiehl, H. S., Sprengel, R., Buzug, T. M., Weese, J., and Kuhn, M. (2001). Landmark-based elastic registration using approximating thin-plate splines. *IEEE Transactions on medical imaging*, 20(6) :526–534.

[99] Rosenbaum, D. A. (2009). *Human motor control*. Academic press.

[100] Rychlowska, M., Jack, R. E., Garrod, O. G., Schyns, P. G., Martin, J. D., and Niedenthal, P. M. (2017). Functional smiles : Tools for love, sympathy, and war. *Psychological science*, 28(9) :1259–1270.

[101] Saha, P., Bhattacharjee, D., De, B. K., and Nasipuri, M. (2018). Facial component-based blended facial expressions generation from static neutral face images. *Multimedia Tools and Applications*, 77(15) :20177–20206.

[102] Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *CoRR*, abs/1606.03498.

[103] Saragih, J. M., Lucey, S., and Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International journal of computer vision*, 91(2) :200–215.

[104] Schachter, S. and Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5) :379.

[105] Schaefer, S., McPhail, T., and Warren, J. (2006). Image deformation using moving least squares. In *ACM transactions on graphics (TOG)*, volume 25, pages 533–540. ACM.

[106] Schmidt, K. L., Bhattacharya, S., and Denlinger, R. (2009). Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. *Journal of nonverbal behavior*, 33(1) :35–45.

[107] Seaward, B. L. (1997). Principles and strategies for health and wellbeing.

[108] Sederberg, T. W. and Parry, S. R. (1986). Free-form deformation of solid geometric models. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 151–160.

[109] Shapiro, F. (2001). Eye movement desensitization and reprocessing : Basic principles, protocols and procedures, 2nd edn. new york. *NY : Guilford Press.[Google Scholar]*.

[110] Shapiro, F. (2017). *Eye movement desensitization and reprocessing (EMDR) therapy : Basic principles, protocols, and procedures*. Guilford Publications.

[111] Sless, D. (1981). *Learning and visual communication*. Halsted Press.

[112] Song, L., Lu, Z., He, R., Sun, Z., and Tan, T. (2018). Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635.

[113] Stoiber, N., Breton, G., and Séguier, R. (2010). Modeling short-term dynamics and variability for realistic interactive facial animation. *IEEE Computer Graphics and Applications*, 30(4) :51–61.

[114] Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile : a nonobtrusive test of the facial feedback hypothesis. *Journal of personality and social psychology*, 54(5) :768.

[115] Suzuki, K., Yokoyama, M., Yoshida, S., Mochizuki, T., Yamada, T., Narumi, T., Tanikawa, T., and Hirose, M. (2017). Faceshare : Mirroring with pseudo-smile enriches video chat communications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5313–5317. ACM.

[116] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., and Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6) :183–1.

[117] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face : Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395.

[118] Tian, C., Li, H., and Gao, X. (2018). Photo-realistic 2d expression transfer based on fft and modified poisson image editing. *Neurocomputing*, 309 :1–10.

[119] Valstar, M. and Pantic, M. (2010). Induced disgust, happiness and surprise : an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC) : Corpora for Research on Emotion and Affect*, page 65.

[120] VanSwearingen, J. M., Cohn, J. F., and Bajaj-Luthra, A. (1999). Specific impairment of smiling increases the severity of depressive symptoms in patients with facial neuromuscular disorders. *Aesthetic plastic surgery*, 23(6) :416–423.

[121] Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., Albohn, D., Allard, E., Benning, S., Blouin-Hudon, E.-M., et al. (2016). Registered replication report : strack, martin, & stepper (1988). *Perspectives on Psychological Science*, 11(6) :917–928.

[122] Wang, S., Yi, X., and Chen, Y. (2017a). Piecewise affine warp based frontal face synthesizing and application on face recognition. In *Control And Decision Conference (CCDC), 2017 29th Chinese*, pages 3098–3102. IEEE.

[123] Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., and Catanzaro, B. (2017b). High-resolution image synthesis and semantic manipulation with conditional gans. *CoRR*, abs/1711.11585.

[124] Wang, X., Li, W., Mu, G., Huang, D., and Wang, Y. (2018). Facial expression synthesis by u-net conditional generative adversarial networks. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 283–290. ACM.

[125] Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in my insula : the common neural basis of seeing and feeling disgust. *Neuron*, 40(3) :655–664.

[126] Wolberg, G. (1998). Image morphing : a survey. *The visual computer*, 14(8-9) :360–372.

[127] Wolpe, J. (1968). Psychotherapy by reciprocal inhibition. *Conditional reflex : a Pavlovian journal of research & therapy*, 3(4) :234–240.

[128] Wood, A., Rychlowska, M., Korb, S., and Niedenthal, P. (2016). Fashioning the face : sensorimotor simulation contributes to facial expression recognition. *Trends in cognitive sciences*, 20(3) :227–240.

[129] Woodbury-Fariña, M. A. and Schwabe, M. M. R. (2015). Laughter yoga : benefits of mixing laughter and yoga. *Journal of Yoga & Physical Therapy*, 5(4) :1.

[130] Wu, H., Zheng, S., Zhang, J., and Huang, K. (2019). Gp-gan : Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2487–2495.

[131] Wu, X., Xu, K., and Hall, P. (2017). A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology*, 22(6) :660–674.

[132] Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., and Yang, M.-H. (2017). Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 251–260.

[133] Xue, Z., Li, S. Z., Lu, J., and Teoh, E. K. (2000). Bayesian model for extracting facial features. In *Sixth International Conference on Control, Automation, Robotics & Vision, ICARCV 2000, Dec., Singapore*.

[134] Xue, Z., Li, S. Z., and Teoh, E. K. (2001). Facial feature extraction and image warping using pca based statistic model. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 2, pages 689–692. IEEE.

[135] Yan, X., Yang, J., Sohn, K., and Lee, H. (2016). Attribute2image : Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer.

[136] Yeh, R., Liu, Z., Goldman, D. B., and Agarwala, A. (2016). Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv :1611.09961*.

[137] Yoshida, S., Sakurai, S., Narumi, T., Tanikawa, T., and Hirose, M. (2013a). Incendiary reflection : evoking emotion through deformed facial feedback. In *ACM SIGGRAPH 2013 Emerging Technologies*, page 8. ACM.

[138] Yoshida, S., Tanikawa, T., Sakurai, S., Hirose, M., and Narumi, T. (2013b). Manipulation of an emotional experience by real-time deformed facial feedback. In *Proceedings of the 4th Augmented Human International Conference*, pages 35–42. ACM.

[139] Yu, H., Garrod, O. G., and Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3) :152–162.

[140] Zaied, S., Soladie, C., and Richard, P.-Y. (2019a). Person-specific joy expression synthesis with geometric method. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 504–508. IEEE.

[141] Zaied, S., Soladie, C., and Richard, P.-Y. (2019b). Personalized expression synthesis using a hybrid geometric-machine learning method. In *International Conference on Image Analysis and Processing*, pages 24–34. Springer.

[142] Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., and Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273 :643–649.

[143] Zhao, G., Huang, X., Taini, M., Li, S. Z., and PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9) :607–619.

[144] Zhou, Y. and Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. *arXiv preprint arXiv :1708.09126*.

[145] Zhu, Y. and Gortler, S. J. (2007). 3d deformation using moving least squares.

[146] Zimmermann, G., Quartier, V., Bernard, M., Salamin, V., and Maggiori, C. (2007). The 20-item toronto alexithymia scale : Structural validity, internal consistency and prevalence of alexithymia in a swiss adolescent sample. *L'encéphale*, 33(6) :941–946.

