Modeling emotionnal facial expressions and their dynamics for realistic interactive facial animation on virtual characters

Nicolas Stoiber

N° d'ordre : **4273** **ANNÉE 2010**

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du Signal et Télécommunications

Ecole doctorale Matisse

présentée par

Nicolas Stoiber

préparée à l'unité de recherche SCEE-SUPELEC/IETR UMR 6164
Institut d'Electronique et de
Télécommunications de Rennes
Composante universitaire: S.P.M.

Modeling Emotional Facial Expressions and their Dynamics for Realistic Interactive Facial Animation on Virtual Characters

Thèse soutenue à Rennes
le 10 décembre 2010

devant le jury composé de :

Gérard BAILLY
Rapporteur
Directeur CNRS, GIPSA-Lab, Grenoble

Patrice DALLE
Rapporteur
Professeur, Université Paul Sabatier, Toulouse

Pascal HAIGRON
Examinateur
Maitre de Conférences HDR, Université Rennes 1

Jacques PALICOT
Directeur
Professeur, Supélec/IETR, Rennes

Gaspard BRETON
Co-Directeur
Ingénieur-Docteur, Orange Labs, Rennes

Renaud SEGUIER
Co-Directeur
Professeur adjoint, Supélec/IETR, Rennes

He was only a machine, and if it were only more visible that he were it wouldn't be so frightening. Or if his expression would change. It just stayed there, nailed on. You couldn't tell what went on behind those dark eyes and that smooth, olive skin-stuff.

Isaac Asimov, *Satisfaction Guaranteed*, Robots Series, 1951.

Ce n'était qu'une machine, et si seulement cet état mécanique avait été un peu plus visible, elle aurait ressenti moins de frayeur de sa présence. Ou s'il avait changé d'expression. Mais celle-ci demeurait invariablement la même. Comment deviner ce qui se passait derrière ses yeux sombres et cette douce peau olivâtre ?

Isaac Asimov, *Satisfaction garantie*, cycle des robots, 1951.

Remerciements

Je souhaite en tout premier lieu remercier Gérard Bailly, Directeur CNRS au GIPSA-Lab de Grenoble, ainsi que Patrice Dalle, Professeur à l'université Paul Sabatier de Toulouse, de me faire honneur d'avoir accepté la charge de rapporteur de cette thèse.

Je remercie également Pascal Haigron, Maître de Conférence HDR à l'université de Rennes 1, d'avoir accepté de juger mon travail en tant que membre du jury.

Ce travail de thèse est né de l'inspiration de ses deux instigateurs et encadrants : Gaspard Breton, Ingénieur R&D à Orange Labs, et Renaud Séguier, Maître de Conférence à Supélec. Je souhaite vivement les remercier de l'opportunité qu'ils m'ont offert de travailler avec eux sur un sujet passionnant, et de m'avoir prodigué leurs conseils et leur soutien sans faille tout au long de cette aventure. Je remercie également Jacques Palicot, Professeur à Supélec, pour avoir dirigé ma thèse et pour m'avoir fait profiter de ses conseils et de son expérience. Merci également à Danielle Pelé, Responsable d'unité à Orange Labs, pour son soutien et les conditions de travail idéales qu'elle a su préserver dans son équipe de R&D.

Je tiens à saluer et remercier chaleureusement les membres des équipes que j'ai côtoyé au quotidien durant ces trois années. Merci à vous Abdul, Olivier, Sébastien, Noémie, Jérôme, Rozenn, Ricardo, Eric, Sosthène, Loïc, Pascal, Philippe, Isabelle, Christophe, Guillaume, Patrick, et tous les autres ... Je garderai un excellent souvenir de l'ambiance chaleureuse et stimulante que nous avons entretenu en tant que collègues de travail.

Merci également à Cédric, mon collègue de bureau virtuel : qui eut cru qu'on puisse partager autant en étant si éloignés ?

Je souhaite remercier ma famille qui a toujours été à mes côtés et m'a donné sans compter son soutien et son amour. Merci à ma mère et mon père pour m'avoir offert un si bel exemple de vie. Merci à mes deux soeurs belles et re-belles pour leur désarmante malice. Merci à ma famille re- mais bien composée, ainsi qu'a ma belle-famille pour la place qu'ils m'ont fait dans leur coeur.

Enfin, je tiens à souligner que rien de tout cela n'aurait été possible sans ma compagne, qui m'a toujours soutenu et m'a épaulé dans les choix et les épreuves rencontrées. Merci, car c'est ta force et ton amour qui me permettent d'avancer.

Nicolas Stoiber.

Contents

Table of Contents i

Thesis Summary in French 3
0.1 Introduction . . . 3
0.2 Representation objective et pertinente des expressions faciales . . . 6
0.2.1 Espace d'apparence faciale . . . 6
0.2.2 Variété ("manifold") des expressions faciales émotionelles . . . 7
0.2.2.1 Extraction de la variété des expressions faciales . . . 8
0.2.2.2 La variété des expressions comme representation visuelle 10
0.3 Manipulation des expressions faciales émotionelles . . . 11
0.3.1 Construction d'espaces apparence faciale pour des personnages virtuels . . . 11
0.3.2 La variété des expressions faciales comme interface de contrôle . 13
0.4 Dynamique des expressions faciales émotionnelles . . . 16
0.4.1 Observation de la dynamique des expressions . . . 16
0.4.2 Modélisation de la dynamique des expressions . . . 17
0.4.3 Performance et évaluation du système . . . 20
0.5 Conclusion . . . 21

1 Introduction 25
1.1 Context and Motivation . . . 25
1.1.1 Virtual Characters . . . 25
1.1.2 Facial Expressiveness . . . 27
1.1.2.1 Channels of Facial Expressiveness . . . 27
1.1.2.2 Emotional Facial Expressions . . . 27
1.1.2.3 Facial Expression Dynamics . . . 28
1.2 Problem Statement . . . 30
1.3 Thesis Organization . . . 31

2 Representation and Animation of Emotional Facial Expressions: State of the Art 33
2.1 Representation of Facial Deformations . . . 35
2.1.1 Low-level Representations of Elementary Deformations . . . 35
2.1.1.1 Generative representations . . . 36

2.1.1.2 Descriptive representations . . . 38
2.1.1.3 Data-driven representations . . . 40
2.1.2 High-level Computational Representations of Emotional Facial Expressions . . . 44
2.1.2.1 Top-down approach (emotion-driven representations) . . . 45
2.1.2.2 Bottom-up approach (expression-driven representations) 50
2.2 Temporal Aspect of Facial Expressions . . . 53
2.2.1 Keyframing . . . 54
2.2.2 Performance-based approaches . . . 57
2.2.2.1 Concatenation of motion data . . . 58
2.2.2.2 Motion data editing . . . 60
2.2.3 Dynamic Models . . . 63
2.2.3.1 Physics-based Dynamic Models . . . 63
2.2.3.2 Learning-based Dynamic Models . . . 65
2.2.4 Motion Variability . . . 68

3 A Data-driven Meaningful Representation of Facial Expressions 71
3.1 Facial Appearance Space . . . 72
3.1.1 Active Appearance Models . . . 73
3.1.2 Facial Database . . . 74
3.1.3 Human Facial Appearance Space . . . 75
3.2 Facial Expression Manifold . . . 77
3.2.1 Dominant Directions . . . 78
3.2.2 Facial Expressions Manifold Extraction . . . 79
3.2.2.1 Dominant Direction Detection . . . 80
3.2.2.2 Expression Space Embedding . . . 81
3.2.2.3 Projection on the Manifold . . . 84
3.2.3 Expression Manifold as a Visual Representation . . . 85
3.2.3.1 Dimensionality . . . 86
3.2.3.2 Representation Results . . . 88

4 Manipulation of Emotional Facial Expressions 91
4.1 Construction of Facial Appearance Spaces . . . 92
4.1.1 Database Bootstrapping . . . 93
4.1.2 Facial Appearance Space for a Virtual Character . . . 95
4.2 Expression Retargeting . . . 96
4.2.1 Background on Expression Retargeting . . . 96
4.2.2 Expression Retargeting in Appearance Spaces . . . 101
4.3 Low-Dimensional Control Space for Facial Expressions . . . 106
4.3.1 Background on Intuitive Control of Facial Expressions . . . 107
4.3.2 Facial Expression Manifold as a Control Space . . . 108
4.3.3 Intuitive Control of Synthetic Faces . . . 109

5 Dynamics of Emotional Facial Expressions 115
5.1 Facial Dynamics Problematic . . . 116
5.1.1 Dynamics for Real-time Applications . . . 116
5.1.2 Global State-spaces . . . 118
5.1.3 A Local Approach . . . 120
5.2 Modeling Motion Dynamics . . . 121
5.2.1 Motion Signals Analysis . . . 121
5.2.2 Computational Models of Motion . . . 122
5.2.2.1 Nonlinear Dynamics . . . 122
5.2.2.2 Motion Variability . . . 124
5.2.3 Coordination . . . 126
5.3 Results and Evaluation . . . 128

6 Conclusion 133
6.1 Summary . . . 133
6.2 Valorization . . . 134
6.3 Publications . . . 139
6.4 Perspectives . . . 141
6.4.1 Research Perspectives . . . 141
6.4.2 Industrial Perspectives . . . 143

Thesis Summary in French

Contents

0.1 Introduction 3
0.2 Representation objective et pertinente des expressions faciales 6
0.2.1 Espace d'apparence faciale 6
0.2.2 Variété ("manifold") des expressions faciales émotionelles . . . 7
0.2.2.1 Extraction de la variété des expressions faciales . . . 8
0.2.2.2 La variété des expressions comme representation visuelle 10
0.3 Manipulation des expressions faciales émotionelles 11
0.3.1 Construction d'espaces apparence faciale pour des personnages virtuels 11
0.3.2 La variété des expressions faciales comme interface de contrôle 13
0.4 Dynamique des expressions faciales émotionnelles 16
0.4.1 Observation de la dynamique des expressions 16
0.4.2 Modélisation de la dynamique des expressions 17
0.4.3 Performance et évaluation du système 20
0.5 Conclusion 21

0.1 Introduction

Ces dernières décennies, la synthèse d'image par ordinateur est devenue une source principale de contenu pour toute une variété d'applications. En marge des industries classiques comme la production cinématographique et les jeux vidéo, l'image de synthèse a plus récemment ouvert de nouvelles perspectives dans des domaines tels que les télécommunications et les interactions homme-machine.

Dans ces mondes virtuels, une des tâches les plus ardues est l'intégration de personnages virtuels réalistes. Les personnages virtuels se sont cependant imposés comme essentiels, car ils simulent les formes de vie avec lesquelles nous sommes habitués à échanger dans le monde réel. En effet, interagir avec d'autres humains est un processus naturel pour lequel nous avons été entrainés toute notre vie. Malheureusement, c'est

cette même familiarité qui complique la tâche: l'expérience que nous avons à intéragir avec nos semblables fait de nous tous des experts de l'observation des comportements humains, et rend la création d'humains synthétiques réalistes extrêmement difficile.

Dans ce contexte, le visage est probablement l'aspect le plus important d'un personnage. Il apparait que dans les mondes virtuels comme dans la réalité l'attention des observateurs se focalise instinctivement sur les visages. Ce fait n'est pas surprenant, étant donné que les visages concentrent les canaux de communication les plus importants. La conséquence de cela est que l'obtention de personnages virtuels réalistes passe inévitablement par la reproduction d'une l'expressivité faciale convaincante.

Expressivité faciale

La richesse de l'expressivité faciale provient des multiples canaux de communication qu'elle met en jeu, et l'un des plus connus est porté par ce que l'on appelle les expressions faciales. Plusieurs classifications des expressions faciales existent, mais dans la plupart des contextes applicatifs on en distingue trois principaux types: les visemes, les expressions conversationnelles et les expressions émotionnelles. Dans cette étude, nous nous intéressons exclusivement aux expressions faciales émotionnelles. Sous ce terme nous regroupons les expressions manifestant des états émotionnels, ou humeurs (joie, tristesse), ainsi que des réactions émotionnelle plus spontanées (surprise, dégoût). Nous nous focalisons sur ce type d'expressions car il est une composante essentielle de la communication non-verbale, et son rôle primordial dans les interactions humaines a été reconnu dans de nombreuses études. Les expressions émotionnelles aident à sentir le contexte, le ton d'un message et la signification de certains comportements. Elles semblent ainsi indispensable dans le cadre de la création de personnages virtuels réalistes, d'autant plus que des études ont montré que leur absence était perçue comme artificielle, et parfois même dérangeante par des observateurs humains.

Un des aspects que nous cherchons à souligner particulièrement dans cette étude est la *dynamique* des expressions faciales. S'agissant d'humains synthétiques, l'hypothèse de l'"uncanny valley" [Mor70] symbolisait déjà l'importance du mouvement dans la perception des entités vivantes. Des études de psychologie et de sciences cognitives ont depuis plus spécifiquement souligné l'importance de la composante temporelle dans l'analyse des expressions.

Dans cette thèse, notre objectif est de développer une meilleure compréhension pratique des expressions faciales naturelles, et de faciliter leur manipulation sur les visages de personnages virtuels. Techniquement parlant, il est clair que les expressions faciales sont la manifestation extérieure des processus cognitifs complexes régissant les états émotionnels. Cependant, nos travaux ne considèrent pas cet aspect, et se concentrent exclusivement sur le réalisme visuel des expressions issues de ces processus. Nous insistons en particulier sur l'aspect dynamique, qui est selon nous un élément primordial de ce réalisme.

Nature du problème et contributions de la thèse

Plusieurs méthodes d'animation permettent déjà la manipulation des expressions fa-

ciales sur les personnages virtuels. Très schématiquement, on peut distinguer deux catégories de méthodes: les approches paramétriques, et les approches basées données. Les approches paramétriques disposent d'un jeu de paramètres permettant à un animateur de manipuler les déformations faciales. Elles sont généralement très précises, mais imposent un travail de longue haleine et un sens artistique développé pour créer des expressions faciales réalistes. Les approches basées données s'affranchissent de l'expertise d'un animateur en capturant directement les séquences à animer sur un acteur réel. Les animations obtenues bénéficient ainsi du réalisme naturel de mouvements réels, cependant elles sont non-paramétrables, spécifiques à une action particulière, et ne peuvent pas être généralisées à des contextes évolutifs.

Notre objectif est de proposer un système d'animation faciale présentant les avantages des deux approches: La précision et la flexibilité des méthodes paramétriques, et le réalisme visuel des méthodes basées données. Plus précisément, le système devra répondre aux considérations suivantes:

- **Expressions réalistes.** Les expressions synthétisées doivent être visuellement réalistes, tant sur le plan statique (spatial) que sur le plan dynamique (temporel).
- **Contrôle intuitif.** Nous cherchons à rendre la création d'animations faciales plus simple et adaptée aux non-experts, afin de doter le plus possible de personnages d'expressions faciales réalistes.
- **Adapté aux systèmes temps-réels.** Les contraintes de temps et l'interactivité rendent plus difficile la création d'animations réalistes dans les applications temps-réel. C'est pourtant dans ces cas là que la crédibilité d'un personnage est cruciale.

L'approche que nous proposons dans cette thèse consiste à utiliser des techniques de traitement du signal pour modéliser les comportements naturels des visages. Les représentations et modèles ainsi extraits peuvent ensuite être utilisés pour la génération de nouvelles séquences d'animation possédant les mêmes caractéristiques. En pratique, notre étude est segmentée en deux aspects: l'aspect "spatial" des déformations faciales, et l'aspect dynamique, ou temporel. Cette division du problème se reflète dans la présentation de nos travaux.

La première contribution que nous présentons est une méthode permettant d'extraire un nouvel espace de représentation des expressions faciales émotionnelles (partie 0.2). La méthode apprend les modes de déformation des visages en analysant des données réelles, et en extrait automatiquement une représentation simple en basse dimension. La topologie simple de la représentation permet de l'associer à des interprétations sémantiques des expressions. Nous étudions également comment cet espace de représentation peut être étendu à d'autres visages, typiquement ceux de personnages virtuels (partie 0.3). Cela permet de l'utiliser comme interface pour la manipulation intuitive des expressions faciales sur n'importe quel personnage. Contrairement à d'autres approches antérieures se basant sur des interfaces préétablies, l'intérêt ici est qu'étant issue de données réelles l'espace de représentation proposé est cohérent avec la vraie topologie des déformations faciales.

La seconde contribution s'adresse plus particulièrement à l'aspect temporel des expressions (partie 0.4). Nous proposons de piloter les expressions en temps-réel à l'aide d'une collection de systèmes dynamiques. Contrairement aux systèmes classiques, ces systèmes modélisent une signature dynamique générique pour les expressions faciales, et peuvent donc être utilisés pour générer n'importe quelle séquence d'animation. Ils contiennent également une composante stochastique produisant des mouvements non-déterministes, et moins répétitifs lors d'interactions longues.

0.2 Representation objective et pertinente des expressions faciales

Les expressions faciales sont causées par la contraction de nombreux muscles dont l'interaction avec les tissus du visage produit des déformations visibles. La réalité physique de ce phénomène est complexe, aussi il est intéressant d'un point de vue pratique d'encapsuler cette complexité dans une représentation plus simple. Une représentation pertinente et cohérente permettrait de faciliter la manipulation de ce canal de communication important dans des contextes applicatifs.

Plusieurs travaux antérieurs ont poursuivi cet objectif et ont pour la plupart eu recours à des représentations préétablies. Le travail consistait a associer de manière supervisée l'espace de représentation avec les données réelles. Typiquement dans le cas des expressions émotionnelles, les espaces utilisés étaient des modèles théoriques de représentation des émotions humaines, tels la roue de Plutchik [Plu80], le disque de Cowie [CES+00], ou l'espace plaisir-stimulation-dominance de Mehrabian [MR74]. Notre approche, quant à elle, consiste à faire émerger objectivement une représentation à partir de données réelles, de telle sorte que la représentation reste cohérente avec le phénomène qu'elle est censée représenter. Cette approche peut être qualifiée de "bottom-up" car elle extrait une abstraction de haut-niveau partant de données bas-niveau.

0.2.1 Espace d'apparence faciale

L'extraction de la représentation est basée sur des données d'expressions, ce qui implique la capture et la paramétrisation de déformation faciales sur des visages humains. Dans cette thèse, les déformations faciales sont mesurées par des paramètres d'apparence, issus de la méthode des modèles actifs d'apparence (AAM). Les paramètres d'apparence mesurent conjointement les déformations géométriques ainsi que les changements photométriques, et sont choisis ici pour leur robustesse et leur capacité à représenter fidèlement les variations d'objets déformables comme les visages. Les paramètres d'apparence considérés ici forment les dimensions de ce que nous appelons l'espace d'apparence facial. Un vecteur de cet espace décrit ainsi une configuration faciale particulière.

Les paramètres d'apparence sont déterminés par l'analyse statistique d'un corpus de données (utilisation de l'analyse en composante principale, ou PCA). Dans cette étude, nous utilisons des corpus individuels, c'est à dire contenant les expressions faciales d'un seul individu. Notre corpus principal de travail contient un total de 4365 images

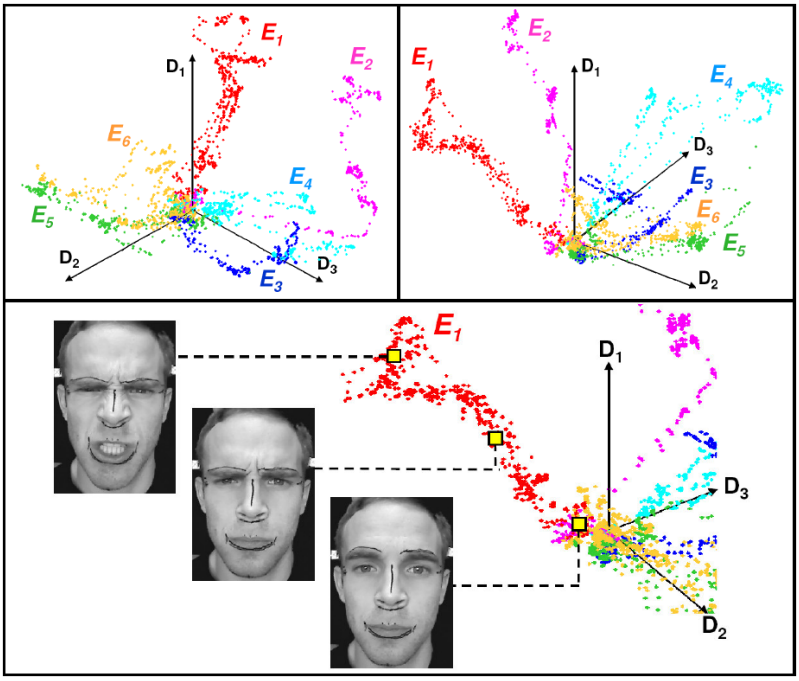

Figure 1: **Haut:** deux vues de la distribution du corpus dans l'espace d'apparence (seulement les trois premières dimensions sont représentées). **Bas:** les expressions sont concentrées autour de directions dominantes dans l'espace. Les expressions neutres occupent le centre et les expressions extrêmes la périphérie du nuage de point.

d'expressions faciales naturelles. Parmi elles se trouvent des expressions extrêmes et subtiles, typiques d'une émotion particulière ou mélanges indéfinis. Il est important de mentionner que les corpus considérés ne sont pas annotés, et que l'extraction de la représentation se fait de manière non-supervisée.

L'espace d'apparence faciale est intéressant d'un point de vue calculatoire. Il réduit considérablement la taille des données (images d'expressions) à des vecteurs de dimensionnalité plus réduite (environ 30 dimensions) et réalise des interpolations locales cohérentes des déformations faciales. Cependant la manipulation d'expressions dans un espace de plusieurs dizaines de dimensions n'est pas pratique. De plus, les paramètres considérés, issus d'une analyse statistique, ne correspondent généralement pas à des modes de déformations naturels ou interprétables.

0.2.2 Variété ("manifold") des expressions faciales émotionelles

Les paramétrisations telles que les AAM imposent un nombre important de paramètres car les expressions faciales sont des déformations non-linéaires des visages. Une majorité de chercheurs considère d'ailleurs aujourd'hui que les expressions faciales naturelles for-

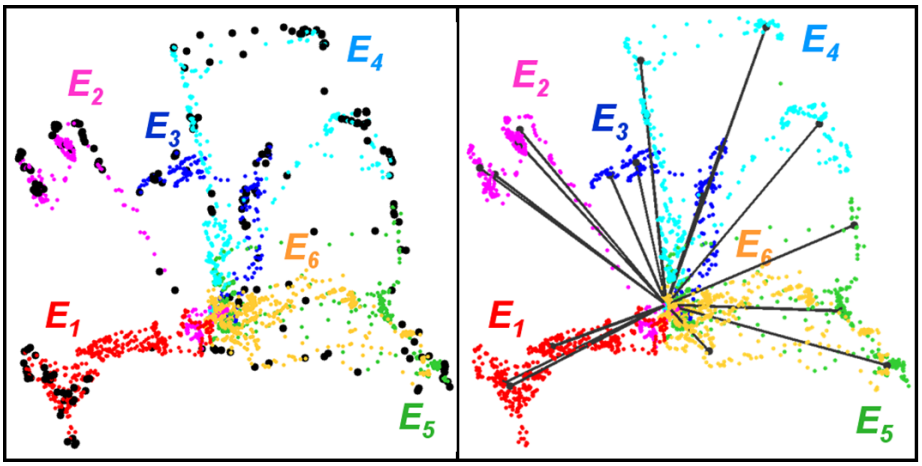

Figure 2: **Gauche:** détection des expressions de l'enveloppe convexe du corpus (points noirs). **Droite:** les directions dominantes détectées (lignes noires).

ment un espace continu non-linéaire, modélisé généralement par une variété géométrique ("manifold" en anglais).

Plusieurs études se sont grandement intéressées à cette variété des expressions faciales car elle constitue un espace de représentation naturel pour les expressions. Des méthodes génériques d'extraction de variété, telles l'IsoMap [TdSL00] ou le Locally Linear Embedding [RS00], ont été plusieurs fois proposées pour extraire la variété à partir de données. Cependant, ces méthodes génériques ne prennent pas en compte les spécificités des données faciales, notamment le fait que les corpus d'expressions forment un échantillonnage relativement peu uniforme de la variété.

La figure 1 illustre la distribution des échantillons d'un corpus d'expressions. Si l'échantillonnage n'est pas uniforme, d'autres caractéristiques intéressantes sont néanmoins visibles: Les expressions neutres sont situées au centre du nuage, et les expressions extrêmes en périphérie. De plus, la densité des échantillons est concentrée autour de directions reliant le centre et les expressions extrêmes. Le long de ces axes, on observe des expressions à niveaux d'intensité croissants à mesure que l'on s'éloigne du centre du nuage. Nous appelons ces directions les *directions dominantes* du corpus. Il est intéressant de noter que cette structure en direction dominante n'est pas spécifique à nos travaux, mais se retrouvent dans de nombreuses études.

0.2.2.1 Extraction de la variété des expressions faciales

La nature des directions dominantes mentionnées plus haut révèle une structure interessante de l'espace d'apparence faciale. Contrairement aux approches antérieures qui utilisent des méthodes génériques, nous nous basons sur ces caractéristiques particulières pour extraire une représentation pertinente de la variété des expressions faciales.

La partie intéressante de l'espace d'apparence est la partie située entre les ex-

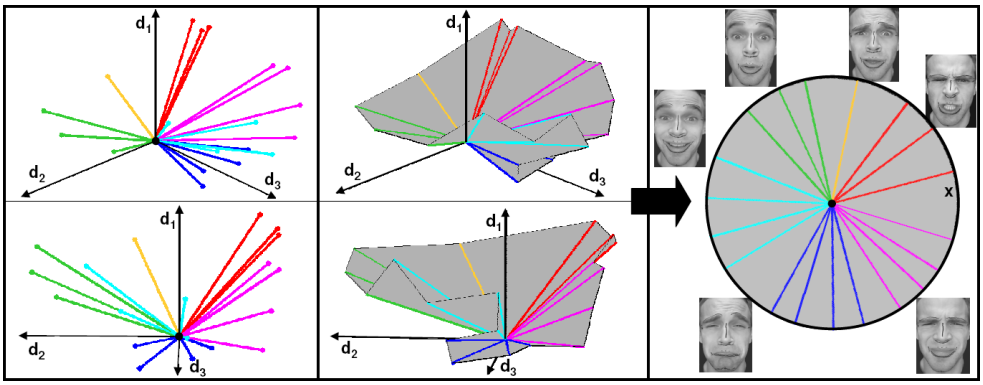

Figure 3: Extraction d'une approximation 2D de la variété des expressions faciales.*Gauche:* directions dominantes détectées. *Centre:* Association des directions déterminées par optimisation type TSP. *Droite:* Représentation "dépliée" de la variété sur forme de disque 2D.

trémités des directions dominantes et le centre du nuage de données. Cet espace forme topologiquement une *hypersphère* centrée sur l'expression neutre, chaque direction dominante formant un "rayon" de cette hypersphère. Nous considérons que la variété des expressions faciales est constituée de cette hypersphère et de son intérieur (formant ainsi une boule). Suivant cette logique, une formulation bidimensionnelle (2D) de la variété est englobée par une hypersphère de dimension 1, soit un cercle passant par chacune des directions dominantes. La variété elle-même forme une surface 2D topologiquement similaire à un disque. Cette logique s'étend de la même manière aux dimensions supérieures. Des approximations de la variété à différentes dimensionnalités peuvent ainsi être formulées.

La méthode d'extraction de la variété que nous proposons consiste en deux étapes (illustrées figure 3 pour le cas 2D):

- **Détection des directions dominantes.** Les directions dominantes du corpus sont automatiquement détectées dans l'espace d'apparence.
- **Association des directions en variété.** Une approximation de la variété est constituée en reliant les directions dominantes pour en former l'enveloppe.

Détection des directions dominantes

Les directions dominantes ont été définies comme les segments reliant le neutre et les expressions extrêmes du corpus. Si le neutre peut être désigné manuellement, les directions dominantes doivent être identifiées de manière objective telles qu'elles englobent l'ensemble des données du corpus. En pratique, les directions dominantes désignées sont les échantillons situés sur l'enveloppe convexe du nuage de données (voir figure 2).

Association des directions en variété
L'approximation de la structure de la variété formée par l'association des directions dominantes est censée refléter leur organisation. Pour une représentation fidèle de la variété, les directions liées doivent correspondre à des directions proches dans l'espace d'apparence. Nous avons présenté ce processus d'association sous la forme d'un problème d'optimisation visant à minimiser une fonction coût. La fonction coût choisie est la somme des "angles" formés par les directions reliés. En effet, les angles représentent une mesure de la proximité, il s'agit ainsi de relier les directions formant un angle faible.

Le processus peut être facilement illustré dans le cas 2D: Si l'on suppose que n_d directions $\mathbf{d_i}$, $i \in \{1, ..., n_d\}$ ont été détectées, l'approximation de l'hypersphère (un cercle topologique ici) est un chemin reliant les n_d directions. La fonction d'optimisation s'écrit alors:

$$cost_{2D} = \sum_{k=1}^{n_d} \Theta_{i_k, i_{k+1}} \tag{1}$$

où $\{\mathbf{d_{i_1}}, \mathbf{d_{i_2}}, ..., \mathbf{d_{i_{n_d}}}\}$ est l'ordre selon lequel les directions sont liées le long du chemin ($i_{n_d+1} = i_1$), et $\Theta_{i,j}$ est l'angle entre les directions $\mathbf{d_i}$ et $\mathbf{d_j}$. Les chemins respectant l'organisation initiale des directions accumulent des angles faibles et sont donc favorisés par la fonction coût.

Il est interessant d'observer qu'en 2D, le processus d'optimisation s'apparente à un problème bien connu. En effet, nous cherchons a ordonner les directions en une suite qui minimise la distance totale accumulée par les directions consécutives. C'est l'exacte formulation du problème du voyageur de commerce (ou "Traveling Salesman Problem", TSP). En conséquence, notre problème d'optimisation est NP-Complet, et peut être résolu par des algorithmes adaptés aux problèmes de type TSP. Notons que le même type de problème d'optimisation peut être formulé pour des approximations de la variété de dimensionnalités supérieures (3D, 4D, *etc*). L'extraction d'une approximation 2D de la variété des expressions faciales émotionnelles est illustrée figure 3.

0.2.2.2 La variété des expressions comme representation visuelle

L'intérêt porté à la variété est justifié par les gains de performance qu'elle peut générer, en reconnaissance d'expression par exemple, mais aussi car elle forme une représentation visuelle naturelle des expressions faciales. Ceci est particulièrement vrai dans les cas basse-dimension (2D, 3D) pour lesquels la représentation obtenue est très simple. En 2D, l'approximation de la variété forme un disque topologique, que l'on peut représenter par un simple disque unité en 2D. Une représentation similaire sous forme de sphère unité peut être produite pour le cas 3D. Ces représentations peuvent être vues comme un "dépliement" de la variété identifiée.

La simplicité de ces espaces de représentation en basse dimension permet d'associer *a posteriori* des interprétations sémantiques aux différentes zones de la variété (voir figure 4). L'agencement des différentes expressions émotionnelles peut d'ailleurs être comparé aux modèles théoriques des émotions utilisés dans les travaux antérieurs. Bien entendu, les deux approches sont différentes, car ne se basant pas sur les mêmes critères;

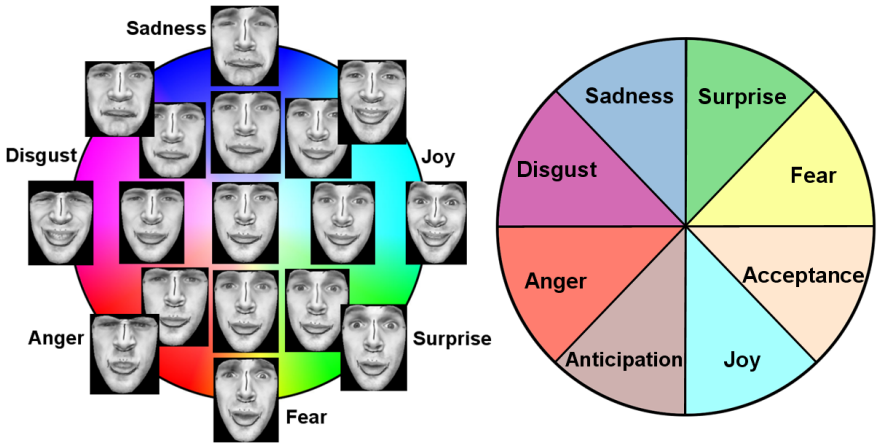

Figure 4: Comparaison entre notre représentation de la variété des expressions faciales (gauche) et une représentation théorique des émotions humaines [Plu80] (droite).

la représentation présentée ici ne prétend pas formuler de conclusions sur les caractéristiques des émotions. Néanmoins, étant extraite de données elle est plus à même de représenter fidèlement les déformations faciales expressives.

0.3 Manipulation des expressions faciales émotionelles

La méthode présentée dans la partie précédente cherche à extraire une représentation pertinente des expressions d'un individu particulier. Cette représentation peut cependant être réutilisée pour d'autres visages, typiquement ceux de personnages virtuels. Dans cette partie, nous présentons l'utilisation de l'espace de représentation des expressions dans un contexte d'animation faciale. Les différents aspects abordés dans cette partie sont illustrés figure 5.

0.3.1 Construction d'espaces apparence faciale pour des personnages virtuels

L'espace d'apparence faciale, formé par les paramètres AAM, est la base de l'extraction de la variété des expressions faciales. Les paramètres d'apparence étant déterminés par l'analyse statistique d'un corpus de données, une telle paramétrisation est envisageable pour des personnages virtuels des lors qu'un corpus d'expressions peut être constitué. Cependant, les expressions faciales sur des visages de synthèse, 2D ou 3D, sont modélisées à la main par des animateurs. Il est donc nécessaire de réduire au maximum le nombre d'échantillons du corpus.

Construction de l'espace d'apparence par "bootstrapping"
Confronté au problème de constitution d'un corpus pour un nouveau visage, les travaux

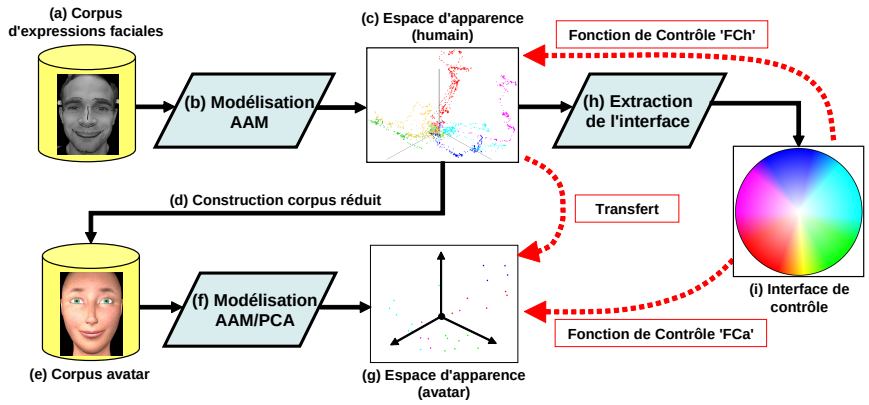

Figure 5: Schéma de l'utilisation de la variété des expressions faciales pour l'animation d'un personnage virtuel.

antérieurs ont eu majoritairement recours à des considérations empiriques. Nous proposons une méthode plus objective, qui sélectionne les éléments d'un corpus réduit à l'aide d'une méthode itérative de "bootstrapping". En effet, un corpus complet tel que celui de 4365 images présenté en 0.2.1 présente de nombreuses redondances. Il s'agit d'extraire le sous-ensemble d'expressions (corpus réduit) qui modélise l'expressivité faciale d'un individu aussi bien que le corpus complet.

A partir d'un corpus réduit à un seul échantillon (expression neutre), nous intégrons une nouvelle expression à chaque itération de l'algorithme. L'élément sélectionné pour intégrer le corpus réduit est à chaque itération celui le moins bien modélisé par le modèle courant. La procédure est plus clairement décrite en pseudocode (algorithme 1). A l'aide de cette méthode les éléments du corpus réduit sont désignés objectivement. La construction du corpus dépend finalement du nombre d'éléments retenus, ce qui relève d'un compromis entre richesse du corpus réduit et quantité de travail manuel nécessaire pour le constituer. Nous avons identifié qu'environ 35 éléments sont suffisants pour assurer une erreur de reconstruction négligeable sur les éléments du corpus complet. Un espace d'apparence faciale peut ainsi être construit facilement pour n'importe quel visage réel ou synthétique, dés lors qu'il est possible de fournir les 35 échantillons d'expressions du corpus réduit pour ce visage.

Transfert d'expressions

Un intérêt des espaces d'apparence crées à partir de la méthode ci-dessus est que les espaces peuvent être mis en correspondance sur la base des expressions du corpus réduit. A partir de cette correspondance discrète entre certains vecteurs des espaces d'apparence, on peut construire un lien analytique entre eux permettant le transfert d'expressions.

Etant donnée que les paramétrisations utilisées dans les espaces d'apparence sont linéaires (PCA, AAM), le lien entre les paramètres d'apparence vise à maintenir cette

Algorithm 1: Sélection des éléments du corpus réduit d'expressions.

Entree : **CorpusComplet**, Le corpus d'expression complet.
Sortie : **CorpusReduit**, Le corpus d'expression réduit formé par les éléments de **CorpusComplet**.
Variable: *nbElementsCorpus*, le nombre courant d'éléments dans **CorpusReduit**.
Variable: *maxNbElements*, le nombre d'éléments maximal de **CorpusReduit**.

```
// Procedure:
Initialisation de CorpusReduit à un seul élément (expression neutre) ;
while ( nbElementsCorpus < maxNbElements) do
    Générer ModelePCACourant avec CorpusReduit ;
    Projeter et reconstruire les éléments de CorpusComplet avec
    ModelePCACourant ;
    Calculer les erreurs de reconstruction des éléments de CorpusComplet ;
    Identifier l'élément ayant la plus grande erreur de reconstruction ;
    Ajouter cet élément à CorpusReduit ;
end
```

chaîne linéaire:

$$(\mathbf{c_a} - \mathbf{c_a^n}) = \mathbf{A}.(\mathbf{c_h} - \mathbf{c_h^n}) \tag{2}$$

$\mathbf{c_h}$ et $\mathbf{c_a}$ représentent les vecteurs dans les espaces d'apparence source et cible respectivement. $\mathbf{c_x^n}$ désigne dans chacun des cas l'expression neutre. $\mathbf{A}$ est la matrice (application linéaire) assurant la transformation d'une expression $\mathbf{c_h}$ en une expression cible $\mathbf{c_a}$. $\mathbf{A}$ est obtenue par régression linéaire sur les échantillons des corpus réduits.

L'utilisation typique de ce schéma de transfert est la génération d'expressions pour un avatar (la cible) à partir d'un visage réel (la source). Quelques exemples de transfert sont présentés figure 6.

0.3.2 La variété des expressions faciales comme interface de contrôle

En 0.2.2 il a été détaillé comment une représentation simple de la variété des expressions peut être extraite à partir de données dans l'espace d'apparence. Cet espace peut servir de représentation visuelle, mais il est également possible de l'utiliser comme interface pour manipuler les expressions faciales de manière intuitive.

Le recours à des interfaces intuitives pour la manipulation des expressions a été proposé par plusieurs études. Elles reposent généralement sur l'utilisation des représentation théorique des émotions humaines déjà évoquées en 0.2. Cette approche est discutable car rien ne garanti que ces représentations, basées sur des considérations psychologiques, soient fidèles à la topologie des modes de déformations des visages. Il a notamment été observé que, dans ces espaces, des vecteurs d'émotions proches pouvaient correspondre à des expressions faciales très différentes (c'est généralement le cas des représentations de la colère et la peur par exemple). L'espace de représentation décrit en 0.2 ne présente pas ce défaut car il reflète la variété des expression faciales, et

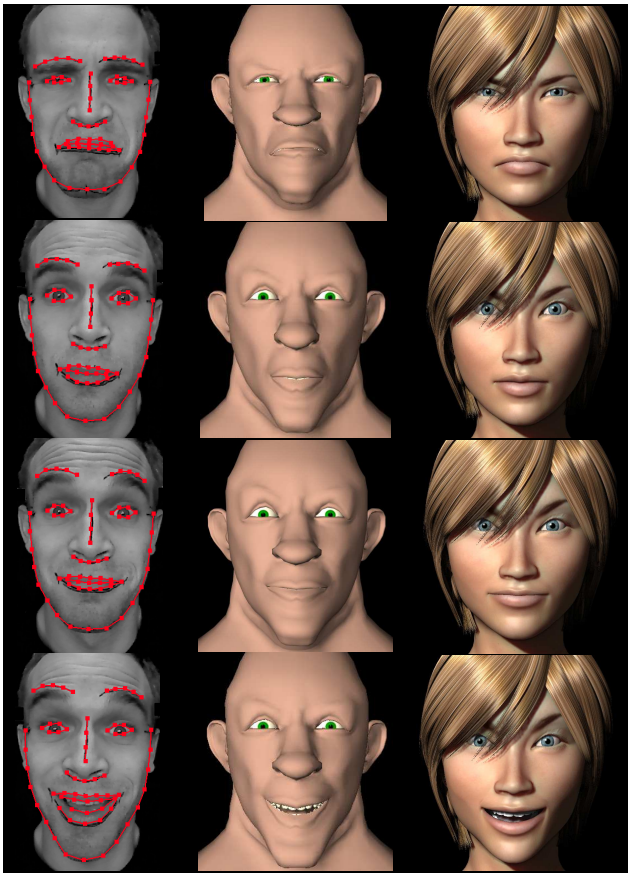

Figure 6: Exemples de transfert d'expressions. Les expressions capturées sur le visage réel (gauche) sont transférées sur les visages de deux personnages virtuels 3D (centre et droite).

est donc topologiquement cohérent avec les modes effectifs de déformation des visages. De plus, comme nous l'avons observé, sa structure simple en basse dimension permet de l'associer à des interprétations sémantiques similaires à celles des espaces émotionnels.

Fonction de contrôle

Reste à bâtir un lien analytique ("fonction de contrôle") entre l'interface et l'espace des déformations faciales. L'extraction de l'espace de représentation des expressions faciales (partie 0.2.2.1) repose sur certaines expressions particulières de l'espace d'apparence: les directions dominantes. Ces directions dominantes forment un lien discret entre la représentation visuelle "dépliée" de la variété et sa structure originale dans l'espace d'apparence. En construisant une fonction analytique sur la base de ces correspondances, il est possible de naviguer dans l'espace d'apparence en manipulant simplement l'interface basse dimension de la variété.

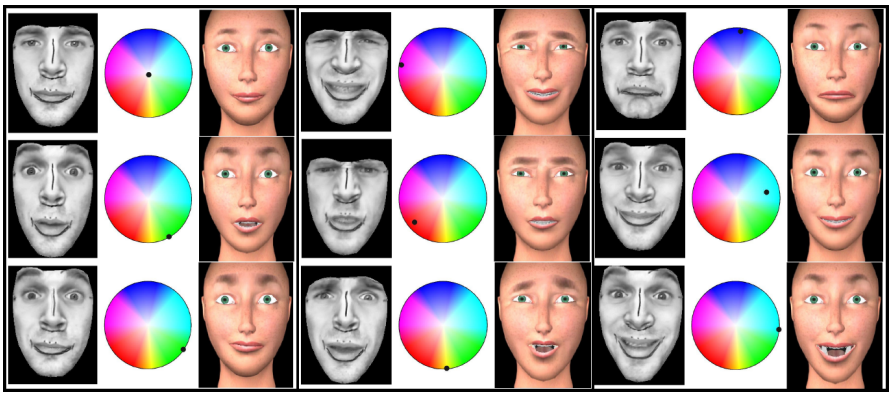

Figure 7: Exemples de manipulation d'expressions par l'interface de contrôle 2D. Sélectionner un point sur la variété des expressions (au centre sur chaque exemple) déclenche le calcul et la synthèse de l'expression faciale correspondante sur un visage réel (gauche) comme sur un visage virtuel (droite).

En pratique, la fonction est calculée par régression, en se basant sur les correspondances entre interface et espace d'apparence fournies par les directions dominantes. Concrètement, une fois la fonction de contrôle calculée, il est possible de transformer n'importe quel point de l'interface (représentation basse dimension de la variété) en un vecteur de l'espace d'apparence. L'expression faciale correspondante peut ensuite être synthétisée. Cette fonction vise ainsi à "contrôler" les expressions faciales en naviguant dans l'espace simple, typiquement 2D ou 3D de représentation de la variété des expressions. Notons qu'une telle fonction peut être construite pour l'espace d'apparence faciale humain aussi bien que pour l'espace d'apparence d'un avatar. Ces deux cas de figure sont illustrés respectivement par les flèches "FCh" et "FCa" sur la figure 5.

La figure 7 illustre la manipulation d'expressions sur des visages réels et synthétiques en naviguant dans l'interface de contrôle que constitue la variété des expressions faciales. On peut observer que cette interface, provenant des données, est fidèle aux caractéristiques observées dans les espaces d'apparence: la composante radiale (le long des directions dominantes) exprime l'intensité des expressions, tandis que la composante angulaire (différentes directions dominantes) permet de naviguer parmi les différentes expressions. La cohérence de la représentation vis-à-vis des modes de déformations faciales permet non seulement de retrouver les expressions du corpus, mais également de former de nouvelles expressions issues de l'interpolation et extrapolation des modes de déformation représentés.

L'utilisation pour un personnage virtuel d'une variété des expressions faciales apprise sur un visage humain constitue un transfert analyse/synthèse intéressant. Il est néanmoins un aspect crucial des expressions faciales que ce système ne considère pas: la dynamique des expressions émotionnelles. Cela s'observe directement sur les animations synthétisées par navigation sur une trajectoire de l'interface: les expressions

sont spatialement cohérentes, mais présentent une dynamique non-naturelle. Pour des animations convaincantes, la composante temporelle doit être considérée explicitement.

0.4 Dynamique des expressions faciales émotionnelles

Bien qu'un consensus semble émerger sur l'importance de l'aspect dynamique des expressions, relativement peu d'études l'ont abordées du point de vue de la synthèse d'animation. Dans cette thèse, nous proposons une nouvelle approche de l'animation des expressions reposant sur une collection de systèmes dynamiques apprenants et reproduisants la signature temporelle des expressions faciales naturelles.

0.4.1 Observation de la dynamique des expressions

Les méthodes d'animation classiques (keyframing, motion capture) ne constituent pas une modélisation explicite du mouvement naturel. Cela rend la synthèse de nouvelles animations fastidieuse. En revanche les recherches plus récentes ont introduit l'emploi de "modèles de mouvement", c'est à dire des formulations de systèmes dynamiques "boites noires" permettant de modéliser et reproduire l'évolution temporelle de signaux de mouvement. Cette approche à été motivée par l'observation d'invariants dynamiques, ou "signature dynamique", sur les corps et les visages [SC01, SC03].

Une approche locale
La plupart des travaux relevant de cette approche utilise le formalisme des systèmes dynamiques linéaires ou leurs extensions non-linéaires. Les paramètres de ces systèmes sont déterminés par apprentissage sur des mouvements réels, et les systèmes peuvent ensuite être utilisés pour générer de nouvelles séquences de mouvement obéissant à de nouvelles contraintes. Un point important est que ces systèmes modélisent systématiquement l'entité visage dans son ensemble, ceci afin de factoriser les corrélations entre les différents éléments faciaux. Cependant; nous avons observé que les comportements temporels ainsi que les coordinations des éléments faciaux sont changeants, ce qui crée des variations dynamiques non-linéaires, impossible à modéliser efficacement avec les systèmes globaux existants. Cela explique que les modèles actuels soient spécialisés dans la représentation d'une seule action spécifique (sourire par exemple).

Par opposition à ces approches globales, nous proposons de modéliser la dynamique faciale par une *collection* de systèmes dynamiques. Le rôle de chacun de ces systèmes est de modéliser seulement le mouvement d'un point local du visage. Une telle approche locale permet d'éviter des non-linéarités complexes observées sur des séquences variées d'expressions. Elle est par ailleurs motivée par l'identification de signatures dynamiques génériques sur les mouvements locaux, comme expliqué ci-dessous.

Analyse des signaux de mouvement
Nous nous basons ici sur l'observation des signaux de mouvements 2D de certains points faciaux d'intérêt lors de transitions entre différentes expressions faciales. Il est intéressant de constater qu'une fois le mouvement rigide de la tête éliminés, les trajectoires des

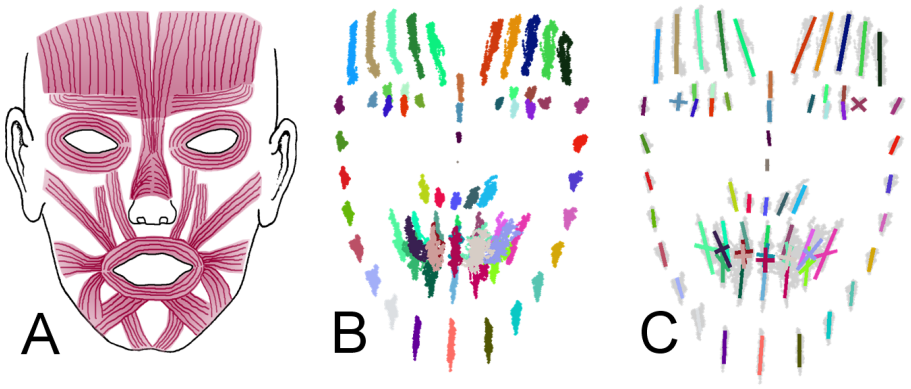

Figure 8: Trajectoires spatiales des différents marqueurs faciaux. **A**: Principaux muscles faciaux. **B**: Déplacements observés des marqueurs faciaux 2D dans un corpus d'expressions. **C**: Apprentissage des axes principaux de déplacement.

points d'intérêts sont contraints selon certains axes bien définis (voir figure 8). Dans la suite, nous considérons uniquement les signaux de déplacement des marqueurs faciaux le long de ces axes. Cela permet de s'affranchir du mouvement spatial pour se concentrer sur la composante temporelle.

Nous avons observé que les déplacements locaux mis en jeu dans les expressions faciales présentent un comportement temporel intéressant (voir figure 9). Malgré des échelles différentes, les éléments faciaux présente le même comportement accélération-inflexion-décélération, d'ailleurs déjà observés par d'autres études [SC01, SC03]. Si l'on considère la position finale du marqueur comme valeur cible, son comportement temporel est typique d'une réponse indicielle telle qu'observée sur des systèmes asservis en automatique. Les réponses indicielles décrivent la réponse d'un système à une entrée constante, et sont souvent utilisées pour identifier les paramètres dynamiques de ce système.

La "signature" dynamique observée ci-dessus est stable pour un marqueur donné; il parait ainsi judicieux de modéliser le déplacement de chaque marqueur par un système dynamique asservi dont la valeur cible, ou commande, est l'expression finale à réaliser. Les paramètres des systèmes, quant à eux, sont appris sur les signaux réels. La nature de ces systèmes est détaillée dans le paragraphe suivant.

0.4.2 Modélisation de la dynamique des expressions

Dynamique non-linéaire

Les déplacements des éléments faciaux résultent de la contraction musculaire dont le caractère non-linéaire a été souligné par de nombreuses études [Zaj89]. Bien que considérer uniquement les mouvements locaux des marqueurs réduise la complexité de la modélisation, les réponses indicielles observées présentent tout de même des signes de non-linéarité. Elles ont en particulier des caractéristiques dépendantes de l'amplitude,

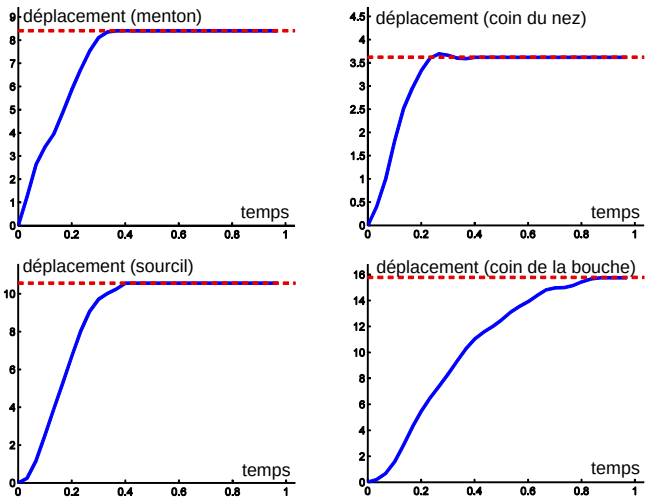

Figure 9: Profil temporel du déplacement des marqueurs le long de leurs axes (courbes bleues).

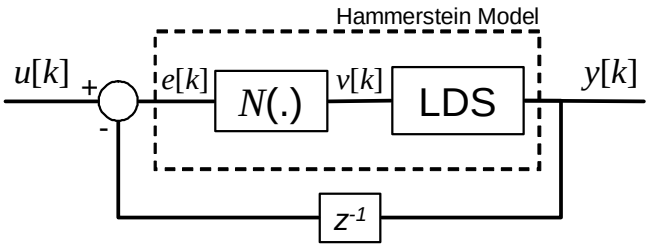

Figure 10: Système dynamique de contrôle du déplacement d'un marqueur facial.

ce qui est incompatible avec la formalisme d'un système linéaire. Ainsi, nous proposons l'utilisation d'un système non-linéaire capable de s'adapter aux non-linéarités observés: Le modèle de Hammerstein, dont les équations s'écrivent:

$$v[k] = N(e[k]) \tag{3}$$

$$y[k] = \sum_{i=1}^{p} a_i.y[k-i] + \sum_{j=1}^{q} b_j.v[k-j] \tag{4}$$

e and y désignant respectivement les signaux d'entrée et de sortie du système. Le modèle de Hammerstein gère les dépendances à l'amplitude en séparant le système en deux composantes: une composante non-linéaire statique N, suivi d'un système dynamique linéaire statique. Nous intégrons ce système à un formalisme d'asservissement en boucle fermé présenté figure 10.

Le formalisme proposé fourni une modélisation satisfaisante des signaux de mouvement réel, notamment par rapport à un système linéaire (voir figure 11).

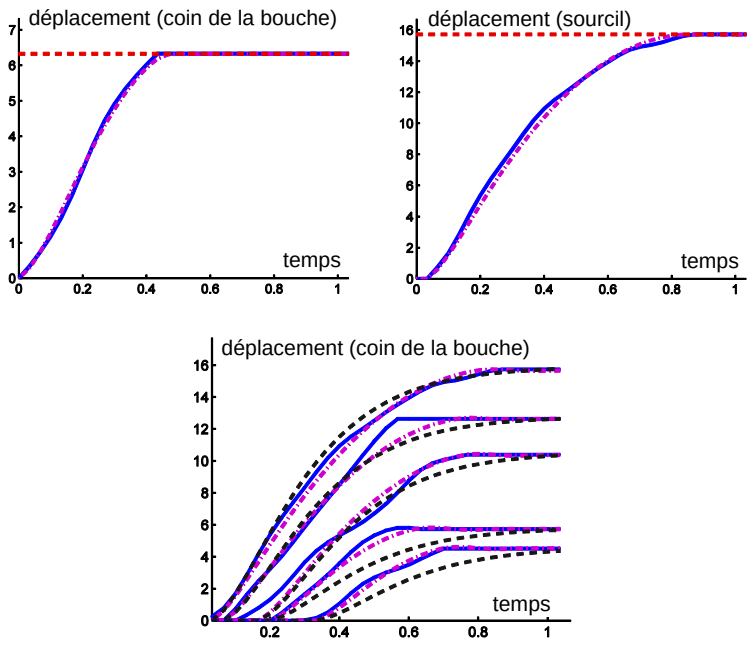

Figure 11: **Haut:** Modélisation des signaux de mouvement. Signaux de mouvements réels *(courbes bleues)*, déplacement généré par le système dynamique *(courbes pointillées magenta)* et valeur cible du déplacement *(courbes pointillées rouges)*. **Bas:** Comparaison entre modélisation linéaire *(courbes pointillées noires)* et non-linéaire *(courbes pointillées magenta)* des signaux de mouvement *(courbes bleues)*.

Variabilité des mouvements

Malgré la validité du modèle proposé, en tant que tel il n'explique pas l'intégralité des phénomènes observés sur les données. En effet, en plus du déplacement principal, les mesures revèlent des variations dynamiques non-déterministes (bruit de mesure mis à part); les signaux de mouvement ne sont jamais identiques, même pour des déplacements similaires. Cette composante reflète ainsi la variabilité inhérente au mouvement humain. Il est, selon nous, important de prendre cet aspect en compte dans un contexte d'animation car il contribue au naturel de l'expressivité humaine.

La variabilité du mouvement a été étudiée dans le domaine biomédical, et l'hypothèse à été envisagé qu'elle proviendrait de perturbations ("bruit") subies par les signaux neuronaux contrôlant les muscles [HW98]. En nous inspirant de cette hypothèse, nous proposons de modéliser la variabilité par l'ajout d'un bruit $n[k]$ venant perturber le signal de commande u du système dynamique. La formulation complète des systèmes dynamiques utilisés s'écrit finalement:

$$v[k] = N(e[k]) \tag{5}$$

$$e[k] = (u[k] - y[k-1]).(1 + n[k]) \tag{6}$$

On peut noter que les paramètres stochastique du bruit n peuvent être appris sur les

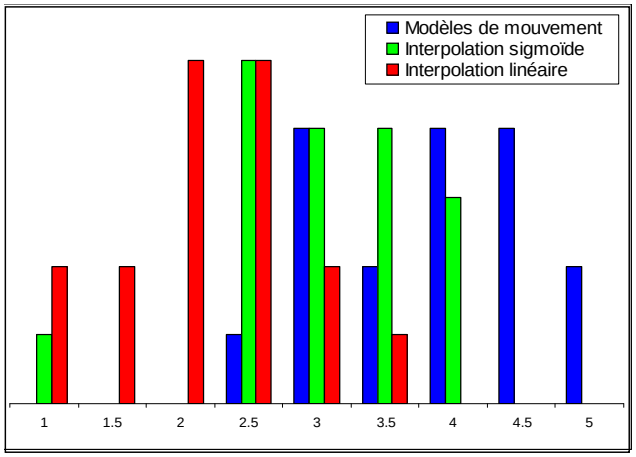

Figure 12: Histogramme des notes attribuées à trois méthodes d'animation. Les notes attribuées allaient de 1.0 (mouvement artificiel) à 5.0 (mouvement réaliste). Les participants ont montrés une préférence pour nos modèles de mouvement *(bleu)*, moyenne de 3.87, par rapport aux interpolations sigmoïde *(vert)* et linéaire *(rouge)* avec des moyennes de 3.04 et 2.18 respectivement.

données réelles, de manière à reproduire statistiquement la même variabilité que celle observée dans la réalité. Au final, la composante stochastique apporte de la variabilité. De plus, contrairement aux approches ajoutant simplement un bruit *a posteriori* [Per95], le fait qu'elle fasse partie intégrante du système de génération du mouvement garanti que le résultat final respecte la dynamique apprise sur les données.

0.4.3 Performance et évaluation du système

Une fois les paramètres dynamiques et stochastiques des modèles de mouvement appris, ceux-ci peuvent être utilisés pour animer les expressions faciales de n'importe quel personnage virtuel (par transfert d'expression). Par rapport à des approches basées-données, ce système est avantageux car il sauvegarde la signature naturelle des expressions humaines en utilisant que peu de ressources (équations simples, et peu de paramètres à mémoriser). Il est, de plus, plus flexible car la signature modélisée se généralise à la synthèse de n'importe quelle expression, et pas seulement celles ayant servi à l'apprentissage.

En termes pratiques, les animations sont générées en sélectionnant une expression de départ, une expression cible (servant de commande aux systèmes dynamiques), et en laissant le système simuler le mouvement de transition correspondant. Les animateurs réalisant cette tâche peuvent également avoir recours à l'interface présentée en 0.3 pour sélectionner les expressions sources et cibles (voir figure 13). On peut noter que la formulation entrée/sortie du système d'animation permet de varier les expressions cibles

Figure 13: Exemple de génération d'une nouvelle animation. A partir d'expressions source *(haut gauche)* et cible *(haut droite)* le système d'animation calcule le mouvement de transition correspondant.

(autrement dit les commandes des systèmes dynamiques) à la volée. Ceci assure une réactivité immédiate du système aux contraintes temps-réel.

Du point de vue de la qualité des animations, une évaluation du système sur des utilisateurs atteste de l'intêret de l'approche proposée par rapport à des techniques répandues d'animation temps-réel (figure 12). En moyenne, les participants ont attribué une meilleure évaluation à notre approche, en remarquant notamment que les méthodes classiques devenaient répétitives et prévisibles au bout de quelques secondes. Il semble que l'introduction d'une composante de variabilité cohérente renforce l'aspect naturel des animations, en particulier lors d'interactions prolongées.

0.5 Conclusion

Bilan

Comme indiqué en introduction, les travaux de thèse présentés ici ont été axés autour de deux aspects complémentaires: la représentation pertinente des déformations faciales expressives, et la gestion de l'aspect temporel des expressions faciales. De nouvelles approches ont été proposées pour appréhender ces deux aspects. La notion de variété des expressions faciales, ainsi qu'une méthode pour en extraire une représentation pertinente, ont été mis en avant pour faciliter la manipulation d'expressions sur des visages réels et synthétiques. D'autre part, un système d'animation basé sur les modèles locaux de mouvements faciaux a été proposé pour piloter explicitement la dynamique des expressions faciales.

Nous pensons que ces contributions, en encapsulant une partie de la complexité des expressions faciales, peuvent faciliter la création d'animations et la rendre plus accessible aux utilisateurs non-experts. Du point de vue du réalisme visuel, les modèles présentés dans cette thèse se basent systématiquement sur l'information contenue par les données

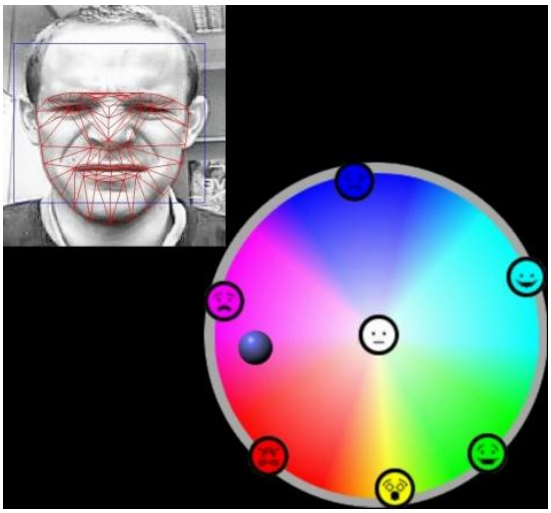

Figure 14: Aperçu de l'application d'analyse des expressions faciales d'un individu en temps-réel.

réelles, afin de bénéficier du naturel des expressions humaines. Cependant, ils ne se contentent pas de répéter ces données mais cherchent à en apprendre les caractéristiques, afin de s'adapter à de nouveaux contextes dans le cadre d'applications interactives.

Plusieurs aspects de ces travaux sont susceptibles de donner lieu à des évolutions dans le futur. Une des limites les plus importantes des méthodes présentées est le recours à des corpus d'expressions d'un seul individu. Que ce soit pour la modélisation des déformations faciales ou leur dynamique, les modèles sont propres à une personne. En théorie, une analyse des expressions multi-individu pourrait améliorer la richesse des modèles et leur capacité à s'adapter à n'importe quel visage. Malheureusement, ce genre d'approche est encore à l'état d'étude dans la communauté de recherche. La problématique réside dans l'identification d'invariants dans les expressions, les différences inter-individus étant souvent plus importantes que les variations dues aux expressions elles-mêmes.

D'autre part, pour la construction des modèles de déformation et de dynamique nous nous basons exclusivement sur l'emploi de données bidimensionnelles. Les données 2D fournissent une bonne approximation des phénomènes d'expressivité faciale, ceux-ci se déroulant principalement dans le plan frontal du visage. Cela réduit également la complexité des modèles et le processus d'acquisition des données qui repose sur de la simple vidéo. Néanmoins pour une précision accrue des modèles, l'utilisation de données 3D doit être envisagée.

Valorisation applicative

Outre les contributions scientifiques, certaines idées présentées dans cette thèse répondent à des besoins des industries de la vision et de l'animation par ordinateur. Nous

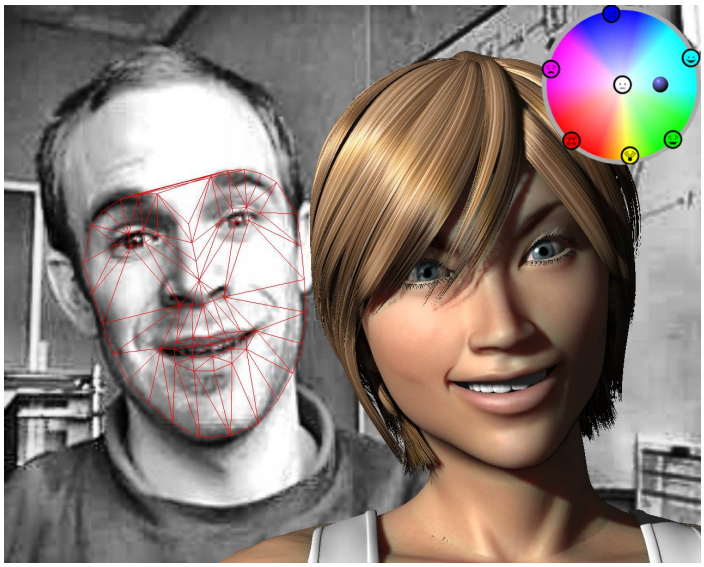

Figure 15: Le jeu du mime : reconnaissance en temps réel et imitation d'expressions faciales émotionnelles.

évoquons ici certains prototypes développés pour illustrer l'intérêt des techniques dans des applications concrètes (voir document en anglais pour plus de détails).

Un des prototypes développés réalise l'analyse des expressions faciales d'un individu en temps-réel. En pratique, les expressions de l'utilisateur sont capturées par une webcam standard, extraites par une segmentation AAM, et projetées sur la variété des expressions faciales présentée en 0.2. Le point correspondant à la projection peut être affiché sur une représentation de la variété (typiquement en 2D ou 3D), ce qui permet une interprétation visuelle de l'expression de l'utilisateur. Cette application est illustrée sur la figure 14.

Une autre application, dénommée "le jeu du mime" ("the Mimic Game"), a été présenté à la conférence SIGGRAPH en Juillet 2010. Une de ses caractéristiques intéressantes est de mettre en jeu les deux aspects principaux de cette thèse: déformation faciales expressives et animation. Plus concrètement, le jeu du mime est un programme temps-réel dans lequel un personnage virtuel "imite" en temps-réel les expressions faciales d'un utilisateur filmé par une webcam. Ce processus repose sur deux étapes: dans un premier temps l'expression faciale de l'utilisateur est capturée et projetée sur la variété des expressions faciales (de la manière décrite dans le paragraphe précédent). Le résultat de l'analyse de l'expression est ensuite passé au système d'animation (partie 0.4.2) qui reproduit l'expression sur le visage du personnage virtuel de manière autonome (voir illustration figure 15. Il s'agit bien d'un système d'"imitation", car à la différence d'un système de transfert d'expression le personnage virtuel utilise ses propres ressources (le système d'animation) pour reproduire l'expression cible. D'autre part, le fait de passer par une étape d'analyse de l'expression de l'utilisateur permet d'envisager

des scénarios plus évolués qu'une simple imitation (un personnage prenant l'air surpris lorsque l'utilisateur exprime la colère par exemple).

En conclusion, il est satisfaisant de constater que les résultats de cette thèse sont susceptible de donner naissance à des applications pratiques. Notre souhait principal, cependant, reste que les travaux présentés ici puissent encourager plus de protagonistes de la recherche et de l'industrie à s'intéresser à l'expressivité des personnages virtuels, afin de lui donner toute l'attention quelle mérite.

- -
Ce chapitre ne constitue qu'un résumé du contenu anglophone présenté dans ce document. Pour plus d'information et de détails, prière de se référer aux chapitres suivants.
- -

Chapter 1

Introduction

Contents

1.1 Context and Motivation 25
1.1.1 Virtual Characters 25
1.1.2 Facial Expressiveness 27
1.1.2.1 Channels of Facial Expressiveness 27
1.1.2.2 Emotional Facial Expressions 27
1.1.2.3 Facial Expression Dynamics 28
1.2 Problem Statement 30
1.3 Thesis Organization 31

1.1 Context and Motivation

1.1.1 Virtual Characters

In the last few decades, computer-generated imagery and virtual worlds have emerged as a primary source of visual content for a variety of applications. Apart from the now classic movie and video game industries, computer graphics have more recently conquered new platforms and opened up new perspectives in fields such as telecommunication and human-machine interaction.

In all these applications, one stimulating task has been the integration of believable virtual characters. As a matter of fact, the use of virtual characters in computer-related applications has tremendously increased over the past decade. A growing number of websites now host virtual character technologies to deliver their contents in a more natural and friendly way. The fields of multimodal human-machine interaction and human-human communication also show a developing interest in avatars for applications such as internet chats, virtual collaborative worlds, virtual classroom teaching, and conversational social agents. Figure 1.1 illustrates the presence of virtual characters in various applicative contexts.

Figure 1.1: Virtual characters in various applications. From left to right and top to bottom: Cherry (conversational agent) [PL06], Characters from the movie *Avatar*, Character from the game *God of War 2*, Greta (conversational agent) [PP01], Character from the game *Read dead redemption*, Character from *Second Life*.

Virtual characters have established themselves as essential components in virtual reality because they simulate the living form we encounter daily in the real world. Observing or interacting with humanized characters is a natural, instinctive process we have unconsciously been trained for our whole life. However, this familiarity is also the cause of the extreme difficulty of creating realistic virtual humans. The experience we all have at dealing with human behaviors makes us experts in the observation of the finer details. It makes the creation of visually acceptable synthetic characters all the more demanding.

One problematic aspect in that matter is that it is difficult to quantify or even identify what creates an impression of realistic or artificial behavior. Consequently, every aspect -from the appearance to the behavior- of a character needs to look coherent to maintain the user's interest and attention.

Above all other features of a character, the face is arguably the most important one. Indeed, in virtual reality applications as in real life, people's attention focuses instinctively on faces. This is hardly a surprise, as faces concentrate the most important channels of human communication [Arg69]. Facial appearance and behavior are able to convey valuable information to an observer or a communication partner. Our observation expertise is thus particularly developed when it comes to faces. The road toward more realistic virtual characters inevitably goes through a better understanding and reproduction of natural facial expressiveness.

1.1.2 Facial Expressiveness

1.1.2.1 Channels of Facial Expressiveness

Facial communication between two or more human beings is a complicated phenomenon we have not yet been able to fully understand. Its richness comes mainly from the multiple communication channels it involves. While some of these channels explicitly deliver the content of a message, other "non-verbal" communication channels are unconsciously accounted for by the human brain to interpret this message.

Facial expressions are a well-known example of non-verbal communication. They enrich the speech with additional information that helps transmit the sense and the tone of a message.

Specialists often identify different categories of facial expressions [FL03]. Some of them, such as eye blinking or breathing, do not specifically carry valuable information and are purely biological reflexes; others are essential channels of non-verbal communication. There is not yet a standardized taxonomy of facial expressions, but most practical applications rely on a rather simple classification that distinguishes between three types of expressions: visemes, conversational displays, and emotional expressions [BHN04, ASHS05].

The *visemes* are the representation of phonemes (speech units) in the visual domain. Practically speaking, they are the labial movements corresponding to the production of speech. Visemes reinforce or replace the verbal communication channel (lip-reading) but usually do not carry any additional information.

Conversational displays are facial movements that augment the message delivered by speech. They can be used to confirm the verbal information (nodding while saying "yes") or to emphasize a particular word or sentence (raising the eyebrows usually indicates that this part of speech is important).

Finally, *emotional facial expressions* reflect a person's emotional state. Under this label can be considered expressions of long-term emotional states or moods (joy, sadness), as well as more spontaneous affects (surprise, disgust).

In this work we specifically focus on emotional facial expressions. We believe that emotional expressions represent the most interesting type of non-verbal facial communication. If conversational displays are important to accentuate and reinforce the message delivered by speech, emotional expressions bring *additional* information that can considerably affect the interpretation of the message. They help the communication partners sense the context of an action, or the meaning of a particular behavior [PBS96, OPS08].

1.1.2.2 Emotional Facial Expressions

The central role of emotional facial expression has been known for a long time, as it was already highlighted by Charles Darwin in 1872 [Dar72]. Darwin was the first to point out the universality of expressions, and their relation to emotional states. He also identified the continuity of those expressions in human beings and animals, which contributed to strengthen his previous theory on species and evolution.

Since that, more recent studies have highlighted the role of emotional expressions in communication and social interaction: In 1968, Mehrabian empirically measured the relative importance of different channels in the communication of feelings and attitudes [Meh68]. According to his study, facial expressions are clearly the most important component, as he estimated that 7% of the emotional message was transmitted through the verbal channel (the words that are spoken), 38% though paralinguistic cues (the way the word are spoken), and 55% though facial expressions.

In the attempt of creating believable virtual characters, it appears crucial to address the problem of the simulation of natural and credible emotional facial expressions.

Indeed, previous studies have shown that synthetic faces displaying no emotional expressions, or incoherent ones, were perceived as unrealistic and even unpleasant by the human spectator [Bat94, VGS+06]. The way a virtual character behaves and emotionally reacts to its environment define the character's personality and credibility.

According to Thomas and Johnston of Disney [TJ81]:

> *From the earliest days, it has been the portrayal of emotions that has given the Disney characters the illusion of life.*

In this thesis, we aim at developing a practical understanding of natural emotional expressions and how they can be brought to virtual characters' faces more efficiently.

Technically, emotional facial expressions are the exteriorization of complex internal cognitive processes driving emotional states and reactions. Several studies have already proposed interesting models of human emotion that can drive the reactions of a virtual character [VGS+06]. However, those studies generally lack a component that correctly exteriorizes the emotions as believable facial expressions. Indeed, too often, mostly in interactive applications, emotional expressions of virtual characters are animated with artificial-looking, repetitive expressions.

As a consequence, in this thesis we do not consider the internal mechanisms of emotion but focus entirely on the *visual* outcome of emotional reactions. We found out that one important aspect is facial expression dynamics, which we introduce in the following section.

1.1.2.3 Facial Expression Dynamics

As stated above, our main concern is to provide virtual characters with realistic display of emotion. Indeed, providing a character with a complex emotional personality is worthless if the emotions are not realistically expressed at the visual level.

It actually turns out that generating visual animation of virtual characters without revealing their artificial nature is a difficult task in itself. This difficulty is often associated with the concept of the "uncanny valley" exposed by Mori in 1970 [Mor70] (see figure 1.2).

A common interpretation of Mori's hypothesis is that if artificial characters look and act *almost* as perfect imitations of human beings, they appear disturbing and even

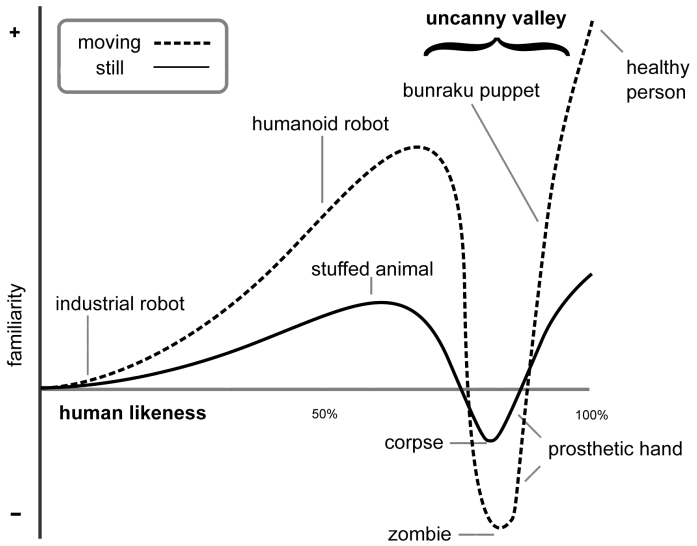

Figure 1.2: The "Uncanny Valley" hypothesis, first stated for the field of robotics [Mor70], symbolizes as curves the emotional response of human observers to artificial characters. The valley from which the name of the representation derives is the region of negative response towards characters that seem almost human.

repulsive to human observers. While the concept of uncanny valley is regarded as just a hypothesis by some, it is frequently referred to to evaluate observers' reactions to virtual characters.

Of particular importance to us is the dashed line in figure 1.2, which models observers' impression on *moving* synthetic characters. In his model, Mori clearly indicated that motion is a key component of lifelike-looking characters. This is illustrated by the rightmost part of the graph in figure 1.2, in which the dashed line (moving character) moves over the solid line (still character): at a given degree of human likeness the introduction of motion increases character familiarity.

The model also shows that motion degrades the impression produced by the synthetic character when the imitation is not perfect (as illustrated by the middle part of the valley). What we gather from all this is that the introduction of motion amplifies the emotional response of human observers to an artificial character.

Some related studies even argue that facial dynamics[1] is more important than the realism of facial appearance, when displaying facial expressions for instance [ESS00, VGS+06].

Until now however, the *dynamic* aspect of facial expression has been studied mostly

[1]Facial dynamics is the *temporal* behavior of a face when performing expressions. It is influenced by parameters such as timing, duration acceleration or speed profile of the different facial elements.

in psychological and cognitive science research. Studies from those fields have notably revealed that human observers are very sensitive to temporal cues in emotional facial expressions [Edw98] and that those play a crucial role in their interpretation [ASC05]. This suggests that humans rely on temporal dynamics in their mental representations of natural facial expressions. The effect is particularly noticeable in the case of low-intensity expressions, which are the ones mostly encountered in real life [BM08].

These observations have motivated the consideration of facial expression dynamics in computational applications. Some facial analysis systems have verified the importance of expression dynamics, as it has been observed that using temporal information improves the rates of facial expressions recognition [Pan09]. Facial dynamics has also recently started to be used as a biometric identification measure [LBF+06, THJ07, BKC+08, PG10].

Emotional facial expressions are resolutely dynamic processes, and should be considered as such. Unfortunately this aspect has been overlooked in most interactive applications featuring virtual character animation. We believe that explicit handling of the dynamic aspect is necessary for the reproduction of believable emotional expressions. We are now at a point where the appearance of virtual characters has become almost indistinguishable from photographs. It is time for character animation to achieve a similar level of realism.

1.2 Problem Statement

As exposed above, several research communities have highlighted the importance of emotional facial expressions in natural interaction scenario. One of its key aspects, as it was identified more recently, is facial expression dynamics. Rigorous facial animation is thus necessary to reach a higher-level of realism for virtual character. Indeed, a rich and genuine facial expressiveness results in more compelling characters in immersive applications such as movies or video games. It has an even stronger impact in the case of interactive applications such as conversational agents or virtual-world communication, as it enriches speech with non-verbal information.

Several facial animation methods developed in the past allow the creation and manipulation of facial expressions on synthetic faces. At a very coarse level, we distinguish two main categories: parameter-based approaches and performance-driven approaches. Parameter-based animation systems, aim at providing a parameterization describing facial deformations. They deliver a set of variables one has to manipulate to generate the desired expressions with a given degree of realism. Parameter-based approaches have been highly successful mainly because they usually provide a very fine control of subtle movements on the face. Unfortunately, the parameter sets are generally large, complex and can even hold conflicting deformation patterns causing unnatural facial expressions. Manipulating the parameters correctly requires hours of practice and only professional animators manage to use them efficiently. Even then, the construction of

a convincing animation sequence is a long and laborious process[2].

Performance-driven approaches on the other hand do not result in a parametric control over faces, but instead aim at animating the synthetic characters with motion signals that have been captured on human actors. The acquisition process usually requires heavy and expensive equipments, but the produced animations greatly benefit from the realism of real human motion and involve less animation expertise than parameter-based approaches. Nevertheless, the obtained animation sequences are delivered as immutable data and cannot easily be generalized to simulate novel facial movements in interactive scenarios.

Our objective with this work is to propose an animation framework that gathers the advantages of both approaches: the precision and flexibility of parameter-based methods, and the realistic look of performance-based methods. More precisely, our requirements are the following:

- **Realistic expressions.** Facial expressions produced by the framework must look natural from the static perspective as well as the dynamic perspective.
- **Accessible and intuitive.** We wish to make it easier to come up with good animation sequences. Typically, the system must be accessible to non-expert users to allow more applications to benefit from realistic facial animations.
- **Suited for real-time interactive applications.** As mentioned above, real-time applications are the ones that lack believable animations the most. In addition to the computing time aspect, interactive applications also impose that adaptive and varied facial expressions can be generated in dynamically-evolving contexts.

Facial expressiveness is a phenomenon involving complex skull-muscle-skin physical interactions. This is the reason why physical simulation approaches, although more pragmatic and rigorous, cannot meet the above requirements at the current level of knowledge and computing resources.

Instead, we choose to study facial expressions from a signal processing perspective. We propose to use techniques such as statistical analysis and signal modeling to model the *external* deformations of faces performing natural expressions. Those techniques allow us to extract more practical representations and characteristics from human faces. The obtained representations can ultimately be used to generate new facial animations in real-time, which display the naturalness of the original faces. This approach is detailed in the following section.

1.3 Thesis Organization

The chapters of this thesis expose how we propose to learn natural characteristics of expressive faces, and use this knowledge to synthesize facial animation. We chose to

[2]The mean animation productivity of an experienced animator is estimated at 3 seconds/day

segment the problem of facial expressiveness into two key aspects: The *facial deformation* aspect and the *dynamic/temporal* aspect. The deformation aspect deals with the way faces deform to produce expressive configurations. Although originally complex, these deformations and their link with emotional expressiveness can be modeled and efficiently manipulated. The dynamic aspect deals with the temporal profile of the movement of facial elements, whose importance has been highlighted in section 1.1.2.3.

This segmentation is not particularly obvious, nor is it common in the community, but we consider it as a good divide-and-conquer approach to realistic facial expressions. Our segmentation of the problem reflects on the chapter organization of this document:

The "state of the art" chapter (chapter 2) presents previous work relevant to the subject of this thesis. In section 2.1, we focus on the *deformation* aspect; we investigate how previous works have constructed representations of expressive facial deformations, and how these representations adapt in computational applications. Section 2.2 then focuses on how previous studies have dealt with the *temporal* aspect of facial expressions synthesis.

Chapters 3, 4 and 5 describe our contributions, and follow the same segmentation logic. *Chapter 3* presents our first contribution: a new representation of expressive facial deformations. We present a method that learns the facial deformation modes by analyzing real facial expression data, and automatically extracts a simpler low-dimensional representation of these modes. The obtained representation forms a continuous low-dimensional space on which facial expressions can be visualized and manipulated. The simplicity of the representation's topology enables it to be linked to meaningful semantic interpretations (emotions).

In *chapter 4* we study how the models and the representation of chapter 3 can be adapted to new faces, typically the faces of virtual characters. This enables the transfer of facial expressions across different faces (typically from a human face to a virtual character). Additionally, the low-dimensional representation of chapter 3 can be used as a control interface to manipulate the facial expressions of any virtual character. This control interface proves to be very useful in a character animation context as it can be used to generate whole sequences of facial animation in a very intuitive way.

Chapter 5 focuses on the second key aspect, the dynamics of emotional expressions. Our contribution in that matter is an animation system composed by a collection of motion models. The models are trained to learn the dynamic signature of human facial expressions, and can be used to animate a virtual character in real-time interactive applications. They contain a stochastic component that produces non-deterministic facial movements and improves the naturalness of the long-term visual behavior.

Chapter 6 concludes this work with a summary and a few comments on the presented contributions. The publications extracted from this work are also presented. Finally, this document ends on the brief description of a few research and industrial future perspectives for this work.

Chapter 2

Representation and Animation of Emotional Facial Expressions: State of the Art

Contents

2.1 Representation of Facial Deformations 35
2.1.1 Low-level Representations of Elementary Deformations 35
2.1.1.1 Generative representations 36
2.1.1.2 Descriptive representations 38
2.1.1.3 Data-driven representations 40
2.1.2 High-level Computational Representations of Emotional Facial Expressions 44
2.1.2.1 Top-down approach (emotion-driven representations) 45
2.1.2.2 Bottom-up approach (expression-driven representations) 50
2.2 Temporal Aspect of Facial Expressions 53
2.2.1 Keyframing 54
2.2.2 Performance-based approaches 57
2.2.2.1 Concatenation of motion data 58
2.2.2.2 Motion data editing 60
2.2.3 Dynamic Models 63
2.2.3.1 Physics-based Dynamic Models 63
2.2.3.2 Learning-based Dynamic Models 65
2.2.4 Motion Variability 68

In this chapter, we reference the previous studies most relevant to the research work on the synthesis of facial expression presented in chapters 3, 4 and 5. Previous surveys and state of the art reports in the facial animation domain [RP06], [DN07] pointed

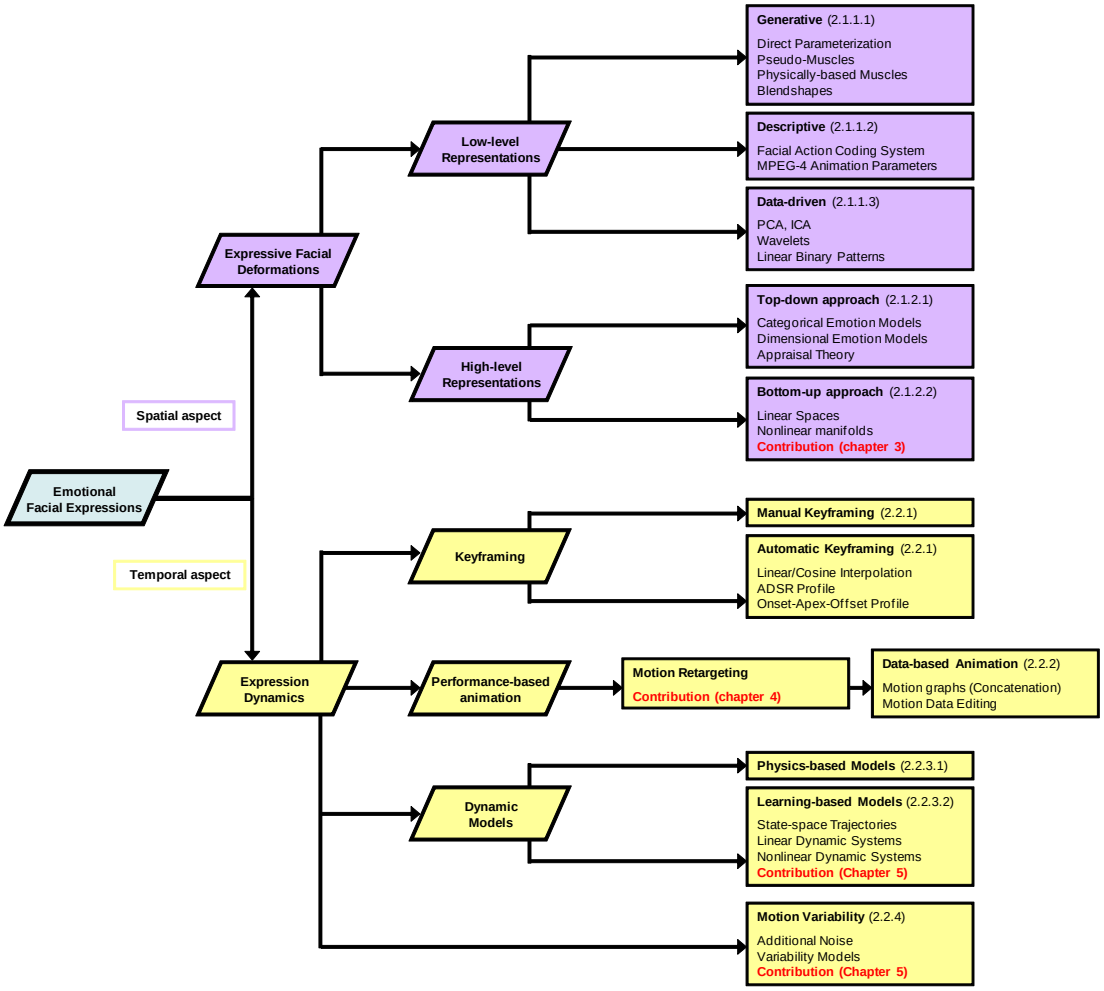

Figure 2.1: Overview of the state-of-the-art content. This figure presents the organization of the state-of-the-art chapter from the abstract classification categories (left) to the specific techniques that are discussed (right). Next to the final categories we indicate the number of their corresponding section in the text.

out the difficulty of organizing all research contributions of the field in well-identified categories. The fast evolution of the animation techniques and growing interactions between them makes it hard to agree on a relevant and stable classification. In the introduction chapter, we identified two important aspects to consider for the synthesis of natural facial expressions: the static deformation aspect, and the temporal aspect. The static aspect deals with the representation and the manipulation of expressive facial deformations, while the temporal aspect focuses on how these deformations evolve in time to display natural dynamic facial expressions. We believe that this segmentation of the facial expression synthesis problem is also appropriate to classify the research

contributions of the facial animation community. Figure 2.1 illustrates the organization of this state-of-the-art chapter as a block diagram. It can be referred to to quickly find the section number of a particular theme, or to consider its position in the segmentation we propose.