Annexe A

Perception kernels interpretations

In chapter 5, we proposed to investigate the way that has each person to perceive the expression of joy on her face and on the faces of the others. We determined for each participant their **Self** and **Actors** mental representation. The MRs are determined based on the reverse correlation procedure on the random faces generated with our visual deformation block. Moreover, using their real smiling video we determined their **Production** kernel. We give in this appendix the results of the analysis conducted by IRCAM. This analysis aims at investigating the link between the 3 Kernels of each participant and studying the three hypotheses introduced previously (chapter 5, section 5.1).

A.1 Results and interpretations

IRCAM conducts the analysis of the determined kernels. At first, (subsection A.1.1) the differences and similarities between the determined MRs are analyzed. Secondly (subsection A.1.2), the first hypothesis is investigated by determining the link between the participants MRs (**Self** and unrelated actors' faces **Actors**). Then (subsection A.1.3), to analyze the second hypothesis the link between the individual manner of smiling (**Production**) and their determined MRs (**Self** and **Actors**) is determined. Finally (subsection A.1.4), the third hypothesis is verified to investigate whether the participants' personalities (traits of empathy/affiliation) affect their perception manner.

A.1.1 Differences and similarities between kernels

The smile description proposed in FACS [28] states that the smile expression is performed with 2 action units (AU6 : check raiser with wrinkling of eyes and AU12 : lip corner puller) which is consistent with the results illustrated in Fig. A.1. Fig. A.1 shows the average of the 3 sets of participants kernels (**Self** and **Actors** and **production**) over the 24 participants. We found that the eyes and the lips are mostly involved to express a friendly smile.

We also noticed differences between the nature of these kernels. The averages of **Self** and **Actors** kernels seem more comparable, while the average of **production** kernels differs. As illustrated in Fig. A.1 which represents the averages of the 3 sets of Kernels, there is a strange asymmetry between the two landmarks 64 and 70, with larger values on the right in the **Self** and **Actors** kernels. However, the **production** kernels are quite symmetric. This results shows that there is a link between participant's MRs (**Self** and **Actors**) which could justify that the participants perceived others as they perceive themselves. However, the manner of producing a friendly expression is different of their perception manner.

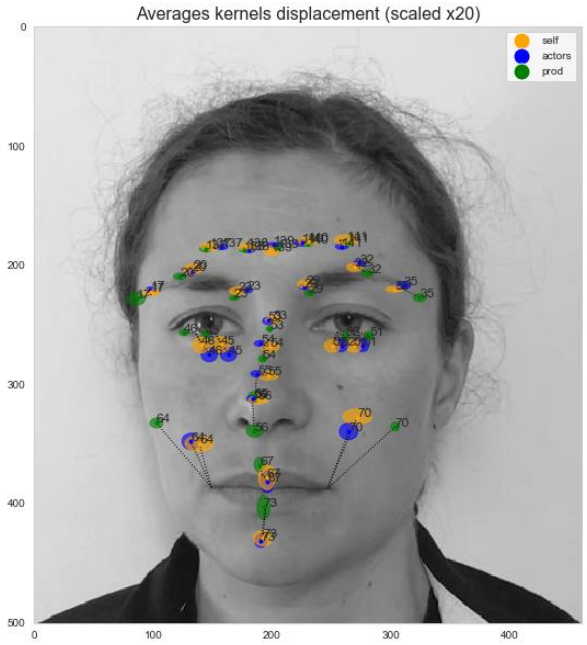


FIGURE A.1 The averages of the 3 determined sets of kernels. As expected, we found that the eyes and the lips are mostly used to express a friendly smile. The **Self** and **Actors** kernels seem to be more linked, while **production** differs.

IRCAM proposed to analyze these kernels by computing higher-level features and comparing them based on repeated-measures ANOVA. A PCA is performed on the determined kernels. The first 8 PCA components are chosen as illustrated in Fig. A.2 to conduct the analysis because they contain the most amount of information. These 8 components of eyes and lips (PCs) conserve 96% and 77% of the energy respectively.

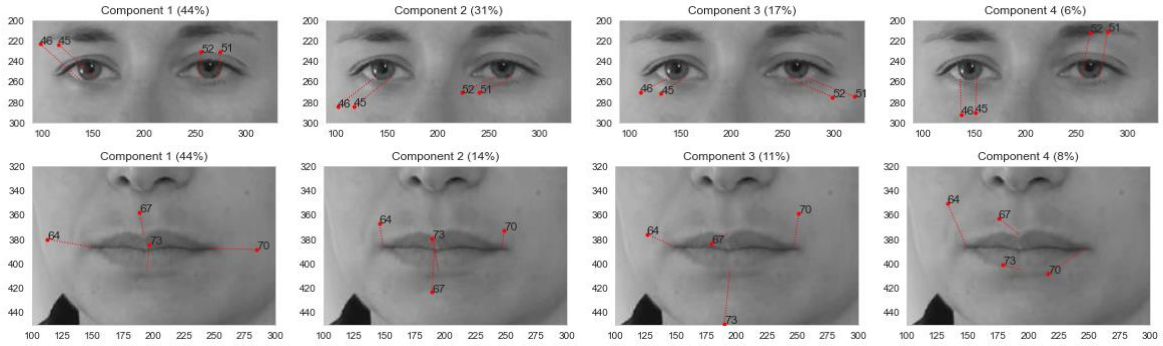


FIGURE A.2 The first 8 PCA components that contain the most information for the lips and eyes. These components (PCs) conserve 96% and 77% of PCA energy respectively.

The repeated-measures ANOVA shows that the kernels significantly differ between **Self**, **Actors** and **Production** for lips stretching (PC1) : F(46,2)=56.09, p=0.0000. We notice that p<α and F>$F_{t(46,2)}$ where α=0.05 is the the probability of the study rejecting the null hypothesis when it is true. The null hypothesis in our study states that there is no differences between the 3 kernels of each participant. Therefore, our results reject the null hypothesis and prove a significant difference between the participant's kernels for the lips.

For the eyes, the repeated-measures ANOVA also shows that there is a significant difference between the participant's kernels as presented in A.1. As shown in Fig. A.1, we notice that the PC1 ("eye closer") is more important in **Production** and PC2 "eye opener" is more important in **Self** and **Actors**. Then, the results show that the PC3 ("eye stretcher") is more important in **Production**. These features are further analyzed in order to investigate the link between the three kernels of each participant. More details are presented in the following subsections.

The results prove at first that the regions involved in a joyful expression are the eyes and lips either in the MRs or in the production of the expression. Secondly, the results show that there is a differences between each participant's kernels. Further analysis is conducted to

TABLE A.1 Statistic results of PCA results for eye landmarks

PCs	RM-ANOVA
PC1 "eye closer"	F(46,2)=56.62, p=0.0000
PC2 "eye opener"	F(46,2)=7.78, p=0.001232
PC3 "eye stretcher"	F(46,2)=12.86, p=0.000037

explore these differences based on 3 posed hypotheses. More details are given in the next subsections.

A.1.2 Is Self kernel more like Actors kernel ?

As a reminder, **the first hypothesis** aims at testing whether the representations are similar for the self-face and for unrelated actor faces. If cognitive judgments of smiling faces involve the same reasoning, then there should be a significant association between the self-face and other-face representations across individuals.