Important themes and keywords are also indicated in the page margins to provide a quick overview of the content of each paragraph.

With most contributions made in the last 15 years, the study of facial animation is still considered a recent area of research. The synthesis of realistic facial expressions, and its temporal aspect in particular, is still largely an open problem. As such, it has been significantly influenced by other areas of research, such as facial expression analysis and recognition, visual speech synthesis, and full-body animation. Facial animation clearly benefited from advances and technical solutions developed for those domains, which have been successfully applied to expressive facial animation. Consequently, in this description of the state of the art we will not only reference facial animation studies, but also relevant studies from related research fields.

2.1 Representation of Facial Deformations

For the vast majority of human beings, the handling of facial expressions comes naturally and does not involve any conscious actions. The cognitive processes involved in the generation and the comprehension of someone's facial movements remain largely unknown. Yet, when trying to reproduce these skills with a computer, it is necessary to provide it with an explicit representation of the information carried by facial expressions, so that they could be objectively manipulated. At the final stage of every display system, expressive faces ultimately consist of pixel intensity values (for traditional 2-D imagery) or of 3-D geometry and textures (for computer-generated imagery) [Par72]. This generic representation is obviously not adapted for an efficient and comprehensive handling of facial expressions. More specific schemes have been developed throughout the years to achieve greater and more accurate control over faces. In the followings sections, we will review the most significant computational representations of human facial deformations. We distinguish between low-level methods, which represent facial expressions as a combination of elementary mechanical deformations, and high-level methods, which consider expressions on a more abstract level of interpretation. These two approaches will be presented in section 2.1.1 and section 2.1.2 respectively.

2.1.1 Low-level Representations of Elementary Deformations

Many studies dealing with expressions have described facial deformations with a *divide and conquer* approach. They generally do not consider facial expressions as a whole, but as a combination of elementary patterns of deformations. We refer to these schemes as low-level representations. Using these methods, facial expressions can be observed or manipulated as a set of low-level parameters. Each parameter describes or controls a given deformation pattern on the face.

The low-level parameter sets can be constructed according to different considerations: how practical the parameters are to manually create or manipulate expressions (generative representations), how well they can be used to visually describe expressions (descriptive representations), or how optimal they are mathematically (data-driven representations). Those different types of parameterizations are presented in the following sections.

2.1.1.1 Generative representations

A natural approach when trying to understand a complex phenomenon is to try to reproduce it, and model its mechanism. Since the pioneering work of Parke in 1974 [Par74], many methods have been proposed in the computer graphics community to generate facial expressions on synthetic heads. Their objective is to provide a parameter set that simplifies the manipulation of the vertices that form the face geometry. They are therefore frequently referred to as *parameterization* of the facial motion. A facial expression is ultimately represented by the parameter values that generate this very expression.

The first parametric model of facial deformations was proposed by Parke. In this study, approximately ten parameters were proposed to manipulate the vertices of a facial mesh efficiently. The number of these parameters as well as their associated deformation patterns were determined experimentally by observing moving faces. Such methods are often referred to as *direct parameterizations*, in that the parameters directly control the translation, rotation and scaling of facial elements. This early model used relatively coarse deformation primitives. Better, more complete models have later been implemented based on Parke's technique [PWWH86, DiP89].

Pseudo-muscles

Expressive faces often display rather complex deformation patterns, and usually the transformation used in direct parameterization schemes cannot reproduce this complexity. From a biomechanical point of view, facial movements are caused by the activation of facial muscles. The action of the muscles and the reaction of other physiological entities such as fat tissues, bones and skin ultimately produce the deformation we visualize. A collection of techniques have tried to conciliate the simplicity of direct parameterization approaches with realistic facial deformations. These approaches are often called *pseudo-muscle* approaches in that they approximate the visual effects of muscle contraction while using only purely geometric techniques. Magnenat-Thalman *et al.* [MTPT88] developed a system based on *Abstract Muscle Actions* (AMA). Each AMA simulates the action of one or more muscles in specific regions of the face. In practice, The AMA consists of empirical geometric deformations of the mesh (volumetric scaling to model the action of the zygomatic for instance). Kalra *et al.* [KMMTT92] used FreeForm Deformation (FFD) of the facial mesh to simulate the movement of the skin caused by each muscle. FFD is a deformation technique in which surfaces are parameterized by a 3-D lattice of control points [SP86]. Nahas *et al.* [NHS88] and Viaud *et al.* [VY92] proposed facial animation systems based on splines (bicubic B-splines and cardinal splines respectively). The deformations of the splines simulate the action of the muscle up to

fine details such as expressive wrinkles. Splines and FFD can successfully emulate a wide range of deformations such as skin stretch and volumetric effects. They are computationally efficient as well, since they ignore the underlying mechanics of the face and run simple mathematical operations directly on the skin mesh. However, creating perceptually valid facial configurations with these techniques is a difficult task. Indeed, the control parameters have to be handled carefully to produce realistic deformations, and only experts manage to use them efficiently. Moreover, pseudo-muscle sets are usually strongly dependent on the topology of the face, and so adapting the controllers to new faces is not straightforward.

Physically-based muscles

An ideal way to generate realistic facial animation would be to perfectly model the biomechanical structure of the face. By accurately modeling the components of a human head (skin, bones, muscles, soft tissues) and their physical properties, facial animation would only consist in activating muscles and letting a physical simulation compute the resulting facial expression. Platt and Badler [PB81] first investigated this orientation by proposing a mass-spring network to model the behavior of the skin. Forces simulating the action of muscle fibers were applied to the network whose deformations were automatically computed with a propagation algorithm. Waters' seminal work [Wat87] later improved this technique by formalizing the action of multiple types of muscles (linear muscles and sphincter muscles) on the skin mesh. The computational framework relied on directional 'muscle vectors' which could be adapted to any specific face topology. Terzopoulos and Waters [TW90] reused this formalism but refined the simulation of mesh deformations by developing a more anatomically-accurate description of human skin. Skin and muscular tissues were modeled by a three-layer spring mesh in which each layer had different dynamic characteristics. The actual dynamic simulation was performed in real-time by numerically integrating the second-order differential equation associated to spring mesh. More recent contributions [KHYS02, SNF05, YSZ09, NPCP09] kept improving the anatomical structure of the synthetic face and the precision of the simulations. Sifakis et al. [SNF05] and Nazari *et al.* [NPP+08] used medical data (MRI[1] and CT[2] data respectively) to create a highly-accurate surfacic mesh for facial bones as well as volumetric meshes for muscles and soft tissues. Both used optimization algorithms based on finite-elements for the simulation of resulting skin deformations. Compared to direct and pseudo-muscle parameterizations, physically-based approaches have the greatest potential for producing realistic facial expressions on virtual characters. Yet the modeling of facial elements is still not perfect, and the computational cost of expression synthesis is much more important than for simpler schemes.

Blendshapes systems

All parameter-based methods presented above significantly facilitate the creation of facial expressions on virtual characters. However, only animation experts manage to use them efficiently; the parameter sets are generally large, complex and can even hold conflicting deformation patterns. In practice, many animation systems set up detailed facial configurations for only a few key expressions and interpolate them to create new ones.

[1] Magnetic Resonance Imaging.
[2] Computed Tomography.

When the 3-D positions of the vertices is directly interpolated, the systems are referred to as *blendshapes* systems. Most systems use simple linear interpolation, but more elaborate schemes have also been developed. Arai *et al.* [AKA96] use bilinear interpolation, which allows them to create a greater variety of variations from a set of key expressions. Joshi [Jos03] interpolates the shape locally in the different regions of the face. The blending approach has also been successfully applied to image-based facial expressions synthesis [PHL+98] based on previously developed image-warping techniques [BN92]. In those blending-oriented approaches, a particular expression is ultimately represented by the set of key expressions, and the associated blending coefficients. These features are however geometry-specific and do not straightforwardly translate to different faces.

Discussion

The schemes presented in this section provide a parameterization of facial deformations that is meant for generation purposes. They are not supposed to be used as a generic representation format for facial expressions, but they act as one when it comes to producing expressions on synthetic faces. Those representations suffer from several disadvantages however. Most schemes are specific to one particular face and, even then, using them requires a good deal of expertise. Moreover, despite good results obtained by several systems, the reliable extraction of expression parameters from data (video [TW90], motion capture [KPL98, SNF05], 3-D scan [LTW95]) remains a complicated problem.

2.1.1.2 Descriptive representations

The schemes presented as generative parameterizations act naturally as representations of facial deformations when they are used to produce facial expressions. Yet the nature of their parameters is highly dependent on the actual deformation methods they use, and these parameter sets are mostly not generic. As an example, the facial animation community often makes the distinction between 3-D geometry-based techniques and image-based techniques [NN98, RP06] because the deformation algorithms significantly differ for both categories. Nevertheless, the ultimate goal is the same: the visual production of some chosen facial expression. In that sense, a representation of expressions that is independent of the underlying generation technique is desirable. Descriptive representations of expressions do not originate from synthesis algorithms, they simply describe the visual changes caused by facial expressions.

The simplest way to describe visual changes is to measure them directly. Early experiments showed that the information carried by facial expressions can be conveyed by the movement of well-chosen facial points [Bas78]. Most optical motion capture systems are based on this characteristic and record the position of facial feature-points at each frame. Other systems rely on image-based measurement of facial configurations, such as optical flow [Mas91] or isodensity maps [KSH+92], for a more dense representation of facial deformations. Using such raw geometric representations is rather straightforward, but not particularly advantageous since among individuals faces have very different morphologies and deformation patterns. For an efficient description scheme of facial deformations, a more generic and rigorous scheme is required.

FACS

The most influential work in the formalism of facial expressions description is the Facial Action Coding System (FACS) invented by Ekman and Friesen [EF78b]. The FACS aims at describing all visually perceptible movements on the face in a systematic and adapted framework. The system decomposed facial motion into 44 elementary displacements called Action Units (AUs). AUs depict local deformations caused by the action of one or more muscles (for example AU10 corresponds to raising the upper lip). In practice, real-world facial expressions are eventually described by a combination of local AUs.

The FACS was originally created to guide humans through the observation of expressive faces. As such, it was not particularly well-adapted to computational applications, mainly due to the important number of AUs (some of which may sometimes interfere) as well as their limited spatial precision. Several extensions of the FACS have therefore been proposed to address the FACS's limitations. Rydfalk [Ryd87] created the CANDIDE system, an interesting 3-D facial parameterization based on the concept of AU. The parameters were associated to actual quantitative deformations of vertices on a 3-D wireframe model. This parameterization could thus be used in practical computational applications [LRF93]. Similarly, Kalra *et al.* [KMMtT91] introduced a pseudo muscle-based facial animation system featuring Minimal Perceptible Action (MPA), an abstraction layer inspired by the action units of the FACS. MPA also describes the visible deformations caused by groups of muscles more precisely than original AUs, with normalized real values. Essa and Pentland [EP97] later developed the FACS+ system, a FACS-like parameterization based on computer vision data. The FACS+ framework also featured quantitative action units as well as dynamic motion models to improve the recognition of expression parameters from video sequences. Tian *et al.* [TKC01] used the geometric movement of facial elements to measure the traditional AU on moving faces, and augmented this description with the measurement of additional cues such as wrinkles and furrows. Another limitation of the FACS is that it is based on local deformations whereas natural expressions usually involve coordinated movement in the entire face. Bassili [Bas79] suggested that emotional expressions could be represented by basic information about the spatial motion patterns of facial features. Yacoob and Black [YD94] later combined some AU parameters from the FACS and global motion patterns identified in [Bas79] for their own representation of expressive facial deformations. Originally, FACS representations were used as static descriptions of expressions. In [YD96], Yacoob and Davis associated a simple temporal signature to facial action units for a more accurate spatiotemporal description of recognizable facial expressions.

MPEG-4

The MPEG-4 compression standard also contains a description of visual facial actions [EHRW98]. Indeed, the standard enabled the combined encoding of real and synthetic visual content, and thus formalized the transmission of facial movement information from an encoder to a decoder [PF03]. Along Facial Definition Parameters (FDPs) used to encode a static geometry of a synthetic face, the standard defines 68 Facial Animation Parameters (FAPs) which signal the expressive configuration of this face within every animation frame. The FAPs actually specify the current position of well-chosen Feature Points (FPs). The magnitude of each FAP is expressed in Facial

Animation Parameter Units (FAPU), which are geometry-independent and encode identical deformations on faces with different dimensions. The concept of FAP is closely related to Ekman and Friesen's facial AU, although more oriented toward computational applications [EG97, KVMK99]. Like the FACS, the MPEG-4 representation does not cover the temporal aspect of local deformations, nor does it account for global coordination between facial movements. Moreover, as a compression standard, MPEG-4 looks to encode as few data as possible; consequently, the precision of the description of facial movements is somewhat limited.

FacEMOTE

Other descriptive representations of facial deformations have eventually been proposed, yet the FACS and the MPEG-4 system (and their derivatives) are undoubtedly the most widespread ones. An interesting contribution though is Byun *et al.*'s FacEMOTE facial animation system [BB02]. FacEMOTE is the extension of EMOTE, a parameterized system for gesture and body posture synthesis [CCZB00]. FacEMOTE and EMOTE partially rely on Laban Movement Analysis (LMA), a language invented for the visual interpretation and description of movement. LMA is interesting for qualitatively representing different styles of execution of a movement, but is not adapted for a precise description of the movement itself. In FacEMOTE, Byun *et al.* use MPEG-4 FAPs to encode the content of facial movements and actually use LMA to create stylistic variations on those movements.

Discussion

Descriptive representations of facial deformations appear as an advantageous alternative to generative methods. They provide coherent and generic parameterizations based on an intuitive approach (an expression is a combination of elementary deformations). Since they rely only on *visual* cues of facial movement, they are well-adapted to the extraction of description parameters from traditional visual data such as images, videos or motion capture sequences. However despite promising early results [BVS+96, LKCL98, DBH+99], the reliable extraction of AU parameters [TKC01, PR04, CBK+06, PP06] and MPEG-4 FAP [Ahl02, TRK+02] remain open problems. Indeed, FACS and MPEG-4 use parameter sets which are intuitive for humans, but computationally not ideal. The deformations associated to facial actions are not independent and often interfere with each other when activated simultaneously. Additionally, although decomposing facial movements into individual components facilitates the description, it also leaves out an important aspect of natural expressions: global correlation between facial elements.

2.1.1.3 Data-driven representations

Descriptive representations, presented in the previous section, are empirical parameterizations derived mostly from experience and subjective analysis. They were originally created out of considerations for the human observer, and helped formalize some intuitive analysis notions. From a computational standpoint however, they are not optimal in that they do not natively account for important characteristics of facial deformations, such as correlation and interactions between facial elements. To obtain a representation that best matches actual facial deformation data, the most effective solution is to

extract the representation from the data itself. Data-driven methods thus aim at identifying the most relevant parameterization, based on the *objective* analysis of a database of facial deformation examples.

Databases consist of raw data describing a collection of facial expressions examples. The expressions are typically represented as vectors formed either by 2-D or 3-D coordinates of facial features (geometry-based representation), or by pixel intensity values of a frontal-face image (pixel-based representation). Mathematical procedures are then executed on these samples to extract general and meaningful deformation patterns, or deformation *modes*. The concerned methods are therefore often referred to as statistical analysis methods.

PCA

Among these, a very popular procedure, Principal Component Analysis (PCA) [Jol86], has been used extensively in data-driven representations of faces [CM90, BVS+96, PC97]. PCA linearly transforms the original variables into a new set of statistically uncorrelated variables, the *principal components*. PCA is popular because it produces a clean and compact representation[3]. Nevertheless, the relevance of PCA eigenmodes has been questioned by several studies. Indeed, PCA optimizes the representation of the variance present in the database, yet the extracted deformation modes can rarely be interpreted physically.

ICA

Other studies preferred the properties of Independent Component Analysis (ICA) [HO00] to extract. Cao *et al.* [CFP03] used it to extract a meaningful parameterization of speech-related facial expressions, and Shin and Lee [SL09] of purely emotional expressions. Instead of searching for modes of largest variance, ICA decomposes the data into a set of deformation modes acting independently on the face. ICA produces less compact facial parameterizations than PCA, but its modes are more likely to have a meaningful physical interpretation (for instance, the independent action of a muscle on facial features). Other data-driven decomposition schemes derived from PCA have been successfully applied to facial databases. Local Feature Analysis [PA96] and Fisher Linear Discriminants [BS95], were investigated for face-based identity recognition, and then later as parameterization of facial deformations [DBH+99].

Pixel transformations

As mentioned above, statistical analysis schemes have been applied to geometric data and to pixel intensity values of pixel-based representations as well. Raw intensity values however may not be the most relevant datatype to encode changes in expressive faces. Lyons *et al.* [LAKG98] transformed pixel values of expressive image into a more attractive format using a Gabor filter bank. The filters consisted of a multi-resolution, multi-orientation bank of Gabor wavelet functions. User evaluations showed the benefit of this representation by highlighting that a significant similarity exists between Gabor coding and human appreciation of facial deformations. Other image-transformation techniques, such as Discrete Cosine Transform (DCT) [CNH04], have also been investigated.

In a similar pixel-processing fashion, Local Binary Patterns (LBP) have been used as an improvement of pixel-based representations. LBP was originally used in texture classification tasks, but proved to be efficient for facial expression representation as

[3]The axes, or components, of the PCA are orthogonal and arranged in descending order of variance.

well [SGM05b]. LBP converts an image into a collection of binary patterns, where pixel grayscale values are mapped to zeros and ones depending on their neighborhood. A histogram of these patterns reveals the nature and the distribution of local patterns, such as edges or flat areas, and thus describes the visual entities in the image. Working with binary values yields an interesting solution to the illumination problem, often encountered with pixel-based data.

Hybrid representations

Early approaches actually used only one type of data in their database, either geometric coordinates of feature points or pixel values of face images. Each datatype has its specific characteristics: images are subject to head orientation and lighting condition issues but record more information, while feature coordinates are more precise but capture only sparse information. Hybrid representations have later been created to take advantage of both data formats. Lanitis *et al.* [LTC97] introduced Flexible Appearance Models combining shape (geometry) and texture (pixel values) parameterizations. Their model was later improved and included in the well-known Active Appearance Model (AAM) formulation [CET98]. Zhang also proposed a hybrid model based on two features [Zha99]: the geometric positions of a set of points on a face, and multi-orientation Gabor wavelet coefficients extracted from the face image at the points' location. These models benefits from precise facial configuration information parameterized by the shape parameters, as well as the information from fine details such as wrinkles, which are encompassed in the texture. One can note that all presented approaches are 'holistic'. However, models that consider each region of the face individually [PC97] or hierarchically [ZG05] have also been proposed. Yet, considering the face as a whole makes sense as far as facial expressions are concerned, because of the important coordination of facial elements involved in those movements.

Nonlinear techniques

Most of the schemes listed above can be referred to as 'linear', since the identified deformation modes are obtained through linear transformation of the original degrees of freedom. Yet, it has been stressed in many studies that facial expressions involve globally nonlinear deformations and lie on a nonlinear low-dimensional manifold[4] [CXH03, CHT03, HCFT04, WHL$^+$04, CHT04, LE05, SGM05a, CK09]. Consequently, parameterizations obtained from linear transformations cannot precisely extract the true deformation modes of expressive faces. Researchers have thus turned to nonlinear dimensionality reduction techniques to extract the space of facial expressions. IsoMap and Locally Linear Embedding (LLE) in particular have been used frequently with facial expression data [CHT03, WHL$^+$04, HCFT04, ZZL05, LEM06, RGM06, PZ07, EL07]. IsoMap [TdSL00] and LLE [RS00] are nonlinear embedding methods that look for the optimal embedded arrangement of training data in a low-dimensional space. This embedded representation constitutes an approximation of the true manifold of facial expressions and thus implicitly reflects the underlying deformations modes. IsoMap is a variant of MultiDimensional Scaling [CC94] that aims at producing an embedding that

[4] A manifold is a mathematical space with more general properties than vector spaces. Manifolds behave locally as traditional $\mathbb{R}^n$ euclidean spaces, but globally describe more complex structures such as continuous nonlinear spaces.

preserves pairwise geodesic[5] distances between all database samples. LLE uses another optimization criterion in which more emphasis is put on local neighborhoods: the embedding is meant to preserve the relative position of each database sample to its direct neighbors in the low-dimensional representation. Other interesting nonlinear reduction techniques, such as Laplacian Eigenmaps [BN02], Kernel PCA [SSM98], artificial neural networks [SYH07] or global alignment of locally linear reductions [RSH02, TR03, Bra03], have also been studied in this context (see [vdMPvdH07] for a review of the most important nonlinear reduction methods).

Good results obtained in facial expressions recognition and synthesis suggest that nonlinear schemes are more likely to extract a faithful representation of the space of facial expressions. They also provide more compact representations, since linear scheme have to approximate non linear variations by several linear deformation modes. On the other hand, they are heavily dependent on the richness and the density of training data, and usually show disappointing results when being used to extrapolate new facial configurations. Additionally they are generally computationally more expensive than linear methods. To cope with these issues, some recent studies have developed linear versions of the optimization processes of nonlinear schemes (for example, Locality Preserving Projections [HN03] and Neighborhood Preserving Embedding [HCYZ05] use the same optimality criterion as Laplacian Eigenmap and LLE respectively).

Discussion

Data-driven representations undeniably produce more efficient parameterizations from a computational point of view. They can objectively detect, measure or control the deformations of expressive faces based on ground-truth experience (the database). Most of them are adapted to specific computational operations such as interpolation or classification of facial configurations. On the other hand, they rarely correspond to intuitive descriptive parameterizations (for instance, most variation eigenmodes extracted by PCA do not correspond to a physically intuitive deformation). Clearly, they are not meant to be manipulated as is by human observers.

Low-level approaches of facial deformations presented in this section are used extensively in face-related applications. They are essential since they convert raw, unpractical information (pixels or vertex coordinates) into a more tractable one (a set of parameters). Yet, as it has been mentioned in many studies, the ideal facial parameterization does not exist. Depending on the application, their respective advantages and drawbacks have to be considered. On table 2.1 we sum up the characteristics of the presented low-level representations of facial deformations. In this work we use a parameterization inspired by Active Appearance Models (AAM) to measure and synthesize low-level facial deformations. Details can be found in section 3.1.1.

[5]The geodesic distance between two points is the length of the shortest path belonging to the manifold that binds these points. In IsoMap, it is approximated by adding up distances between successive neighboring points

Representation type:	Generative	Descriptive	Data-driven
Range of expressions	++	+	-
Ease of use	-	-	- -
Interpretability	+	+	-
Automatic measurability	-	+	++
Computational consistency	-	-	+
Efficiency/Compactness	-	-	+

Table 2.1: Advantages and disadvantages of the presented low-level facial parameterization approaches. The comparison criteria are inspired by some of Pandzic and Forchheimer's requirements for ideal parameterization [PF03]). Since each type of approach has strengths and weaknesses, the choice of a parameterization depends on the application and context. A major issue in all case however is their ease of use. Indeed, it is generally difficult, and unnatural to non-experts, to describe or manipulate expressions using a set of atomic, possibly conflicting low-level deformation parameters.

2.1.2 High-level Computational Representations of Emotional Facial Expressions

The previous section focused on the representation of low-level facial deformations. In this work, we focus specifically on *emotional* facial expressions. In that context, the study of elementary facial movements is obviously important. Before analyzing or synthesizing natural facial expressions, the computer must first understand what they are made of. However, as humans, we do not only comprehend or produce emotional facial expressions as a combination of standardized elementary deformations. We are able to instinctively recognize a joyful face for example, without measuring the amplitude of lip deformations, eyebrows rise, *etc.* Similarly, we can instinctively associate different low-level deformations to one same interpretation, as well as identify a large variety of mixed expressions, even on very different facial morphologies. This leads to believe that emotional facial expressions are conceptualized and interpreted at a higher abstraction level in the human brain.

Such a high-level representation is desirable for the computational processing of facial expressions. Indeed, it would constitute a simpler and more intuitive way to manipulate facial expressions than low-level mechanical deformations, both manually and computationally. Additionally, real applications (in the field of affective computing or natural human-machine interaction) eventually deal with higher-level concepts such as human emotions and behavior. A correspondence between these concepts and their effects on facial expressions seems relevant.

The representations presented in section 2.1.1 work at the *mechanical* level, and do not directly relate to those semantic concepts. Researcher have therefore relied on what we call *high-level representations* of emotional facial expressions to bridge the gap

between raw facial deformations[6] and less precise but more meaningful abstract notions, such as emotions.

In practice, previous works associate the high level (emotion-level) and the low level (expression-level) of description using one of the following two approaches. The "top-down" approach (from emotion to expression) consists of using a predetermined model or human emotion, and force expression data to comply with it in a supervised fashion. On the other hand, the "bottom-up" approach (from expression to emotion) considers the expression level primarily and make more abstract representations arise from the expression data itself. These approaches and their related works are presented in sections 2.1.2.1 and 2.1.2.2 respectively.

2.1.2.1 Top-down approach (emotion-driven representations)

As mentioned above, the interest of a high-level representation in real-world applications lies in the ability to relate it to semantic concepts, typically human emotions. In the psychological literature, studies have come up with interesting structures that help formalize emotions. Many researchers studying the human face have naturally adopted some of these structures to represent the facial expressions associated to emotions. The main advantage of this approach is that it provides applications with a straightforward and intuitive link to semantic interpretation of emotional facial expressions. Emotion-driven representations correspond to a top-down approach: the representations originate from higher, abstract considerations on emotion and are assigned to lower-level, mechanical facial deformations.

In the following, we reference the relevant emotion-driven structures, and how they have been used as high-level representations of facial expressions. Three principal types of structures have been used in previous studies: discrete categorical segmentations of emotions, continuous description as dimensional spaces, and finally the breakdown of emotions as cognitive appraisal processes.

Categorical representations

The categorical interpretation of emotional facial expressions is by far the most widespread one. It consists of classifying facial expressions into a set of categories, or classes of interpretation; the categories generally correspond to common human emotions. This approach is justified by many studies in the field of psychology and behavioral biology, starting by Charles Darwin's contribution of 1872 [Dar72]. He asserted the existence of universal expressions of emotion according to his observations of faces of humans and animals. Later, the work of Izard [Iza71] and Ekman and Friesen [EF71] confirmed this assertion by identifying facial expressions of emotions that can be universally recognized by humans of different cultures, background and nationalities. The concept of universal expressions was later formalized by Ekman [Ekm82], who isolated six universal expressions associated to basic emotions: joy, sadness, fear, disgust, anger and surprise.

[6] High-level representations generally rely on a low-level parameterization (such as those described in section 2.1.1) to manage elementary facial deformations.

This strong link between some facial deformation patterns and emotional interpretation constitute a straightforward high-level representation of emotional facial expressions. It is particularly adapted to facial expressions recognition systems, which classify facial configurations according to a set of output classes. An exhaustive listing and accurate description of automatic facial expressions recognition systems would be too long, and lies outside the scope of this thesis. Please refer to [FL03, PB07, LTZ09] for more detailed studies on that particular subject.

Ekmanian classes

An overwhelming majority of those systems use Ekman's six basic emotions as output classes [KH92, PC97, BY97, LKCL98, LBA99, Bar03, HCFT04, ADD04, KP07, SK08]. By virtue of their universality, they lend themselves well to basic recognition tasks. As a matter of fact, modern classification systems generally claim a recognition rate of above 90% on standard databases [KCT00]. The introduction of temporal information in modern classification systems even improved recognition rates [CGH00, ZLGS03, AK06, SGM06].

Yet, the full-blown expressions of Ekman's emotions -such as those typical modeled in the cited studies- rarely occur in practice. In the real world, the majority of emotional facial expressions are more mixed and subtle, leading to somewhat ambiguous impressions. Recent recognition systems tend to reflect this reality by associating unknown expressions to a probability distribution along emotional classes [HCFT04]. The universal facial expressions themselves represent a rather restrictive concept: an expression of joy can display significant interpersonal and even intrapersonal variations. Moreover the universal expressions have been determined based essentially on human observations and impressions, not on actual facial data. Despite their undeniable intuitive nature, the six basic categories are not optimal from a computational point of view.

Facial expressions are inherently a continuous phenomenon. The category-based representation, on the other hand, is discrete. It reduces the richness and diversity of emotional facial expressions to a few class labels. To perform a more precise analysis, synthesis or editing of expressions, *dimensional* representations of emotional facial expressions are presumably more appropriate.

Dimensional representations

The major drawback of the categorical approach is the reductive aspects of a discrete representation. Dimensional representations, on the other hand, opt for a continuous representation of emotional facial expressions: each expression is represented by a point in a multidimensional space. This is more in accordance to the rich and dense nature of facial deformations. Dimensional representations are theoretically able to represent stereotypical as well as mixed expressions, handle different expressive intensities, and account for the subtle variations often displayed in natural facial expressions.

Weighted categories

The simplest structure used by researchers is a straightforward dimensional extension of Ekman's basic emotions. Du and Lin [DL02, DL03] relied on such as representation to construct a facial expression synthesis system driven by an emotion vector.

Each dimension of the vector reflects the contribution of one basic emotion (among joy, sadness, surprise, anger, disgust and fear). The mathematical association between this space and actual facial expressions is computed by parametric regression on a large set of emotion-annotated examples. Zhou and Lin [DL02, DL03, ZL05] later extended the approach; their system used a more elaborate facial deformation model (close to AAM) for a better adaptation to new facial morphologies. Zalewski and Gong [ZG05] proposed to capture facial expressions performed by a webcam user and decomposed them as a weighted combination of the six basic emotional expressions. The corresponding expressions can be synthesized on a virtual character using a set of morph targets, each representing one of the basic emotions. Littlewort and his colleagues developed a similar application, an expression mirror that retargets emotional facial expression based on a 7-Dimensional emotion score (the quantitative contributions of each basic emotion). Although fairly straightforward, the dimensional extension of the universal expressions paradigm is questionable. It is highly doubtful that humans comprehend natural facial expressions as mixes of extreme emotions, and that representation probably does not reflect a reality.

In search for alternative dimensional representations, researchers from the image analysis and synthesis communities have once again drawn their inspiration from the psychology literature.

Schlosberg proposed that emotional concepts could be adequately described in a circular arrangement [Sch41]. This arrangement actually originated from the observation of emotional facial expressions: perceptually close categories of expressions could be grouped in a circular pattern. More recently, Plutchik [Plu80] exhibited a circumplex-like emotional space based on similarity ratings of emotion-related words. Circumplex arrangements were not truly emotion spaces yet, but they provided a symbolic representation of emotion closeness (adjacent emotion on the circumplex) and opposite emotions (see figure 2.2).

2D models

Schlosberg was arguably the first to formalize an actual dimensional space by identifying two quantitative emotional axes that match his initial circular arrangement: the pleasantness-unpleasantness axis and the attention-rejection axis [Sch52]. Years later, Russell [Rus80] derived a similar structure, proposing arousal and valence as the two dimensions and locating the archetypal emotion states in a circumplex in that space. Analogous 2-D emotion spaces were later proposed based on different considerations. A valence-activation space was supported by Whissel [Whi89], who associated these notions to a collection of words. In that context, activation denotes the passive-active nature of an emotion. Cowie *et al.* [CES+00] recently provided the description of a 2-D emotion space formed by activation and evaluation (see figure 2.2). Evaluation (ranging from negative to positive) is close to the previous notion of valence, but is actually inspired by cognitive theories of emotional reactions and states. Several other comparable representations have been suggested in the psychology literature.

3D models

Mehrabian and Russel presented multiple multimodal evidence in favor of a more

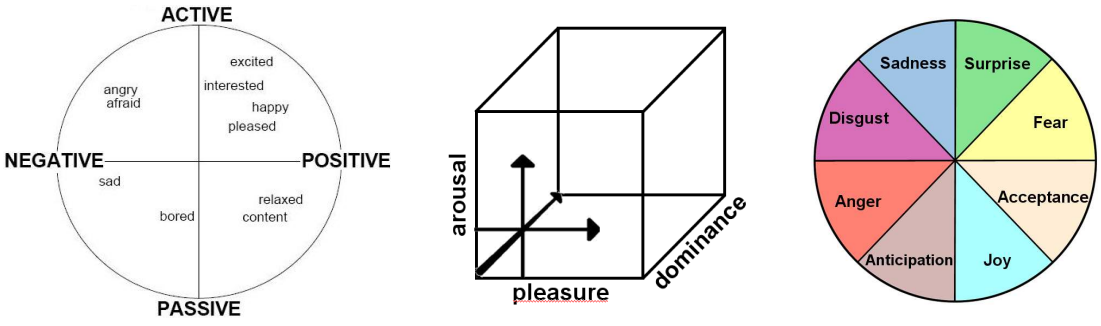

Figure 2.2: Visual illustration of some dimensional emotion models. **Left:** Cowie's activation-evaluation space [CES$^+$00]. **Middle:** Mehrabian Pleasure-Arousal-Dominance space [MR74]. **Right:** Plutchik's circumplex of emotions [Plu80].

complex, three dimensional space of emotion: Pleasure-Arousal-Dominance [MR74, RM77], in which dominance reflect the level of dominance-submissiveness toward the environment. 3-Dimensional spaces were also motivated by the fact that in 2-D spaces distinct emotions happen to display almost equal parameter values. Typical examples are the anger-fear and disgust-fear pairs whose disturbing proximity in terms of valence-arousal values has been reported in several studies [LK00, SDF06]. Mehrabian suggested 'dominance' as a third parameter, but other influential parameters of emotional states have been advanced, such as 'stance' or 'agency'. Those concepts are conceptually close as they represent the level of inwardness-outwardness of the emotion.

Several recent research contributions used these 2D and 3D emotion spaces as high-level representations in concrete applications. They were particularly popular in animation systems, which looked to use these simple yet meaningful spaces as control interfaces to manipulate the facial expressions of real or virtual characters. This topic and its background are developed in section 4.3.1.

Discussion

Dimensional representations of emotion provide applications with meaningful and intuitive structures for fine expression analysis and synthesis. They conform to the rich and continuous nature of facial expressions, and are thus more appropriate than categorical structures. Nevertheless, one can argue whether theoretical spaces comply with the reality of facial deformations modes. Artificially associating facial movements with a discrepant representation would cause errors in expression interpretation and manipulation tasks.

Appraisal-based representations

Recently, some researcher have emphasized the fact that emotional facial expressions are the external display that reflect inner emotional functions, and thus pushed further the cause-consequence relationship between emotional mechanisms and expressions. They

aimed at generating natural expressions by relying on simulations of human emotional reactions.

A widely accepted theory and a central concept of cognitive emotion mechanisms is the theory of appraisal. The theory assumes that emotional states result from cognitive processes that evaluate occurring internal and external events with respect to the individual's objectives, beliefs, attitude, *etc.* The evaluation processes are referred to as *appraisal.*

Most practical implementations of human emotional processes have been inspired by the appraisal theory. Scherer developed a well-known appraisal-based model of emotion in which emotional responses are triggered by the outcome of a determined sequence of evaluation processes [Sch84]. A more practical appraisal-based model is the OCC model [OCC88] developed by Orthony *et al.* It is commonly used in application to simulate the emotional reactions of virtual agents in response to specific interactive events.

The simulation of realistic emotional behavior is part of artificial intelligence processes, and clearly lies outside the scope of this paper. However, Appraisal mechanisms also provide a new comprehension of emotional facial expressions.

In his component process model of emotion, Scherer [Sch87] made predictions about which facial actions are likely to occur with respect to the outcome of successive appraisal processes. Scherer assumes that the induced facial changes combine temporally to ultimately form an emotional expression. This paradigm as later been applied in practice: Wherle *et al.* [WKSS00] generated dynamic facial expressions on a simple face sketch by cumulating appraisal-specific Facial Action Units. Grizard and Lisetti [GL06] proposed another practical application of Scherer's theory to create emotional facial expressions on an expressive robot.

The combination of successively triggered facial movements is not a trivial issue, since some of them can be conflicting. Malatesta *et al.* [MRK06] implemented appraisal-based dynamic facial expressions for a more realistic 3-D virtual agent. They used two methods to treat the combination of appraisal facial actions: additive animation and successive animation. De Rosis *et al.* [dRPP$^+$03] provided an elaborate formalization of the appraisal approach, using Dynamic Bayesian Networks (DBN) to model the evaluation processes. In accordance to their previous work [PP02], they included an additional DBN to specifically handle the combination of resulting facial signals and produce a consistent emotional expression.

An interesting characteristic of the appraisal approach is that, in opposition to other emotion-driven representations, it inherently deals with the temporal aspect of expressive behavior. We have already mentioned that temporal considerations are crucial for realistic expressive faces, and this will be further discussed in section 2.2. Nevertheless, appraisal-related studies still primarily focus on pure emotional aspects. The associated expressive components are secondary. The consequence is that, in many approaches, powerful and precise emotion models are hindered by the insufficient expressiveness of their facial animation systems. Kshirsagar *et al.* [KMt02] and Egges *et al.* [EKMt04]

developed elaborated OCC-based models for natural dialogue applications. The actual expressive output however, is limited to the simple display of a combination of Ekman's prototypical expressions.

Discussion

Appraisal is interesting from the behavioral point of view, and arguably provides the best framework to describe believable emotional behaviors. However, realistic emotion simulation alone does not guarantee the realism of the corresponding expression. External manifestations of emotions, such as expressive deformation and movement of faces, are also undeniably linked to the biomechanical aspects of the human body. In that sense, semantic considerations are desirable but not sufficient to realistically represent facial expressions. In this thesis we focus principally on the realism of expressions, and not on the emotional mechanisms that cause them. Our work can thus be considered more as a complementary aspect to appraisal approaches. Dimensional representations, such as those presented in the previous section, emphasize the functional nature of facial expressiveness, and thus constitute a more appropriate paradigm considering our objectives.

2.1.2.2 Bottom-up approach (expression-driven representations)

Emotion-driven high-level representations of facial expressions produced encouraging results in the referenced studies. Dimensional representations are particularly relevant structures considering the continuous nature of facial expressions. Their main advantage is that they manipulate semantic dimensions: facial expressions can be intuitively manipulated in semantic terms, which is interesting from an application perspective. However, it is important to note that such systems rely on supervised associations between emotional parameters and facial deformation parameters. The emotion space is known *a priori*, and associating facial expressions with it forces the facial deformation modes to comply with its structure. It is not obvious that the emotional considerations are consistent with the physical characteristics of facial expressions.

To avoid distortion between the facial deformation data and their high-level representation, an interesting idea is to make this representation arise from the data itself. Relying primarily and objectively on expression data guarantees the consistency of the high-level representation. On the other hand, the link to more abstract notions is not straightforward anymore. Contrary to emotion-based representations where semantic meaning is at the origin of the construction of the space, here the significance of the represented facial deformations has to be interpreted *a posteriori*. The objective is thus to extract a structure from low-level deformations that is simple, intuitive enough to allow higher-order semantic dimensions to be identified. This approach can be thought of as a bottom-up approach.

Linear spaces

Most of low-level schemes presented in section 2.1.1 do not qualify as meaningful representations. Indeed, they represent facial movements as a combination of multiple elementary deformation patterns, and thus produce high-dimensional parameter spaces. Such space are not only hard to manipulate, they are also hard to interpret from a semantic point of view. The chosen representation must first and foremost have a simple

structure, suitable for intuitive interpretation. We have previously mentioned that Principal Component Analysis has the property of concentrating data variation on its first parameters. Ruttkay [RNtH03] used this property to construct a meaningful representation of facial expressions, the 'emotion squares', based on the first 4 principal components of their MPEG-4 FAP database. The expressions could be manipulated easily by navigating in two 2-Dimensional spaces (the aforementioned emotion squares) formed by the first and second eigenvectors and the third and fourth eigenvectors respectively. Only these four principal components held 84% of the database variance and provided a simple representation space. However, each eigenmode could only describe a simple, linear deformation pattern. Consequently, the scope of represented expressions was small, and the role of each eigenvector hardly inspired any high-level interpretation. In their talking head animation system, Chuang *et al.* [CB02] separated structurally emotional expressions from speech-related expressions using a bilinear framework. Emotional expressions were parameterized by a 3-Dimensional PCA-like parameter space. Expressive 'styles' from their database such as anger and happiness could be extracted or synthesized on a talking head as 3-D vectors. Deng *et al.* [DBNN04] introduced expressive intensity in addition to their PCA-based expression synthesis system. The intensity was automatically or manually adjusted by scaling the facial deformation vectors. Again in those studies, the expressive parameters were not intuitive in nature. In practice, the manipulations were limited to the reproduction of expressions from their database.

Independent Component Analysis looks to isolate statistically independent components in a database of different configurations. The components therefore tend to make more semantic sense, to a certain extent, than principal components which are purely variance-driven. Cao *et al.* [CFP03, CTFP05] successfully isolated expressive deformation patterns, such as eyebrows motion or smiling lips. Yet, like PCA, ICA is technically limited to linear deformations modes, whereas emotional expressions often involve multiple nonlinear deformations. Simple data analysis schemes provide high-dimensional structures, which do not form intuitive representation spaces for emotional expressions. Multilinear approaches [VT02, EL04, CB05, VBPP05, LE05, MVBV06] manage to explicitly separate different influential factors such as head orientation, illumination or identity, but suffer from the same limitations regarding the linearity of facial expression parameters.

Nonlinear spaces (manifolds)

It is now widely accepted in the research community that facial expressions involve nonlinear deformation patterns. As mentioned in section 2.1.1.3, nonlinear reduction techniques aim at extracting the nonlinear space of facial expressions. Finding the right structure for this space is still a research theme, but most studies focused on low-dimensional manifolds. Indeed, the advantage of non linear manifolds is that they encompass much richer information in fewer dimensions than linear structures. A reduced dimensionality reduces algorithmic complexity for classification systems [HCFT04], but it also facilitates comprehension and interpretation. Indeed, from a geometrical point of view, *a manifold is an abstraction.* Nonlinear analysis method are liberated from the geometrical constraints typically associated with linear schemes, and adapt to the data in a more empirical and natural way. A graphical illustration of this point is provided

in section 3.2.

The obtained manifolds represent a compact and consistent simplification of the facial expression space. Most studies use very low dimensionality (2-D or 3-D manifolds), which can be more easily analyzed and interpreted. Exploring those spaces generally allows researchers to identify some meaningful tendencies in the manifold dimensions. Roweis and Saul [RS00] identified two clear variation modes in the 2-D embedding of approximately 2000 small facial images: head pose change (yaw) and expression change (smile). Brand [Bra03] recognized pose-, expression- and scale-changes as the 3 degrees of freedom of their nonlinear embedding of Roweis' data. In the nonlinear dimensionality reduction of speech-related facial deformations, Pei and Zha [PZ07] and Saul and Roweis [SR03] exhibited modes such as lips protrusion-retraction and mouth opening in their IsoMap-based 2-D embedding. Facial data are frequently used, but meaningful manifold extraction techniques have obviously been applied to other problems. Tenenbaum *et al.* [TdSL00] and Weinberger and Saul [WS04], among others, detected low-dimensional manifold structures in various types of image collections such as objects with rotated point of view or handwritten digits with varying shape. Pless [Ple03] demonstrated the results of an IsoMap-based reduction technique on images showing hand configurations. The resulting 2-D space clearly isolates two meaningful actions: rotation of the wrist and opening-closing of the hand. ElGammal and Lee [EL07] identified and parameterized different arm-moving gestures on a 2-Dimensional manifold. The manifold results from a Locally Linear Embedding (LLE) on a database of images showing upper body movement.

Discussion

The pixel-intensity changes induced by the moving or deforming entities are highly nonlinear, yet they form easily-identifiable trajectories on low-dimensional manifolds. Consequently, high-level, semantic interpretations emerge from the simplicity of the representation, something which would not have been possible with linear schemes.

Expression-driven representations are less rigorously connected to semantic notions than emotion-driven ones, yet meaningful patterns can be derived *a posteriori* from their simple and intuitive structures. Additionally, their structure is topologically more consistent with the phenomenon they are supposed to represent (low-level facial deformations), since they originate from real data. Such a bottom-up approach is more likely to draw near the true space of emotional facial expressions.

In chapter 3, we present one of the contribution of this thesis: the construction of an expression-driven high-level representation of emotional facial expressions. The obtained representation benefits from the consistency of a data-driven method and the intuitive structure of dimensional emotion models. A specific manifold-fitting technique has been developed to cope with the particularities of facial data.

2.2 Temporal Aspect of Facial Expressions

As mentioned in the introduction chapter, the second part of our study focuses on the temporal aspect of facial expressions. Correctly representing and manipulating expressive facial deformations (section 2.1) is an essential ingredient of realistic expressions, but human beings are also very sensitive to facial dynamics. Indeed, studies have proved that humans are highly efficient at noticing temporal inconsistencies in human behavior, and that distorted dynamics produces a negative effect on human observers (see introduction chapter). The studies referenced in the first part of this chapter were mainly focused on the representation, and for some of them the synthesis, of expressive facial configurations. Yet, when it eventually comes to synthesizing facial animation, these deformation models need to express how their parameters behave over time, in order to guarantee natural-looking movements. For instance, the dynamic behavior of Facial Action Units (FAUs) can be the difference between a realistic and a synthetic-looking expression.

The important role of the dynamic factor has been well understood by pattern recognition specialists. In facial expressions recognition tasks, it has clearly been identified that facial movements display a valuable dynamic signature. While early systems were meant to recognize spatial configurations only, most of today's classifiers use temporal patterns in order to improve their recognition rates. Some recognition systems relied on artificial neural networks to do so [KH93], but, nowadays Hidden Markov Models [OO98, CGH00, CSG+03, AK06] or Dynamic Bayesian Networks [ZLGS03, SGM06] are the most popular description models when it comes to encode dynamic patterns of facial expressions. More recently, dynamic models have been also used to successfully track the activation of individual Facial Action Units [Val07, TLJ07].

Incorporating temporal signatures into classification systems reportedly boosted recognition performances [Pan09]. However, these recognition systems usually rely on purely descriptive models of facial dynamics. The models (whether HMM, DBN, or others) encompass the dynamic patterns learned from training data, and new, captured sequences can be matched with the models to recognize particular movements. The dynamic signature that carries the naturalness of the movement is not explicit in those models. They are not particularly adapted to the *generation* of new sequences of facial movements with realistic dynamics, which is what we focus on in this study.

The importance of correct timing in animation is now widely accepted, yet only a few studies have investigated this aspect in details so far. In the following, we discuss how this essential aspect has been handled in previous studies[7].

Historically, the handle of timing in animation was performed manually via keyframing techniques (section 2.2.1). Then, with the progress of imaging technologies and computing power, the capture of temporal sequences of motion became reliable and affordable. Although keyframing is still used today, most people thus turned to performance-

[7] The temporal realism of character animation is obviously not limited to the face. Body animation studies cope with a similar challenge, and have provided interesting solutions for the modeling of dynamic signatures. The discussion will therefore reference these contributions as well.

based animation (section 2.2.2), in which the motion data is used directly to drive virtual characters over time. Real sequences guarantee the naturalness of motion, but cannot be generalized to generate new animations. Therefore researchers have developed dynamic models (section2.2.3), able to learn the natural dynamics of human movements and can synthesize believable animations in varying contexts. Along with realism and adaptability, the notion of motion variability is essential, particularly in the case of interactive applications. This aspect and relevant studies will be presented in section 2.2.4.

2.2.1 Keyframing

Manual keyframing

Moving virtual characters first appeared in animated cartoons of the early 20^{th} century[8]. At that time, the characters were entirely animated with *manual* techniques: the animators had to draw each and every frame of a temporal sequence to create the illusion of natural movements. Most people are actually familiar with this process, as it is used to create flip books. Animation studios soon developed more efficient workflows with the use of the *keyframing* technique: A lead-animator first designed the structure of a sequence by drawing a sparse set of 'key poses' of the character over time, and assistant-animators (also known as inbetweeners) then took care of the intermediate drawings, between key poses, to complete the animation timeline.

Almost a hundred years later, keyframing remains a very popular animation technique, used in many recent productions. However the technique has evolved with the use of modern computers. Roughly, animators still need to design and place key poses on the timeline, but can now use computers to perform interpolation between keyframes. Automatic interpolation schemes have somewhat substituted the manual work of inbetweeners. Despite the help of computers, keyframing animation remains fundamentally a manual process, even a craft industry. Indeed, the determination of key poses and the choice of appropriate interpolation schemes require a unique mix of technical and artistic skills. Some guidelines can be derived from the experience of animators. Among others, Lasseter [Las87] put forward a few principles inherited from traditional animation that formalize some important realism and entertainment aspects. More recently, Sloan *et al.* investigated artistic practice of creating perceptually believable animation. In [SCR09] in particular, they evaluated the perceptual impact of movement and coordination of facial elements.

When handled correctly, keyframing animation is extremely flexible and powerful. In most case, it is the method that offers the best precision and versatility. For planned, precalculated animations, impressive results can be achieved, however the production of high-quality animation with the keyframing technique requires hours of tedious work and relies on the creative skills of experienced animators. This kind of workflow is not adapted to real-time applications. By definition, these applications deal with live

[8] *Gertie the Dinosaur*, created in 1914, is considered the first cartoon to feature a fictional character with an appealing personality. This makes it the predecessor to later popular cartoons such as those by Walt Disney. *Gertie the Dinosaur* also records as the first occurrence of keyframing animation. `http://en.wikipedia.org/wiki/Gertie_the_Dinosaur`

constraints and uncertain events, consequently animations cannot be manually tuned to produce the best-looking motion. Nevertheless, many interactive animation systems rely on real-time keyframe-like schemes. The particularity is that key poses and interpolations are automatically handled in real-time to comply with dynamically evolving situations.

Automatic keyframing

An intuitive and typical approach of 'automatic keyframing' animation consists of dynamically generating key poses (imposing an expression of happiness whenever some positive event occurs for instance), and using generic interpolation functions to blend between these poses. This mechanism is conceptually similar to the successful work of Cohen and Massaro for the synthesis of visual speech. In [CM93], Cohen and Massaro automatically placed visemes[9] along the timeline at the location of their corresponding phonemes. Phoneme-specific weighting functions accounted for the temporal influence of each viseme, and each frame the resulting facial configuration was calculated by blending all active visemes.

Equivalent strategies have been proposed for emotional facial expressions. The most trivial scheme consists of using simple linear interpolation between keyframes. This basic approach can be justified in systems with very low computing power, however linear movements are unnatural and look extremely robotic. More appropriate interpolation schemes can improve the visual quality, even for computationally-limited systems. In his seminal work, Parke used an animation system based on cosine interpolation [Par72]. Sigmoidal[10] temporal profiles, like cosine functions, have since been used extensively in real-world animation system. In spite of their simplicity, they bring a natural sense of smoothness during transitions from an expression to another. More elaborate spline-based interpolation scheme are often used too. They provide more control over temporal profiles, but basically rely on the same principle.

The above strategies are simple to understand, and computationally light, but they hardly reflect the reality of facial movements. In many situations, motion patterns are much more complex than simple abstract curves.

To provide adapted and believable animations, it is important to consider the mechanisms that motivate expression changes in the first place. In the case of emotional expression models of human emotional mechanisms can be helpful; particularly the *appraisal* model, (see section 2.1.2.1) which explicitly deals with temporal notions. Contrary to purely abstract interpolation functions, it characterizes a more rational model of the evolution of expressions over time. However the link between appraisal processes and expressions rely solely on high-level psychological assumptions. The more practical considerations of timing and temporal profile of expressions are not explicitly part of the appraisal theory.

ADSR profile

To bridge the gap between high-level emotional processes and actual facial animation, researchers have come up with empirical dynamic profiles of expressive move-

[9] A viseme is the representation of speech units (phonemes) in the visual domain. `http://en.wikipedia.org/wiki/Viseme`

[10] Sigmoids, sometimes refer to as S-shaped curves, illustrate a temporal behavior consisting of an initial acceleration, an inflexion at the point of maximal speed, and a final deceleration.

ments. Ekman and Friesen [EF78a] proposed to describe emotional expressions with a generic temporal envelope, the Attack-Decay-Sustain-Release (ADSR) profile[11]. The ADSR model, or variations of it, has been adopted in many high-level animation systems [KMMtT91, DRSV02, PB03, EKMt04, Tan06] to describe the amplitude of the whole emotional expression over time. The duration of each phase of the envelope is however adjusted empirically to produce natural-looking animation. In [PB03], Pelachaud and Bilvi relied on empirical observations of expression to set adapted temporal duration for each type of emotion (for instance the "sadness" expression has a long release time, since this emotion takes more time to disappear, while the "surprise" expression has a short attack). Some systems even adjust the envelope parameters dynamically, to adapt expressions to the current context. Tanguy [Tan06] relied on an ADSR dynamic model of human emotions. For each emotional event, the temporal envelope was dynamically adjusted using temporal filters, depending on the emotional context. Those temporal effects had a direct impact on the animation of corresponding facial expressions.

Handling the whole facial expression with one single motion pattern is a simplification that puts a strong limitation on realism. Even if strong correlations between facial elements are observed in emotional expressions, they are never perfectly synchronized in the real world. Moving facial elements with one single animation curve therefore produces a very robotic impression. Some studies preferred to rely on separate temporal behaviors for each facial element.

Onset-apex-offset profile

A widespread representation of localized facial movements is the Facial Action Coding System (FACS). However, as pointed out by Essa and Pentland [EP97], there is no time component of the description of movement in the FACS. Essa and Pentland thus defined the "FACS+" system which features a temporal profile for their Facial Action Units. From the study of Electromyography (EMG) signals, they derived that most facial actions present three distinct phases: *application*, *release* and *relaxation*. Yacoob and Davis also identified a recurrent temporal pattern in their own study of facial dynamics [YD94]. They described its parts as *beginning*, *epic* and *ending*, but it is now mostly referred to as the *onset-apex-offset* profile. This approach is very popular in expression recognition studies, in particular for the temporal modeling of individual Facial Actions [ZJ03, Val07].

A few studies have successfully translated this kind of temporal pattern to facial animation synthesis. Latta et al. proposed a framework designed to link emotional concepts with realistic expressive behaviors for a virtual agent or a robot. The individual facial actions were animated with an *onset-apex-offset* profile [LAAB02]. A consequence of working with local actions is the necessity to ultimately combine them to produce the final expression. Bui *et al.* [BHN04], broke down facial expressions into atomic facial movements, each of them characterizing the local action of facial muscles. These atomic actions were modeled temporally by individual *onset-apex-offset* values. The system

[11] *Attack* is the transition between the absence of expression and the maximum expression intensity, *Decay* is the transition between maximum intensity and a stabilized intensity, *Sustain* is the duration of the stabilized expression, and finally *Release* is the transition back to an absence of expression [KMMtT91].

then combined the local actions, using a fuzzy rule-based system for the resolution of possibly conflicting movements.

Discussion

Reproducing the complex phenomenon of human movement with such simple pattern as the one described above is not trivial. While very good results can be obtained with manually tuned keyframing, they are rarely convincing in real-time contexts, where animation cannot be finely manufactured beforehand. The preset interpolation patterns often fail at displaying the natural temporal signature of human expressions, and result in artificial-looking animation.

Making abstract, synthetic models look alive is a daunting task, but it was the only way to go with reduced computing power. With the help of today's computing power and tools, human motion can be efficiently measured on real subjects with sufficient precision and time resolution. Most people now turn to performance-based animation, which inherently provides the realism of human motion.

2.2.2 Performance-based approaches

A sure way to display realistic motion in animation is to use real motion data. The recent evolution of computing power and dedicated hardware has enabled the reliable capture of human motion, and its use to animate virtual characters. Motion capture, often abbreviated as *MoCap*, is now a very common process in modern CGI-films and video game productions. This technology has evolved into an industry, now involving standard tools, organized workflows, and trained MoCap specialists. High-quality visual productions rely massively on the motion capture technology[12] to make sure that every animation of a movie or a video game benefits from a realistic, genuine look.

Since movements are not explicitly designed, as they are in keyframing techniques, performance-based approaches require less expertise on pure animation and artistic skills. However, performance-based animation has to cope with other issues which make it a non-trivial process as well. An important point is the quality of the capture itself. MoCap sequences only benefit from the realism of human motion if the capture is accurate and produces consistent data. Post-processing on the data is often used in practice to filter noise and "clean" occasional inconsistencies. Historically, several types of professional capture systems have been used: optical-fiber systems, mechanical systems, electro-magnetic systems, and -probably the most popular today- camera-based optical systems. Each of them has advantages and disadvantages in terms of reliability, accuracy, complexity and cost [13]. Nowadays consumer applications can also benefit from lighter and less costly capture systems based on simple video (with or without markers). In addition of its practicality, it enables the use of motion capture in real-time applications. In this work, we rely on such a video-based capture to record facial motion data (see details in section 3.1.1).

[12] 12 hours of MoCap were gathered for *Fahrenheit* (Quantic Dream, 2005), and over 100 hours for *Pro Evolution Soccer 2011* (Konami, 2010).

[13] `http://en.wikipedia.org/wiki/Motion_capture`

Independently from the capture itself, the major issue of performance-based animation is the transfer of motion data to the animated entity. This process is also referred to as motion *retargeting* or *expression mapping* in the case of faces. This aspect matters because inter-personal morphological differences greatly influence body movements and facial expressions. Careful processing must be applied to the raw motion data to ensure that coherent movements are displayed by the target character. This particular issue is presented in more detail in section 4.2.1.

With an efficient retargeting system, animating virtual character with MoCap data guarantees realistic-looking movements and expressions. This solution is nowadays widely adopted for offline applications. The issue with motion capture, is that it delivers movement sequences that can hardly be modified. In many applications however, typically real-time and interactive applications, the animation needs to be *contextualized*: the movements of virtual characters are triggered and articulated depending on the real-time context, internal events or user input. Additionally the animation associated to these actions is subject to constraints from the environment (contact with other objects, timing constraints, *etc*). The corresponding movements thus need to be planned at runtime, and cannot be entirely choreographed beforehand. Raw motion data, on the other hand, is not flexible, and is thus only relevant in one specific situation.

To bring the realism of natural human motion to interactive applications, researchers have come up with systems that dynamically adapt MoCap data to real-time evolving constraints. Those systems produce contextualized animations either by concatenating existing motion data segments in time, or by blending motion segments together to create new ones. The most relevant contributions of those two approaches are referenced in the following paragraphs, and a more exhaustive bibliography on the topic can be found in [Gle08].

2.2.2.1 Concatenation of motion data

Concatenation-based animation systems represent the most straightforward way to conciliate the requirements of real-time applications and MoCap animation. The principle is to record a large database of motion sequences, and ensure that for each situation likely to occur in the application, a corresponding animation sequence can be retrieved from the database and played in real-time.

The first basic systems of this kind looked to play whole sequences of recorded animation when a real-time event requested a particular response from a virtual character. Most of the time, the association between events and animations, as well as the transitions between the different sequences themselves, were parameterized manually. If the sequences were too long, the animation system lacked reactivity and controllability, yet on the other hand using too short sequences resulted in substantial manual work. Data-driven animation studies have since introduced automatic processing of motion databases. The animation systems are now able to break a raw database of MoCap data into small pieces, often called *motion segments*, and automatically assemble them in the context of a real-time application.

This technique was used within the video rewrite framework[14] to create lip synchronization animation for unknown audio tracks. In [BCS97], Bregler *et al.* recovered the consecutive visemes associated with the detected phoneme-string, and assembled them in time using image-morphing. For non-speech animation however, the task of selecting the right motion segments is not as simple, since no fine-grain animation guideline, such as a phoneme list, is provided.

Motion graphs

The main challenge of concatenation-based approaches is the construction of an efficient data structure to store motion data, and allow easy retrieval and combination of appropriate motion segments. In the last decade, one specific representation has emerged as the dominant form of concatenation-based animation systems: *motion graphs* [KGP02, AF02, LCR+02]. The motion graph approach consists of preprocessing raw motion databases to automatically organize them as a mathematical graph. At runtime, the graph structure is used to efficiently look for motion segments that meet the real-time constraints and goals. Most of the time, the structure is the following: the edges of the graph represent motion segments extracted from the database, while the nodes represent possible transitions between those segments. Some studies have considered the problem the other way around, with node representing segments and edges representing potential transitions [AF02]. In both case the principle remains the same, possible transitions between motion segments are automatically identified and allow the system to create new animation sequences by assembling compatible motion segments. The length of motion segments considered in motion graphs is not trivial: it is a compromise between the flexibility of the resulting animations, and the processing complexity of the graph. In some case, the granularity of motion segments goes down to a single frame [ZSCS04].

Edges are created by comparing the poses at the edges of motion segments with an adapted distance metric. If the distance measured on segments is below a defined threshold, natural transition between those segments is possible. Practical systems measure the distance on joint angles and velocities of edge poses [LCR+02], and sometimes even joints acceleration [AF02]. Direct comparison of raw vertex coordinates is also possible [KGP02]. In the case of cyclic movements such as walking, running or swimming, the motion segments can be compared based on the composition of their frequency spectrum [PB02]. Apart from graph construction issues, the main differences between motion graphs approaches lie in the real-time retrieval and combination of motion segments.

At the real-time animation stage, the retrieval of appropriate motion segments consists of searching a path in the graph that meets the application requirements. Several search strategies can be considered. Arikan and Forsyth [AF02] used global randomized search on a hierarchical abstraction of the motion graph to extract an animation satisfying the constraints. For more time-efficiency, local search strategies have also been proposed. Kovar *et al.* [KGP02] used dynamic programming, namely the branch and bound algorithm, to explore multiple paths simultaneously and quickly discard

[14] Video rewrite consists of assembling video frames or groups of video frames from a video database to create new visual sequences.

non-optimal solutions. Lee *et al.* [LCR+02] relied on a higher-level layer, the cluster trees, which encode local transitions possibilities between groups of frames. The trees were used as a more efficient structure for motion retrieval, given application constraints. Retrieval identifies motion segments which are likely to produce a natural-looking animation, however the edges of the motion segments rarely correspond exactly. Motion-graphs systems adapt the motion signals in real-time to produce smooth transitions. Smoothness can be achieved by merging successive motion segments over a short overlapping time interval using linear interpolation [KGP02]. Some researcher have also proposed forcing endpoints of consecutive motion segments to match, by directly editing values of the motion signals [LCR+02, PB02].

Face graphs

The efficiency of motion graphs has first been demonstrated on full-body movements. Later, Zhang *et al.* [ZSCS04] introduced face graphs, the adaptation of motion graphs to facial animation. In practice, they formed a graph by connecting the frames of a motion database, with edge weights representing transition likelihood. From user- or application-defined constraints (facial expressions to display at specific times) their systems generated a continuous animation by constructing a minimum-cost walk on the graph.

Apart from the popular motion graphs, other structures have been proposed for concatenation-based animation systems. Some of them took inspiration from texture synthesis methods [PP97, EL99, SSSE00] to generate original motion sequences that statistically resembles motion examples for a training database [SBS02]. Deng *et al.* [DBNN04] introduced such an approach for facial animation, by using a patch-based random sampling algorithm inspired by [LLX+01]. At fixed time intervals, candidate motion segments are selected from a database and inserted in real-time following a probabilistic law, to create novel facial motion.

Discussion

Concatenation-based animation systems are very successful in practical applications and the academic world, and continue to be improved today [ZNKS09]. In many studies however, the focus has somewhat shifted toward the handling of higher level task such as motion planning [TLP07, KS08]. Motion graphs transform the animation synthesis problem into a discrete graph search problem, which significantly speeds up the motion generation process. However, a major drawback is that the resulting animation is necessarily limited to motions present in the database.

2.2.2.2 Motion data editing

A drawback of concatenation-based animation systems, is that they need a recording of every single movement that is to be displayed in the final application. This becomes problematic when a system must account for many varied situations, and still provides fine-grain control. A system, animating a jumping body for instance, must theoretically possess a sequence of every jumping styles of the character. Similarly different sequences are needed for every height the character eventually needs to jump to, since the movement is not exactly the same. In terms of capture cost and storage space, the pure concatenation-based approach is clearly not optimal. Instead, other

schemes proposed to actually edit motion capture data to adapt to different situations than the one for which the data was recorded. These schemes are often combined with concatenation-based frameworks for finer and more efficient control of animation.

Motion blending

MoCap data generally take the form of a massive amount of high-dimensional vectors, each representing a facial expression or a body configuration. Directly editing the values of these vectors to make meaningful modifications of the corresponding motion is almost impossible in general. Most motion editing methods actually work by example. Different variations of a motion are recorded and used to create a fine, continuous parameterization of that motion. In practice, this approach takes its roots in *motion blending*. Motion blending consists of mixing existing sequences of motion data to create new ones. Motion blending in practical systems is done by interpolating between motion sequences that represent the same action, but may vary in terms of spatial trajectory, speed or style. Cleverly adjusting the interpolation weights allows animators to create motion sequences that satisfy a whole new range of constraints, and more precisely adapt to different situations. To a certain extent, motion blending can be thought as a parameterization that enables generalization of motion data.

Motion blending is an idea that was used empirically for a long time. Rose *et al.* [RCB98] first provided a formalization of this tool, and its application to fine animation control. They classified motion sequences using a verb/adverb formalism: a verb symbolize the action that is being done (such a walk, jump, *etc.*), while adverbs represent the different variations that were recorded for one verb. These variations can be of multiple origins: speed, emotion or influence of the environment. An infinite range of new motion sequences could be generated for a verb by mixing the adverbs together.

Interpolation is not the only way to create new sequence. In some cases, transformations can be applied directly on motion data [EM04]. Simple operations like translation, rotation or scaling on trajectories coordinates help adjust the motion for different spatial configurations. The start- and end-points of a jump can be moved, and the motion trajectories warped accordingly to adapt the size of the jump to application requirements. These transformations however are very limited, and must be carefully handled since incoherencies can appear for large modifications.

Editing the motion signals directly (vertex positions or joint angles) is not particularly recommended since they show significant correlation in natural human motion. For instance, movement of the skin on the cheeks is strongly correlated to the lip-corner in the case of a smile; so each of these components should not be edited separately. Most edition systems rely on procedure like PCA to work in a reduced configuration space, in which correlations between motion signals are accounted for, and where motion blending is relevant [SBS02, CDB02, DBNN04, EM04]. Chuang *et al.* [CDB02] learned a bilinear decomposition of PCA space for expressive faces, in which they separate content (speech related facial animation) and style (behavior or emotional state). This efficient model allowed to turn any speech animation sequences into a happy speech sequence, an angry speech sequence, or into a little bit of both with an interpolation of the style parameters. Similar approaches have been proposed with more elaborate configuration

spaces, typically using nonlinear dimensionality reduction [WHL+04, EL04, GMHP04]. Wang *et al.* [WHL+04] performed a decomposition of expressive facial animations into style and content, and relied on Locally Linear Embedding (LLE) to construct a very low-dimensional configuration space.

Motion alignment

As far as the interpolation is concerned, simple methods such as linear interpolation are commonly used. For larger examples databases, more specialized methods can be used to merge only the most relevant movements. Among others, RBF networks [RCB98, PSS02] and k Nearest Neighbors interpolation [WH97, KG03] are used. The actual concern when merging motion segments, is to make sure that they are structurally similar. They must represent compatible actions and also be coherent temporally. Blending asynchronous movements generally produces bad results and temporal artifacts. Ideally, sequences must be correctly "aligned" in the time-domain before being interpolated. When working with atomic facial expressions, researchers synchronize the beginning and the end of expressions, and a simple scaling of time is applied [WHL+04]. It is more complicated for body movements, for which some intermediate constraints need to be respected (for example the feet that touch the floor in a walk sequence). Real "timewarping" techniques have been extensively used in motion editing systems, to ensure the temporal alignment of motion segments. Early systems performed semi-automatic time alignment based on user specification of some relevant timepoints [RCB98, PSS02]. Later, automatic alignment schemes based on Dynamic TimeWarping (DTW) were proposed, and successfully associated to motion merging techniques [BW95, KG03].

Modern motion editing techniques provide a rather straightforward parameterization of motion capture data, which enables interactive applications to modify motion signals and adapt them to real-time situations. They are mostly used in full-body animation to adapt very specific, short-term gestures to context, such as the parameterization of a grab movement depending on the position of the object to grab, or the modification of a run sequences depending on how fast the character has to run [PSS02]. To provide both long-term coherence and fine animation control, top-of-the-range systems nowadays combine concatenation-based techniques and motion editing techniques into fully automatic animation systems [KG03, KG04]. For a specific runtime situation, a sequence of relevant motion segments can be retrieved from a natural motion database, and finely adapted to the present constraints using motion editing techniques.

Discussion

Performance-based animation has proved its efficiency, and is today's choice for top-of-the-range productions, yet this approach suffers from certain limitations. Motion editing techniques provide a better adaptation of motion capture data to varied situations, but in some cases motion can hardly be interpolated or extrapolated, and using editing techniques causes to lose the naturalness of human motion. Additionally, despite the precision they add to concatenation-based systems, they are still limited to combinations of database motions. Consequently those systems cannot adapt to unknown situations, in which the required movements are outside the scope of the motion database. To ensure diversity and adaptation to diverse real-time contexts, those

systems require a large database of varied movements, leading to high capture costs. Additionally, the database needs to be available at runtime which imposes important memory requirements

2.2.3 Dynamic Models

With sufficient time and resources, very convincing results can be obtained from keyframing and performance-based systems for planned, precalculated facial movements. They are however not optimal in the context of interactive applications: keyframing approaches often lack realism, while performance-based systems impose high requirements, in terms of cost and storage space, for a limited adaptability. Both approaches are not successful at adapting natural human motion signature to the various situations an interactive application can find itself in. Performance-based systems actually do hold this dynamic signature intrinsically in the stored motion data, but they do not *explicitly* model it. This prevents them from generalizing this natural motion to new, unknown situations.

Instead of relying on recorded motion data, some researchers have proposed animation systems that truly synthesize motion signals. They focus on modeling and regenerating the natural dynamics of human movements with the help of mathematical models, which is why we refer to them as *dynamic models*. Obviously, they can hardly outperform captured motion signals on realism, but on the other hand they benefit from a greater flexibility and adaptability. Indeed, since the mathematical formulation explicitly models motion dynamics, it can theoretically be applied to any situation, which is valuable in interactive applications.

Dynamic models can arguably be segmented into physics-based models (section 2.2.3.1), and learning-based models (section 2.2.3.2). In a nutshell, physics-based models produce new animations by simulating the internal mechanical phenomena at the origin of motion, while learning-based models aim at simulating only the visible effects of motion with more computationally-efficient systems. Important contributions of these two areas of research are presented in section 2.2.3.1, and section 2.2.3.2 respectively.

2.2.3.1 Physics-based Dynamic Models

The parts of the human body that are responsible for its movements can be seen as mechanical systems, and as such their behavior is fundamentally described by the universal laws of physics. The natural dynamic signature of human motions, and facial expressions in particular, is therefore a consequence of the structure and physical properties of body mechanics. A perfect modeling of the physical system's components would theoretical enable the prediction of physically-valid motion signals, provided that all influential factors, whether internal (nervous signal activating the muscles) or external (contact with other entities), are known. In this regard, researchers have proposed to recreate human movements through simulation of the mechanical phenomena at the origin of motion.

From an animation point of view, the advantage of such systems is their flexibil-

ity: physics-based models can theoretically adapt to any requirements and constraints within the humanly possible, and emulate the natural reactions of a human body. This is illustrated in the work of Witkin [WK88]. His animation system looks to fulfill *spacetime*-requirements given by an animator, while respecting constraints imposed by the physical model of a body. This formulation constitutes a constraint optimization problem, which is solved numerically to provide an optimal, physically-valid animation.

The physics of human body is far too complicated to be modeled exactly, but its behavior can be approximated by a few mathematical equations, mostly with forces and differential equations from the classical mechanics framework. More precisely, systems typically model the elements of the human body (muscles, articulations, *etc.*) using standard mechanical components, such as joints or springs, whose dynamic behavior is easily computed. The resulting behavior of the global system is obtained by applying Newton's laws or Lagrange's equations of motion to the global systems.

The physical simulation itself is an automated process, but its outcome still depends on the parameters that describe the mechanical elements of the system. These parameters typically consist of masses, friction coefficients for the joints or spring stiffness coefficients. It has often been reported that these parameters significantly impact the resulting motion, and thus its realism. Yet, defining their correct values is not trivial because they influence motion only indirectly. Liu *et al.* [LHP05] also relied on those parameters to perform physical simulation in a spacetime optimization framework, and in addition proposed a method to automatically determine their values. The values of the mechanical parameters, such as joint stiffness representing muscle/tendon elastic properties, were estimated from MoCap data via an optimization process.

Purely physics-based models continue to be investigated in today's research. Recent contributions still rely on the same principles, but have improved the formulation of the optimization problem for a better integration of real-time external events and animation control [JYL09]. Most of the research effort in that area has been devoted to full-body movements. The main drawback is their non-universality, in that they are generally meant to synthesize motion signals of only one type of movement, typically human locomotion [LHP05, SCCH09].

Interesting physical models for human faces have been proposed as well [SNF05, NPP+08]. As a matter of fact many physics-based formulations of the facial anatomy have been presented, often referred to as muscle-based animation systems (see section 2.1.1.1). Yet, those systems use physics-based models of facial structure mostly to obtain very accurate simulations of facial deformations. An explicit focus on the *temporal* aspect of facial animation is missing from those studies, mainly because the emphasis is put on simulation and not on *control*. Indeed, the reaction of facial tissues to muscle activation is precisely described, but not how these activations need to be managed for a realistic animation of facial expressions. The control aspect, that has to do with muscle contraction dynamics [Zaj89] and coordination of the different muscles, is unfortunately usually overlooked.

A few studies did account for the role of muscle contraction dynamics in facial animation. Sifakis *et al.* [SNF05] exhibited an animation system based on a very complete

anatomical model of skull, musculature and skin tissues. A finite element method is used to simulate the physical contractions of muscles, which in turn deform the skin tissues of the face to produce expressions. Sifakis *et al.* relied on a volumetric simulation technique of skeletal muscle behavior introduced in [TBTHF03]. Nazari *et al.* [NPP+08] more recently proposed a similar animation system. The accurate computations used in those studies reportedly improve the quality of facial motion simulation. As far as control of animation is concerned, muscle contraction is driven by muscle activation values, which need to be determined. So far animation systems have relied on captured facial motion data, from which they extract appropriate activation patterns [SNF05]. To our knowledge no generic model has been proposed to synthesize activation signals for the generation of facial animation.

Discussion

Physics-based dynamic models provide a good framework for accurate and flexible animation systems. However, the dynamics of the human body is a complicated phenomenon. Simplified models usually target only specific situations, and not always produce believable movements. This is especially true for faces, for which the interactions of skin, muscles and bones are particularly crucial and complex. It appears that the autonomous control of physics-based systems for interactive facial animation is a too complicated problem at the current stage of research.

2.2.3.2 Learning-based Dynamic Models

Physics-based models theoretically give the best results, since they look to be as close as possible to the reality. Unfortunately, due to the complexity and lack of knowledge of some aspects of the human biomechanics, the models rarely display perfectly realistic movements. Additionally the more accurate ones rely on intensive computations and are usually not meant for real-time applications.

The purpose of learning-based dynamic models is to be able to generate realistic animation using simpler and more time-efficient computational models. Their goal is conceptually similar to what pseudo-muscle models looked to accomplish for the simulation of skin deformations (section 2.1.1.1): reproduce only the *visible* effects of the physical phenomenon, without having to handle the complexity of the underlying biomechanical simulation. In this regard, they are sometimes referred to as "black-box" models. As their name imply, learning-based dynamic model look to *learn* how the visible facial elements behave in time when displaying natural facial expressions, and reproduce this behavior on virtual character given the requirements of interactive applications.

Dynamic models generally consist of a set of mathematical equations or functions that describe the visible dynamic behavior, directly in terms of vertex coordinates or joints angle values. During a training phase, the models' parameters are adjusted so that the behavior matches the dynamic patterns observed on real motion sequences. Here the parameters directly influence the animation output, which makes parameter learning easier than in the case of physics-based systems. Once trained, dynamic models can be used to generate natural motion signals that adapt to real-time situations. Typically,

the systems are asked to interpolate real-time generated keyframes while retaining the learned dynamic patterns.

Trajectory in state-space

Learning visual patterns from examples is not new. In section 2.1.1.3, we described parameter spaces such as PCA, in which natural facial deformations can be manipulated more easily. Each point of those spaces represents a facial configuration, which explains why they are sometimes called *state-space*. Some studies treat the generation of movements as a trajectory synthesis problem in the state-space [EGP02, BCT02, LE06]. Ezzat *et al.* [EGP02] used such an approach to synthesize mouth movements synchronized with an audio input. The motion sequence is generated by interpolating the mouth configurations corresponding to the successive visemes. The use of a consistent data-driven state-space (morphable model parameters) ensures the coherence of the transitions, but the temporal aspect is not explicitly treated. Govokhina *et al.* proposed a more precise temporal control of the trajectory with the use of HMM synthesis [GBBB06]. The HMM trajectories were learned and synthesized in a state-space of articulatory parameters to ensure a good compliance with the actual phonemes. Motion capture segments were then recovered based on the obtained trajectory and concatenated to display the resulting animation.

Such interpolation techniques yield compelling results for visual speech, because phoneme-lists provide a very fine sampling of the timeline (typically 200ms between successive phonemes), whereas emotional expressions have much less guidance (transition between emotional expressions can go up to one second). In their study Bettinger *et al.* [BCT02] used trajectory synthesis to generate short motion segments of expressive facial animation. Their system interpolates coherent expressive configurations, but the animated face often displays a visually unnatural dynamics. In the case of emotional expressions the understanding and correct modeling of facial dynamics is essential.

Linear dynamic systems

More recent approaches of motion modeling usually use a model with an explicit notion of temporal dynamics. Several systems use simple Linear Dynamic Systems (LDSs) as a generic motion model [GCT+04, DM08]. LDS efficiently model systems for which the temporal evolution of an entity linearly depends on the previous configurations. It was observed however, that realistic human motion acts for the most part as a nonlinear system. Some studies argued that this nonlinear behavior could be approximated in a piecewise linear fashion [PRM00, LWS02]. Such models are often called Switching-LDS (SLDS) since they decompose the global motion model into local linear systems. Li *et al.* [LWS02] developed a SLDS in which they modeled short movement units by LDS and connected these units with a transition matrix accounting for the long-term motion. A new sequence is generated simply by selecting a sequence of movement units corresponding to the application's need and by sampling the corresponding dynamics systems frame by frame.

Chai and Hodgins [CH07] formalized more precisely the relationship between the motion model and the application's time constraints to produce the desired movement. The motion model, derived from motion capture data, acts as a motion prior. Together with the constraints, it forms a maximum a posteriori estimation problem which is

solved by a nonlinear optimization procedure. The motion priors are modeled by a LDS, but sufficient and carefully handled constraints manage to reproduce flexible and natural results.

Nonlinear dynamic systems

The nonlinear nature of human motion somewhat forced researchers to use more empirical approaches, which explains the diversity of investigated models. Brand *et al.* [BH00] represented motion segments as a sequence of discrete states using Hidden Markov Models. This limits the model to specific movements, but their model supports the representation of stylistic differences in those movements. Many artificial neural network architectures have been proposed to encode the dynamic signature of motion for recognition tasks. However, some have been used for synthesis as well [THR06]. Min *et al.* [MCC09] decomposed the motion synthesis problem into two components: the successive spatial configurations on one side, and the temporal sampling on the other. In practice, the spatial model consists of a trajectory synthesis in state-space, and the temporal model is represented by trained timewarping functions. While this approach can be justified to model one stereotypical type of movement, in the general case spatial and temporal aspects are interdependent, and can hardly be treated separately. Another interesting nonlinear approach has been proposed by Wang *et al.* [WFH08], who used a dynamic model inspired by recent advances on Gaussian processes. The resulting model is nonparametric: it does not rely on only one specific, potentially incorrect, nonlinear system form, but instead accounts for a wide range of nonlinear behaviors. This type of non-parametric nonlinear models have received a lot of attention lately in the field of motion synthesis [LSZ09, YL10].

Discussion

From the theoretical standpoint, learning-based dynamic models offer the best compromise between realistic and flexible motion generation. They learn the right temporal dynamics from motion data, but do not explicitly reuse this data once they have been trained, in opposition to performance-based systems. The trained models are represented by only a few parameters, which is clearly preferable in terms of memory occupation. They are meant to generalize the patterns of training motion signals to comply with new constraints (typically keyframes), but are computationally more efficient than physics-based models. This makes them more suited for real-time applications.

A drawback of the learning-based approach is that the complex nonlinearities of human motion cannot be represented by a single generic model. In practice, learning-based models are designed for one specific type of movement, such as walk or jump. In this regard, they have the same range as motion editing techniques (section 2.2.2.2), with the difference that motion data does not need to be available at runtime. As far as facial animation is concerned, the models have been largely inspired by full-body animation studies; Similarly, they aim at modeling only specific expressions, such as happiness or surprise [CH07, DM08].

In this work, we introduce a new type of dynamic model, specifically developed to control the temporal dynamics of emotional facial expressions. The model is a black-box model trained on a database of facial movements in order to learn the dynamic

signature of human facial expressions. Contrary to many dynamic models, it is meant to be generic, and can generalize the learned dynamic signature to any desired emotional facial expression, including those outside of the scope of the database. Details on this system can be found in chapter 5.

2.2.4 Motion Variability

An important characteristic of human motion is its uniqueness. Indeed, many researchers have observed that human bodies and faces never show the exact same motion twice, even when looking to repeat the same action. When analyzed more precisely, motion signals do display variations, which we refer to as "motion variability" in this work. This variability is an unconscious phenomenon but it is inherent to natural human motion, and we believe it to be an important component of the richness and the expressiveness of human behavior; this component is however largely overlooked, even in modern animation systems.

Motion variability is a particularly crucial point in the context of real-time interaction, for which the virtual characters can be lead to reproduce similar actions multiple times. Identifying repetitive animations immediately emphasizes the artificial nature to human users, no matter how realistic the animations and the character look like. Some applications such as embodied conversational agents are even more demanding as they imply medium to long-term interactions with the user. Today's virtual agents are often criticized for being repetitive, predictable, and this clearly impairs the user's interest for the agent. To prevent the agents from becoming monotonous and dull, we think that not only do the animations have to be realistic, they also need to show variability.

Additional noise

In previous works, the variability in human motion has traditionally been modeled independently from the motion itself. The generic idea was to add some kind of stochastic noise to the deterministic motion signals. Perlin [Per95] used noise signals with specific spectra to drive the joints and simulate a lifelike unconscious behavior. Byun and Badler [BB02] later extended this approach by perturbing MPEG-4 FAPs streams using parametric modulation functions. A user could modify the look of an animation by acting on a few subjective parameters. The variety was seen as a user-editable feature, and not a part of the motion generation mechanism. These approaches are rather straightforward and computationally efficient. However they rely on generic noise functions, and are therefore not dynamically consistent with the eventual conscious movements.

Modeling variability

In reality, variability is not an external factor but truly belongs to the motion generation process. For realistic variability results, it should be treated as a core part of the character animation system. For dynamic motion models, variability can be seen as a perturbation of the trajectory in state-space[15]. In practice, two main aspects influence

[15]Parameter space describing the spatial configuration of the body or the face. State-space often consists of parameter space obtained by PCA or other dimensionality reduction techniques.

this trajectory: the coordination between the different moving elements of the entity (face or body), and the own temporal dynamics of each facial elements.

In some studies, variations have been handled directly within the trajectory synthesis operation. Bettinger *et al.* [BCT02] generated new sequences of facial animation by constructing a statistical model of trajectories in an AAM state-space. Assuming a Gaussian distribution, they sample this model to obtain new variations of a movement. This generation technique does not take the temporal sampling of trajectories into account, and can result in unnatural dynamics. Min *et al.* [MCC09] modeled dynamic body animation by separating the aspects of spatial variations (trajectory synthesis) and temporal variations (timewarping of the trajectory). For a specific movement with rather stable coordination of the limbs both aspects can be treated separately; yet for more general cases they strongly interfere and this segmentation is not justified anymore.

Lau *et al.* [LZK09] used a more general probabilistic framework, in which they model all causes of variability in state-space at once. They relied on Dynamic Bayesian Networks to synthesize new variants of movements that are statistically similar to example motion, but vary in terms of spatial poses and timing of the movements. This type of modeling is more robust and efficient, however due to the complexity of motion variability in state-space it is meant for only one type of motion.

Discussion

Most facial motion models, inspired by body motion models, treat the face globally as one dynamic entity. This makes the modeling of temporal dynamics and motion variability very complex, as the facial behavior in a global state-space is highly nonlinear. The dynamic model of facial animation we present in chapter 5 proposes to dynamically control different parts of the face with separate dynamic models. This allows the system to animate *any* expression that can be decomposed into movements of these facial parts. Additionally, these innovative motion models lend themselves particularly well to motion variability, which is treated as a stochastic component within the motion generation process of each facial element. Stochastic variations are thus coherent with the generated motion dynamics, and their properties can still be learned from training data. Besides, our results partly validate conclusions of biomedical research on motion variability [HW98]. More details on how we propose to handle motion variability can be found in section 5.2.2.2.

Chapter 3

A Data-driven Meaningful Representation of Facial Expressions

Contents

3.1 Facial Appearance Space 72
3.1.1 Active Appearance Models 73
3.1.2 Facial Database 74
3.1.3 Human Facial Appearance Space 75
3.2 Facial Expression Manifold 77
3.2.1 Dominant Directions 78
3.2.2 Facial Expressions Manifold Extraction 79
3.2.2.1 Dominant Direction Detection 80
3.2.2.2 Expression Space Embedding 81
3.2.2.3 Projection on the Manifold 84
3.2.3 Expression Manifold as a Visual Representation 85
3.2.3.1 Dimensionality 86
3.2.3.2 Representation Results 88

The purpose of this chapter is to introduce a new representation of expressive facial deformations. As mentioned in the introduction, human facial expressions are a phenomenon caused by the contraction of numerous muscles whose interaction produces a complex facial deformation pattern. Although we do not fully understand the physical mechanism of facial expressions, as humans we are naturally capable of recognizing them and give them a meaning. From a computational perspective, it is an interesting challenge to find a way to encapsulate this complexity, and provide applications with a tool to comprehend and interpret facial expressions more easily.

Understanding a phenomenon has a lot to do with how we represent it. Indeed, representations structure the formal knowledge as well as the qualitative and quantitative information we have about a system. In that sense, our work looks to generate a consistent and meaningful representation of emotional facial expressions that enables the manipulation of this important communication channel. This can be particularly interesting for improved human-machine interaction or affective computing applications.