To analyze this hypothesis, the correlation between the features (8 PCs) of **Self** Kernels and **Actors** kernels is determined. As illustrated in Fig. A.3 PC1, PC2, PC3 of lips are significantly correlated (PC1 : R=0.588, p=0.0025. PC2 : R=0.555, p=0.0049. PC3 : R=0.518, p=0.0095). However, we noticed a low correlation between the eyes features. This suggests that the subjects judge the **Self** and **Actors** as friendly with the same lip configuration. We can say that each participant implies the same reasoning which is specific to him for the judgments of the smiling faces whatever it is for his face or for others. So, a significant association exists between the **Self** and **Actors** mental representations. The link between these 2 MRs and the **Production** Kernel for each subject is tested in the following subsection.

A.1.3 Is Production kernel more like Actors or Self kernel ?

The second hypothesis consists of inspecting if these representations (**Self** and **Actors** Kernels) are linked to the individual manner in which participants produce these expressions (**Production** kernel). The same metrics is used to evaluate the link of PCA features of **Production** kernel either with **Self** kernel or with **Actors** Kernel. Results are given in the following subsections.

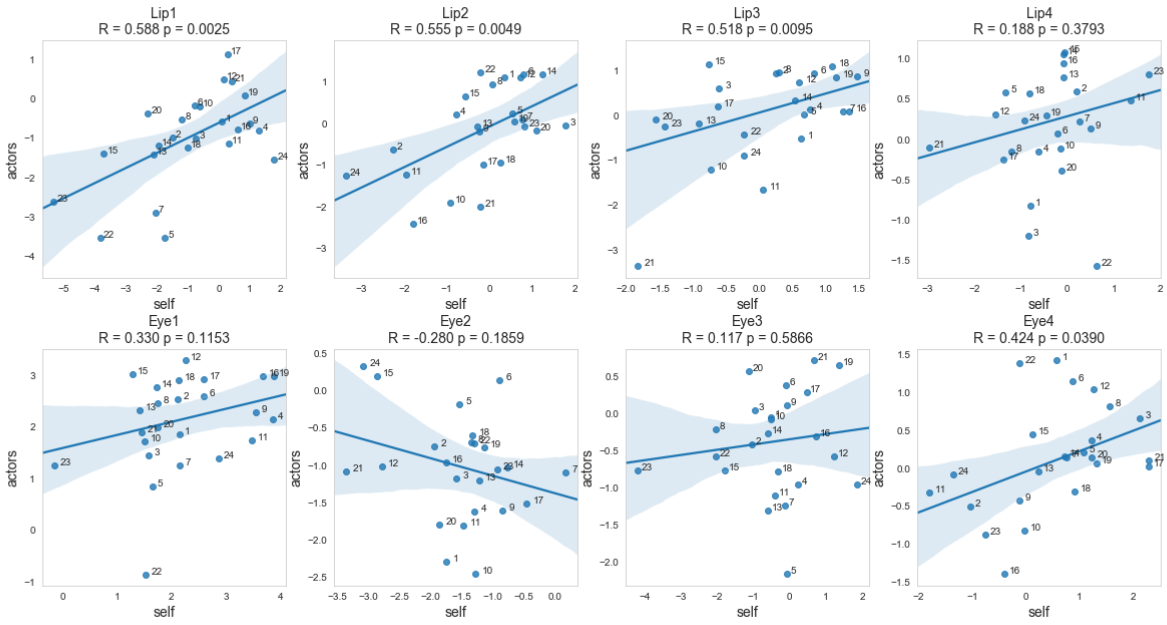


FIGURE A.3 Correlation coefficient between the PCA features of **Actors** kernel and **Self** kernel. The lip features PC1, PC2, PC3 are significantly correlated, unlike the eyes features. the results show that the subjects judge the **Self** and **Actors** as friendly with the same lip configuration. Therefore, each participant implies the same reasoning which is specific to him for the judgments of the smiling faces whatever it is for his face or for others.

A.1.3.1 Production Vs Self

The correlation between **Production** and **Self** kernels is determined to investigate the link between the participant's smile manner and the perception manner of their own faces. As illustrated in Fig. A.4, the correlation shows that self-representations are generally not correlated to how smiles are really produced. Moreover, all the produced smiles have a larger stretch than in perception, and we didn't find a relation between the degree of stretching between both kernels. These results prove that the perception of a joyful expression on the participants own faces is independent of how they produce their joyful expression.

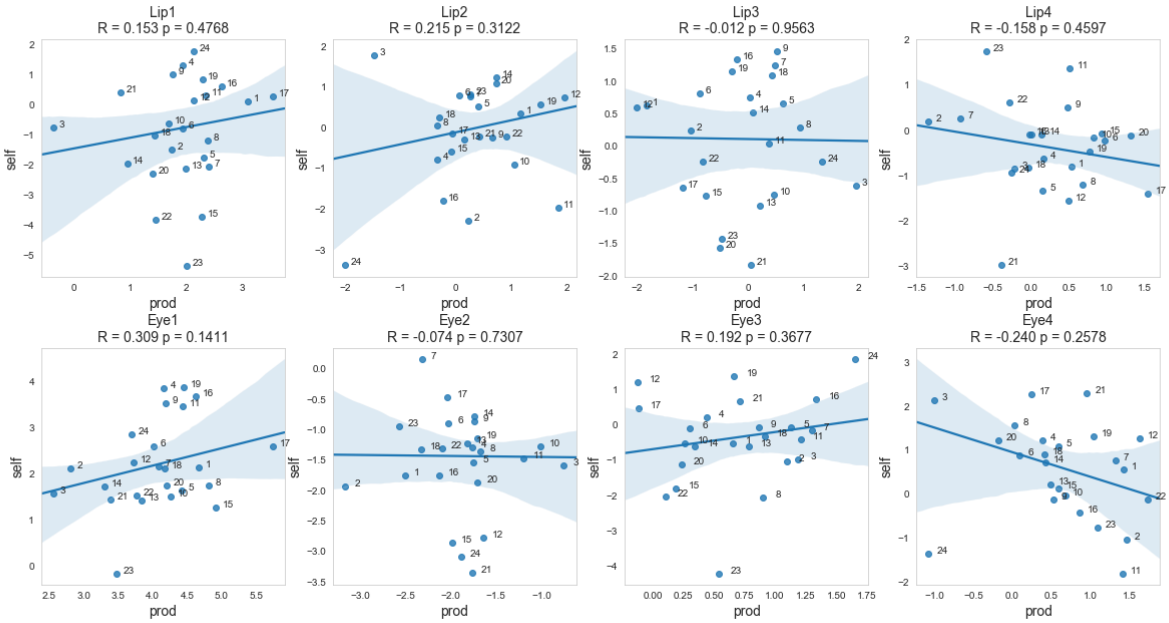


FIGURE A.4 Correlations between the PCA features of **Actors** kernel and **Self** kernel. The results show that the participants perceive the expression of joy in their faces independently of how they produce it.

A.1.3.2 Production Vs Actors

Using the same tools, IRCAM also analyzed the link between **Production** and **Actors** features for each subject. The correlation shows that **Production** and **Actors** representations are not correlated. All produced smiles have a larger stretch than in perception. Fig. A.5 shows the correlation between the eyes and lips features for the two types of kernels. These results prove that the perception of a joyful expression on others faces is independent of how

participants produce their joyful expression.