Our method does rely on one low-level scheme, as introduced in section 2.1.1, as an intermediate step to efficiently measure facial deformations, but it ultimately aims at producing a high-level representation of emotional facial expressions. Contrary to most approaches[1], we do not use a representation known *a priori*, but look to make the representation arise from real data. This can be classified as a bottom-up approach[1]. The resulting representation is consistent with low-level deformations, yet it has a simple and intuitive structure that can be linked to semantic notions.

The emergence of the representation will be described in the sections of this chapter. In section 3.1, we describe the capture of facial expression data, and the construction of the facial appearance parameter that account for low-level expressive facial deformations. Afterward, section 3.2 depicts how a simpler, high-level representation can be extracted from facial expression data of the appearance space.

3.1 Facial Appearance Space

The method we present to extract a representation of facial expressions is described as data-driven. As such they obviously imply the capture and recording of expressive facial deformation data. As investigated in section 2.1.1.3, low-level facial data can be represented by different parameter spaces. In many studies, the data consist of frontal images of expressive faces. This textural description is not ideal since pixel values form rather massive databases, and carry significant redundancy. Other studies rely solely on geometrical information (location of the main facial features like the eyes, the mouth, etc.). Although more compact, this description leaves out some important expressive aspects such as skin deformation, wrinkles or dimples.

We choose to describe the facial expressions using both geometrical and textural information of the face, based on the modeling procedure of Active Appearance Models (AAM). AAM propose a generic method to model the appearance of deformable entities, and they have been applied on several occasions to the modeling of facial identity [BV99] and facial expressions [CB02]). Geometrical and textural variations are modeled jointly and described by so-called appearance parameters. Put together, facial appearance parameters form what we call the facial appearance space. In the next section, we provide a description of the AAM modeling stage used in this thesis.

[1]Refer to the state-of-the-art in section 2.1.2 for more information.

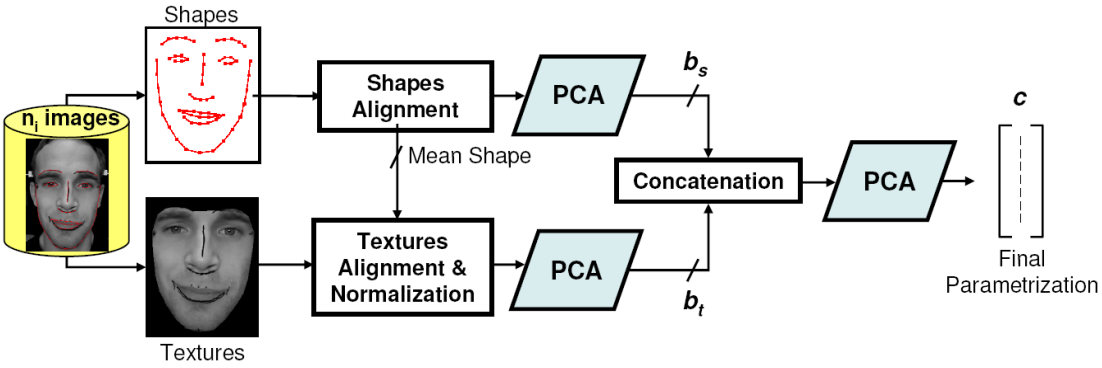

Figure 3.1: The AAM Modeling stage.

3.1.1 Active Appearance Models

Active Appearance Models in their current form have been introduced as a computer vision scheme by Cootes in 1998 [CET98]. They include a modeling step that leads to the parameterization of the observable appearance modifications of a deformable object. The purpose is to facilitate the tracking of a similar object in an image by the use of the obtained parameterization. In this work, we focus on the modeling phase (see figure 3.1).

The modeling procedure starts with a database of n_i images of the considered object, displaying the appearance modifications the object can be subject to. Several remarkable points are annotated on the object for each image of the database. Together they form the object's *shape*. The pixel intensities contained in the area spanned by the shape is called the *texture*. In the database, the object is presented with varying shapes and textures. The role of the model is to identify the principal variation eigenmodes of the shape and the texture of the object when deformations occur. These variation modes then serve as an efficient parameter set to describe the object's appearance changes on the images of the database. To detect the variation modes, the modeling uses Principal Component Analysis (PCA), for both the shape and the texture. The shape of the i^{th} element of the database is a collection of N_s-Dimensional points ($N_s = 2$ or 3), and its texture is generally a collection of pixel values, but both can be treated as vectors $\mathbf{s_i}$ and $\mathbf{t_i}$ and feed the PCA routine:

$$\mathbf{s_i} = \bar{\mathbf{s}} + \mathbf{\Phi_s}.\mathbf{b_i^s} \tag{3.1}$$

$$\mathbf{t_i} = \bar{\mathbf{t}} + \mathbf{\Phi_t}.\mathbf{b_i^t} \tag{3.2}$$

Where $\bar{\mathbf{s}}$ and $\bar{\mathbf{t}}$ are the database mean shape and texture, $\mathbf{\Phi_s}$ and $\mathbf{\Phi_t}$ the matrices formed by the PCA eigenvectors, and $\mathbf{b_i^s}$ and $\mathbf{b_i^t}$ are the decomposition of $\mathbf{s_i}$ and $\mathbf{t_i}$ on the identified eigenmodes. Note that before being compared to each-other, the shapes and textures undergo an *alignment* procedure. In the case of shape, the procedure removes translational, rotational and scaling differences that the shape might have to ensure that only non-rigid deformation of the entity are modeled by PCA. Textures

go through a similar alignment step, in which they are warped to a common shape (usually the mean shape). Additionally, texture coefficients go through a photometric normalization to remove the influence of lighting changes.

A third PCA is performed on the mixed vector $\mathbf{b_i} = \left[w_s.\mathbf{b_i^s} \mid \mathbf{b_i^t} \right]$:

$$\mathbf{b_i} = \mathbf{\Phi}.\mathbf{c_i} \tag{3.3}$$

Where $\mathbf{\Phi}$ is the matrix formed by the eigenvectors. Its role is to identify the correlations between shape variation $\mathbf{b_i^s}$ and texture variation $\mathbf{b_i^t}$ and take advantage of them to reduce the size of the final parameter vector $\mathbf{c_i}$. w_s is a scaling factor that ensures shape and texture parameters have comparable variances. The vector $\mathbf{c_i}$ represents the final parameterization of the i^{th} element of the database. It contains the contribution of the identified eigenmodes of both shape and texture for this element. The variation modes $\mathbf{\Phi_s}$, $\mathbf{\Phi_t}$ and $\mathbf{\Phi}$ are then used to depict any appearance change of the modeled object within the scope of the database.

3.1.2 Facial Database

The entry point of the AAM modeling system is an image database. The content of the database is important, because it reflects the visual variations that the AAM will be able to parameterize. In this work, it is formed by the frames of a video of an actor performing facial expressions without rigid head motion. The database was constructed to contain a total of 4365 images of natural expressions, both extreme and subtle, stereotypical of one emotion, or mixed. However, it is focused on the expressions of only one individual, making it a *person-specific* database. Indeed, it has repeatedly been observed that facial expressions can be visually very variable from a person to another. They are influenced by morphology, facial biomechanics, and even psychological factors like personality.

Methods have been proposed in past studies to deal with individual differences. Low-level schemes like multi-level AAM fitting are able to capture person-independent facial deformations [PBMD07]. At a higher level, approaches like multilinear analysis [VT02] [EL04] [LE05] or manifold alignment techniques [CHT03] [WHL+04] [CVTV05][SGM05a] have tried to isolate generic components of human faces that would universally describe expressive deformations. At the current stage of research however, we feel that no satisfying unification has been proposed. Additionally in many multi-individual studies, we observe that individual differences tend to hinder personal subtleties in the models. Here, we try to avoid this, since we believe that these subtleties are part of human expressiveness. Consequently, the following sections deal primarily with person-specific data. We set aside the problem of individual differences, and focus on the challenge of faithfully representing the facial expressiveness of one person.

Another aspect worth mentioning is that the captured expressions forming the database do not carry any explicit class label. More generally, all database analysis and representation generating processes described in this study are completely unsupervised. This ensures that the outcome is free of any bias introduced by supervision and conveys the reality of the data.

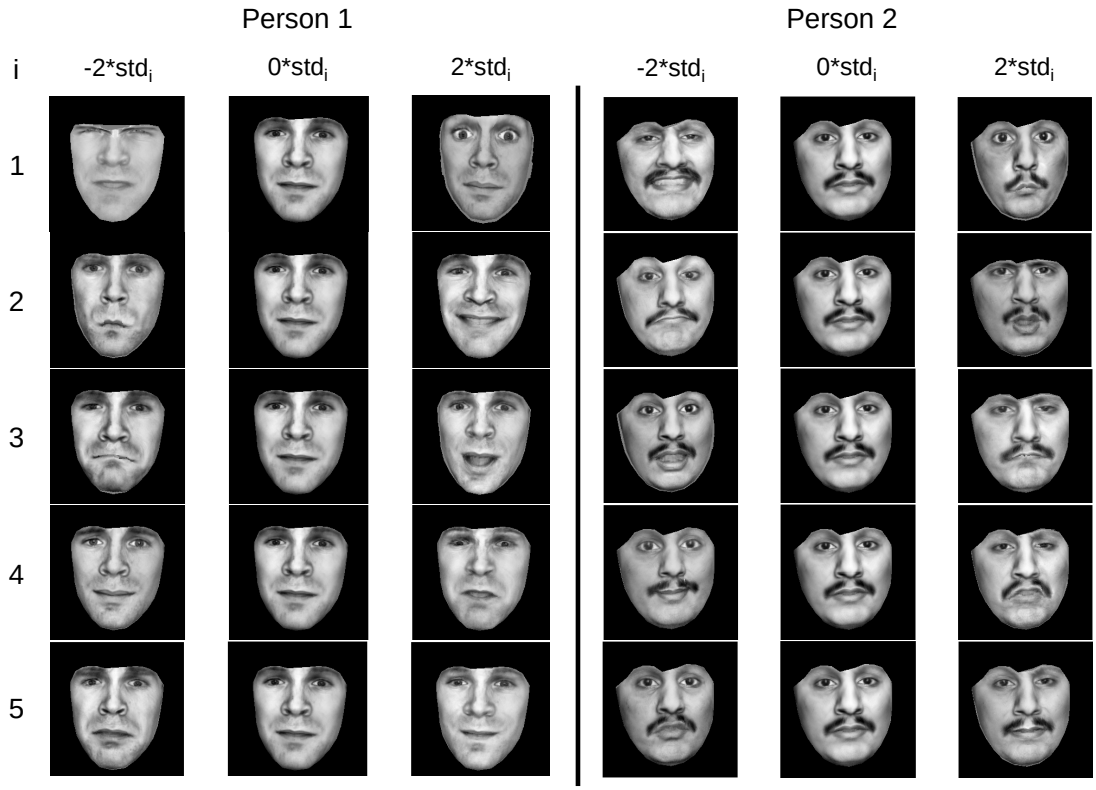

Figure 3.2: Visual illustration of the first five modes of person-specific AAMs (from top to bottom). Each mode is depicted by 3 samples scatterered along its axis. The samples correspond respectively to -2 times the standard deviation (STD), zero times the STD (mean face), and 2 times the STD. It is interesting to notice that although both person-specific models were constructed on equivalent databases, the deformation patterns of each mode do not always correspond.

3.1.3 Human Facial Appearance Space

In this work we use the AAM formalism described in section 3.1.1 to parameterize the visual deformations of expressive faces. This corresponds to a data-driven low-level parameterization of expressive facial deformations, as presented in section 2.1.1.3).

In practice, the image database introduced in section 3.1.2 are first annotated semi-automatically[1], and the AAM modeling is performed on the image and annotation data. The modeling delivers a reduced set of parameters, which represent the principal variation patterns detected on the human face (see figure 3.2). According to the model

[1]Some examples were annotated by hand, and an AAM search procedure used these examples to automatically annotate the rest of the database.

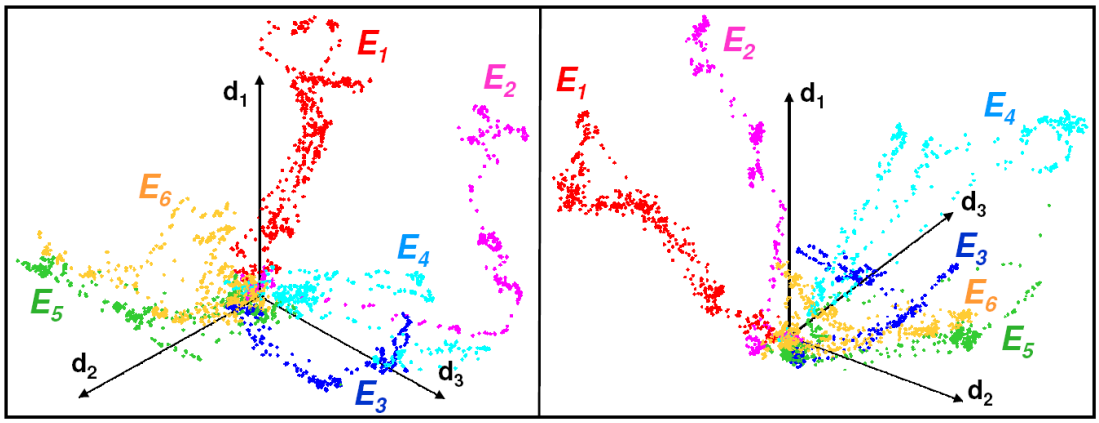

Figure 3.3: Two different views of the database samples distribution in the appearance space (4365 samples). Only the first three dimensions (d_1, d_2, d_3) of the appearance space are drawn. The sample coloring is based on an emotional label subjectively assigned to each facial expression (Blue=Sadness, Cyan=Joy, Green=Surprise, Yellow=Fear, Red=Anger, Magenta=Disgust).

of section 3.1.1, the parameters are obtained with the following equations:

$$\mathbf{b_i^s} = \mathbf{\Phi_s^T}.(\mathbf{s_i} - \bar{\mathbf{s}}) \tag{3.4}$$

$$\mathbf{b_i^t} = \mathbf{\Phi_t^T}.(\mathbf{t_i} - \bar{\mathbf{t}}) \tag{3.5}$$

and

$$\mathbf{c_i} = \mathbf{\Phi^T}.\left[\ w_s.\mathbf{b_i^s} \mid \mathbf{b_i^t}\ \right] \tag{3.6}$$

The dimensions of the final vector c form a parameter space of dimensionality N_a we refer to as the appearance space.

Once the AAM model is trained, meaning matrices $\mathbf{\Phi_s}$, $\mathbf{\Phi_t}$ and $\mathbf{\Phi}$ of equations 3.1, 3.2 and 3.3 have been determined, every facial image can be *projected* onto this N_a-Dimensional parameter space, and be represented by a point of the appearance space (see figure 3.3). Note that this process is invertible: it is always possible to map a N_a-Dimensional point of the appearance space back to a facial configuration, and thus synthesize the corresponding facial expression as a facial image.

The AAM parameter space is very interesting from a computational point of view. It significantly reduces the dimensionality of facial expression data and is locally consistent. Indeed, facial expressions can be modified and interpolated simply by acting on their appearance parameters. This however is valid only locally, as certain areas of the appearance space produce inconsistent expressions. Careless extrapolation of the local consistency of AAM parameters can lead to incoherent results (see figure 3.4). Besides, working in the appearance space is not particularly intuitive, because of the significant number of parameters ($N_a \approx 30$ dimensions). Additionally these parameters arise from objective statistical analysis and do not generally correspond to natural, interpretable

Figure 3.4: Careless manipulation and extrapolation of AAM parameter can lead to incoherent configuration. **Left:** neutral expression. **Right:** uncontrolled extrapolation of AAM parameters.

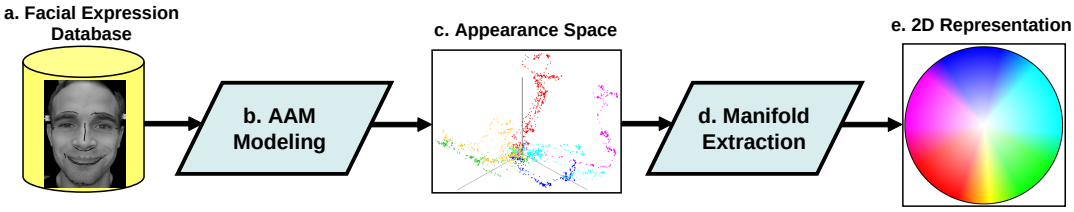

Figure 3.5: Overview of the manifold extraction workflow.

expressions. Alone, the appearance space is not compact and consistent enough to enable intuitive manipulation of facial expression in high-level applications.

3.2 Facial Expression Manifold

Low-level description schemes require many parameters to precisely describe the diversity of natural facial expressions, mostly because of the complex nonlinear deformation induced by expressions. Due to the continuity and consistency of facial movements, a majority of researchers now agree on the fact that human facial expressions form a nonlinear continuous space often depicted as a manifold. This manifold of facial expressions carries all natural facial configurations and encompasses the nonlinear modes of deformation of expressive faces. Since it models the natural nonlinear modes of deformations, the manifold of facial expressions is generally to be simpler and have a lower dimensionality than traditional data-driven representations. In this section we detail how we propose to extract an approximation of this manifold and use it as an intuitive, high-level representation of emotional facial expressions.

As discussed in the state-of-the-art report in section 2.1.1.3, previous studies have used data-driven nonlinear analysis to extract a manifold-like description of facial expressions. They heavily rely on widespread generic nonlinear embedding techniques such as Isometric Mapping (IsoMap) and Locally Linear Embedding (LLE). While the efficiency of those techniques has been validated on artificial benchmark data, results with real experimental data such as facial expressions have not been very convincing. In particular, since they are based on the detection of spatial neighborhoods, they implicitly require that the database form a relatively uniform sampling of the underlying manifold. However, as noted by Chang *et al.* [CVTV05], the whole expression space of an individual is large, containing many blended expressions, and only a small portion of this space can be sampled. Consequently, generic nonlinear reduction on facial expression data mostly results in inconsistent representations.

In the following, we detail our scheme of extraction of the nonlinear manifold of facial expressions (see figure 3.5). Contrary to classic approaches, the scheme is adapted to the specificities of facial expression data.

3.2.1 Dominant Directions

On figure 3.6, we can get a feeling of how the facial expression database is structured. The figure displays the distribution of expressions in the appearance space. Obviously, the samples cannot be visualized as N_a-Dimensional points, but we can visualize them as 3-Dimensional points by drawing their 3 most important principal components. To highlight the peculiarity of the structure on the figure, the database samples have been manually colored according to their resemblance to one of Ekman's basic expressions (groups E1 to E6 on figure 3.6). It is important to notice, however, that this subjective labeling was not used in the modeling process, but only for the convenience of the display.

Neutral facial expressions are located at the center of the point cloud while highly expressive faces are located on the edges of the cloud. Intermediate expressions are located in the continuous space between neutral and extreme expressions. We observe that the point cloud possesses a few dominant directions which are identified as the segments separating the neutral expression from extreme ones. Most of the natural expressions are distributed either along these directions to form a given expression at different intensity levels, or between these directions to form transitional expressions between the dominant ones.

It is worth noting that this special structure is not specific to our database or the appearance model used here. Characteristics such as non-uniform sampling and dominant directions can be identified in other studies of facial expression data [HCFT04, SGM05a]. This structure can be explained to a certain extent. Databases of natural facial expressions are not exhaustive and do not finely sample the entire space of possible expressions. However it generally contains expressions with varying intensity levels, which explain the radial organization of the samples. Additionally, the fact that one specific database does not contain all possible combinations of facial actions but only a few expressions explains the presence of identifiable directions.

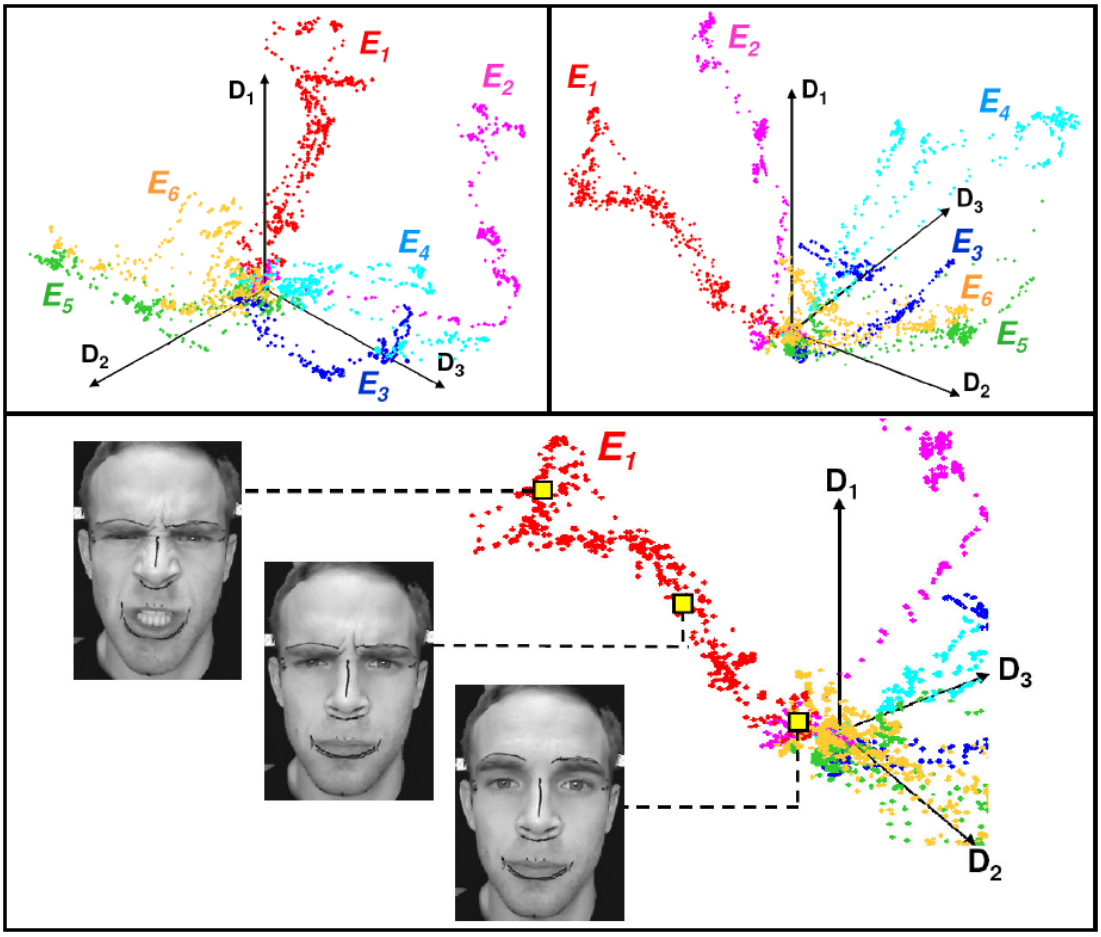

Figure 3.6: **Top:** two different views of the database samples distribution on the first three dimensions of the appearance space. **Bottom:** the expressions of varying intensities are concentrated on dominant directions in space. Neutral expressions are located in the center, and extreme expressions on the edge of the point cloud.

Instead of raw facial expression samples, in the following we look to extract the manifold of facial expressions based on the structure described by those dominant directions.

3.2.2 Facial Expressions Manifold Extraction

The interesting part of the appearance space is the space spanned by the dominant directions. It is tempting to use this structure as a skeleton for the geometrical formulation of the expression manifold. Based on our observations of the appearance space, we chose to consider that the manifold of emotional facial expressions is spanned by an

hypersphere[2] centered on the neutral expression. The (normalized) dominant directions identified in the last section represent the radii of this hypersphere. We can construct an approximation of continuous surface of the hypersphere by connecting neighboring dominant directions. The manifold itself is represented by the "surface" of the hypersphere and its interior.

This approach can be clarified by considering the example of a 2D manifold of facial expressions. Following our logic, the 2D manifold is spanned by a 1D hypersphere, which is simply a circle. The circle we are looking for can be approximated by a line connecting all dominant directions. The manifold approximation consists of a collection of triangles, forming a "triangle fan". Each triangle is composed by two vertices corresponding to dominant direction, the last one always being the vertex corresponding to the neutral expression.

More generally for a manifold of dimensionality n, its structure will consist of a "fan" of connected $n-$simplices. Each simplex is formed by $n+1$ vertices, with n of them corresponding to dominant directions and the last one being the neutral expression vertex. Ultimately, the "simplex fan" forms piecewise linear approximation of a nonlinear manifold. Its goal is to encompass the interesting parts of the appearance space in a simple structure.

The quality of the manifold approximation depends on how dominant directions are connected. The structure has to be consistent, and reflect as well as possible the original spatial organization of the dominant directions. Note that what actually matters is the topological arrangement of directions within the fan, not the measurement of geometrical distances between them. We can therefore formalize this process as a non-metric manifold embedding problem, which differentiates it from metric problems such as IsoMap and LLE.

In concrete terms, the objective and automatic construction of this space involves two principal tasks: dominant directions detection and expression space embedding. Those processes are detailed in the following sections.

3.2.2.1 Dominant Direction Detection

We have defined the dominant directions as the segment between the neutral expression and the extreme expressions in the database. The database sample representing the neutral expression can be easily designated manually. On the other hand, the extreme expressions have to be objectively chosen so that the majority of the significant part of the appearance space is encompassed by the dominant directions. In that sense, the most appropriate candidates are the sample located on the convex hull of the sample cloud (see figure 3.7). Unfortunately, the hull detection algorithm can become rather costly at high dimensionalities. However, since the dimensions of the appearance space

[2] An hypersphere, or n-sphere, in mathematics is the generalization of the traditional sphere to higher dimensions.

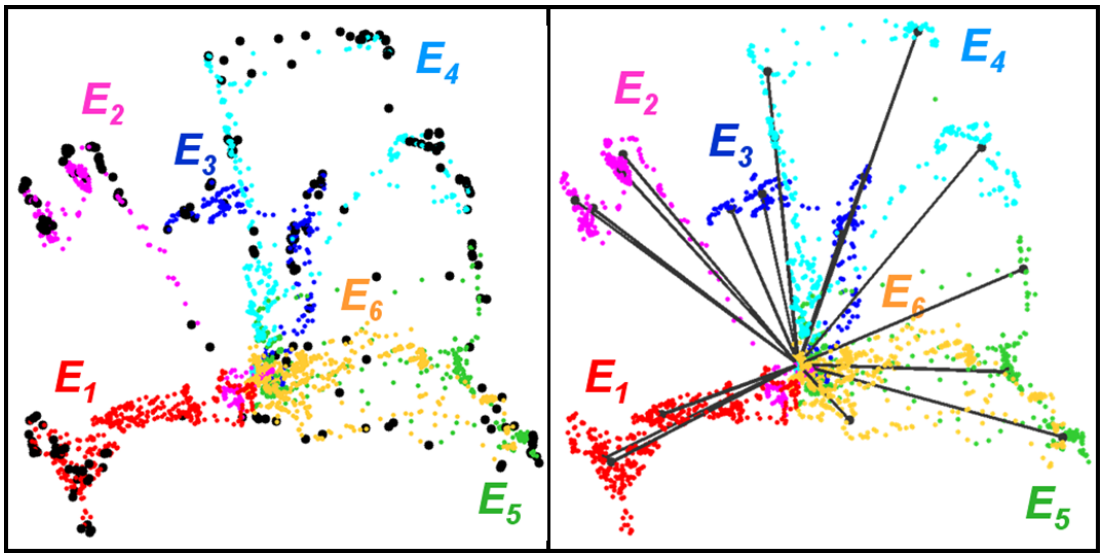

Figure 3.7: **Left:** extreme expressions are detected as the convex hull points (black dots). **Right:** the dominant directions (black lines) are identified as the segment between the neutral expression and a few selected extreme points.

are determined by a PCA routine (see section 3.1), they are ordered by their percentage of participation to the global database variance. In other words, the first dimension of the appearance space has the most important variance, and the variance decreases when we move from the first to the last dimension. We can use this property to perform the hull detection on only a subset of dimensions, and therefore avoid high computation times. In practice, using the first 8 dimensions (75% of the total database variance) is enough for stable hull detection. Besides, the result of the hull detection process usually contains an important redundancy: the convex hull often intersects the sample cloud at several neighboring points. These neighboring points, however, represent one single dominant direction of the database distribution. They have to be merged into a single representative hull sample. This can be achieved by running a mean-shift algorithm on the set of detected hull points, so that groups of neighboring samples are merged into one mean-shift mode while isolated hull samples remain unchanged.

3.2.2.2 Expression Space Embedding

The goal of the embedding is to associate a set of discrete directions to form a continuous approximation of the hypersphere. For instance, in the 2D case the embedded directions will be embedded on a circle (1-Dimensional hypersphere) and in the 3D case on a sphere (2-Dimensional hypersphere). More generally, the dominant direction will be embedded on a n-Dimensional mesh. The final manifold is supposed to reflect the original organization of the appearance space, so the association must be connected carefully. To avoid distortion of the manifold as much as possible, the connected direc-

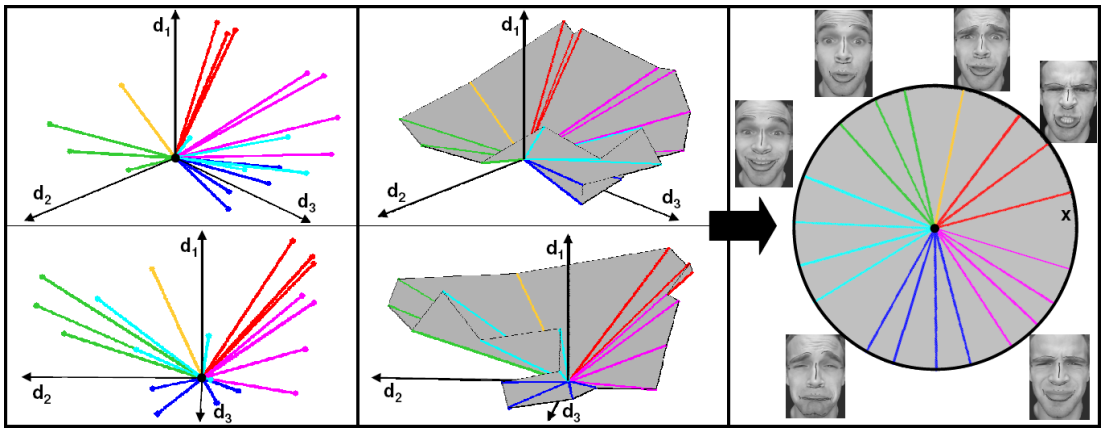

Figure 3.8: Appearance space embedding on a 2D manifold.
Left top and bottom: two views of the detected dominant directions in the appearance space (only the first three dimensions (d_1, d_2, d_3) are drawn). Directions coloring is based on the database emotional labeling. *Middle top and bottom:* two views of the path (gray surface) determined by the traveling salesman optimization routine. *Right:* unfolded view of the embedded 2D manifold (disk).

tions must correspond to directions that were spatially close in the original appearance space.

We have formulated this task as an objective optimization problem, which aims at minimizing a cost function. As a cost function, we use the total sum of "angles" between connected directions. Indeed, angles between directions are a measure of their closeness in the appearance space.

This can be clearly illustrated in the 2D case, if we suppose that n_d directions $\mathbf{d_i}$, $i \in \{1, ..., n_d\}$ have been detected, the approximation of the hypersphere (a circle in that case) is a path, connecting all n_d directions. In that case, the cost function reads:

$$cost_{2D} = \sum_{k=1}^{n_d} \Theta_{i_k, i_{k+1}} \tag{3.7}$$

where $\{\mathbf{d_{i_1}}, \mathbf{d_{i_2}}, ..., \mathbf{d_{i_{n_d}}}\}$ is the order in which the directions are connected along the path ($i_{n_d+1} = i_1$), and $\Theta_{i,j}$ is the angle between high-dimensional directions $\mathbf{d_i}$ and $\mathbf{d_j}$. If connected directions have small inter-direction angles in the appearance space, the resulting cost will be low. Embeddings that respect the original organization of the directions in the appearance space are therefore rewarded.

In the 3D case, the hypersphere is a surface approximated by a triangle mesh. Neighborhoods connecting dominant directions are now bidimensional, so the angles of

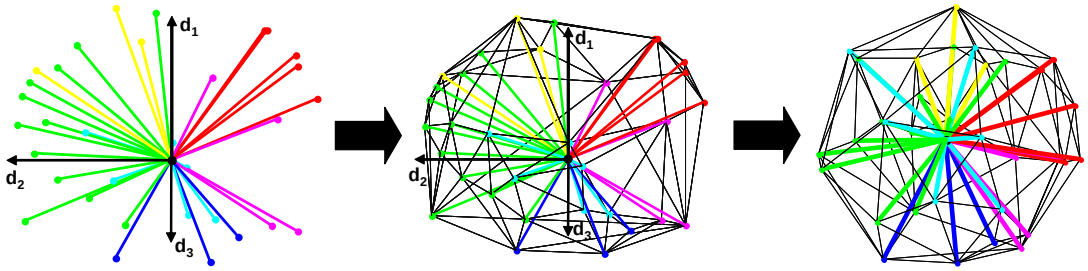

Figure 3.9: Appearance space embedding on a 3D manifold. *Left:* dominant directions in the appearance space. *Middle:* 2D mesh of dominant directions determined by the 2D extension of TSP optimization. *Right:* unfolded view of the embedded 3D manifold (ball).

equation 3.7 are replaced by solid angles. The cost function becomes:

$$cost_{3D} = \sum_{k=1}^{n_c} \Omega_{i_1^k, i_2^k, i_3^k} \tag{3.8}$$

Where the triangle mesh has n_c cells, each formed by three high-dimensional directions $\{\mathbf{d_{i_1^k}}, \mathbf{d_{i_2^k}}, \mathbf{d_{i_3^k}}\}$ and spanning the solid angle $\Omega_{i_1^k, i_2^k, i_3^k}$. The cost function can be expressed similarly in higher dimensionalities. For a n-dimensional approximation of the manifold of facial expressions, the hypersphere is approximated by the cells of a $(n-1)$-Dimensional mesh, spanning $(n-1)$-Dimensional solid angles. The cost function is equal to the sum of those solid angles, and the goal is still to find the mesh configuration that minimizes this cost function.

An interesting fact of this optimization task is that, in 2D and 3D, it degenerates into well-known problems. Indeed, in 2D, we search to order directions $\mathbf{d_i}$, $i \in \{1, ..., n_d\}$ in a sequence so that the accumulated angle between consecutive directions is minimized. This is the exact formulation of the famous Traveling Salesman Problem (TSP). Indeed, the TSP is a NP-complete problem whose purpose is to determine the shortest path running through a set of cities, when the distance between them is known. By replacing the cities by the N_a-Dimensional directions and the inter-city distance by the inter-direction angle, we can solve our specific optimization task by using any generic method solving the traveling salesman problem.

In 3D, the problem is a higher-order extension of the TSP where the accumulated area of the surface connecting the directions has to be minimized. This logic can be extended to the n-Dimensional manifold embedding problem.

As a result of the above, the embedding problem can be solved in the general case by any algorithm adapted for such "TSP-like" optimization problems. However, because of the NP-Completeness of the TSP, the optimization task becomes excessively costly when the number of detected directions grows. In practical cases, researchers often turn to

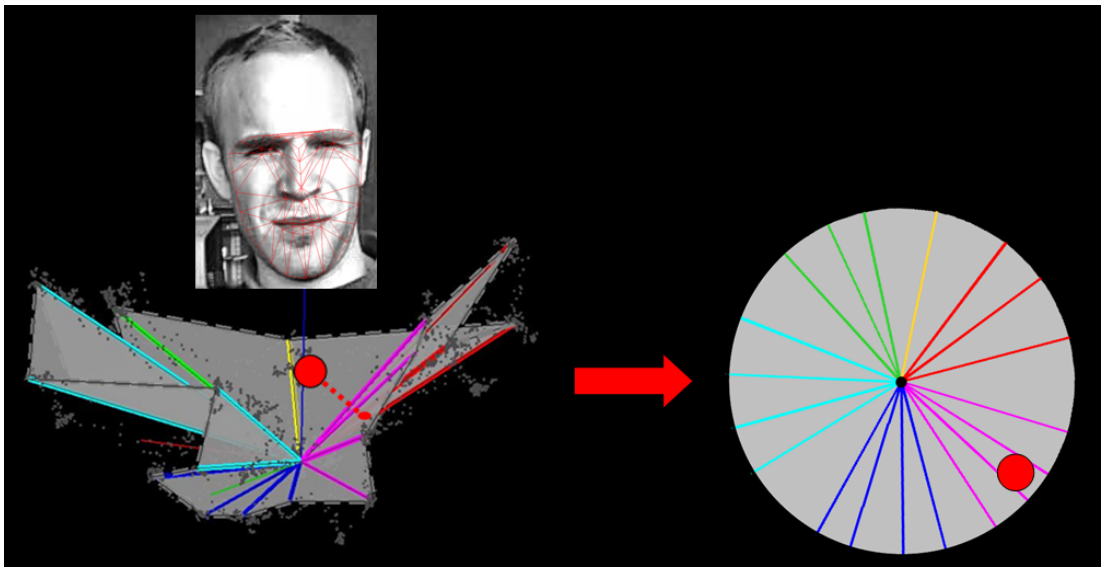

Figure 3.10: Symbolic illustration of an unknown expression on the 2D facial expression manifold. The red point corresponds to the vector formed by the AAM coefficients of the considered expression. Projecting this point on the manifold allows identifying the closest manifold expression, and thus measuring the representation error for this particular expression

heuristic methods to keep the optimization solving time short. In this study, we used a stochastic method based on simulated annealing. Our experiments showed that, with a reasonable number of directions, the TSP optimization always reaches a stable possibly-optimal solution in acceptable time (less than 2 minutes for 30 directions). Once solved, the TSP optimization delivers the sequence or the mesh of directions (approximation of the hypersphere) that minimizes the overall distortion (see figure 3.8 for the embedding of a 2D manifold, and figure 3.9 for the embedding of a 3D manifold).

3.2.2.3 Projection on the Manifold

No matter the dimensionality we choose, our formulation of the manifold of facial expressions remains an approximation. Intuitively we imagine that the higher the dimensionality the better the approximation, which will be confirmed by figure 3.12. However even for high dimensionalities, the approximation cannot be perfect. For one, the facial expression database is not (and cannot be) exhaustive. Other, unaccounted for types of expressions might not fall exactly into the scope of the manifold. Additionally, even when ignoring capture noise, certain subtleties of the database expressions are also not covered exactly by the manifold. In those cases, it is interesting to identify what we call the projection of the expression on the manifold. The projection of an expression is defined here as the point on the manifold that is the closest to the considered expression

in terms of AAM parameters (see illustration on figure 3.10). Knowing the projection of an expression on the manifold is interesting mainly for two reasons: it determines the closest "known" expression (which can be thought of as expression recognition), and it also determines the projection error, which measures the relevance of the manifold to the considered expression.

When considering traditional, "linear" vector spaces, the straightforward and prevailing form of projection is the orthogonal projection. With nonlinear spaces, the projection task is more complex, as the space cannot be entirely described by a single set of basis vector. In general unconstrained nonlinear spaces, empirical methods are often use such as relying on neighboring samples with known projections to determine the projection of an unknown sample. In our approach, the considered projection space is a piecewise linear manifold. Indeed, as described in section 3.2.2, our facial expression manifold is composed of adjacent geometrical simplices. This makes the projection process somewhat better defined than in an unstructured manifold scenario. In practice, our projection procedure performs traditional orthogonal projection onto the simplices of the manifold. For a given unknown expression, this procedure can be broken down into the following steps:

- Simplex identification: identify the simplex on which the projection is to be performed. This simplex is simply the closest to the considered expression.
- Projection on the identified simplex: each simplex behaves locally as a linear space, so orthogonal projection can be used to determine the projected point.

Once the right simplex has been identified, the orthogonal projection matrix is computed from the vectors spanning the simplex. It is however important to note that the simplices are limited in space. After projecting the considered expression on the linear space spanned by the simplex, one must make sure that the projected point lies within the simplex' boundaries.

Figure 3.11 shows the projection of all database samples on the 2D manifold of facial expression described in the previous section. The coloring of the projected samples is similar as the one of figure 3.3. An equivalent graphical illustration can be produced for the 3D manifold embedding. The distribution of the projected samples tends to indicate that the manifold has a coherent structure, as morphologically similar expressions are projected at neighboring locations. Note that the sample coloring is once again used for convenience of the display only, and that all processing use unlabeled data.

3.2.3 Expression Manifold as a Visual Representation

Researchers have been interested in the manifold of facial expressions because it produces more accurate quantitative performances, but also because it offers a more natural way of describing facial expressions. The nonlinear modes of manifolds can more flexibly adapt to facial deformation data while retaining simple topologies. The latter aspect is particularly desirable for visual representations. In this section, we elaborate on how our

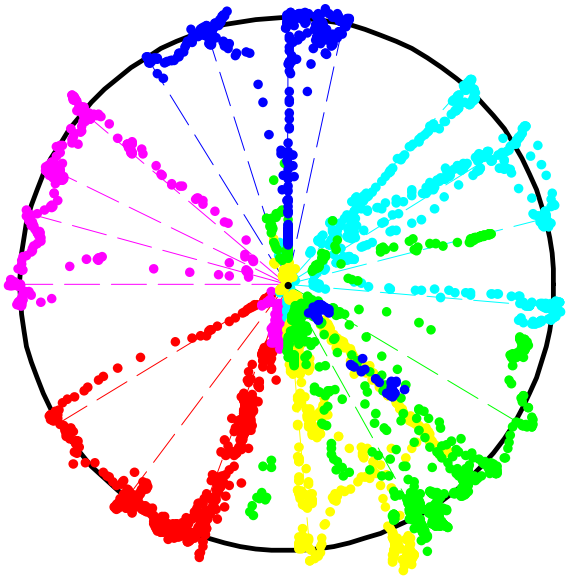

Figure 3.11: Projection of all database samples on the 2D manifold of facial elements. The sample coloring is based on an emotional label subjectively assigned to each facial expression (Blue=Sadness, Cyan=Joy, Green=Surprise, Yellow=Fear, Red=Anger, Magenta=Disgust). The distribution of the projected expressions illustrates the consistency and coherence of the extracted manifold.

manifold formulation can act as an attractive representation of the space of emotional facial expressions.

3.2.3.1 Dimensionality

Previous sections emphasized 2D and 3D manifold extraction examples, as they can be illustrated more easily. However, as we have previously stated, the extraction method described above can be generalized to higher dimensionalities. While the manifold extraction process is fairly similar for all choices of dimensionalities, this choice significantly affect the properties of the output. Choosing a dimensionality n implies that we can only address expressions located on an n-D manifold in the appearance space. For instance, a dimensionality 2 implies that we address only a 2D manifold in the facial appearance space (so a topological disk). With an additional dimension, a 3D Manifold is able to address more facial configurations more accurately. This logic goes on for higher orders, however adding dimensions complicates the structure of the manifold.

The choice of the right dimensionality actually becomes a tradeoff between representation accuracy and topological simplicity. It also depends on one's motivations: systems focused on precision and performances (such as recognition systems) will work with higher dimensionalities, while studies focused on visual representations (see next section) will favor 2D or 3D spaces.

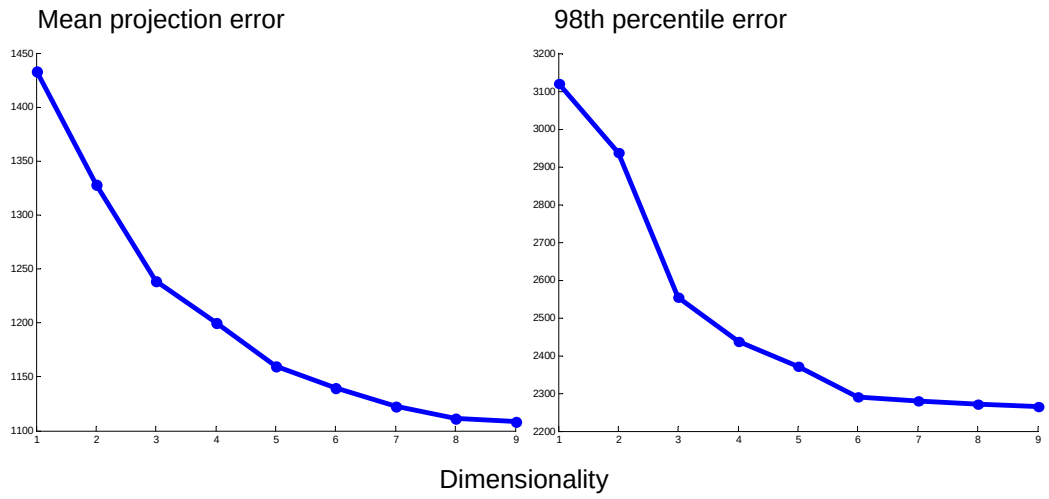

Figure 3.12: Trend of the projection errors for different dimensionalities of the facial expression manifold.
Left: Mean projection error. *Right:* 98^{th} percentile of the projection error[4].

Obviously, the improvement of precision when increasing the dimensionality is not constant, and not infinite. In particular, if we assume that the "true" space of facial expression is an n-Dimensional space, then using an (n+1)-D manifold -or higher- to approximate it is not relevant. In that perspective, it is interesting to measure this property quantitatively. The error between expression from the database and their projection on the manifold (see section 3.2.2.3) can be used to evaluate the quality of representation for increasing dimensionalities. The result is presented on figure 3.12[3].

From the figure we observe that the mean error over all database samples decreases monotonically, yet the error reduction becomes smaller with each added dimension. Since the data contain noise, we reduce the influence of noise overfitting by also presenting the 98^{th} percentile[4] of the projection error on figure 3.12. This graph also shows a monotonic decrease of the projection error, but highlights a clear stabilization of the error at dimensionality six. For dimensionalities above that, the error reduction is significantly less interesting, which tends to indicate that the space of facial expressions can be reasonably well approximated by a manifold of dimensionality six.

To our knowledge, the aspect of facial expression space dimensionality has never been properly investigated. Zhang [Zha99] did look for the appropriate dimension to represent facial expressions in their recognition system. The investigated question was how many hidden layer neurons enabled a good recognition rate with their two-layer perceptron. It turned out that five to seven hidden units were satisfactory to represent

[3]On figure 3.12, the "1D" value corresponds to a projection space formed by the discrete collection of dominant directions (which can be considered as directions connected by a mesh of order zero).

[4]the 98^{th} percentile of the error means that 98% of the samples have a smaller errors than this value.

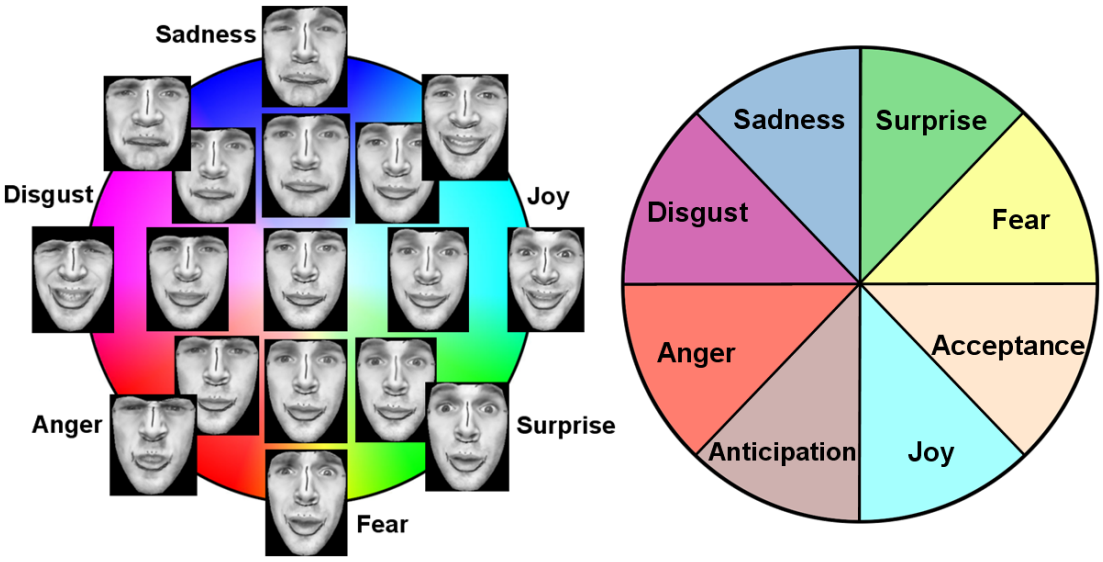

Figure 3.13: Comparison between an emotion model and our Data-based representation.
Left: interpretation of the expressions' semantic of our data-driven representation. *Right:* the relations between 8 basic emotions according to Plutchik's theory [Plu80].

the space of facial expressions. Our observations tend to confirm those results, as the trend of projection errors seems to stabilize with manifold dimensionality equal to six.

Apart from the purely quantitative evaluation, the quality of the approximation can also be observed qualitatively. Indeed, the error between an expression and its projected version can be visualized as a change in facial expression appearance. Early qualitative tests have been conducted on a few individuals to assess the level of projection error that produces visible expression change. Those tests indicate that at dimensionality 3, the manifold's approximation errors are seemingly invisible. However, no clear test protocol has been defined for this study so these results should only be taken has a hint to choose the right dimensionality. In the future, a more comprehensive user study of the visual impact of manifold projection would be needed. This task is however not straightforward: in practice it is difficult to differentiate slight facial appearance changes from capture noise in the error term. Moreover, the error-appearance relation does not seem to be stationary over different expressions.

3.2.3.2 Representation Results

As mentioned above, the interest we have in extracting the expression manifold is that it constitutes an attractive representation of the otherwise complex phenomenon of facial expressions. Indeed, especially in the 2D and 3D cases, our formulation of the manifold of emotional facial expressions has a rather simple topology (a 2D disk and a 3D ball respectively). These topologies can be presented more intuitively as low-dimensional

entities: in the 2D case, the high-dimensional triangle fan can be symbolically represented by a 2D triangle fan, where the dominant expressions of the appearance space have been distributed on the radii of the unit circle. Similarly, the 3D approximation of the manifold can be embedded in the unit sphere, respecting the mesh identified by the optimization procedure (section 3.2.2.2). These low-dimensional embeddings can be seen as a visual "unfolding" of the manifold of facial expressions. The unfolded representations form a simple and intuitive image of the distribution of expressive faces in the appearance space (see figure 3.13, left). When considered as visual interfaces, those representations exhibit the simplicity and the interpretability of an emotion model (see section 2.1.2.1). Nevertheless, they originate from an automatic, data-driven extraction process and thus remain coherent with facial deformation modes.

We mentioned in the introduction that, contrary to previous approaches, the new facial expressions representation does not rely on a theoretical emotional space like Cowie's activation-evaluation space or Plutchik's emotion wheel [CES+00, Plu80]. It is interesting to observe that we can associate a semantic meaning to the data-driven representation by interpreting the facial expressions *a posteriori*. We obtain the formulation of a *morphology-based* emotional space (figure 3.13, left). The difference with previous emotional spaces is that this one is derived from actual facial deformation data, and not from theoretical considerations. The disk-based structure, with the emotion type as the rotational parameter and the emotion intensity as the radial one reminds us particularly of Plutchik's emotion wheel (figure 3.13, right). Plutchik formalized the relationships between 8 primary emotions according to his psychoevolutionary theory on adaptive biological processes. We observe that the closeness identified by Plutchik between some of the basic emotions still holds from the morphological point of view. The triplet Anger-Disgust-Sadness is common to both representations, as well as the proximity of Surprise and Fear.

Obviously, we do not claim to offer a new formal description of human emotions, but our data-driven expression space enriches the purely theoretical models with considerations on the mechanical aspect of facial expressions. Moreover, since it originates from facial motion data, it can more efficiently be associated with low-level data and avoid the typical distortions of previous representations (in which close representations would correspond to very different expressions, and vice versa).

This chapter focused on the analysis of natural facial expressions, and presented construction of the appearance space and the high-level representation of facial expressions. In chapter 4 we present interesting uses of these tools for the facial animation of virtual characters.

Chapter 4

Manipulation of Emotional Facial Expressions

Contents

4.1 Construction of Facial Appearance Spaces 92
4.1.1 Database Bootstrapping 93
4.1.2 Facial Appearance Space for a Virtual Character 95
4.2 Expression Retargeting 96
4.2.1 Background on Expression Retargeting 96
4.2.2 Expression Retargeting in Appearance Spaces 101
4.3 Low-Dimensional Control Space for Facial Expressions . . . 106
4.3.1 Background on Intuitive Control of Facial Expressions 107
4.3.2 Facial Expression Manifold as a Control Space 108
4.3.3 Intuitive Control of Synthetic Faces 109

Chapter 3 focused on the analysis and representation of a person's emotional facial expressions. We now study how this knowledge can be applied to new faces, especially the synthetic faces of virtual characters. This chapter puts forward the use of the appearance space (section 3.1) and the high-level representation of facial expressions (section 3.2) in an animation context for facial expression synthesis.

In section 4.1 we describe how various facial appearances spaces can be constructed for real and synthetic faces. In section 4.1.1 we present a procedure that processes an initially large database to identify only the most important samples. Those samples can be used to constitute databases for new faces with minimal effort. This is particularly advantageous for virtual character, for which each sample is manually designed (section 4.1.2). Besides, using databases with corresponding samples provide an efficient framework for expression mapping (see section 4.2). Finally, in section 4.3, we adapt the high-level representation presented in chapter 3 to efficiently navigate in the obtained appearance spaces, and thus control the facial expressions of human and synthetic faces. All those processes are illustrated visually in figure 4.1.

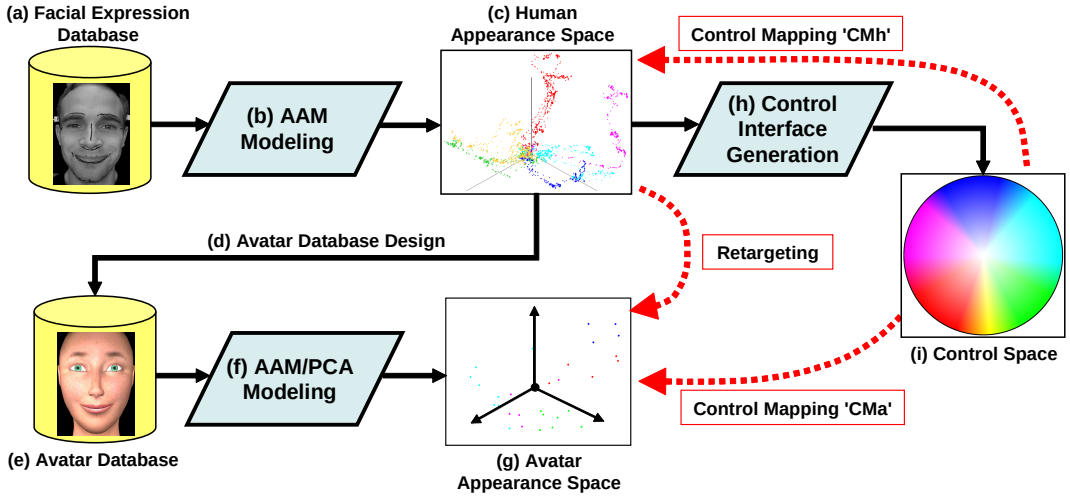

Figure 4.1: Use of the human appearance space and high-level representation of expressions for the animation of a virtual character. Overview.

4.1 Construction of Facial Appearance Spaces

The appearance space is the parameter space delivered by the AAM modeling of natural facial expressions (section 3.1). Although simple, The PCA-based parameter space is very interesting for facial parameterization because it only models the perceptible variations of the face, and not the underlying mechanisms causing these deformations. Therefore, contrary to muscle-based approaches for example, the reduced space forms a compact and continuous space, with few inconsistencies (areas of the parameter space that correspond to unnatural facial configurations). Additionally this space is a basis for high-level representation of facial expressions (section 3.2).

The modeling procedure at the origin of appearance space is often referred to as "statistical" modeling, because it extracts the facial deformation patterns from the analysis of database of examples. For the appearance space of section 3.1 as many as 4365 images of facial expressions were used in an attempt to represent the diversity of natural facial expressiveness.

Such modeling can be carried out for any face, virtual faces included. Just as for real faces, it relies on a database of examples of deformation induced by emotional expressions. However, building a large database represents a significant workload. Relying on thousands of expression samples to ensure the diversity of modeled expressions is unpractical especially in the case of a virtual character: while for real faces, thousands of database samples can be produced with a video camera and a feature-tracking algorithm, the elements of an equivalent synthetic database are manually-designed 2D or 3D facial poses. It is thus desirable to keep the number of required samples as small as possible.

The choice of a reduced set of "representative faces", which will act as a database for is not trivial. Previous studies predominantly rely on empirical considerations to form the database. Ekmanian universal expressions are an intuitive selection, used by a large number of practical systems. In [CB02], Chuang and Bregler have proposed heuristic approaches to determine the right representative faces for their blendshapes database. Based on an already existing large database of visemes and expressions, Chuang and Bregler investigated three selection methods:

- Maximum spread along principle components. This method picks the examples with the maximum and minimum projection values on the axes of a PCA on the database.
- Clustering. This method selects the examples corresponding to the centroid of the segmentation of the database into k clusters.
- Convex Hull. This methods picks the examples located on the convex hull of the database distribution in appearance space.

Although more rational, these techniques do not objectively guarantee an optimal choice of representatives. Clustering produces a good segmentation of the examples in term of classes, but this does not ensure that the diversity of expressions can be *reconstructed* from class representatives. Selecting the elements located on the convex hull of the database points concerns mainly high-intensity expressions, and is also not an ideal strategy. Facial deformation patterns occurring at low-variation levels would not be accounted for, and the model would consequently not be able to resynthesize these deformations correctly. Expressions corresponding to principal component axes would form an optimal set of mathematically orthogonal expressions. However, they generally consist of unnatural combination of facial deformations, which are very unpractical to reproduce on other real or synthetic faces.

In the following, we propose an objective selection of the elements of the reduced database, based on an iterative *bootstrapping* method, and on the measurement of actual reconstruction error.

4.1.1 Database Bootstrapping

Our idea for building the reduced database is to use the human database of section 3.1.2, and extract the expressions that have an important impact on the formation on the appearance space. Indeed, a lot of samples from the human database bring redundant information to the modeling process, and are therefore not essential for the database. In particular one can notice that the modeling scheme presented in section 3.1.1 is linear. Consequently, if a sample of the human database is linearly dependent of other samples, then the corresponding facial configuration can be recovered by linear combination of these samples. The linearly dependent sample is then expendable. Following this logic, we are able to reduce the set of necessary expression to a reasonable size.

Algorithm 2: Pseudocode of the selection of representatives for the reduced expressions database.

Input : **humanDB**, the complete human facial expression database.
Output : **reducedDB**, the reduced database formed by elements of **humanDB**.
Variable: *nbElementsInDB*, the current number of elements in the reduced database.
Variable: *maxNbElements*, the maximum number of elements the reduced database can hold.

```
1 // Procedure:
2 Initialize reducedDB with one single element (neutral expression) ;
3 while (nbElementsInDB < maxNbElements) do
4 |   Compute currentPCAModel on reducedDB ;
5 |   Project and reconstruct all elements of humanDB using currentPCAModel ;
6 |   Compute the error between the elements of humanDB and their
  |   reconstructed version ;
7 |   Identify the element with the highest error ;
8 |   Add this element to reducedDB ;
9 end
```

Figure 4.2: The first elements of the human expression database extracted by the bootstrapping algorithm.

The problem can then be described as follows: for a given database of human expressions, we must identify the smallest group of elements that model the variety of facial expressions as well as the whole database would. This group will constitute a reduced database, used to model the facial deformation on the synthetic face. The elements of the reduced database will be extracted from the human one.

We propose to use an iterative bootstrapping procedure. Starting with a basic model computed on just one element, we integrate the other elements successively. Within each iteration, the model is refined, as a new element is added to optimally improve the modeling power. The procedure is described as pseudocode in algorithm 2 on page 94. The main advantage of this approach is that it produces an optimal model within each iteration. The more iteration we allow, the more elements are added to the

reduced database. The modeling of facial expressions for a new face actually becomes a trade-off between the precision and diversity of the appearance space and the amount of manual work to create the database elements. Alternatively, a stopping criterion to the algorithm could be set based on the measured reconstruction error (algorithm 2, line 6). The procedure would then select the smallest group of database elements needed to reach a given reconstruction precision.

For the human database, we used more than 4000 elements (images of facial expressions extracted from a 3-minute video). Using the bootstrapping procedure, we have identified that the first 35 extracted elements (see figure 4.2) lead to a negligible reconstruction error for the rest of the human database. From this outcome we deduce that the large database, however complete, contains significant redundancy. The preceding procedure has allowed us to remove this redundancy and minimize the effort of creating the avatar database: only the 35 expressions of the reduced database have to be provided to cover the same scope as the full human database of section 3.1.2.

4.1.2 Facial Appearance Space for a Virtual Character

Most of the work and applications of facial animation deal with synthetic faces. An appearance space with similar properties to the one described in section 3.1 can be constructed for a synthetic face, provided that a database of facial expressions of the virtual character is available (database in figure 4.1e). Typically, the representative expressions identified in section 4.1.1 can be used to constitute the database (as illustrated by the arrow (d) in figure 4.1).

One way or another, the representative expressions still need to be manually constructed. This manual intervention might be seen as a weakness of statistical modeling for synthetic faces, as it imposes a database creation workload. Yet, other animation techniques also rely on this type of database: Blendshapes techniques require the setup of multiple facial expressions on a 3D model. This step can actually be seen as an advantage: animation systems based on the appearance face are applicable to any synthetic face. It only requires that the database can be generated, regardless of the method used to generate it (direct editing, pseudo-muscle, motion capture, *etc*).

One can note that relying on an appearance space enables the use of different visual representations of the virtual character, typically 2D (frontal image) or 3D (3-Dimensional geometry). The appearance space for the 2D or 3D synthetic face is build through statistical modeling, similarly to the human appearance space. In the 2D case we use the AAM modeling technique as presented for the human face (section 3.1.1), which to our knowledge has never been done before. Similarly to human faces, this allows the animation system to add skin deformation details, such as wrinkles, to purely 2D geometry. In the 3D case a single Principal Component Analysis (PCA) is used on the 3D geometric data, according to equation 3.1.

4.2 Expression Retargeting

In section 4.1.1 we show how to identify a reduced set of facial postures from the human database so that a coherent appearance space is constructed for a virtual character (figure 4.1g). The purpose of this avatar database creation scheme is that the appearance spaces of the human and the synthetic face have the same semantic meaning, and model the same information. A straightforward use of the correspondence between both appearance spaces is to perform expression mapping, also known as expression *retargeting*[1].

In the following section, we present a short insight on the different types of retargeting methods existing in the literature. We then show how we can take advantage of the correspondence between facial appearance spaces to perform facial parameter mapping. The resulting facial expression retargeting scheme is investigated in section 4.2.2.

4.2.1 Background on Expression Retargeting

Retargeting of facial expressions can be done either directly or indirectly. Direct retargeting means transfering the facial expressions at the geometric level, by adapting the local geometric deformations from a source mesh to a target one. On the other hand, indirect retargeting schemes rely on sets of parameters shared by the source and the target faces, which inherently encode the faces specificities.

Geometric Retargeting
Raw facial motion data usually consists of a set of markers that describes more or less finely the deformation of facial skin. Geometric retargeting works directly in the spatial domain, and locally transforms the recorded deformations to adapt to the geometry of the target face. This process is sometimes called *geometry warping*. These techniques were motivated by the work of Guenter et al. [GGW+98] that successfully replayed 3D motion capture data on a corresponding static face scan. A highly influential contribution in geometric retargeting is the work of Noh and Neumann [NN01]. Their system transfers and adapts vertex "motion vectors" from a source face model to a target model, regardless of the structural differences between source and the target meshes in terms of vertex number and connectivity. Based on an initial correspondence between points on the source and the target models, motion vectors are transferred using the respective barycentric coordinates of each model. The drawback of this otherwise good method is that the run-time computational cost depends on the complexity of the target model. Chai *et al.* [CXH03] later adapted this approach by precomputing all deformation bases of the target model offline, in order to ensure model-independent run-time computation.

The previous techniques have been used extensively in practical applications, however one limitation is that they require motion data recorded as 3D coordinates. With an increasing number of systems relying on video sources for motion capture, it is crucial

[1] Retargeting refers to the action of transferring the posture or animation from a source (typically a human face) to a target (another human face or a synthetic one).

to be able to retarget facial expressions across different formats. Actually, Williams, a forerunner of performance-driven facial animation, first proposed a system to transfer facial expressions from 2D data to a 3D face [Wil90]. Like in [GGW+98], the expressions on the target face were animated by modifying the texture coordinates. Later, Liu *et al.* [LSZ01] proposed a purely 2D method to retarget facial expressions between images, in which the geometrical structure of the expression were transferred with the help of 2D warping function. The authors compensated for the lost dimension by also transferring a 2D facial *texture*, along with the 2D geometry, via Expression Ratio Image (ERI). This approach enabled the successful retargeting of volumetric details like wrinkles. Those features however depend on facial morphology and skin properties, so reusing ERIs on very different faces leads to unnatural results. More recently Song *et al.* [SDT+07] submitted a study that unifies geometrical 2D and 3D representations in a general retargeting framework. Expressions could be mapped from an arbitrary source face onto an arbitrary target face. The transfer of subtle details in the 2D case relies on the same principle as in [LSZ01] (transfer of pixel-based facial details) and thus faces the same limitations.

2D data inevitably loses geometric accuracy when representing facial deformations, so in that case direct geometry-based transfer is not ideal. Other studies have thus adopted approaches in which faces are associated in a more abstract way.

Parameter-based Retargeting

Parameter-based retargeting can be described as *indirect* retargeting. Facial deformations are not transferred as such but via the parameters of a representation model. The expressions displayed by the source face are converted to model parameters, and those parameter values are used to animate the target face accordingly. Since facial expressions are broken down into a common set of parameters, this approach adapts well to different formats of source and target face (2D or 3D, with varying structures).

In order for a facial parameterization to be suited for retargeting applications, it must be adapted to the extraction of parameters from raw motion capture data, and offer an accurate description of facial deformations. Typical parameterizations used for that purpose are what we called descriptive representations of expressions in section 2.1.1.2. Curio *et al.* [CBK+06], performed retargeting between 3D MoCap and a 3D model based on Facial Action Units of the FACS [EF78b]. Each capture expression was decomposed into a weighted combination of references describing the standard Action Units. Those weights, which act as FACS parameters, are then applied to the target face to reproduce the same expression.

Several studies adopted a similar approach, but some of them chose to rely on the MPEG-4 standard [EHRW98]. Tsapatsoulis *et al.* [TRK+02] aimed at producing rich human-agent interaction with the support of a rich facial behavior. They wished to analyze primary and intermediate facial expressions on a human face. MPEG-4 Facial Action Parameters (FAP) were used both for the extraction of facial expressions from a video, and for the synthesis of corresponding expressions on several 3D synthetic faces. Byun and Badler [BB02] proposed a system that also uses MPEG-4 FAP to describe and replay facial movements on different faces. Their system additionally perturbs the

values of the FAP to produce non-repetitive, varied expressions on the target face. Chai *et al.* [CXH03] associated low definition, 2D motion capture data with high-quality 3D animations with the help of 15 custom control parameters describing local facial actions (eyes, mouth, nose and eyebrows). These custom parameters highly resemble Facial Action Units from the FACS. Based on recorded values of these parameters, expressions could be efficiently retargeted from real-time video capture to animation of a virtual character.

As far as generative schemes are concerned (see section 2.1.1.1), retargeting schemes based on muscles activation have produced interesting results. Essa *et al.* [EBDP96] proposed an early system which extracts muscular activations from optical flow analysis on 2D images. These values were then used to clone the expression on a synthetic face. Choe *et al.* [CLK01], and more recently Sifakis *et al.* [SNF05] relied on similar systems, yet using 3D motion capture and synthesis. Muscle parameters are well adapted to retargeting since the muscular system is similar in all human faces, and they yield impressive visual results due to the associated physical simulation of skin [SNF05]. However the reliable extraction of muscle parameters from raw facial motion data (video or 3D motion capture data) is a complicated process, and remains a research problem in itself.

Data-driven representations (see section 2.1.1.3) can be used as well, but : since the nature of their parameters statistically depends on the data they originate from, the parameters they deliver are not necessarily standard to all faces. However some semi-supervised data-driven methods segment the variations of the data into well-defined classes. Parameter values belonging to these classes can be transferred across different faces since they always describe the same components. Vlasic *et al.* [VBPP05] and Macedo *et al.* [MVBV06] presented results of successful transfer of 3D and 2D facial expressions respectively, mostly on real human faces. They both relied on multilinear analysis to separate the contribution of identity, visemes and emotional expressions to the global facial configuration. The expression alone could be transferred with an implicit adaptation to the target's face identity.

Unnatural results due to transfer of incompatible facial deformation are less likely to occur in parameter-based retargeting, since the source and target parameterization are supposed to be adapted to their respective face. However, it remains problematic in the case of atypical morphologies (nonhuman faces), because those approaches implicitly rely on prior knowledge of human facial structure. In the articles cited above, some examples are provided for nonhuman faces, however the structure of those faces was close to a human one. Even more abstract retargeting can be achieved by turning to the examples-based retargeting approach.

Examples-based Retargeting

Examples-based retargeting is another method of *indirect* retargeting, but does not depend on a standardized representation like the FACS or MPEG-4. Instead a customized analytical link is established between the configuration space of the source face and the configuration space of the target face. In practice, when associating given source and

target models, a specific mapping is constructed to transform a source expression to a corresponding expression on the target face. The adaptation of the geometrical and topological particularities between the source and target models is implicitly encoded in the mapping. In general, the mapping cannot be used for other models.

This method is often called *examples-based retargeting* because the construction of the mapping generally requires to set up an explicit correspondence between a few given facial configurations (the *examples*) on the source and target face. Based on these examples (which are grouped in what we call the *database*), the mapping learns how to transform from one specific face to the other, and can then extrapolate to new configurations.

Although the mapping can be built directly on global geometric coordinates of a facial mesh [KPL98], this generally results in bad numerical conditioning, because the number of variables is much higher than the number of equations provided by the correspondences. In practice, examples-based use a more compact representations of the facial configuration space. A popular representation used in early systems is blendshapes. In the blendshapes approach, a facial configuration (whether for the source or the target face) is decomposed into a combination of basis expression that form the blendshapes set. The coefficients of this decomposition form a compact representation of this expression. Pighin *et al.* [PHL+98] used such a method to clone and generate facial expressions from photographs. Natural facial expressions could be decomposed into a reduced set of expressions, and on the synthesis side 3D shape morphing and texture blending mixed the expressions from the database to reconstruct the expression on the virtual face. Kouadio *et al.* [KPL98] clearly formalized a retargeting technique between two matching blendshapes sets (one for the source, the other for the target). Facial deformations for the input face were automatically decomposed into the source blendshapes set, and mapped as such to blendshapes weights for the target model in real-time. Pyun *et al.* [PKC+03] and Park *et al.* [PCNS05] obtained better results by introducing the use of Radial-Basis Functions (RBF) along with linear basis functions in the computation of weights. It enables the construction of a nonlinear mapping of expressions, and a better preservation of the characteristic features of the example-expressions.

Blendshapes systems are intuitive compact representations, yet they rely on small, empirical and potentially non-optimal sets of basis expressions (in [PKC+03, PCNS05], only six expressions are used for emotional states: neutral happiness, sadness, surprise, fear and anger). More objective schemes, such as data-driven statistical models, are more likely to provide efficient and compact representations. Among those, Principal Component Analysis[2] (PCA) is by far the most popular. Principal components, which represent the contributions of eigenmodes of facial deformations, provide a good simplicity/optimality compromise to represent facial configurations.

Nevertheless, as for blendshapes sets, explicit correspondence between the source and target representations is necessary. In the case of PCA, the statistically detected principal components do not necessarily match across faces, so a mapping between principal

[2] A description of PCA can be found in section 2.1.1.3.

components is required to ensure the fidelity of the expression transfer [TMC07]. Deng *et al.* [DCFN06] use a PCA representation of 3D MoCap data to animate a blendshapes-controlled avatar. They manually establish a few well-chosen MoCap/Blendshapes correspondences to train the mapping. Byun [Byu07] presented a system with similar objectives, but with a video-based markerless capture system. The displacement of feature points is automatically detected and reduced to PCA parameters, which in turn are mapped to blendshapes weights to animate the target 3D face. Once again a mixture of linear and radial basis functions is used for the fitting of the mapping on six examples. For video sources, using the texture information in addition to feature points generally improves the quality of the compact representation. Hybrid points/texture models such as Active Appearance Models (AAM) are often used in retargeting from video [ZG05, CRRM07]. Wang *et al.* [WHL$^+$04] proposed a nonlinear reduction technique (Locally Linear Embedding, LLE) to construct the configuration space on which the retargeting is based. However they used a much reduced set of facial movements, and whether LLE can successfully interpolate between very different expressions is a debatable point.

Recent PCA-based schemes obtain impressive results of facial expressions transfer on either human or synthetic faces. For this type of retargeting scheme to be successful however, the configuration spaces of the source and the target must characterize the same scope of expressions. In particular their databases (collection of examples for a face) must correspond.

Most examples-based systems use empirical database, composed of only a few stereotypical expressions to establish the correspondence between the source face and the target face. The six universal expressions of emotions [EF71] (anger, disgust, fear, happiness, sadness and surprise) are used in many studies for that purpose [PKC$^+$03, PCNS05, ZG05, Byu07].

Deng *et al.* [DCFN06] relied on much more examples (approx. 40) for a richer output animation. However the examples were chosen empirically, and partly to ensure good performances of the RBF network. Cosker *et al.* [CRRM07] isolated individual, local facial actions which served as correspondence examples and were included in the blendshapes database of the target avatar. Despite improvements on the expressive range, this imposes increased database creation efforts and loss of natural facial correlation information. In their real-face-to-real-face retargeting system, Theobald *et al.* [TMC07] proposed a method to automatically find correspondences, and compute the mapping between two unlabeled facial databases. Yet, their method implicitly assumes that facial expressions produce similar deformation patterns on different subjects, which is a very crude approximation.

Examples-based retargeting seems to imply more manual work than previous methods, since a database needs to be created for each new face. On the plus side, examples-based retargeting adapt to any kind of target face (including nonhuman ones), provided that a database can be constructed for that face. Moreover, correspondence-based approaches can retarget movements on very disparate structures thanks to the purely

semantic correspondence that is set between examples. Baran *et al.* even showed that one type of movement can be retargeted to control a "semantically different" type of movement on the target model. In [BVGP09], they successfully convert arm movement to an equivalent leg movement on a different target model.

The retargeting mapping is undeniably more reliable when based on explicit corresponding examples. However, find the best set of examples for the correspondence database is an issue that has been largely overlooked. Chuang and Bregler [CB02] rightly based the quality of their retargeting on the choice of the right examples. They proposed three criteria to choose the right examples from a database of human facial expressions: The expressions that form the convex hull of the configuration space, the representatives of automatic clustering of that same space, and the representatives of PCA deformation patterns. In section 4.1.1 we proposed a new method to optimally select the right examples for corresponding facial databases. In the next section we show how this correspondence enables the construction of an efficient expression retargeting scheme.

4.2.2 Expression Retargeting in Appearance Spaces

In this work we choose to rely on an examples-based retargeting framework (see previous section). Examples-based retargeting requires setting up an explicit correspondence between prototypical configurations of the source and the target faces. It involves more manual work but results in improved robustness and accuracy, and more importantly adapts to any kind of target face.

The ideas developed in section 4.1 have lead to the construction of analogous appearance spaces for the human face and the synthetic face. Both spaces are connected, since the construction of the avatar appearance space is based on elements replicated from the human database. It follows that we have a correspondence between points in the human appearance space and points in the avatar space. Our purpose here is to use this sparse correspondence to construct an analytical link between both spaces. It can be noted that the modeling scheme we use, whether AAM or PCA, are linear (equations 3.1, 3.2 and 3.3). Linear variations and combinations are thus preserved by the modeling steps. To maintain this linear chain, we applied a simple linear mapping on the parameters of the appearance spaces:

$$(\mathbf{c_a} - \mathbf{c_a^n}) = \mathbf{A}.(\mathbf{c_h} - \mathbf{c_h^n}) \tag{4.1}$$

where $\mathbf{c_h}$ and $\mathbf{c_a}$ represent the coordinates in the source and the target appearance spaces respectively. $\mathbf{c_x^n}$ represents the coordinates of the neutral expression. $\mathbf{A}$ is the matrix obtained through linear regression on the set of corresponding points. Although Linear regression seems to be perfectly adapted, it has to be handled correctly. Straightforward Ordinary Least Squares (OLS) can be used to compute a matrix $\mathbf{A}$ that minimizes the regression error (equation 4.2). A classical closed-form solution for $\mathbf{A}$ is given

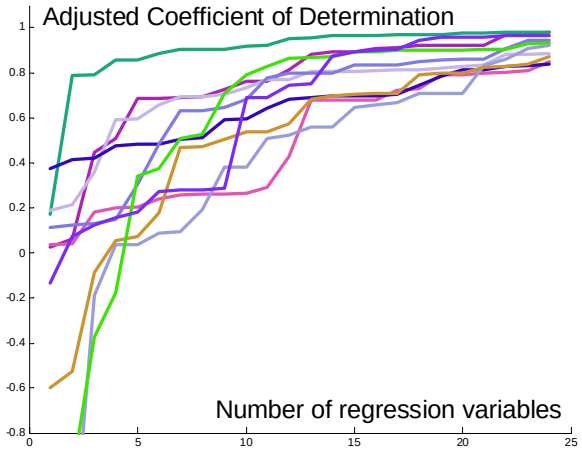

Figure 4.3: Adjusted Coefficient of Determination (ACD) of the regressand for different regressors sets. The regressors set marked n on the x-axis is formed by the first n principal components.

by the pseudo-inverse (equation 4.3).

$$\mathbf{A} = arg\,min\,\|\mathbf{C_A} - \mathbf{C_H}.\mathbf{A}\| \tag{4.2}$$

$$\mathbf{A} = (\mathbf{C_H}^T.\mathbf{C_H})^{-1}.\mathbf{C_H}^T.\mathbf{C_A} \tag{4.3}$$

Where $\mathbf{C_A}$ and $\mathbf{C_H}$ are respectively composed by the concatenation of vectors $\mathbf{c_a} - \mathbf{c_a^n}$ and $\mathbf{c_h} - \mathbf{c_h^n}$ for all database correspondences.

Despite the simple structure and formulation of a linear regression problem, the results are not always as good as expected. It turns out that even in cases where linear regression is appropriate, it must be carefully adapted to the context. Indeed, multiple factors can significantly affect its performances, such as collinearity problems and use of irrelevant regression variables [RPD01]. Collinearity of regression variables is not an issue here: since those variables consist of principal components, they are orthogonal to each other by construction. Regression variable selection, on the other hand, is an issue worth investigating in our case. Indeed, we observed that picking different sets of regressions variables produce sensibly different results. A definite observation is that performing regression on the whole AAM parameter set results in degraded retargeting results. The problem comes from the fact that some parameters are not contributing to facial movement and/or are corrupted by noise. Those variables bias the formulation of the regression problem, and should therefore be excluded from the regression set.

The instability caused by irrelevant variables can sometimes be revealed by surprisingly large regression coefficients. A more to detect irrelevant variables is to use the Adjusted Coefficient of Determination (ACD). The ACD measures the proportion of

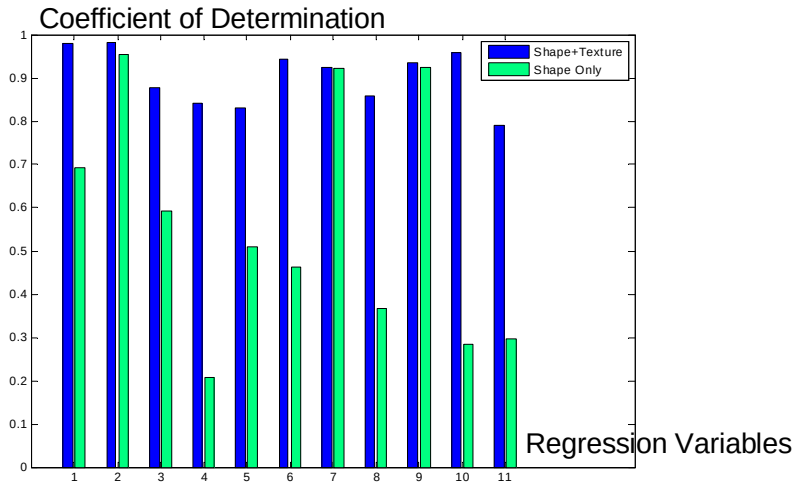

Figure 4.4: Comparison of the regression framework for expression mapping. **Left blue bars:** Adjusted Coefficient of Determination of the regression of AAM coefficients (shape+texture). **Right green bars:** ACD of the regression of PCA coefficients of shape information only. The use of shape and texture improves the ACDs of all regressand. A particularly significant improvement can be observed for the 4^{th} regressand.

regressands' variance that is covered by a given set of regressors[3].

Fortunately, our specific problem does not require testing all possible regressors set. Indeed, the PCA components are ordered by decreasing variance of the facial motion database. Consequently, the first components tend to encompass an important part of facial motion, while the last ones are less likely to bring a relevant contribution. We therefore test different regression models by successively adding regressors according to their ordering in the PCA output. The evolution of the ACD for each regressand is presented on figure 4.3.

The ACDs of the regressands increase strongly when the first principal components are added as regressors, which generally implies an improvement of the regression model. This increase however reaches saturation: above 13 principal components the ACDs do not significantly evolve anymore. Meanwhile the addition of those regressors augments the complexity of the regression model, and the risk of distorted regression coefficients. Therefore, in that specific scenario the choice of the first 13 components as regressors seems reasonable and can be confirmed by qualitative results (figure 4.6). Setting up the right regression equation must theoretically be done for each retargeting scenario; the study of the ACDs constitutes an objective and efficient procedure to solve this issue.

Another interesting observation we can make from the study of ACDs is the impact

[3]Here we use the adjusted version of the coefficient of determination in order to remove the impact of the number of regressors on the traditional coefficient of determination [RPD01].

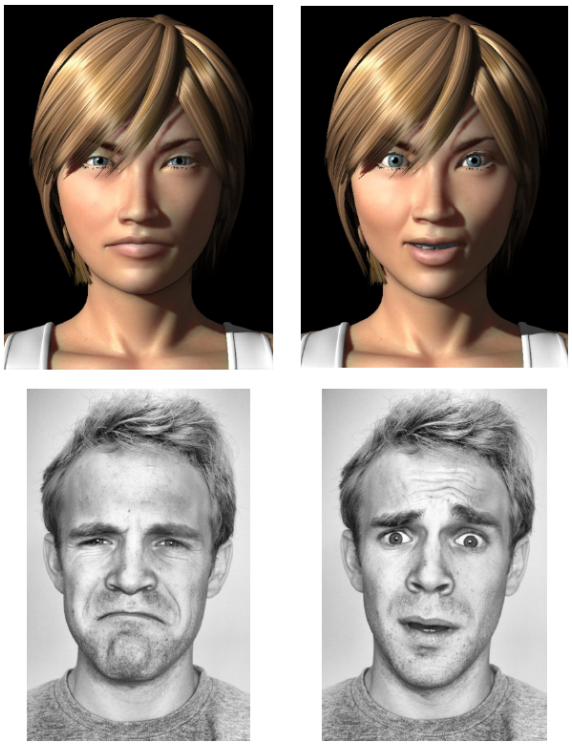

Figure 4.5: Influence of the 4^{th} regressand (PCA component) on the face of the target virtual character. On human faces, this expression creates visible wrinkles on the cheeks and around the eyes. Regression models using texture information perform better expression mapping in that case.

of using texture variations as regressors. A recurring argument in favor of AAM and equivalent models for facial motion is the fact that texture brings additional information that can hardly be captured by discrete markers. For a retargeting framework, these visual cues are completely relevant in that some crucial motion indicators such as wrinkles can only be captured by a textural component. In this perspective, we present an interesting quantitative comparison of two regression framework for expression mapping with the help of their ACDs. The first one uses AAM components mixing shape and texture information, while the second relies on PCA coefficients of shape information only. The results, presented on figure 4.4, confirm the intuition that adding texture information to a sparse marker set significantly improve retargeting performances. This improvement is particularly important for the 4^{th} component of the virtual character's parameterization. On figure 4.5, we see that this 4^{th} parameter controls expressions which theoretically involve cheeks- and eye-wrinkles. This explains the capability of texture information to improve the parameterization of this particular motion.

Retargeting results are illustrated by a few snapshots on figure 4.6. Complete se-

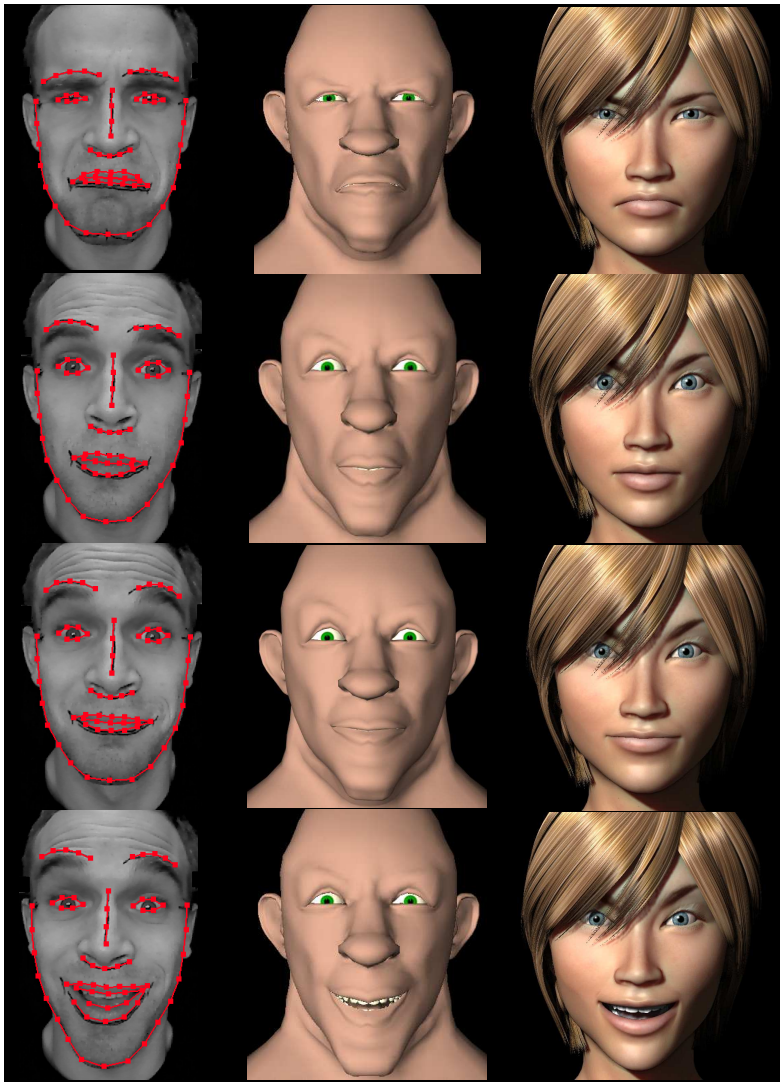

Figure 4.6: Examples of facial expressions retargeting. The expressions captured on the human face (left) are successfully transferred to the faces of two 3D avatar (middle and right) for which optimal databases were created. Dynamic retargeting examples can be found on the accompanying video. Middle character used with permission of the Institute of Animation Baden-Wuerttemberg.

quences of expression retargeting can also be found on the accompanying video[4]. Note that the source and target faces can be 2D or 3D. With the source vector $\mathbf{c_h}$ in equation 4.1, we can allude to the PCA parameters of 3D MoCap signals, or the AAM parameters of 2D video capture. Similarly, the target vector $\mathbf{c_a}$ can refer to the PCA

[4] `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php`

parameters of 3D face geometry, or the AAM parameters of 2D facial image. This can be observed in figure 4.6, in which facial expressions from a 2D video source are mapped on both 2D and 3D synthetic faces.