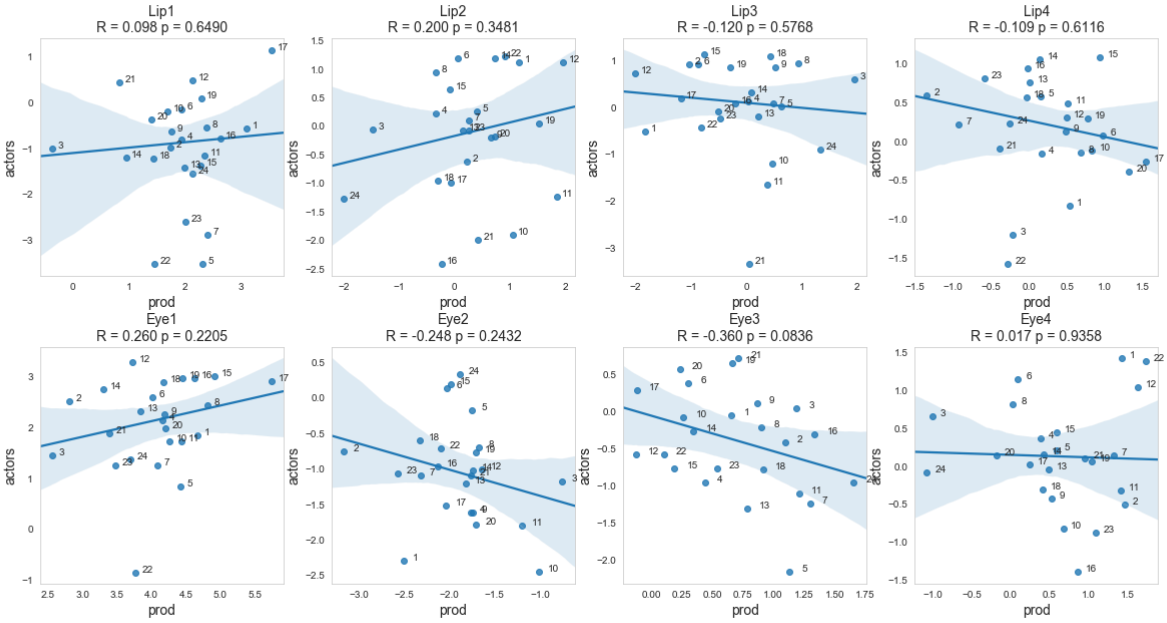


FIGURE A.5 Correlation between the PCA features of **Actors** kernel and **Production** kernels. The results show that there is no correlation between participant's **Production** and **Actors** representations which prove that the perception of a joyful expression on others faces is independent of how participants produce their joyful expression.

The analysis results demonstrate that the first hypothesis is valid. The participant's MRs (**Actors** and **Self**) are linked. However, the second hypothesis is not valid because the results indicate that the participants produce their joyful expression independently of how they perceive it either on their own faces or on the faces of others. The third hypothesis is analyzed in the next subsections.

A.1.4 Personality evaluation

The third hypothesis is tested by analyzing whether these representations are linked to participants' personalities (traits of empathy/affiliation), either for the self-face or for unrelated actors. We expect that the traits of affiliation/empathy may affect how people perceptually represent themselves (viewing themselves as more expressive), and may also affect how they represent others. To do so, each subject is invited to respond to 3 questionnaires. The links between subjects questionnaires scores (TAS20, BESA, IPIP-PIC) and their kernels

are analyzed in subsections A.1.4.1, A.1.4.2 and A.1.4.3 respectively. Then, in subsection A.1.4.4 the correlations between the questionnaires scores are investigated.

A.1.4.1 Toronto Alexithymia Scale TAS20

The alexithymia is mainly characterized by the inability to recognize and verbalize our own emotions that are thought to reflect a deficit in cognitive processing and regulation of emotional states. Zimmermann et al. [146] used a french version of the Toronto TAS-20 alexithymia scale (20 items) for exploring the evolution of alexithymia across age groups in adolescence.

In order to analyze this aspect in our group of participants, the TAS20 questionnaire is employed. The TAS-20 is a self-report scale that is comprised of 20 items divided on 3 sub-scales. The first sub-scale is the Difficulty Describing Feelings (DDF) used to measure difficulty describing emotions with 5 items which are numbered 2, 4, 11, 12, 17. The second sub-scale is the Difficulty Identifying Feeling (DIF) used to measure difficulty identifying emotions with 7 items which are numbered 1, 3, 6, 7, 9, 13, 14. The third sub-scale is the Externally-Oriented Thinking (EOT) used to measure the tendency of individuals to focus their attention externally measured with the reminding 8 items which are numbered 5, 8, 10, 15, 16, 18, 19, 20. Items are rated using a 5-point scale whereby 1 = strongly disagree and 5 = strongly agree. The TAS-20 uses cutoff scoring which are the following : equal to or less than 51 = non-alexithymia, Scores of 52 to 60 = possible alexithymia, equal to or greater than 61 = alexithymia.

Fig .A.6 illustrates the scores of the 24 subjects (Mean TAS20 = 48.71 ; SD=10.14). We noticed that most of the subjects are non-alexithymic and have scores between 32 and 50. However, there is N=3 subjects above 61 (see table A.2) and N=4 in borderline. These results are coherent with the alexithymia analysis on a representative sample of the population [146].

TABLE A.2 The 3 sub-scales scores of the 3 alexithymic subjects. The results show that the subject 15 and 22 have difficulty in identifying their feeling. The subject 3 has a high score for EOT that's mean that he has a difficulty to focus her attention externally.

Subject	DDF	DIF	EOT	TAS20
3	17	22	29	68
15	19	30	18	67
22	18	32	16	66

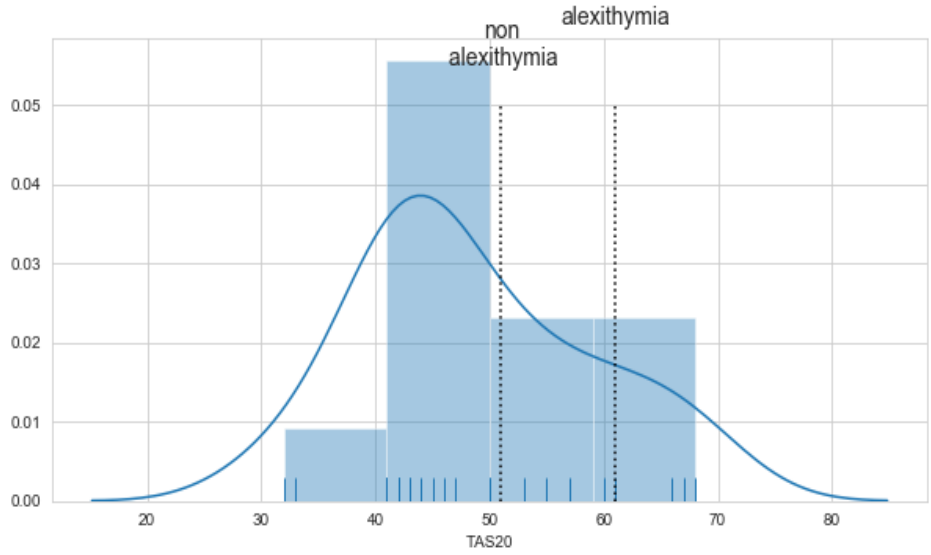


FIGURE A.6 Statistic results on our 24 participants (Mean TAS20 = 48.71 ; SD=10.14). The results hand a good representation of the population which is coherent with state of art results [146].