As far as computational complexity is concerned, the simple form of the retargeting method enables real-time performances. As mentioned earlier, all processes involved in the expression mapping chain are linear. Our system linearly maps appearance parameters $\mathbf{c_h}$ to $\mathbf{c_a}$, which in turns consist of linear transformation of facial features (see AAM equations 3.1, 3.2 and 3.1). The computational complexity is mostly influenced by the number of features of the source and the target face [5]. For a precise AAM source model and a highly complex target 3D face of 56k vertices, the retargeting computation requires approximately 20ms per frame, which corresponds to a real-time performance of 50 frames per second.

The examples of figure 4.6 illustrate the reliability and efficiency of the method described in this section. Expression mapping between appearance spaces also constitutes a cheap and practical solution for performance-based animation, as it enables the retargeting of a large variety of facial expressions from a simple video source.

Appearance spaces are not limited to the transfer of captured data, but can also be used as a navigation space to generate new expressions sequences. One limitation of working in the appearance space however, is that it is not intuitive to manipulate its dimensions. Next section develops the subject of intuitive manipulation of facial expressions based on appearance spaces.

4.3 Low-Dimensional Control Space for Facial Expressions

Real facial expressions are phenomena caused by the contraction of numerous muscles whose interaction produces complex facial deformations. It is thus an interesting challenge to design simple and intuitive control parameters that can encapsulate this complexity. As we have seen above, expressions can be generated by synthesizing the facial deformations corresponding to coordinates in an appearance space. Yet, as mentioned in the state-of-the-art report in section 2.2.2.2, many studies have relied on reduced spaces such as our appearance spaces, and reported a limited edition power. Indeed, appearance spaces generally have a high dimensionality and their dimensions rarely exhibit intuitive facial deformation modes.

In section 3.2 we introduced a compact and coherent structure, the manifold of emotional facial expressions, which was extracted from the human appearance space. We now propose to use this space to manipulate facial expressions in a simple and intuitive way. This approach differs from how previous practical systems have handled high-level expression synthesis, as explained in the following section.

[5] Facial features are the geometric position of vertices, or the intensity value of pixels and the case of AAM modeling.

4.3.1 Background on Intuitive Control of Facial Expressions

Manipulation and control of facial expressions has traditionally been carried out using fine parameterizations of facial deformations. Typically, animators used approaches such as direct parameterization schemes, pseudo-muscles or physics-based muscle systems (see section 2.1.1.1 in the state-of-the-art chapter). Data-driven parameter spaces such as our appearance spaces forms an efficient and accurate facial parameterization.

These systems are based on different philosophies, but ultimately they all deliver a set of low-level parameters one has to manipulate to generate the desired expressions with a given degree of realism. Unfortunately, they lack the simplicity to support the intuitive creation and edition of facial animation. The obtained parameter sets are generally large, complex and can even hold conflicting deformation patterns causing unnatural facial expressions. Manipulating the parameters correctly requires hours of practice and really only professional animators manage to use them efficiently. Even then, the construction of an animation sequence is a long and tedious process.

Some studies have proposed to use simpler structure to allow non-expert users to manipulate facial expressions consistently. They offer to control expressions based on high-level, meaningful concepts rather than on low-level mechanical deformations. In the case of emotional expressions, their idea was to rely on dimensional representation of emotion such as those presented in the state-of-the-art report (section 2.1.2.1). These "emotion spaces" are particularly adapted to manual control by non-expert users, because they use simple structures and deal with high-level, meaningful concepts. In that regard, they can also be used for human-machine interaction or affective computing applications.

In practice, previous works have look to mathematically associate the emotion spaces with low-level parameterizations of facial deformations. The association is typically constructed in supervised fashion. Some key emotions and their corresponding facial expressions are first manually associated; new expressions can then be generated as combination of these key expressions, using the dimensions of the emotion space as interpolation parameters. Ruttkay *et al.* [RNtH03] constructed an intuitive interpolation space for an avatar's facial expressions based on a bilinear interpolation between the six universal emotional expressions [Ekm82]; the expressions were spatially distributed according to Schlosberg's emotion circumplex [Sch52]. Tsapatsoulis *et al.* [TRK+02] proposed a similar application but used a mixture of two emotion spaces for the interpolation, Plutchik's emotion wheel [Plu80] and Whissel's activation-evaluation space [Whi89]. They used the angular coordinate of Plutchik's wheel as a measure of similarity between expressions, and Whissel's activation dimension to interpolate the expression's intensity. For the actual low-level animation of the avatar, Tsapatsoulis *et al.* relied on MPEG-4 Facial Animation Parameters [EHRW98]. Albrecht*et al.* [ASHS05] later extended this approach using Cowie's disk-shaped activation-evaluation space [CES+00] as interpolation space. In their case, the low-level deformations were performed by a physics-based muscle system. Zhang *et al.* [ZWMC07] developed a facial expression synthesis system based on Mehrabian's Pleasure-Arousal-Dominances space (PAD) [MR74].

They trained a mapping between PAD parameters and low-level deformations (MPEG-4 Facial Animation Parameters) using a database of expression examples with known PAD values [LAKG98]. New expressive faces could then be synthesized by varying the high-level PAD parameters. In the latest work to date, Arya *et al.* [ADP09] fine-tuned the idea of facial expression interpolation in a 3-Dimensional emotion space (valence-arousal-agency in their case). They relied on a user experiment to associate a larger number (≈60) of full-blown and mixed emotions to low-level facial actions for a more precise animation system. Based on 3D emotional coordinates, the facial actions were combined using a fuzzy-rule system.

These straightforward approaches obtain interesting results, yet the simplicity of the interpolation and the reduced number of examples cannot fully encompass the complexity of facial deformation patterns. More importantly, they rely solely on abstract models of human emotions. Yet, these emotional spaces were constructed from psychological and social aspects only, with no concern for their mechanical aspect. It is not guaranteed that the topology of the emotion model matches the one of the facial parameters space (whether MPEG-4 FAP, muscle system, *etc*). In particular, close instance in the emotional space may correspond to very different parameter values in the facial parameters space. This is illustrated in the emotional space used in [DBW+07] in which the representatives of "anger" and "fear" as well as "sadness" and "disgust" share almost the same emotional coordinates, whereas they correspond to very different facial configuration parameters. The design of a simple mapping between emotional coordinates and unstable parameters like muscle contractions may lead to important distortions.

4.3.2 Facial Expression Manifold as a Control Space

Instead of constraining the facial deformation systems to adapt to a possibly incompatible control space, we propose to use a control space that arises directly from low-level considerations. In this regard, the visual representation presented in section 3.2.3 constitutes an interesting interface to *interact* with emotional facial expressions. Indeed, it originates from real data and is thus intrinsically coherent with mechanical deformations implied by expressions. Yet, its simple structure is also adapted for high-level, meaningful comprehension and manipulation of facial expressions. Compared to state-of-the-art solution, this approach provides a more accurate control as well as a more consistent framework for animation and interpolation of facial expression.

The visual representations of section 3.2.3 derive from the non-metric embedding scheme we introduced in this thesis. The outcome consists of a one-to-one correspondence between dominant directions in the human appearance space, and their symbolic low-dimensional representations (see figure 3.13). It is thus only a discrete link between the low-dimensional interface and the high-dimensional space it represents. For synthesis and control purposes, it is necessary to construct a dense, analytical link between the interface and the appearance space. We build this link using a regression method: based on the discrete correspondence between low- and high-dimensional directions, as

well as the correspondence of neutral expressions, we fit a mapping from the interface to the appearance space. This control mapping is illustrated as "CMh" in figure 4.1.

Many fitting schemes can be considered for this purpose, like traditional regression methods (linear, polynomial, *etc*). A piecewise linear mapping linking the simplices of the interface and their high-dimensional counterparts is a good solution. Another interesting option is the Thin-Plate Spline (TPS) approximation, which can be viewed as an extension of traditional splines [Boo89] to higher dimensionalities. The attractive characteristic of TPS is to present a tradeoff between the accuracy of interpolation of the known samples and the smoothness of the mapping. Since the control space is supposed to be continuously navigated, the smoothness criterion is an important one in our case. In addition, due to the local coherence of the appearance space, a 100% exact interpolation of the samples is not necessarily relevant. Piecewise linear approximation would perfectly interpolate the known samples. However it produces visible variation discontinuities on the target face when navigating in the space (typically when switching from a simplex to another). With a TPS mapping, we allow a limited and practically invisible interpolation error in order to accentuate the smoothness of the mapping, and improve the comfort of navigation in the control space. Experimentations have been made with a more elaborate regression method, namely gaussian process regression. Despite the advantages of a non-parametric approach, the computational burden it implies in real-time contexts eclipses the flexibility that it brings to the regression process.

The correspondence between high- and low-dimensional dominant directions on which the control mapping is built concerns the human appearance space (figure 4.1c). When a point is selected in the control space, its coordinates are mapped to a high-dimensional vector in the human appearance space. One can note that this vector actually lies on our manifold of facial expressions. The corresponding expressive facial image can subsequently be synthesized by reverting the AAM equations. This allows users to manipulate the face the interface originates from (figure 4.1a). This point is illustrated in figure 4.7.

Our interest is obviously to allow users to control other faces with the interface. A straightforward solution would be to connect this control framework to the retargeting system described in section 4.2. After mapping a control point to a vector in the human appearance space, this vector would be mapped to the target appearance face using equation 4.1, and finally the expression would be synthesized on the target face. Although valid, this system is computationally not optimal. Equivalent results can be obtained addressing the target's appearance space directly. In facial animation context, the target is usually a synthetic character.

4.3.3 Intuitive Control of Synthetic Faces

The control interface presented in the previous section is based on a correspondence with the dominant directions of the human appearance space. In section 4.1 we showed how an equivalent appearance space could be constructed for a synthetic face from a reduced expression database. If the facial expressions that correspond to human

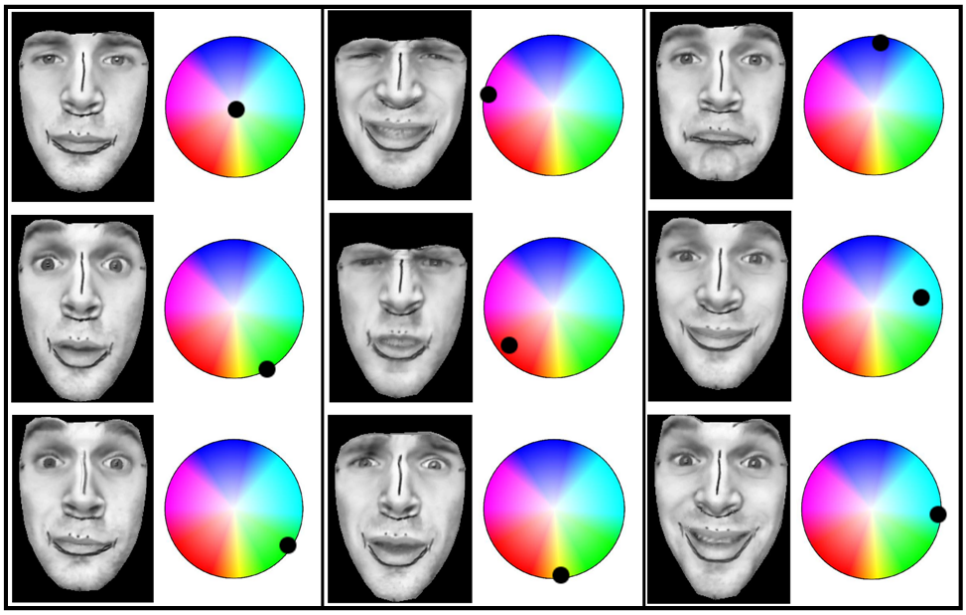

Figure 4.7: Facial expression synthesis on a human character. By selecting a point in the visual representation of the facial expression manifold (right side on each example), the system addresses a point in the human appearance space and can synthesize the corresponding expression.

dominant directions are integrated in that database, the embedded directions of the control space will have corresponding directions in the *avatar* appearance space as well. A mapping similar to the one described earlier can be constructed between the visual manifold interface and the appearance space of the avatar. It is illustrated by the "CMa" arrow in figure 4.1.

Once the control mapping "CMa" is constructed, any point of the low-dimensional control space addresses a corresponding mapped point in the avatar's appearance space, and can then trigger the synthesis of a facial expression on the virtual character. In figure 4.8, we illustrate the synthesis of new facial expressions on the synthetic face by selecting points in the 2D control space. As mentioned in section 3.2.3.2, the dominant expressions of the appearance space have been distributed on the unit circle, thus the angular parameter represents the type of expression being synthesized. By varying the radial distance at a fixed angle, we adjust the intensity of the displayed expression between extreme (on the edge of the disk), and neutral (at the center of the disk). The same observation can be made for the 3D control space: the different expressions are distributed on the surface of the unit sphere, and the intensity is represented by the radial distance. For clarity of representation, the 3D control space is only illustrated in the accompanying video[6].

[6] `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php`

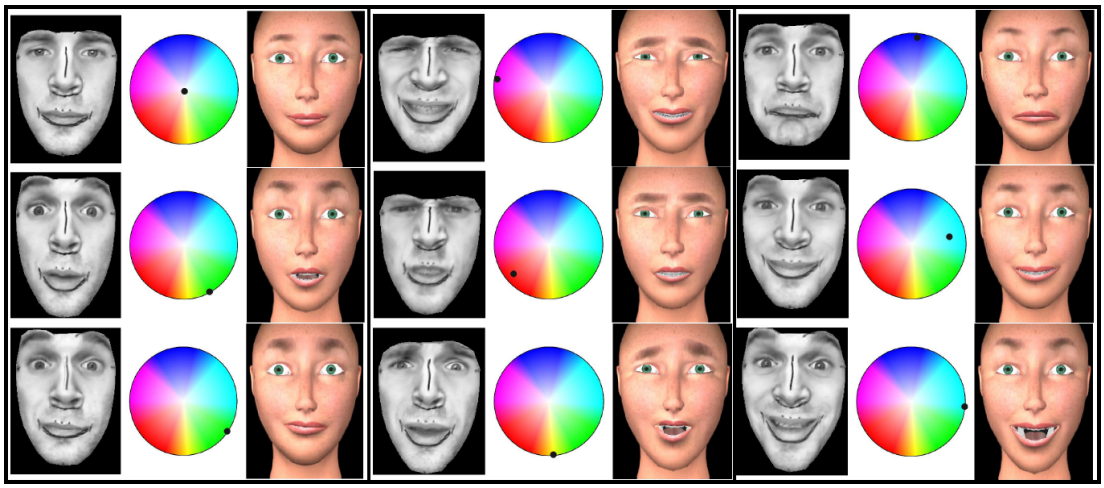

Figure 4.8: Several examples of facial expression synthesis from the 2D control space. Selecting a point in the control space (located the middle on each example) triggers the synthesis of the corresponding facial expression on both the human and the 2D virtual character (trough the control mappings). The nature of the expressions on both entities corresponds perfectly, meaning that the knowledge of the behavior of the human face has been successfully transferred to the synthetic one. Expressions from the database as well as unseen, mixed facial expressions can be successfully generated.

The interesting aspect of this framework is that it reuses the formulation of the facial expression manifold derived from a well-known face, and adapts it to other faces. Indeed, the detection of the dominant directions (section 3.2.2) as well as the embedding of the dominant directions in the 2D space (section 3.2.3) is performed in the human appearance space, which benefits from a much richer database of examples. Yet, once the embedded directions are known, they can be mapped on the directions of any appearance space, typically those of an avatar. This can be thought of as the transfer of knowledge: The knowledge of the nature and organization of the dominant expressions is transferred from the studied human face to the synthetic face. This connection between a human and a synthetic face with the expression manifold as common is well illustrated in figure 4.8.

Obviously stereotypical expressions present in the databases are well reproduced, but it is important to note that unseen, mixed facial expressions that were not included in the database can be synthesized as well. This generalization capacity is linked to the PCA and AAM modeling's ability to learn from a few examples and successfully extrapolate the appearance changes for unseen configurations. Unlike approaches which control underlying mechanisms like muscle contractions [ASHS05], for which mixing parameter values can lead to unnatural facial configurations, the manifold we use is very well adapted to the continuous representation of natural facial expressions. The overall coherence of the control space makes our system not only relevant for the synthesis of

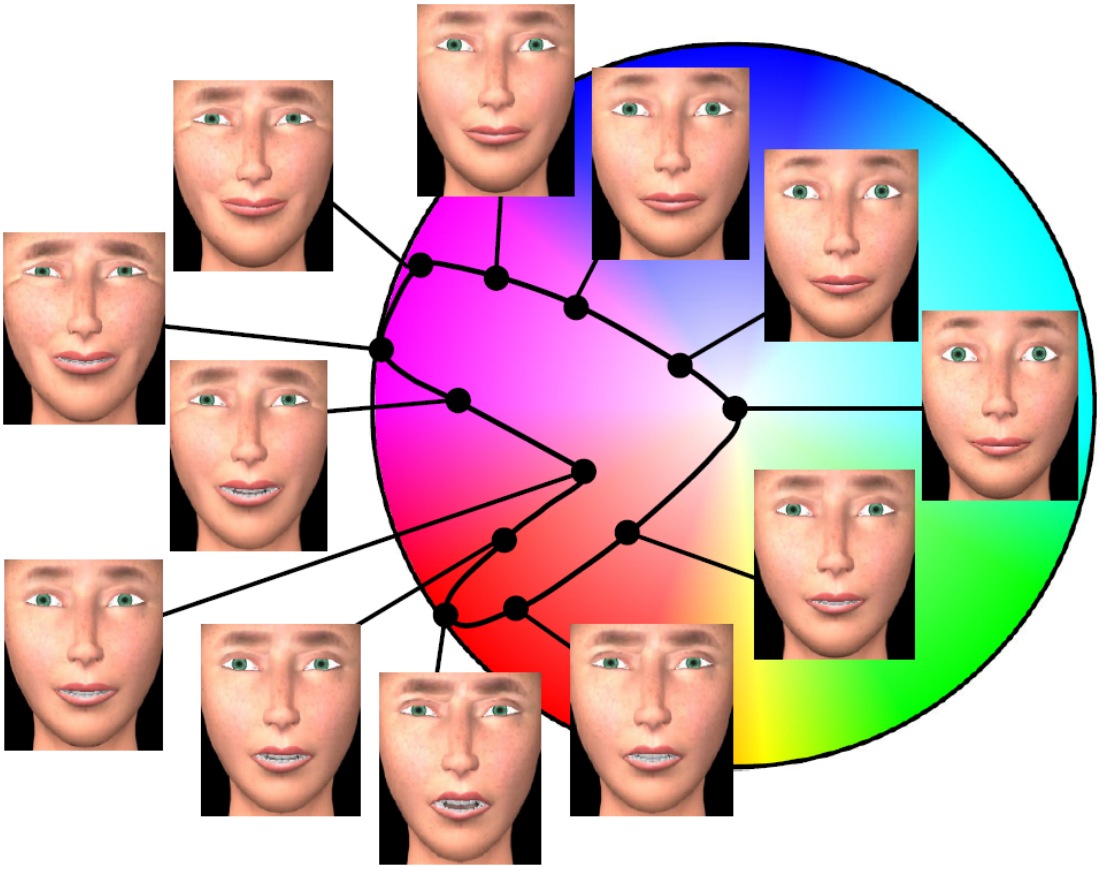

Figure 4.9: Synthesis of facial animation sequences. The continuous nature of the control system allows the synthesis of coherent expression sequences. Along paths in the control space, the face is animated with no perceptible discontinuity, and no unnatural facial expressions. This characteristic offers an important advantage over traditional facial control systems.

still facial configurations, but also for the generation of dense facial animation sequences where the frames correspond to a trajectory in the control space. This use of the system is presented in figure 4.9 and in the video[7].

As an animation tool, our control interface based on the manifold of facial expressions combines several advantages. It provides users with the flexibility and the precision of parametric animation systems, with a more consistent interface than traditional emotion-driven spaces. Additionally since it originates from the analysis of real facial data, it holds the naturalness of human motion.

One drawback of the control system is that the low-dimensional control space obviously cannot address all possible facial configurations of the high-dimensional appear-

[7] `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php`

ance space. Indeed, the 2D control disk addresses a 2D manifold in the facial appearance space. The 3D control sphere addresses a 3D Manifold which allows the animator to reach more varied facial expressions. The catalog of accessible facial expressions could be extended further by using a control space of even higher dimensionality, but then the simplicity of use of the control interface would be lost.

Another severe limitation of this scheme as is, is that the appearance space and the facial expression manifold do not explicitly encompass temporal information. As observed in the introduction however, dynamics is a crucial component of the naturalness of emotional facial expression. Unfortunately, this aspect cannot be left untreated: simple trajectories of the control space such as the one in figure 4.9 do not in general produce correct facial dynamics, and this clearly impairs the naturalness of the resulting animation.

As this state, the control interface is already an interesting tool for animators. They can use it to efficiently navigate though the facial expression space, and draw trajectories to setup facial animation sequences. Facial animation thus becomes a more straightforward process, accessible to non-expert users. However, for perfect results the animators will have to manually edit the temporal sampling of the trajectory, in order to ensure realistic dynamics. This task is actually very similar to their traditional keyframing workflow, only this time the creation and edition of keyframes is greatly facilitated by the use of our interface.

There are applications -typically real-time interactive application- where such fine manual control is not possible. For those applications, an actual objective formulation of facial expression dynamics is necessary. Next chapter focuses on the detailed temporal analysis of facial expressions dynamics, and how natural dynamic expressions can be synthesized in real-time on virtual characters.

Chapter 5

Dynamics of Emotional Facial Expressions

Contents

5.1 Facial Dynamics Problematic 116
5.1.1 Dynamics for Real-time Applications 116
5.1.2 Global State-spaces 118
5.1.3 A Local Approach 120
5.2 Modeling Motion Dynamics 121
5.2.1 Motion Signals Analysis 121
5.2.2 Computational Models of Motion 122
5.2.2.1 Nonlinear Dynamics 122
5.2.2.2 Motion Variability 124
5.2.3 Coordination 126
5.3 Results and Evaluation 128

The preceding chapters have studied the *spatial* aspect of facial expressions. In that sense, they focused on what natural facial expressions imply in terms of spatial configuration of facial elements, and how these configurations can be represented and manipulated. In chapter 4, we proposed a new structure based on the natural manifold of facial expressions to easily manipulate these configurations. In that same chapter however, we observed that good manipulation of facial deformations is not sufficient to guarantee realistic facial animation. The generated animations of section 4.3.3 consisted of sequences of spatially coherent expressions, but they missed the natural dynamic signature of human motion. This often resulted in robotic-looking animations. That verdict is not groundbreaking: numerous studies, mostly in the psychology and medical fields, have already reported the crucial role played by *temporal* dynamics in our perception of facial expressions. They observed that human observers can perceive subtle dynamic nuances in facial movements, and also differentiate natural from unnatural dynamic expressions [Edw98, BM08].

Although a large consensus seems to arise on the importance of the temporal aspect of facial expressions, this particular problem has seldom been tackled from a computational perspective. This chapter is dedicated to our study of the analysis and the synthesis of natural facial dynamics. We begin by exposing the problematic of facial expression dynamics in section 5.1. We subsequently present our contribution: in section 5.2, we describe a new facial animation system adapted to the requirements of interactive applications. The system relies on a set of human motion models, each of them controlling a specific part of the face. The models are trained on a database of captured expressive facial movements and learn the dynamic signature of human facial expressions. They also contain a stochastic component that produces non-deterministic facial movements and improves the naturalness of the long-term visual behavior. Using a retargeting framework (section 4.2), the resulting motion can be efficiently transferred to any synthetic face in real time.

5.1 Facial Dynamics Problematic

5.1.1 Dynamics for Real-time Applications

Most of today's parametric animation systems remain essentially focused on the spatial aspect. Regardless of the parameter space used, whether traditional ones (blendshapes, pseudo-muscles, *etc*) or more intuitive ones such as the control space presented in the preceding chapter, the temporal aspect of facial expression is not addressed. The dynamic consistency of a generated animation in that case is an issue left to the animator. Artistic skills and experience allow those animators to reproduce lifelike movements, at the expense of a long and tedious manual work. In recent years, to ensure realistic dynamics more reliably, modern productions have massively turned to motion capture data to animate virtual characters. This more objectively guarantees the realism of the result, as the natural temporal signature of human motion is encompassed in the recorded data. However, this moves the focus away from understanding the dynamics of facial expression.

Specifically designing animations destined to predetermined situations is not a suitable approach in all cases. Interactive applications require generating contextualized motion sequences at runtime, and thus cannot entirely rely on manufactured or recorded animation. An approach promoted by both the research and the industrial worlds to cope with this issue is performance-based animation. In our state-of-the-art report of chapter 2, we presented how performance-based systems adapt motion data to the varied situations an interactive application can encounter. They do so by assembling and editing small motion data segments to cope with the applications real-time needs. This continues to preserve natural motion dynamics, as the temporal signature is intrinsically contained in all motion segments. Nevertheless, this approach still does not explicitly model nor does it investigate the nature of the temporal signature of facial expressions. Since performance-driven schemes are limited to movements contained in their database, they obviously cannot exhaustively cover the large set of human motion.

Moreover, dealing with a rich scope of varied movements in a database quickly leads to significant capture costs and data storage issues.

Recent studies have favored another approach by using mathematical models of motion. This time the models are explicitly meant to model the temporal evolution of motion signals. They generally still require a motion capture database to *learn* the faces' dynamic behavior, but once trained they can be used single-handedly to generate a wider variety of movements.

This approach tends to be legitimated by interesting temporal characteristics observed on gait and facial motion. Schmidt and Cohn, for instance, conducted in-depth studies of facial expression dynamics, focusing on the dynamics of the smile [SC01, SC03]. Their study of electromyography signals identified both recurring temporal patterns in facial motion on multiple individuals. In particular, smile onset and offsets phases displayed stable dynamic properties in terms of profile and duration. They also observed repetitive patterns in the coordination of different facial actions during a smile. These results tend to indicate that certain dynamic invariants exist. Those invariants much likely form a dynamic signature that we, humans, are used to observing on real face, and associate to natural behavior. Learning-based approaches make sense in that perspective. They look to learn this signature and ensure that it is preserved in the animation mechanism and the generation of new sequences.

Typical motion models used in the literature rely on the formalism of Linear Dynamic Systems (LDS) [GCT+04, CH07, DM08]. LDSs model the dynamic evolution of an entity, assuming that its current state is linearly dependent on its previous states. If the state of an entity at a discrete time instant n is described by vector $\mathbf{x_n}$, the action of an LDS can be concisely expressed by the following equation:

$$\mathbf{x_n} = \sum_{i=1}^{m} A_i.\mathbf{x_{n-i}} + B.\mathbf{u_n} + \mathbf{v_n} \tag{5.1}$$

where m is the order of the LDS, $\mathbf{x_n}$ is the state of the entity (body, face, or individual elements thereof), $\mathbf{u_n}$ is the control input, and $\mathbf{v_n}$ is an independent noise at timestep n. The matrices $(A_i)_{i\in\{1,\dots,m\}}$ actually model the dynamic behavior of the entity. The temporal evolution of the entity is also influenced by external control values $\mathbf{u_n}$ through matrix B. This component is used to adapt the motion to new constraints imposed by the real-time applications' needs.

Learning-based motion models offer the most promising approach for interactive motion generation. They learn the motion's temporal signature from real data, but formalize it to a simple mathematical process, that can be used in more general cases. Consequently, they theoretically combine the realism of motion data and the flexibility required for interactive applications. Linear Dynamic Systems are used extensively in that regard, for the most part because of their conciseness, their stability and the simplicity of their training procedure. However, the question whether such simple state

equations can efficiently model the dynamics of human movement has been raised repeatedly over the last few years. These doubts recently lead to the conception of more complex motion models (see section 2.2.3.2). In this study, we additionally question the use of global state-spaces to efficiently describe a large range of motion and its inherent variability. This specific issue has received significantly less attention, yet represents an important aspect of motion modeling. We develop this point in the following section.

5.1.2 Global State-spaces

As mentioned in the state-of-the-art report of section 2.2.3.2, many previous studies use dynamic system formulations like equation 5.1 as global models for the movements of their entity of interest. In the case of faces, it means that $\mathbf{x_n}$ describes the global facial configuration at time instant n. In practice, instead of directly addressing the positions of the vertices of the facial mesh, or pixel values of the facial image, researchers often promote working in a reduced space that accounts for correlations and redundancies in the original parameterization. In most cases, the components of $\mathbf{x_n}$ are formed by PCA components of the facial geometry [CH07], or by AAM appearance parameters [GCT+04]. An AAM-based parameter space was used in chapter 3 for instance.

The important point is that in those equations $\mathbf{x_n}$ represents the state of the entire facial configuration at once, which is why we refer to them as *global state-space* formulations. Global formulations are attractive because they account for the natural dynamic correlations between facial elements. Indeed, natural movement and expressions involve recurrent coordination patterns between individual elements[1].

In the studies presented in section 2.2.3.2, global state-space dynamics has proved efficient for the adaptation of a specific action to new temporal and spatial constraints. Yet, human movements, particularly faces, form a continuum of motion and not just a collection of isolated actions. Being able to represent changing motion patterns is thus desirable from a motion model, especially in the case of real-time applications. Unfortunately, global dynamic models generally have very narrow modeling scope. A reason for that is the complex spatiotemporal nature of the problem:

Faces consist of multiple elements whose movements are certainly correlated, such as mouth and cheek movements in the above example, but they are obviously not entirely constrained to each other. Their behavior and temporal coordination therefore varies, even for analogous actions. Assuming that a motion sequence is described by a sequence of PCA or AAM configurations vectors $\mathbf{c(t)}$.

$$\mathbf{c(t)} = \begin{pmatrix} c_1(t) \\ \vdots \\ c_l(t) \\ \vdots \\ c_P(t) \end{pmatrix} = \begin{pmatrix} t_{1,1} & \cdots & t_{1,k} & \cdots & t_{1,Q} \\ \vdots & \ddots & \vdots & & \vdots \\ t_{l,1} & \cdots & t_{l,k} & \cdots & t_{l,Q} \\ \vdots & & \vdots & \ddots & \vdots \\ t_{P,1} & \cdots & t_{P,k} & \cdots & t_{P,Q} \end{pmatrix} . \begin{pmatrix} s_1(t) \\ \vdots \\ s_k(t) \\ \vdots \\ s_q(t) \end{pmatrix}$$

[1] For instance a Duchenne smile, which is believed to indicate a spontaneous smiling expression, involve both raising the corners of the mouth and raising the cheek region around the eyes.

where the c_i represent the principal components, s_i the original visual features (facial shape or texture), and $\mathbf{T} = (t_{ij})$ is the PCA or AAM transformation matrix. This multidimensional trajectory $\mathbf{c}(\mathbf{t})$ can be modeled by a dynamic system such as the ones mentioned above. Global timing modifications of the facial feature do not affect the purely spatial trajectory of the motion in state-space:

$$\mathbf{T}.\begin{pmatrix} s_1(w(t)) \\ \vdots \\ s_k(w(t)) \\ \vdots \\ s_q(w(t)) \end{pmatrix} = \begin{pmatrix} c_1(w(t)) \\ \vdots \\ c_k(w(t)) \\ \vdots \\ c_p(w(t)) \end{pmatrix} = \mathbf{c}(w(t))$$

where the purely temporal variation can be modeled by a global time warping function $w(t)$. The trajectory $\mathbf{c}(\mathbf{t})$ remains the same spatially, only with a different temporal sampling. Now, in natural expressions, we might expect that the coordination of facial elements is not immutable. The temporal dynamics of each individual element then varies differently from the others. Mathematically, this can be represented by a different time warping function $w_i(t)$ for each component:

$$\mathbf{c}'(\mathbf{t}) = \begin{pmatrix} c'_1(t) \\ \vdots \\ c'_k(t) \\ \vdots \\ c'_p(t) \end{pmatrix} = \mathbf{T}.\begin{pmatrix} s_1(w_1(t)) \\ \vdots \\ s_k(w_k(t)) \\ \vdots \\ s_Q(w_Q(t)) \end{pmatrix}$$

The temporal synchronization between elements s_i is broken, which this time induces changes on the *spatial* trajectory $\mathbf{c}'(\mathbf{t})$ in state-space. Semantically speaking however, the action is still the same. Substantial and diverse variations of this sort cause multiple non-trivial changes in spatiotemporal trajectories, and it is doubtful that a single model can cover them all.

Although clearly favored in recent studies, the global state-space approach presents notable limitations. As we have seen, modeling one specific action and its different variations alone requires complex nonlinear models. In that perspective, finding and modeling a *generic* global dynamic signature that would generalize to all possible facial movements seems very unlikely.

As an alternative to the global approach, we propose to describe the dynamic properties of human facial motion with a *collection* of dynamic systems. The role of each of those systems is to model the movement of a single facial point directly in visual coordinates. The origin of this approach comes from the identification of an interesting dynamic signature on the motion signals of individual facial elements (see section 5.2.1). This approach prevents us from dealing with complex nonlinearities of global state-space models. Additionally, it enables us to use the modeled signature for the synthesis of *any* global movements, even those for which no training example was provided.

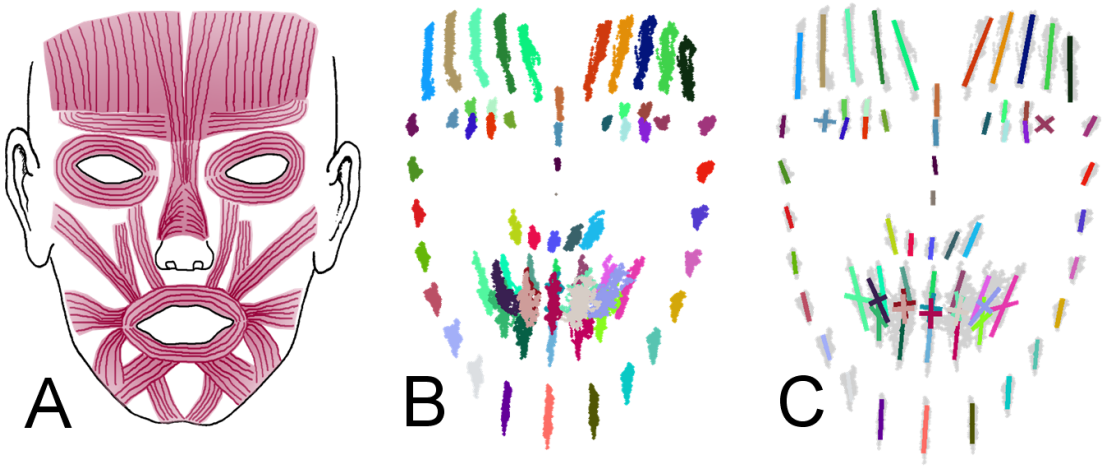

Figure 5.1: The constraints of expressive facial deformations. **A**: Locations of the principal facial muscles. **B**: Observed displacements of the 2D facial markers in the facial expression database. **C**: Identification of the dominant axes of marker displacements.

5.1.3 A Local Approach

Instead of global facial description, we are interested in observing the temporal behavior of local facial elements in straightforward visual coordinates (2D trajectories of selected facial markers, measured in the image plane). Our interest is in observing the short-term dynamics controlling the transition from one expression to another. To this end, we have collected a 2D database of varied dynamic emotional facial expressions performed by an actor and tracked the movement of a set of markers over time. Our goal was to efficiently model a wide variety of dynamic transitions, thus the database is not limited to on-off transitions such as what can classically be found on other databases (transition from neutral to expressive states, then back to neutral). We included numerous transitions between different expressive faces (sadness-to-anger for instance).

Interestingly, once rigid head motion has been canceled out, the visual trajectories of the markers are organized around rather well defined axes (see figure 5.1B). These translation axes correspond to the indirect effect of local muscle contractions (figure 5.1A). The majority of markers have their movement constraint on one dominant axis/muscle; others, mainly around the mouth, are under the influence of several muscles, so their movement has to be decomposed on two independent axes. In the following, we use these axes (displayed on figure 5.1C) to parameterize the displacement of the markers. The complexity of facial movements is then reduced, as it is considered as composition of local translational movements. This simplification will allow us to derive a *generic* animation system, which will not be limited to a small number of global facial actions.

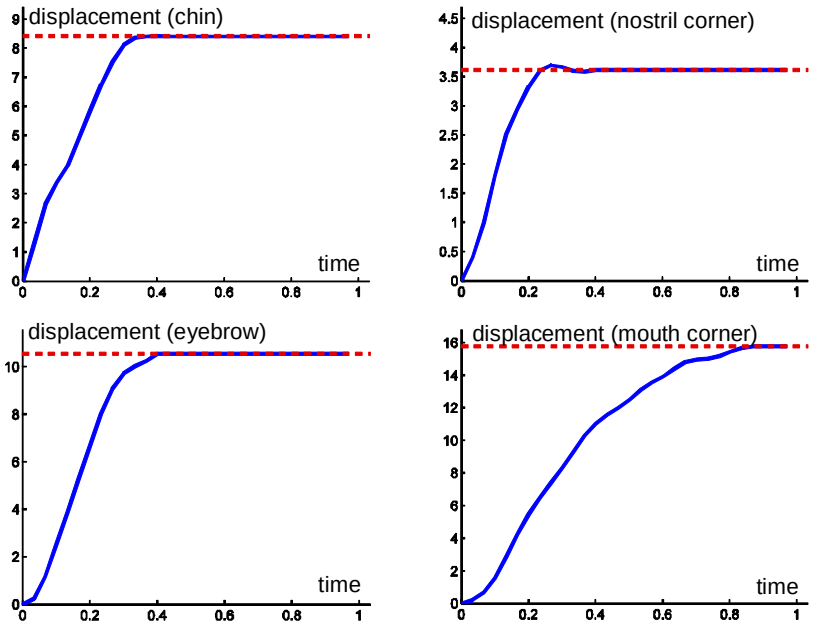

Figure 5.2: Temporal profile of marker displacements along their dominant axes *(blue solid line)*. The dashed line shows the final displacement value.

5.2 Modeling Motion Dynamics

From this point on, we observe the marker displacements along the axes identified in section 5.1.3.

The displacements recorded in the database exhibited interesting temporal behavior, which are presented in the following section. This motivated the use of computational dynamic systems inspired from automatic control science to model human facial motion (section 5.2.2). The proposed approach models the natural dynamic signature of emotional expressions as well as the variability of this dynamics (see section 5.2.2.2).

5.2.1 Motion Signals Analysis

During the transitional phases of facial expressions, we noticed an interesting temporal signature (see figure 5.2). The displacement signals exhibit an initial acceleration, an inflection point, and a damping before reaching the final position. The different markers have different speeds and displacement amplitudes, but they all display the same general behavior. Note that this is not specific to our subject or our data format, as such patterns have already been observed in other studies of facial dynamics [SC01, SC03]. In [SC03], Schmidt *et al.* even suggested that the dynamics of the smile had the properties of automatically controlled systems.

If we consider the final position to be a target value, the behavior of the displacement signals are typical of a step response as described in automatic control science. Step

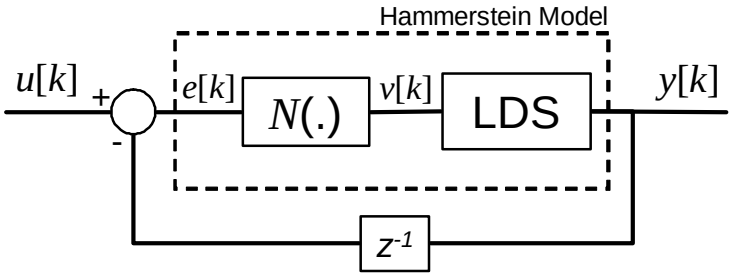

Figure 5.3: Nonlinear dynamic control model of facial marker displacement.

responses depict how a dynamic system responds to a constant input different from its initial state. It is often used for the identification of the system's parameter because it reflects its inherent dynamics. It therefore makes sense to represent the marker displacement along each axis by an individual dynamic system. The parameters of our systems will then be adjusted for each marker to produce step responses that match the specific data for that marker. This motion modeling approach is developed in the next section.

5.2.2 Computational Models of Motion

5.2.2.1 Nonlinear Dynamics

Even though breaking the global facial movement into atomic marker displacements significantly reduces the complexity, the step responses still display nonlinear characteristics. In particular, the data shows an amplitude-dependent settling time which is incompatible with linear dynamic systems for which the dynamic properties are independent of the input amplitude. This amplitude-dependent behavior is consistent with similar observations made in the past [SC03].

The displacements of the marker are a consequence of facial muscle contraction. We can thus draw a parallel between facial movements and the dynamics of muscle contraction. Nonlinearity in the muscular force production has long been identified in biomedical studies. Attempts at creating a valid physical model of this mechanism have introduced empirically measured nonlinearities in their equations [Zaj89]. In the present work, we favor a black box model, which facilitates the parameter identification on input-output data, and is computationally more efficient. A popular black-box model in biomedical engineering is the Hammerstein model [BCC86]. As a matter of fact, Hammerstein models have already been used successfully in the modeling of isometric muscle dynamics [Cho06]. They can be concisely described by the following equations:

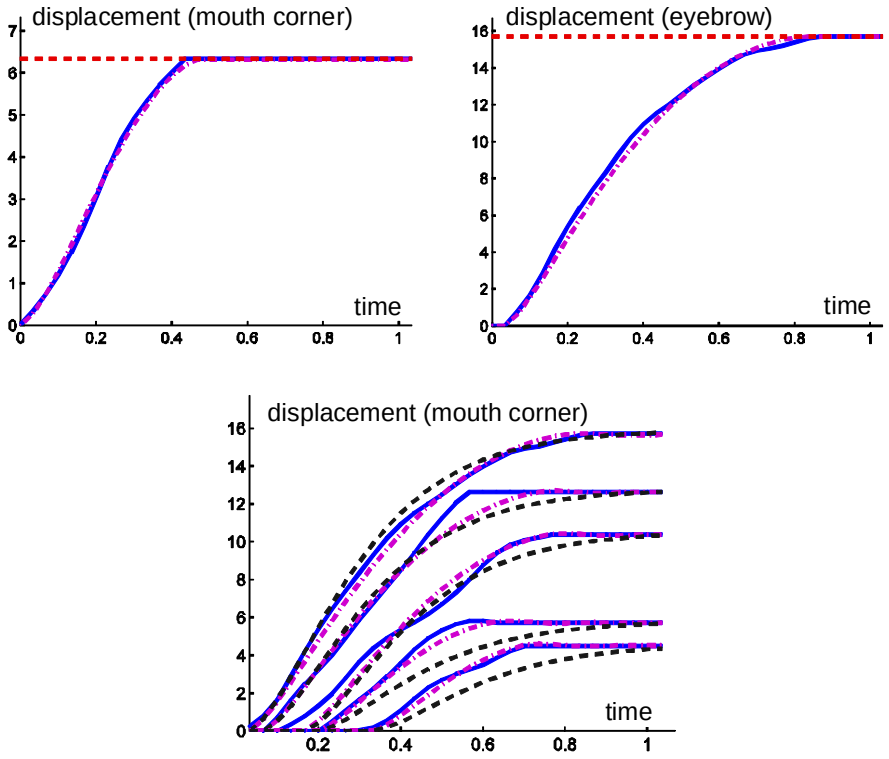

Figure 5.4: **Top left and right:** Marker displacement modeling with a Hammerstein-based control system. Original displacement data *(blue solid line)*, displacement generated by the dynamic system *(magenta dash-dot line)* and final displacement value *(red dashed line)*. **Bottom:** Linear *(Black dashed line)* versus nonlinear *(Magenta dash-dot line)* modeling of the original marker displacements *(blue solid line)*.

$$v[k] = N(e[k]) \tag{5.2}$$

$$y[k] = \sum_{i=1}^{p} a_i.y[k-i] + \sum_{j=1}^{q} b_j.v[k-j] \tag{5.3}$$

where e and y are the system input and output respectively. Hammerstein models explicitly address amplitude-dependent behaviors by splitting the global nonlinear dynamics into a static input nonlinearity N (equation 5.2) followed by a linear dynamic system (equation 5.3). In our solution, we use Hammerstein models as the movement production entities in the dynamic control models of facial marker displacement. The final marker movement system, presented in figure 5.3, is a closed-loop control system based on a Hammerstein model. Equation 5.2 becomes:

$$v[k] = N(u[k] - y[k-1]) \tag{5.4}$$

The control signal u specifies the desired displacement for a specific marker axis, and

the system evolves with its own dynamics until it reaches that control value. This simple Hammerstein-based system provides a very realistic representation of the facial markers' movement (figure 5.4), and its conformance with the captured data is greatly improved compared to simple linear systems (figure 5.4). In addition to its convenient formalism and satisfying results, we can also give a coarse interpretation of the system in figure 5.3: the Hammerstein component can be interpreted as the muscular actuator, and the signal $u - y$ is an abstraction of the neural command signal exciting the muscle.