The hypothesis in our study is that the participant's **Self** MR can be influenced by alexithymia. If a participant doesn't recognize her emotions well, she may have an unstable representation of her own smile (e.g she smiles slightly, but she sees herself very smiling, or she has no idea about what she is saying, etc.). To investigate this hypothesis, correlation between participant's TAS20 score and their **Self** kernel is determined. However, there is no significant correlation between **Self** kernel features and participant's TAS20 score. To go further analysis, the correlation between participant's TAS20 score and the differences between their **Self** and **Production** kernels is determined. As illustrated in Fig.A.7 the correlation of the component LIP1 (the difference between 2 kernels) and TAS20 ($R = 0.453$, $p = 0.0260$) is important. This proves that an important difference in lips stretching (Lip1) between participant's **Production** and **Self** kernels can be a sign that this subject is alexithymic, which is the case of the subjects 3, 15 and 22.

A.1.4.2 Basic Empathy Scale in Adults (BESA)

BESA is an empathy questionnaire [7] that aims at studying the ability to feel the emotions of the others, or to physically feel the pain of the others. BESA is traditionally analyzed

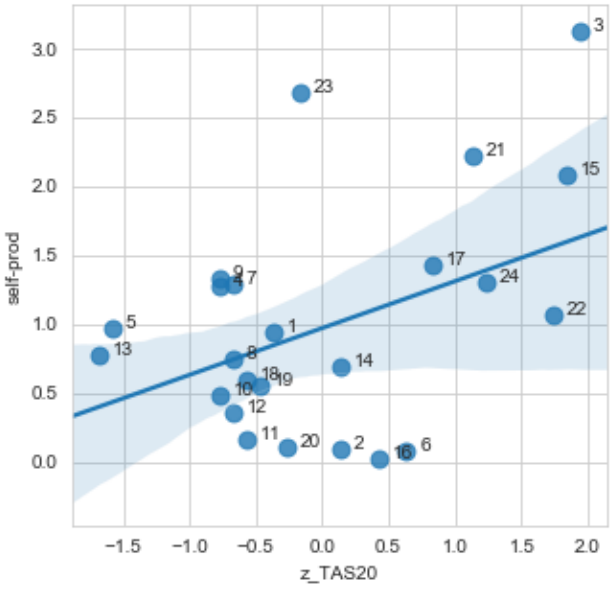


FIGURE A.7 Significant correlation of the component LIP1 and TAS20 (R = 0.453, p = 0.0260). This proves that an important difference between lips stretching between **Production** and **self** for a subject proves that this subject is alexithymic such for the subjects 3, 15 and 22.

with 2 factors composed also of 20 items. In the two-factors model [55], 9 items assess cognitive empathy (COG Items are numbered 3, 6, 9, 10, 12, 14, 16, 19, 20), and 11 items assess affective empathy (AFF Items are numbered 1, 2, 4, 5, 7, 8, 11, 13, 15,17, 18). In the two-factors model, the BESA included seven reversed items and the global score could range from 20 (deficit in empathy) to 100 (high level of empathy).

Fig. A.8 presents the scores calculated with the participant's responses. The mean of scores is 72.4 with an SD=8.9. We noticed that there are no apathy participants.

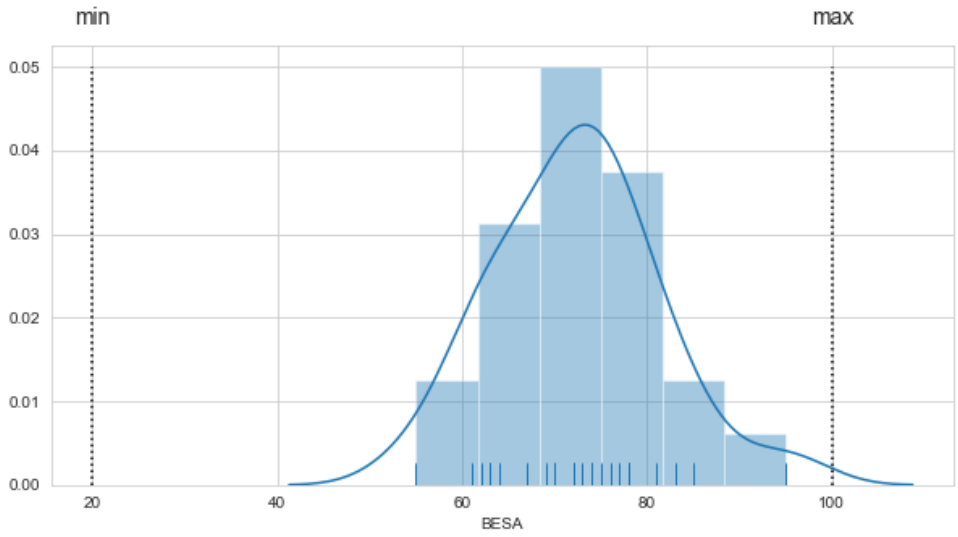


FIGURE A.8 The representation of the subjects' responses on BESA questionnaire. The mean of scores is 72.4 with an SD=8.9. We notice that there are no apathy subjects.

Our hypothesis in this study is that the others' perception (**Actors**) can be influenced by empathy. If a subject has no empathy, her **Self** MR is probably represented completely independently of those of others (**Actors**); if she is very empathetic, she doesn't make a difference between her smile and those of others. The analysis shows that the participant's BESA score is not correlated either with their **Actors** or with **Self** kernels. To this aim, the relationship between the participant's empathy score and the similarity between their **Actors** and **Self** kernels is investigated. The results show that there is a significant negative correlation on the component LIP4 (LIP4 vs BESA : R = -0.550 p = 0.0053) as presented in Fig. A.9. Therefore, the more the subject is empathetic, the more important is the similarity between their **Self** and **Actors** representations. Moreover, our previous analysis (subsection

A.1.2) proves that the participants MRs (**Actors** and **Self**) are linked. This justifies that all the participants are empathetic.

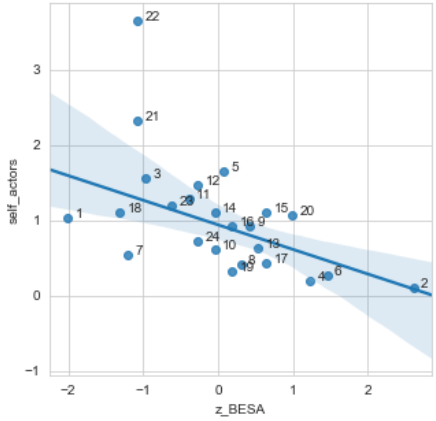


FIGURE A.9 Lack of correlation between most of the features. However for LIP3 there is a significant negative correlation (LIP4 vs BESA : R = -0.550 p = 0.0053) as presented in Fig. A.9. The correlation is Limited only for LIP4 (8% variance !), but that is coherent with the predicted direction : people with more empathy have more similarity between their self and other representations.

A.1.4.3 Interpersonal circumplexe (IPIP-PIC)

The third questionnaire (IPIP-PIC) is a social behavior questionnaire composed of 32 items, which judges how generally the person is affiliative/warm on the one hand, and dominance on the other [74]. It is a simplified scale assessing the interpersonal circumplex. Analyzing the responses of the 24 subjects, we obtained for Warmth : (M=0.00, SD=2.23) and for Dominance (M=0.00, SD=3.72) as illustrated in Fig. A.10.

Our hypothesis here is that the properties of the produced smile may reflect affiliation or dominance e,g. if the subject is very affiliative/warm, maybe she smiles more intensely and vice versa. To investigate this hypothesis, the correlation between BESA scores and **Production** kernel is determined. The results show limited correlation between warmth and PCA features. However, these results indicate that dominance is negatively correlated to LIP1 (R = -0.424 p = 0.0391) as shown in Fig. A.11. This shows that dominance subjects didn't

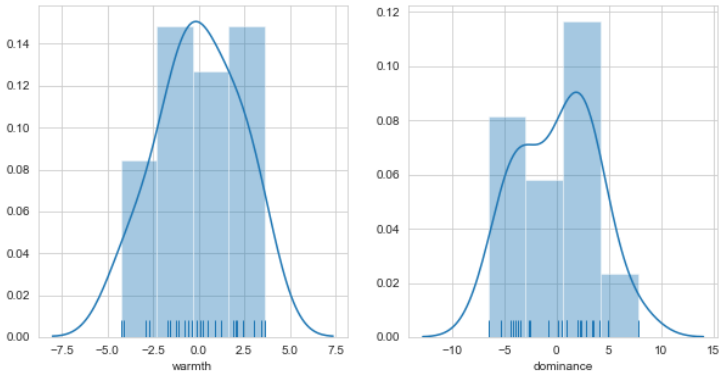


FIGURE A.10 The two representations results of warmth and dominance. The mean of the Warmth axis is 0 with an SD=2.23. The mean for Dominance is 0 with an SD=3.72.

stretch (LIP1) a lot their lips in their real smile. That means that the real smile can reflect the subject's dominance.

A.1.4.4 Questionnaires scores correlations

The correlation is used to analyze the participant's scores. Analysing the link between the different determined scores of the 3 questionnaires shows that there is a significant internal correlation between :

- DDF and DIF :R = 0.451, p = 0.0268
- DDF and TAS20 : R = 0.592, p = 0.0023
- DDF and DIF : R = 0.451, p = 0.0268
- DDF and TAS20 : R = 0.592, p = 0.0023
- DIF and TAS20 : R = 0.834, p = 0.0000
- EOT and TAS20 : R = 0.554, p = 0.005
- AFF and BESA : R = 0.915, p = 0.0000
- COG and BESA : R = 0.465, p = 0.0220

One external correlation is found between Dominance and EOT : R = 0.495, p = 0.0139. As shown in Fig. A.12 this correlation shows that if the subject is dominant so probably she has a difficulty to focus her attention externally but this does not prove that she is alexithymic because the correlation TAS20 and dominance is : R = 0.287 p = 0.1735.

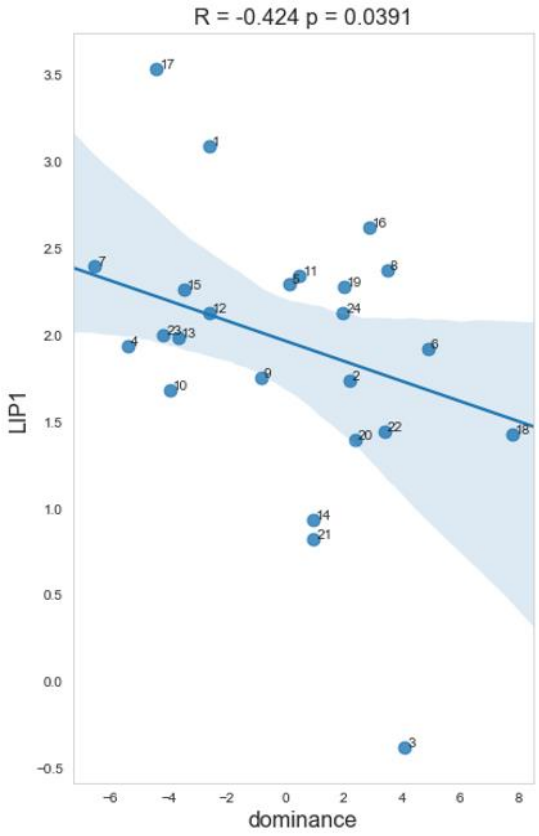


FIGURE A.11 A negative correlation between dominance and LIP1 (Lip stretching). The results shows that dominance subjects didn't stretch their lips in smiling (LIP1).

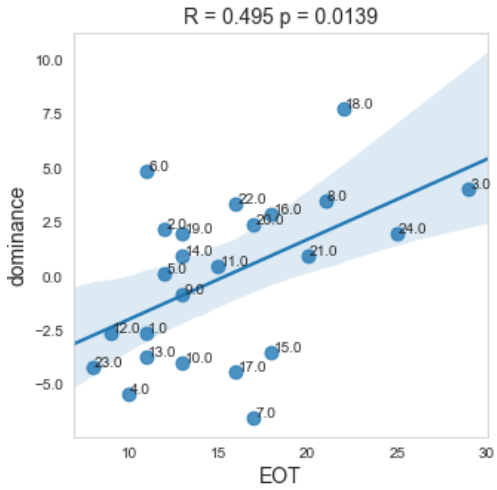


FIGURE A.12 The correlation between dominance vs EOT shows that if the subject is dominant so probably she has a difficulty to focus her attention externally but this does not prove that she is alexithymic (TAS20 vs dominance : R = 0.287 p = 0.1735).

A.2 Conclusion

In this Appendix, we presented the analysis and the results conducted by IRCAM on the determined kernels. The results show that there is a correlation between perception of self-face and others faces which are not linked to how the subject really produce their expressions. Based on these results, we conclude that the perception manner is not related to how we really produce our own expressions. Analyzing the participant's personalities shows that the alexithymia is related to the difference between the **Self** and **Production** kernels. We consider that the participant who doesn't recognize their expression well is alexithymic. Then, empathy is related to the similarity between **Self** and **actors**. Therefore, the participant is empathetic when their **Self** MR is probably represented dependently of those of others (**Actors**).

Annexe B

Quantitative results on the rest of the mouth landmarks

In this Appendix, we present the quantitative results found on the rest of the model landmarks. As presented in the previous chapters, we selected the 10 landmarks corresponding to the corners lips and eyelids to learn the deformation model between neutral X_n and smiling X_s faces. The mouth landmarks have the indexes 64, 65, 75 for the left corner and 69, 70, 71 for the right corner. The eyes indexes are (45, 46) and (51, 52) for the left eye and right eye respectively. In chapter 4, we gave the results only on the landmark 64 which corresponds to the left mouth corner. The results on the remaining model landmarks are presented in the following sections. The eyes landmarks displacement is quite small, so that we cannot estimate their trajectories. Therefore, we give in this appendix the results on each of the mouth landmarks.

B.1 Quantitative results : Efficient smile shape (d=1)

To evaluate our geometric method, we propose to calculate the angles between the determined trajectories (Chapter 4, subsection 4.3.1.1). The angle is determined by :

$$\theta = \tan^{-1} \left| \frac{s_{GT} - s}{1 + s_{GT} s} \right| \tag{B.1}$$

Where s_{GT} is the slope of the real smile expression (of the second smile video) and s is the slope of the generated expression. Table B.1 gives the results of the calculated angles with the 3 methods on each of the mouth landmark. As proved with the landmark 64 in chapter 4

(subsection 4.3.1.1.2, Table 4.2), the mean of the angles generated with our method is the smaller mean and the closest one to the real mean angles. That means that we generate smiles which seem to be closer to the ground truth than the other two methods.

TABLE B.1 Mean and standard deviation of angles calculated with the 3 methods on the 2 databases MMI* and UVA-NEMO for each of the mouth landmarks.

Landmark	Method	MMI*		UVA-NEMO	
		mean	SD	mean	SD
64	Real smile1 and smile2 (baseline)	10.91	12.53	15.01	13.86
	Arias et al. [5]	12.00	13.70	15.65	13.93
	Wang et al. [124]	15.15	13.34	13.40	15.26
	Our geometric method	**7.50**	**9.68**	**12.54**	**14.91**
65	Real smile1 and smile2 (baseline)	13.87	11.89	19.51	22.04
	Arias et al. [5]	20.19	12.06	20.19	17.82
	Wang et al. [124]	21.15	17.04	19.80	18.85
	Our geometric method	**14.25**	**13.09**	**19.31**	**22.63**
69	Real smile1 and smile2 (baseline)	18.07	14.68	19.09	18.80
	Arias et al. [5]	17.19	18.85	20.05	19.50
	Wang et al. [124]	23.36	18.44	21.46	18.89
	Our geometric method	**17.11**	**15.04**	**18.63**	**17.95**
70	Real smile1 and smile2 (baseline)	14.30	13.33	14.78	12.24
	Arias et al. [5]	16.48	13.78	17.24	114.11
	Wang et al. [124]	22.39	22.55	16.30	15.02
	Our geometric method	**15.76**	**18.58**	**15.22**	**13.23**
71	Real smile1 and smile2 (baseline)	17.48	16.27	19.92	19.59
	Arias et al. [5]	21.38	16.97	22.91	22.33
	Wang et al. [124]	21.23	22.17	21.71	18.78
	Our geometric method	**19.17**	**16.37**	**18.67**	**17.42**
75	Real smile1 and smile2 (baseline)	18.28	16.15	19.83	19.89
	Arias et al. [5]	20.20	20.36	20.36	19.94
	Wang et al. [124]	20.37	20.10	21.29	20.94
	Our geometric method	**19.87**	**16.63**	**18.47**	**18.53**

B.2 Quantitative results : Linear and dynamic ways of intensity manipulation

As explained before (Chapter 4, subsection 4.3.1.2), our method generates different expressions intensities with two different ways. The first way is based on the variation of the coefficient d ; to synthesize intermediate expressions (d<1) between the neutral and the learned Apex or an amplified smile expression (d>1). As the correlation is high between the smile videos of the same subject, we propose a second way to manipulate the intensity (dynamic way) which consists of learning the dynamic smile from one video and test it on another neutral frame. The test frame is extracted from a second video of that subject. We give the qualitative and quantitative results of the two ways on the mouth landmarks in the following subsections.

B.2.1 Quantitative results : Linear intensity manipulation

Having the angles of each landmark trajectory, we calculate the angles mean and the standard deviation of these angles on the 4 databases. The mean represents the error between the ground truth (GT) trajectory and a generated trajectory so it is expected to be as small as possible.

Table B.2 shows the statistical results of mouth landmarks. The results show that our geometric method has the smaller mean for the majority of the model landmarks (mean closer to 0, which justifies that the generated trajectories are closer to the ground truth). For MMI* the mean should be closer to the real angles mean between the 2 smiles (see table B.1) because we learn the model on a first smile video and test it on the neutral of a second video of the subject. In addition, we can consider that our method is more stable than the other two methods because of the low value of the standard deviation (SD). We validate also that the GAN influences the resolution of the images, that is why the results of the hybrid method are less good than those of our geometric method, but the generated expression remains closer to the GT than the ones generated with the other two methods [5] and [124].

B.2.2 Quantitative results : Dynamic intensity manipulation

Table B.3 presents the results that aim at evaluating the generated smiles with the different models. The results show that the refined videos of the dynamic model are the closest in dynamics to $V2_{GT}$ for the most of the landmarks. We notice that we keep almost the same correlation as between the real videos and we approach of the real $V2_{GT}$.

TABLE B.2 Mean and standard deviation of angles calculated with the 3 methods on the 3 databases for each of the mouth landmarks. For CK, Oulu-CASIA and MMI, we generate smiles close tho the GT which is justified by the smaller angles mean for the majority of the model landmarks. For MMI* the mean should be closer to the real angles mean between the 2 subject's smiles (baseline) because we learn the model on a first smile video and test it on the neutral of a second video of the subject.

Landmark	Method	CK		Oulu-CASIA		MMI		MMI*	
		mean	SD	mean	SD	mean	SD	mean	SD
64	Real smile1 and smile2 (baseline)	N/A	N/A	N/A	N/A	N/A	N/A	10.91	12.53
	Arias et al. [5]	12.10	9.73	17.81	15.94	12.00	13.70	12.00	13.70
	Wang et al. [124]	16.16	12.24	19.06	18.21	15.26	12.66	15.26	12.66
	Our geometric method	**6.65**	**7.56**	**5.58**	**5.37**	**4.38**	**5.27**	**7.50**	**9.68**
	Our hybrid method	7.65	8.25	6.85	7.26	5.83	7.01	9.20	10.78
65	Real smile1 and smile2 (baseline)	N/A	N/A	N/A	N/A	N/A	N/A	13.87	11.89
	Arias et al. [5]	18.74	3.65	27.73	9.89	20.19	12.06	20.19	12.06
	Wang et al. [124]	18.05	5.11	17.47	7.67	22.87	20.15	22.87	20.15
	Our geometric method	**2.82**	**3.44**	**16.16**	**8.82**	**2.47**	**10.41**	**14.25**	**13.09**
	Our hybrid method	5.29	4.89	18.27	7.79	4.88	11.90	15.06	13.57
69	Real smile1 and smile2 (baseline)	N/A	N/A	N/A	N/A	N/A	N/A	18.07	14.68
	Arias et al. [5]	5.92	8.69	30.26	9.60	17.19	18.85	17.19	18.85
	Wang et al. [124]	14.54	4.83	18.70	7.58	24.15	21.80	24.15	21.80
	Our geometric method	**6.72**	**4.60**	**16.55**	**7.90**	**12.24**	**10.29**	**17.11**	**15.04**
	Our hybrid method	7.33	5.69	17.89	10.27	13.45	13.68	19.01	18.20
70	Real smile1 and smile2 (baseline)	N/A	N/A	N/A	N/A	N/A	N/A	14.30	13.33
	Arias et al. [5]	13.39	5.75	29.61	10.15	16.48	13.78	16.48	13.78
	Wang et al. [124]	9.14	6.44	19.03	7.47	18.20	13.45	18.20	13.45
	Our geometric method	**4.71**	**7.70**	**13.50**	**9.63**	**12.16**	**11.72**	**15.76**	**18.58**
	Our hybrid method	6.96	9.58	19.68	10.20	13.33	12.50	16.21	19.20
71	Real smile1 and smile2 (baseline)	N/A	N/A	N/A	N/A	N/A	N/A	17.48	16.27
	Arias et al. [5]	12.64	4.47	25.78	9.50	21.38	16.97	21.38	16.97
	Wang et al. [124]	8.28	6.86	18.77	8.10	21.86	15.33	21.86	15.33
	Our geometric method	**10.39**	**6.53**	**15.58**	**6.29**	**21.68**	**17.62**	**19.17**	**16.37**
	Our hybrid method	12.23	7.48	16.70	8.99	23.89	18.78	19.89	18.45
75	Real smile1 and smile2 (baseline)	N/A	N/A	N/A	N/A	N/A	N/A	18.28	16.15
	Arias et al. [5]	12.49	2.41	23.69	7.99	22.20	20.36	22.20	20.36
	Wang et al. [124]	10.06	4.45	19.16	9.60	20.75	19.68	20.75	19.68
	Our geometric method	**3.70**	**1.90**	**17.46**	**6.91**	**14.38**	**8.57**	**19.87**	**16.63**
	Our hybrid method	7.64	2.60	19.05	7.25	16.85	10.67	19.95	21.50

TABLE B.3 Quantitative results on texture : Mean error and correlation results on MMI* and UVA-NEMO databases. We notice that $\widehat{V2_D}$ and $\widehat{V2}_{RD}$ conserve the real correlation between the real smile videos.

Landmark	Method	MMI*		UVA-Nemo	
		MSE	correlation	MSE	correlation
64	$(V2_{GT},\widehat{V2_M})$	14.67	0.82	83.75	0.96
	$(V2_{GT},\widehat{V2_L})$	12.94	0.82	38.38	0.95
	$(V2_{GT},\widehat{V2_D})$	12.04	0.85	30.62	0.98
	$(V2_{GT},\widehat{V2_{RL}})$	11.66	0.84	34.41	0.96
	$(V2_{GT},\widehat{V2_{RD}})$	**10.76**	**0.85**	**29.25**	**0.98**
65	$(V2_{GT},\widehat{V2}_M)$	9.59	0.82	83.61	0.94
	$(V2_{GT},\widehat{V2}_L)$	10.66	0.81	29.59	0.93
	$(V2_{GT},\widehat{V2}_D)$	8.73	0.80	27.13	0.96
	$(V2_{GT},\widehat{V2}_{RL})$	8.29	0.82	26.30	0.95
	$(V2_{GT},\widehat{V2}_{RD})$	**7.11**	**0.82**	**25.65**	**0.96**
69	$(V2_{GT},\widehat{V2}_M)$	10.00	**0.89**	64.28	0.95
	$(V2_{GT},\widehat{V2}_L)$	11.39	0.88	26.04	0.95
	$(V2_{GT},\widehat{V2}_D)$	10.28	0.78	25.75	0.95
	$(V2_{GT},\widehat{V2}_{RL})$	8.94	0.77	25.76	0.94
	$(V2_{GT},\widehat{V2}_{RD})$	**8.37**	0.78	**25.03**	**0.95**
70	$(V2_{GT},\widehat{V2}_M)$	13.89	0.76	93.63	0.95
	$(V2_{GT},\widehat{V2}_L)$	12.96	0.77	36.21	0.94
	$(V2_{GT},\widehat{V2}_D)$	12.29	0.74	29.72	0.95
	$(V2_{GT},\widehat{V2}_{RL})$	10.66	0.72	31.70	0.95
	$(V2_{GT},\widehat{V2}_{RD})$	**9.93**	**0.78**	**28.56**	**0.95**
71	$(V2_{GT},\widehat{V2}_M)$	8.97	**0.79**	48.05	**0.91**
	$(V2_{GT},\widehat{V2}_L)$	9.90	0.78	31.47	0.91
	$(V2_{GT},\widehat{V2}_D)$	11.21	0.61	29.53	0.89
	$(V2_{GT},\widehat{V2}_{RL})$	**8.18**	0.62	30.17	0.90
	$(V2_{GT},\widehat{V2}_{RD})$	9.49	0.63	**28.50**	0.90
75	$(V2_{GT},\widehat{V2}_M)$	8.65	**0.90**	50.59	0.88
	$(V2_{GT},\widehat{V2}_L)$	12.59	0.79	22.61	0.89
	$(V2_{GT},\widehat{V2}_D)$	9.72	0.80	21.17	0.91
	$(V2_{GT},\widehat{V2}_{RL})$	10.44	0.80	21.73	0.90
	$(V2_{GT},\widehat{V2}_{RD})$	**7.57**	0.81	**20.37**	**0.91**

DOCTORAT BRETAGNE LOIRE / MATHSTIC

CentraleSupélec

Titre : Synthèse personnalisée de l'expression de joie à l'aide d'une méthode hybride géométrique-apprentissage automatique et analyse de représentations mentales de la joie.

Mot clés : synthèse d'expressions, modèle spécifique à la personne, méthode hybride, représentation mentale.

Résumé : De nombreuses méthodes géométriques ou d'apprentissage machine ont vu le jour récemment pour synthétiser des expressions faciales. Les techniques géométriques sont très efficaces pour modifier la forme du visage, mais ont du mal à générer des détails de texture. D'un autre côté, les méthodes d'apprentissage machine telle que le GAN génèrent des expressions photoréalistes incluant les détails de texture ; mais ces méthodes ne permettent pas de générer des déformations personnalisées, c'est-à-dire conforme à la façon précise qu' a le sujet d'exprimer son émotion. Dans cette thèse, nous proposons une méthode hybride, alliant les avantages des deux techniques pour synthétiser des expressions photoréalistes de joie personnalisées. Plus précisément, notre approche combine une méthode de warping 2D et un réseau adversarial génératif (GAN), pour générer à la fois une forme et une texture d'expression faciale pertinentes. Afin de conserver la façon d'exprimer l'émotion propre au sujet, nous apprenons préalablement cette information et créons un modèle paramétrique pour chaque sujet. Ce modèle est alors utilisé par notre système pour personnaliser la déformation. Les résultats quantitatifs et qualitatifs montrent que cette méthode hybride permet de générer des expressions de joie personnalisées, plus proches des expressions réelles du sujet que les méthodes de l'état de l'art. Cet outil de synthèse de sourire est aussi utilisé dans le cadre d'une collaboration en neurosciences. Nous proposons un nouveau système qui vise à déterminer la représentation mentale du sujet sur un visage donné. Ce système nous permet d'étudier la manière dont chaque personne perçoit l'expression de joie sur son propre visage et sur les visages des autres. Nos résultats montrent que la manière de percevoir n'est pas liée à la façon dont nous produisons réellement nos propres expressions. Nous constatons aussi qu'il existe une variabilité des représentations mentales entres les individus, qui est indépendante de l'acteur et qui est liée à certain traits psychiatrique du sujet.

Title: Personalized Expression Synthesis Using a Hybrid Geometric Machine Learning Method and Mental Representation Analysis of Joyful Expression.

Keywords: Expression synthesis, Person-Specific model, Hybrid method, Mental representation

Abstract: Actually, various Geometric and Machine Learning methods are employed to synthesize expressions. The geometric techniques offer high-performance shape deformation but lead to images which are lacking in texture details such as wrinkles and teeth. On the other hand, the machine learning methods (GAN) generate photo-realistic expressions and add texture details to the images but the synthesized expressions are not those of the person. we propose in this thesis a hybrid geometric-machine learning approach to synthesize photo-realistic and personalized joy expressions while keeping the identity of the emotion. Our approach combines a geometric technique based on 2D warping method and a generative adversarial network. It aims at benefiting from the advantages of both paradigms and overcoming their own limitations. Moreover, by adding a previous knowledge of the way of smiling of the subject, we personalize the synthesized expressions. Qualitative and quantitative results demonstrate that our person-specific hybrid method can generate personalized joy expressions closer to the ground truth than two generic state-of-the-art approaches. Our smile synthesis system is also used in a collaboration work in neuroscience. We propose a new system which aims at determining the mental representation of the subject on a given face. This system allows us to study how each person perceives the expression of joy on their face and on the faces of others. Our results show that the way we perceive is not related to the way we actually produce our own expressions. We also note that there is a variability of mental representations between individuals, which is independent of the actor but is linked to psychiatric traits of the subject.