5.2.2.2 Motion Variability

Although the captured data are well approximated by our nonlinear dynamic system (figure 5.3), it does not explain the data entirely. The displacement signals exhibit subtle dynamics variations that do not correspond to a deterministic phenomenon. These variations illustrate the stochastic nature of human movements: Even for similar facial expressions -which correspond to comparable displacements- the markers always show slightly different behaviors (regardless of measurement noise).

We believe that this factor is a very important part of the richness of human behavior. The variability breaks the repetitiveness of facial actions and contributes to overall motion naturalness. Consequently, if natural human faces never show the exact same motion twice even when producing the same expression, not only do facial animations have to be realistic, they also need to show variability.

Traditional facial animation methods are not well suited for such interactions since they would require a large database of precalculated motions to provide both flexibility and variety. Besides, since the variations are correlated with the motion elicitation process, approaches that consist of simply adding generic noise to the displacement signals generally produce unnatural motion. In the following, we present an innovative way to handle variability: a variability component directly integrated to the motion generation system.

Once again, results in medical research provide an interesting lead to cope with this aspect. Harris and Wolpert [HW98] explained the variability of human movements by the assumption that neural signals controlling muscle excitation are corrupted by noise; this assumption has since been backed up by numerous studies. In a similar fashion, we introduced a stochastic factor in the displacement dynamics by adding noise to the control signal u. By inverting the input-output relationship of our dynamical system, we can even learn the properties of this noise from the motion data itself: figure 5.5 shows the alterations of the command signal u that would exactly reproduce the recorded marker displacement with our dynamic control system. From these signals we can build a noise model that would account for the non-deterministic variations in facial motion.

In [HW98], Harris and Wolpert further assumed that the noise level linearly increases with the amplitude of the control signal. This is consistent with our observations of

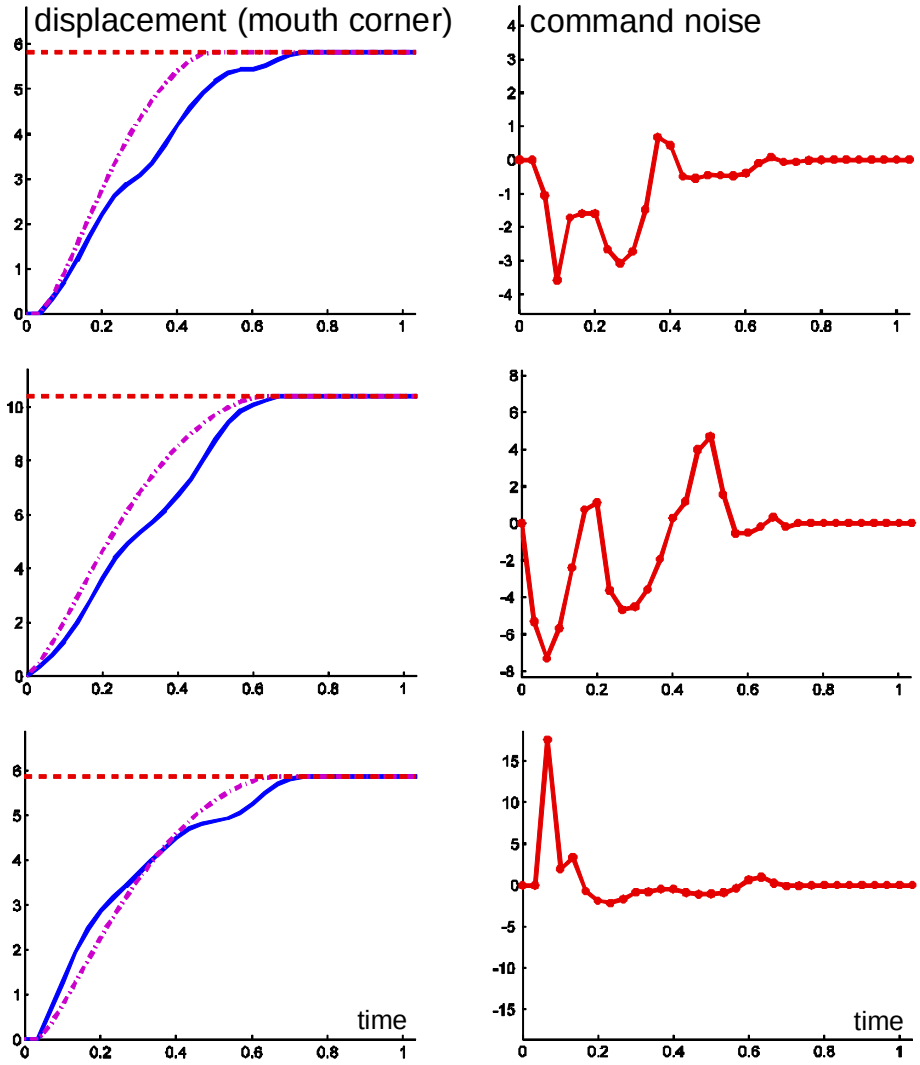

Figure 5.5: Motion variability seen as a noise corrupting the command signal. **Left**: displacement generated by our dynamical system *(magenta dash-dot line)* and observed motion data *(blue solid line)*. **Right**: command perturbations that would actually produce the observed motion data on the left with our dynamic system

figure 5.5: the standard deviation of the observed noise is linearly dependent upon the amplitude of the command signal $u - y$ (see figure 5.6).

After amplitude normalization, the command noise is assumed stationary. We can use the measured noises to construct a simple model that produces signals with the same stochastic characteristics. We proceed according to the Box-Jenkins methodology [BJR70], which is often used to identify simple stochastic processes that fit experimental data. In our case, autoregressive processes of order three provide a satisfying ap-

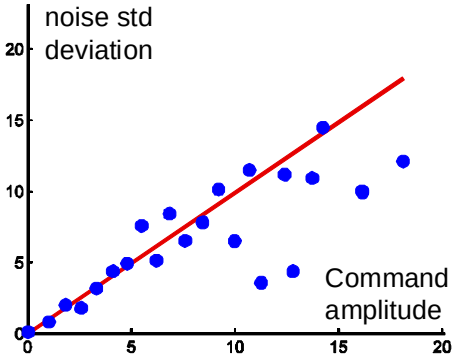

Figure 5.6: The command noise responsible for the variability in the captured motion data is signal-dependent. As assumed in [HW98], the measured noise standard deviation is linearly increasing with the command amplitude.

proximation of the spectral attributes of the measured noise. Realistic synthetic marker displacements are finally generated with the dynamic system presented in 5.2.2.1, where equation 5.4 is rewritten as

$$v[k] = N(e[k]) \tag{5.5}$$

$$e[k] = (u[k] - y[k-1]).(1 + n[k]) \tag{5.6}$$

where n is the noise described above. The stochasticity of the movement is entirely encompassed in the input noise. This framework has an important advantage over traditional methods that directly add noise to the motion signals. Indeed, our noisy commands are ultimately filtered by the dynamic systems; consequently, while displaying a non-deterministic behavior, the displacement signals preserve the characteristic dynamics which has been learned from data.

5.2.3 Coordination

The drawback of our local approach versus global state methods is that it does not account for the natural coordination between facial elements. The models of section 5.2.2 ensure that marker dynamics remains coherent at any time, yet it does not specify how the different models synchronize with each other.

The coordination between our different markers is obviously not random. One essential factor is the physical constraint imposed naturally by the structure of the face. Neighboring markers (such as two eyebrows markers for instance) are strongly coordinated, since they are linked by a short segment of skin and are influenced by the same muscles. This aspect of the coordination problem is actually straightforward to solve in this context. Indeed, in our system markers' movements are constraint to dominant axes (section 5.1.3). Also, the dynamic properties of each marker are necessarily coherent with the ones of its neighbors, since they were learned from real, dynamically

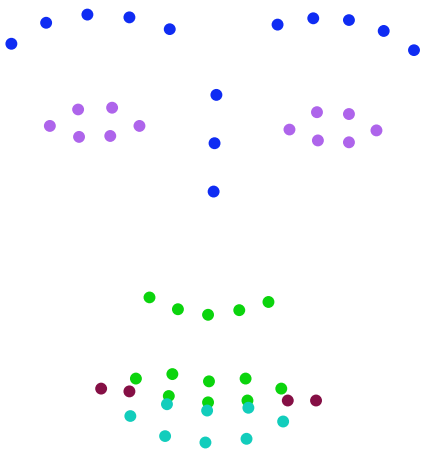

Figure 5.7: Distribution of the noise clusters. The facial markers whose movements are synchronized are grouped into clusters. At runtime an independent command noise will be generated for the markers of each cluster.

coherent data. If the respective control signals $u[k]$ for different markers are handled coherently, their natural coordination will implicitly be respected.

The factor that can actually influence $u[k]$ and marker coordination in our animation system is the control noise (equation 5.6). In particular, it does not make sense to use completely independent control noise for strongly coordinated markers, as it would results in different, possibly conflicting control signals for those markers. A naive solution would be to use the same control noise for all markers. Yet, a detrimental consequence of this is that the movement of all markers would always be synchronized. This would clearly restrict the panel of variations we can generate, and impairs animation realism. Indeed, observations of facial expressions have revealed that faces display different coordination patterns of facial elements, even for different instances of a similar expression. For instance, the coordination of eyebrow raise and mouth opening varied in the observed surprise expressions.

To satisfy both coordination coherence and animation variability we proposed to handle the noise component as follows: The markers that are strongly correlated are grouped in what we call "noise clusters". At runtime, the animation system uses the same control noise for all markers of a noise cluster to guarantee perfect local coordination. On the other hand, different control noises are used for each cluster.

The structure of the noise clusters has been determined objectively based on natural motion data. After measuring the synchronism of markers' motion, we performed spectral clustering to isolate group of markers whose movements are strongly coordinated in the given data. On figure 5.7 we display the five noise clusters identified by

Figure 5.8: Example of generation of new facial animation sequences. The starting *(top left)* and ending *(top right)* expressions are manually or automatically chosen using an intuitive 2D interface [SSB09]. Using the described motion models, our system computes the movement between these expressions.

the automatic clustering process. It is interesting to see that the clustering process naturally grouped markers that are physically close, although no topological constraints were imposed on the clustering operation.

Consequently, the animation system has to generate five independent noise signals at runtime, one for each noise cluster. Within a cluster, this guarantees that the involved markers will remain synchronized at all times. Besides, the coordination independence between clusters ensures the generation of varied, non-repetitive facial animations. Examples of the type of variations produced by the stochastic component for a single expression can be observed in one of the accompanying videos[2].

5.3 Results and Evaluation

In section 5.2.2 we presented motion models that can generate movements for the 68 facial markers of our motion database. From a start and target configuration, the models compute the temporal trajectory of the required displacement.

As such, these "human face" movements are not very useful since we originally wanted to animate an avatar. However, with the help of the retargeting framework presented in section 4.2 we can easily transfer those computer-generated movements to a synthetic face. The motion models thus indirectly control the facial deformation dynamics of any virtual character.

[2] `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php`

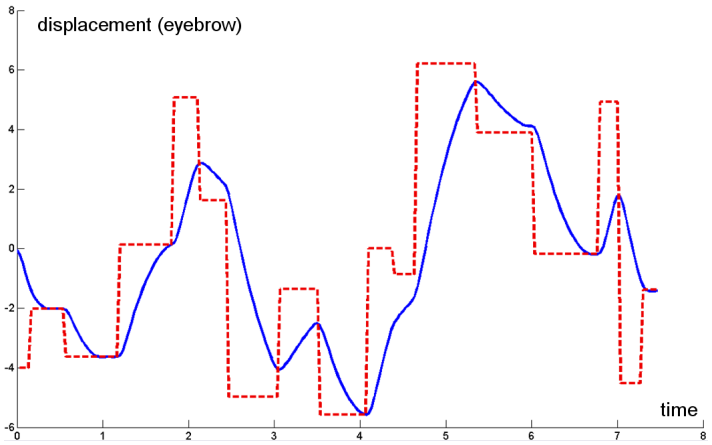

Figure 5.9: Example of a real-time animation scenario in which the target expression is modified on-the-fly. The red dashed line is the command value that represents the target expression. The blue solid line shows the resulting displacement of the marker. The animation system looks to reach the requested target expressions in real-time, yet always maintains the dynamic properties that have been learned on data. Note that command noise has been switched off on this example.

New sequences of expressive facial animation can easily be generated by selecting target expressions and letting the system compute the intermediate motion. For instance, an animator could manually setup a sequence of desired target expressions and let the system interpolate between them in a dynamics-aware keyframing fashion. This task can even be facilitated by using the intuitive facial expressions interface introduced in section 4.3 to select the target expressions (see figure 5.8 and the accompanying video[3] for illustration).

For real-time interactions, the target expressions can be requested automatically according to the application's needs, and the natural facial motion be computed at runtime[4]. More importantly, the system is not limited to static motion calculations. Target expressions, which act as a command value for the motion models, can be changed on-the-fly. The input-output formulation thus ensures a good reactivity of the animation system to dynamically evolving contexts and target modifications. Figure 5.9 as well as the accompanying video[5] illustrate this interactive use of our animation system.

An additional noteworthy characteristic of our system is that it relies on motion data to learn the short-term temporal signature of facial movements, but does not reuse the data itself. Consequently, contrary to most motion capture-based approaches, the

[3] `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php`

[4] The dynamic simulation itself is time-efficient (approx. 30μs per frame); the time consuming task is actually the animation retargeting, which depends on the complexity of the 3D model (approx. 20ms per frame for a highly complex 3D model of 56k vertices).

[5] `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php`

Test sequences	1	2	3	4	5	6	7
Motion capture	2.61	3.29	2.90	1.82	3.73	2.82	3.11
Motion models	3.68	3.32	3.39	3.45	3.13	3.68	3.29
p-value	$9.5e^{-6}$	0.91	$27e^{-3}$	$3.8e^{-11}$	$7.9e^{-3}$	$7.1e^{-4}$	0.45
Stat. significant (5%)	yes	no	yes	yes	yes	yes	no

Table 5.1: Mean grades obtained by motion capture animation and our motion model-based technique during user evaluation. For each sequence, the statistical significance of the results is indicated. Significance tests were conducted using a two-sided t-test, and a level of significance of 5% was assumed.

system can synthesize an unlimited range of realistic facial movements even if they were not included in the database. In section 5.2.2.2, we highlighted the importance of variability in natural human motion. Even with a perfect dynamic behavior, deterministic and repetitive facial animation eventually emphasizes the artificial nature of the virtual character in extended man-agent interactions. The introduction of a stochastic component in our motion models brings variety and ensures that every generated facial animation will be different, while preserving the correct dynamics. The video shows the subtle but noticeable variations of the facial animations generated for similar target expressions.

We conducted two user evaluations to measure the quality of our animations, and see how they compare to genuine human motion (first evaluation) and other animation techniques (second evaluation).

The first evaluation involved thirty-one non-specialist participants. Short facial movements were played multiple times using both ground-truth motion capture data and our animation system. The participants were then asked to grade the realism of the presented animations from 1.0 (unrealistic motion) to 5.0 (lifelike motion). The mean grades are presented in table 5.1.

Our technique matches the grades of motion capture data on almost all sequences. These results indicate that our technique successfully captures the dynamic properties of natural motion. It also turns out that our technique outscores motion capture on some sequences. This can be explained by the fact that raw human motion sometimes displays abrupt variations. Although these variations look natural on a human face, people tend to expect smoother movements from virtual characters. Our dynamic models produce motion signals that are smoother than real-life movements, and might be considered more appealing on synthetic faces. This result tends to back up the uncanny valley hypothesis.

The second user evaluation was conducted in order to compare the presented animation system to traditional interpolation methods. Seventeen participants were presented

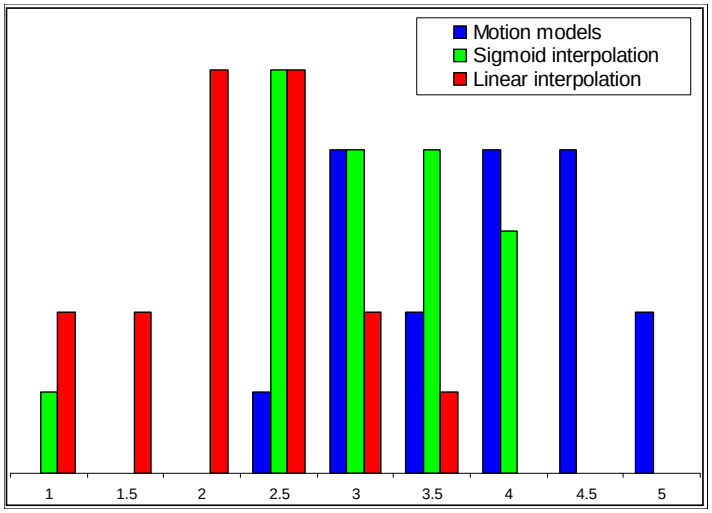

Figure 5.10: Histogram of the grades obtained by three keyframing animation methods. The participants showed a preference for our motion models *(blue)*, with a mean grade of 3.87, over sigmoid interpolation *(green)* and linear interpolation *(red)* with mean grades of 3.04 and 2.18 respectively. These results are statistically significant: Two-sided t-test were conducted, resulting in p-values of $3.2e^{-2}$ and $6.3e^{-5}$ when comparing motion models to sigmoid and linear interpolation respectively.

longer sequences of keyframing animation, and were again asked to grade the animations from 1.0 to 5.0. The keyframes were the same for all evaluated methods (changing only the interpolation method). The results are presented in figure 5.10.

The participants had a preference for our animation system over the other evaluated methods. They generally graded the animations according to their global perception of the animated face, however they found it difficult to explain their choices in detail. Their comments suggest that the impact of such short-term dynamic phenomena is visible, but difficult to measure quantitatively. Several participants mentioned that they were attracted by the more genuine look our animation technique produces. On the other hand, the other animations sequences were often perceived as monotonous and predictable after just a few movements, quickly revealing the artificial and simplistic nature of the interpolation. These impressions can be explained by the adapted action of our individual dynamic systems: In our system, all facial elements are animated according to their own dynamic characteristics, whereas simpler schemes interpolate the whole face simultaneously with the same pattern. Moreover, the variability introduced by the stochastic component of our system also helps preventing the movements from looking repetitive during long-term interactions.

Chapter 6

Conclusion

Contents

6.1 Summary 133
6.2 Valorization 134
6.3 Publications 139
6.4 Perspectives 141
6.4.1 Research Perspectives 141
6.4.2 Industrial Perspectives 143

6.1 Summary

In this section, we sum up the contribution and results presented in this thesis. As mentioned in the introduction chapter, the document was structured around two key aspects of facial expressions: the handling of facial deformations and the dynamics of facial movements.

In chapter 3 we presented an innovative low-dimensional representation of human emotional facial expressions. We described a method to automatically extract this representation from the processing of facial expressions data. Ultimately, we showed how it can be used to easily map the space of natural facial expressions on a simple interface, and manipulate them for the analyzed face. The simplicity of this representation also made it possible to relate deformation modes to meaningful semantic concept, and thus manipulate expressions as meaningful emotional messages. The important difference with previous research is that the representation is not predefined, but objectively extracted from real data. Thus, in addition to its simplicity, it remains consistent with the actual facial deformation modes.

This representation constitutes an interesting way to bridge the gap between low-level considerations (basic description of facial deformations) and high-level considerations (emotional interpretation of facial expressions).

Chapter 4, presented methods to adapt the models developed in chapter 3 on real human faces to new faces, including synthetic ones. This enabled transferring the characteristics learned from human faces to the face of virtual characters. Synthetic faces could, for instance, be animated via direct retargeting of facial expressions. More importantly, the low-dimensional representation of facial expressions described above could be used as a control interface, providing an efficient and innovative way to manipulate facial expressions. Its intuitive structure makes it easy to synthesize still expressions as well as expression sequences, even for non-expert users.

The second part of the thesis (chapter 5) dealt with the dynamic aspect of emotional facial expressions. We proposed an animation system based on a collection of motion models that were trained on real expression data. The motion models were meant to learn the dynamic signature identified on data, and reproduce this natural signature when generating new facial movements.

Previous animation systems relying on motion learning considered the face as one single dynamic entity. Due to the complexity and variability of facial motion, such approaches can model only very *specific* actions ("smile" for instance). By closely observing facial motion data, we were able to identify more *generic* motion signatures in local facial movements. These natural dynamic signatures could be learned by individual local models. Contrary to global approaches, this framework can be used to synthesize *any* dynamic facial expressions, even out-of-sample expressions[1], at real-time speed.

Another important contribution of this framework is the straightforward handling of the variability of human motion, as a stochastic component of the motion models themselves. The animation framework therefore guarantees nonrepetitive yet coherent facial behaviors, which improves the perceived naturalness in long-term interactions.

The research contributions summed up in this section were meant to address the issues pointed out in the introduction chapter ("Problem statement", section 1.2). We feel that the proposed animation methods present an interesting compromise between parameter-based approaches and performance-based approaches. Indeed, all our systems are based on captured expression data to benefit from the naturalness of human expressiveness; yet we do not settle for simple reuse of this data, but use learning methods to construct generalized, controllable systems. Additionally, the systems' control can easily be related to intuitive, high-level notions of human emotions; this makes the presented approaches interesting for non-expert animators, and for affective computing applications that usually work at the semantic level.

6.2 Valorization

The research contributions of chapters 3, 4 and 5 all aimed at one objective: propose models that capture the naturalness of human emotional expressions, and whose knowledge can be used to generate natural-looking animations on synthetic faces. Apart from

[1] Out-of-sample expressions: expressions that were not included in the training database for the motion models.

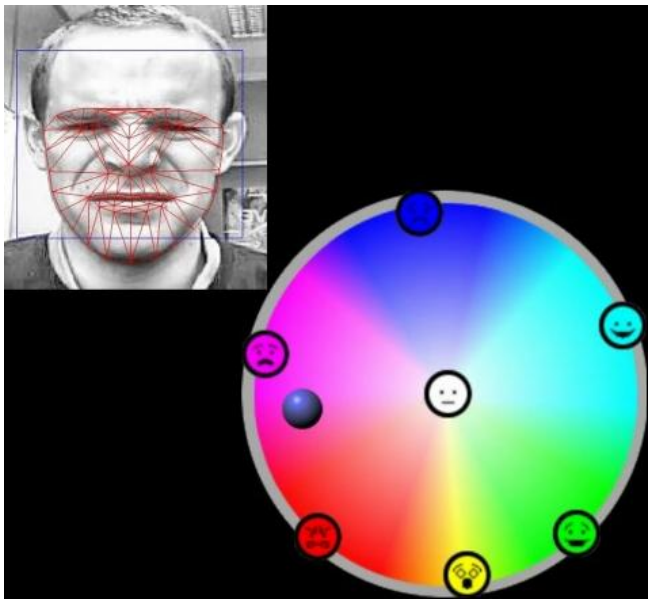

Figure 6.1: Snapshot of the real-time expression interpretation application. A user is filmed by a standard webcam, and his expressive facial deformations are extracted by an AAM search procedure (left) and projected onto the facial expression manifold. The location of the projection -displayed on the low-dimensional interface (right)- enables the interpretation of the captured expression, a disgust expression here. See the accompanying video at `http://www.rennes.supelec.fr/ren/perso/nstoiber/report.php` for more examples.

scientific interest, studying this problem addresses really concrete needs in the computer vision and animation industries. This section puts forward a few prototypes that show how the ideas developed in this thesis can be used as tools in actual applications.

Real-time expression interpretation

In chapter 3, we presented a method that automatically extracts a low-dimensional representation of emotional facial expressions. When presented visually in 2D or 3D, this representation can intuitively be related to interpretations in terms of emotions (section 3.2.3.2). The following part of the thesis concentrates mostly on using this representation for expression synthesis, however in section 3.2.2.3 we pointed out that newly captured expressions could be projected on the facial expression manifold. Captured expressions could thus be displayed as a point on the low-dimensional representation, as illustrated in figure 3.10.

We developed an application in which this projection/representation of a facial expression is performed in real-time. In practice, the expressions of a user are captured by a standard webcam. The facial deformations are then extracted by an AAM search procedure and projected onto the facial expression manifold. Finally, the projected point is

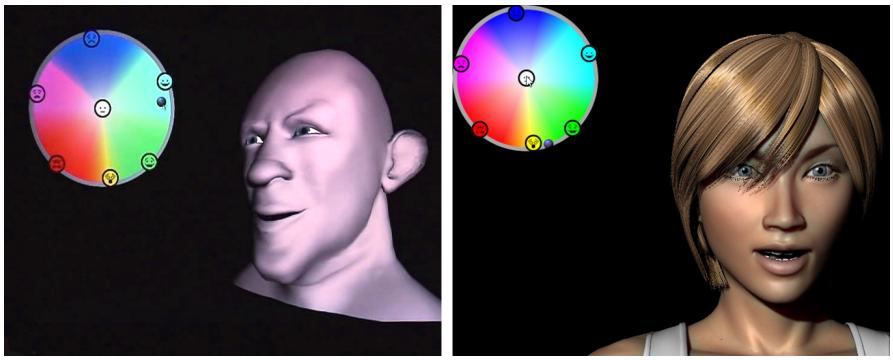

Figure 6.2: Snapshots of facial expression manipulation using the low-dimensional control interface (2D). As explained in section 4.3.3, the intuitive control interface can be adapted to any face. Left character used with permission of the Institute of Animation Baden-Wuerttemberg.

displayed on the 2D or 3D representation of the manifold, from which an interpretation of the user's expression can be formulated. The application is illustrated in figure 6.1.

The scope of expressions that can be interpreted is obviously not infinite, but is spanned by the expressions of the training database. An interesting quality of the system, compared to classification approaches, is to be able to continuously interpret mixtures of expressions, as well as expression intensity. As such, the system depends on a person-specific model, and thus can only interpret the expression of a single individual correctly. However, the system can be adapted to new individuals via a calibration phase. The calibration consists of quickly creating a new person-specific model using the method presented in section 4.1.

Intuitive expression edition tool

The use of the low-dimensional representation as a facial expression manipulation interface has already been highlighted in section 4.3. Prototypes have been developed to demonstrate the utility of such a tool in facial expression edition tasks. The interface received good feedback from professional animators, as it offers a simple but accurate control of facial configurations. The use of the interface could additionally be extended to any virtual face, provided it is possible to construct a database of facial expressions for this character (figure 6.2).

Real-time expression retargeting

Another application that can be derived from our result is real time expression retargeting (also called "expression mapping"). Our retargeting method was described in detail in section 4.2. Expression mapping was based on mapping the parameters of what we called the facial appearance space of the source and target characters. Although not groundbreaking, this application proved reliable and flexible, as it could be adapted to multiple source and target faces. The application was demonstrated in real-time

Figure 6.3: Snapshot of the real-time expression retargeting application presented at the JISE technical exhibition, May 2009.

in the technical exhibition JISE (*Journée Image et Systèmes Embarqués*) in May 2009 (figure 6.3).

The contributions presented in chapters 4 and 5 deal with complementary aspects of facial expressions: the deformation aspect and the dynamic aspect. When put together, they form a complete solution to the problem of realistic facial expressions for virtual characters. The results presented in chapter 5 already displayed how the intuitive control interface and the motion models can be used to create animation sequences in a dynamics-aware keyframing fashion.

We also mentioned in that chapter that the animation system can react to dynamically-evolving expression requirements. This allows us to use the animation system in a completely autonomous way: Depending on the context, the application can autonomously trigger expression requests (using the semantics of the control interface for instance) and the animation system will correctly compute the transition to the desired expression. This kind of autonomous animation is particularly interesting for interactively controlled characters such as game characters or conversational agents, as highlighted by the following prototypes:

Companion project
The techniques presented in chapters 4 and 5 were integrated in a prototype developed for the european project "Companion". The goal of the project is to create an embodied

conversational agent with believable behavior and reactions in a human-agent interaction context. Realistic facial expressions were considered an important part of this naturalness.

The agent's behavior is driven by an interaction engine that handled emotional labels. It is meant to display different types of emotional states while interacting with the user. In the prototype, the emotions are intuitively translated to facial expressions using the low-dimensional expression control interface presented in section 4.3.3. Simply put, emotional states required by the interaction engine are produced by imposing to reach a given point on the facial expression interface. Additionally, the dynamic transitions are handled by the motion models we presented in chapter 5 thus ensuring realistic movements.

Due to the continuous nature of the control interface, various blending of emotional expressions as well as expressions of varying intensity can be produced in real-time, according to the needs of the interaction engine. The simplicity of the control interface allows rich and numerous new behaviors to be created easily. Besides, relying on this interface ensures that facial animation remains consistent at all times. The conversational agent thus benefits from a realistic, reliable and highly-tunable facial expressiveness.

The mimic game: real-time imitation of facial expressions

The last application we developed is entitled "the Mimic Game", and was described and demonstrated in a talk at the SIGGRAPH conference in July 2010. The mimic game is a real-time program in which a virtual character imitates the facial expressions of a real person filmed by a webcam (see figure 6.4). It is particularly interesting in a valorization perspective, because it combines pretty much all the ideas presented in this thesis. The system relies on three steps:

- **Facial expression capture**: the system first captures the facial deformations and head movements of the user (both are captured simultaneously using an extension of the Active Appearance Model procedure called 2.5D AAM [SALGS07]).
- **Facial expression analysis**: the system interprets the captured deformations as an actual emotional expression. This is done by projecting the captured facial configuration on the facial expression manifold (this is similar to the "real-time expression interpretation" prototype described above).
- **Character facial animation**: finally, the face of the avatar is animated to produce an expression analogous to the one that was interpreted on the user's face. The animation system takes only the *target* expression as input, and the character's face is animated autonomously using the motion models of chapter 5.

One may argue that facial expression imitation can be done by simple expression retargeting (expressions transferred frame by frame on the avatar's face). The most interesting point about this application however, is that it is not a "retargeting" scheme. Indeed, the system works just like a human who is trying to imitate someone: It first analyzes the expression on the person's face, and then replicates it *autonomously*, using

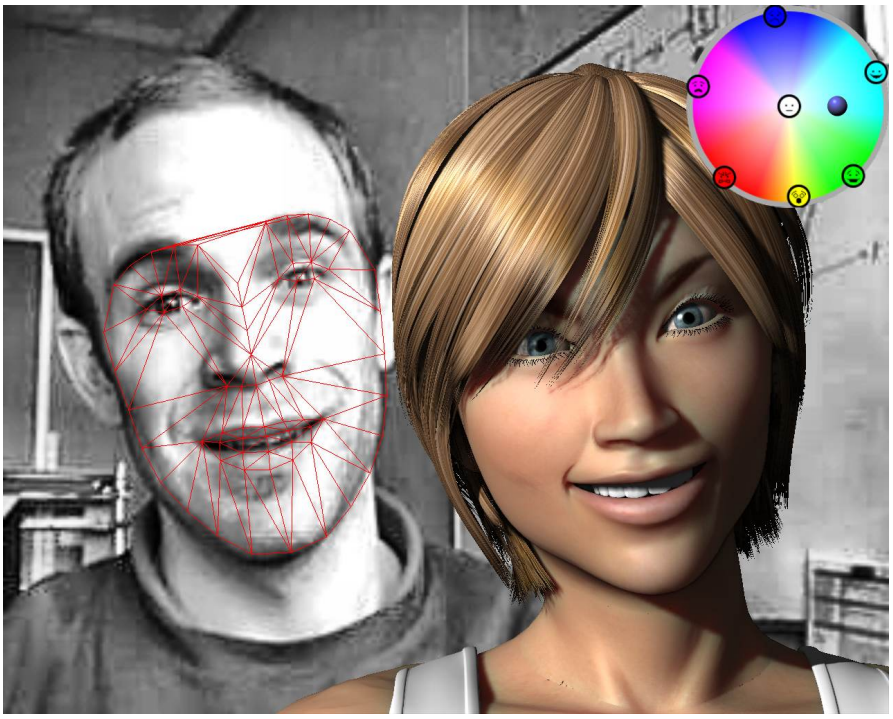

Figure 6.4: The Mimic Game: real-time imitation of emotional facial expressions. The application captures head movements and facial deformations on a human face, interprets these movements as an emotional expression and animates the face of a virtual character to imitate this expression. See the accompanying video for a better demonstration.

its own motion mechanics. In that sense it is really a "mimic" system. The advantage of this approach is that it enables more elaborate behavior from the virtual character. The prototype focused on imitation because it is easier to illustrate; yet more interesting scenarios can be imagined. For instance, if an angry face is detected on the user, the character could react with a surprise expression (the character wonders why the user is angry).

Additionally, on a more pragmatic level, the presented structure improves the realism of the virtual character's behavior. Indeed, with noisy capture results (as we have here), a simple retargeting approach would inevitably produce noisy animation, and impair visual realism. In our structure, the final animations are produced by motion models that are independent from capture noise. The animation component always generates smooth movements, and thus preserves the visual credibility of the virtual character.

6.3 Publications

This section lists the publications extracted from this 3-years work.

Journals

- **Nicolas Stoiber**, Gaspard Breton, Renaud Seguier. Modeling Short-term Dynamics and Variability for Realistic Interactive Facial Animation. *Computer Graphics and Applications. Special issue: Digital Human Faces from Creation to Emotion*, 2010.
- **Nicolas Stoiber**, Renaud Seguier, Gaspard Breton. Facial Animation Retargeting and Control based on a Human Appearance Space. *Computer Animation and Virtual Worlds*, vol. 21, pp. 39-54, 2010.
- Abdul Sattar, **Nicolas Stoiber**, Renaud Seguier, Gaspard Breton. Gamer's Facial Cloning for Online Interactive Games. *International Journal of Computer Games Technology (special issue on Cyber Games and Interactive Entertainment)*, 2009.

International Conferences

- **Nicolas Stoiber**, Olivier Aubault, Renaud Seguier, Gaspard Breton. The Mimic Game: Real-time Recognition and Imitation of Emotional Facial Expressions. *SIGGRAPH Talks*, Los Angeles, CA, USA. July 25-29, 2010.
- **Nicolas Stoiber**, Renaud Seguier, Gaspard Breton. Automatic design of a control interface for a synthetic face. *Proceedings of the 13th international Conference on Intelligent User Interfaces.* Sanibel Island, FL, USA. February 8-11, 2009.

Book Chapter

- **Nicolas Stoiber**, Renaud Seguier, Gaspard Breton. A Data-Driven Meaningful Representation of Emotional Facial Expressions. To appear in *Multimedia Information Extraction* (Mark Maybury, eds. 2010).

Public Presentations

- **Nicolas Stoiber**, Renaud Seguier, Gaspard Breton. Représentation objective et pertinente des expressions faciales émotionnelles. *Présentation à la journée "Visage, geste, action et comportement" du GDR-ISIS.* December 8, 2009.
- **Nicolas Stoiber**, Olivier Aubault, Renaud Seguier, Gaspard Breton. Le jeu du Mime : Reconnaissance en temps réel et imitation d'expressions faciales émotionnelles. *Présentation à la réunion du Paris ACM SIGGRAPH.* May 27, 2010.

Patents

- **Nicolas Stoiber**, Gaspard Breton, Renaud Seguier. Procédé et système de génération d'une interface de contrôle des expressions faciales d'un avatar. *Patent WO2010037956.* April 8, 2010.
- **Nicolas Stoiber**, Gaspard Breton, Renaud Seguier. Procédés et dispositifs de modélisation et de synthèse d'expressions faciales pour personnages virtuels. *Patent FR 06971.*

6.4 Perspectives

6.4.1 Research Perspectives

Person-specific models

The most important limitation of our work is probably that it concentrates person-specific models. Whether for the modeling of facial deformations (chapter 3) or of dynamics (chapter 5), the models account only for a single individual. As an example, it is not possible to correctly analyze the expressions of a person[2] without first constructing a model of that person's facial expressiveness (as explained in section 4.1). Moreover, the expressions' type and diversity of a person-specific model is limited to the expressive scope of its database.

In theory, a more general multi-individual analysis would improve the richness and adaptability of the models. Yet, the methods we described in this thesis reach their limitations in that case. When analyzing different faces simultaneously, it appeared that inter-personal differences were often more important than expressive variations of faces. Similar observations had been made in previous studies [CVTV05, MD05]. In practice, this is illustrated in our case by the formation of clusters of data for each individual in the appearance space. This makes it impossible to extract relevant generic facial patterns with the method developed in 3.2.

A solution would be to use judicious invariant features in facial description to ensure an accurate alignment of similar visual variations across individuals. Interesting ideas have already been proposed, such as using multilinear analysis to separate identity and expressions [VT02, EL04, LE05], or creating analytical correspondences between expression manifolds [WHL+04, CVTV05, SGM05a]. This task however, is still considered an ongoing research problem by the community. Investigating such approaches in order to extend our methods to multi-individual facial analysis probably constitute the most interesting research perspective for our work.

Models accuracy

In this study we rely exclusively on 2D data for the construction of the appearance spaces and motion models. For facial movements, 2D data forms a good approximation

[2] see "Real-time expression interpretation" or "The mimic game" in section 6.2

because these movements are essentially concentrated in the frontal plane of the face. Additionally, working in 2D greatly simplifies the acquisition and processing of motion data and does not require the use of heavy and expensive equipment.

For a more precise representation of facial deformations variance though, our approach would need to be extended to 3D data (use 3D position for the facial markers). This would be particularly interesting for the modeling of dynamics. Indeed, the motion models we propose aims at learning the displacement caused by muscle contraction, which are inherently organized in 3D. A 3D approach would certainly provide even more accurate models, yet at the price of a higher structural and computational complexity.

Beside the dimensionality aspect, the nature of our motion models can be questioned as well. It provides good approximations of actual data, and is consistent with the identified nonlinear nature of muscle contraction and control (see section 5.2.2). However, despite convincing visual results, the Hammerstein-based black-box system we propose still represents a coarse approximation of the actual physical phenomenon. Structural model improvements can be considered to improve accuracy. More elaborate formulations could, for instance, take into account specific muscle properties such as length/velocity dependence, time-varying properties or fatigue. It is important though, to preserve the essence of the approach, which is to consider time-efficient simulation, and models suitable for real-time applications.

Other components of expressiveness

The presented animation system, while remaining computationally simple, produces smooth, natural-looking animations that are consistent with the captured examples of human facial expressions. The evaluation results of section 5.3 highlight a good perceptual feedback from observers, and a visual improvement over traditional methods. Yet, natural facial expressions alone do not guarantee the realism of a synthetic face. As exposed in the introduction chapter, they greatly contribute to the overall credibility of the character, but other influential components such as visual speech synthesis, head movement, gaze and face rendering are also essential. Previous studies emphasize that none of these aspects can be neglected to achieve a truly realistic facial behavior [VGS+06].

If the algorithm and methods presented in this document were applied to emotional expressions only, it would probably be interesting to investigate their potential on other problems. For instance, other types of expressions such as visemes or conversational facial expressions also need to be realistically animated. The proposed motion models (chapter 5) might prove efficient in modeling their dynamic properties as well. Also, extracting a low-dimensional representation of the viseme manifold as done in chapter 3.2 for emotional expressions may be useful. Projecting and interpolating that manifold could constitute an interesting solution to the frequently encountered coarticulation problem, or provide an innovative way to perform lip-reading.

The problem we have addressed in this thesis is only a small part of the "believable character" equation. In future studies we hope it will be possible to integrate the multiple components of facial expressiveness together, and finally interact with a globally realistic and naturally expressive virtual character.

6.4.2 Industrial Perspectives

The prototypes developed to valorize this thesis (section 6.2) fulfill actual needs of the computer vision and the computer animation industries. Indeed, the extraction of semantic information from expressive faces is an application for which no efficient solution has been deployed in the industry yet. Similarly, intuitive tools to realistically animate virtual faces in real time can significantly improve the efficiency of professional animators, and bring facial animation solutions to non-expert users.

It is satisfying to note that this work gave birth to practical applications, and will provide animators and developers with new innovative tools. Above all though, we actually hope that this work will encourage more industry and research protagonists to take an interest in facial expressiveness of virtual characters, and give it the attention it truly deserves.

Acknowledgment

This work was funded by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

Glossary

AAM: Active Appearance Model.
ACD: Adjusted Coefficient of Determination.
AMA: Abstract Muscle Action.
AU: Action Unit.
DBN: Dynamic Bayesian Network.
DTW: Dynamic TimeWarping.
ECA: Embodied Conversational Agent.
EMG: Electromyography.
FACS: Facial Action Coding System.
FAP: Facial Animation Parameter.
FAU: Facial Action Unit.
FFD: Free-Form Deformation.
HMM: Hidden Markov Model.
ICA: Independent Component Analysis.
kNN: k Nearest Neighbors.
LDS: Linear Dynamic System.
LMA: Laban Movement Analysis.
LLE: Locally Linear Embedding.
PCA: Principal Component Analysis.
RBF: Radial-Basis Function.
STD: STandard Deviation.
TPS: Thin-Plate Spline.
TSP: Traveling Salesman Problem.

Bibliography

[ADD04] Bouchra Abboud, Franck Davoine, and MÃť Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19(8):723–740, 2004.

[ADP09] Ali Arya, Steve DiPaola, and Avi Parush. Perceptually valid facial expressions for character-based applications. *International Journal of Computer Games Technology*, 2009.

[AF02] Okan Arikan and David A. Forsyth. Interactive motion generation from examples. *Proceedings of ACM Siggraph*, pages 483–490, 2002.

[Ahl02] Joergen Ahlberg. An active model for facial feature tracking. *EURASIP Journal on Applied Signal Processing*, 2002(1):566–571, 2002.

[AK06] P.S. Aleksic and A.K. Katsaggelos. Automatic facial expression recognition using facial animation parameters and multistream hmms. *IEEE Transaction on Information Forensics and Security*, 1(1):3–11, 2006.

[AKA96] Kiyoshi Arai, Tsuneya Kurihara, and Ken-ichi Anjyo. Bilinear interpolation for facial expression and metamorphosis in real-time animation. *The Visual Computer*, 12(3):105–116, 1996.

[Arg69] Michael Argyle. *Social Interaction*. Tavistock Publications, 1969.

[ASC05] Zara Ambadar, Jonathan Schooler, and Jeffrey F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005.

[ASHS05] Irene Albrecht, Marc SchrÃűder, JÃűrg Haber, and Hans-Peter Seidel. Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8(4):201–212, 2005.

[Bar03] M.S. Bartlett. Real time face detection and facial expression recognition: Development and applications to human computer interaction. *Conference on Computer Vision and Pattern Recognition Workshop*, 5, 2003.

[Bas78] John N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3):373–379, 1978.

[Bas79] John N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37(11):2049–2058, 1979.

[Bat94] Joseph Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[BB02] Meeran Byun and Norman I. Badler. Facemote: qualitative parametric modifiers for facial animations. *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 65–71, 2002.

[BCC86] Leonas A. Bernotas, Patrick E. Crago, and Howard J. Chizeck. A discrete-time model of electrically stimulated muscle. *IEEE transactions on biomedical engineering*, 33(9):829–838, 1986.

[BCS97] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: driving visual speech with audio. *Proceedings of ACM Siggraph*, pages 353–360, 1997.

[BCT02] Franck Bettinger, Timothy F. Cootes, and Christopher J. Taylor. Modelling facial behaviours. *British Machine Vision Conference*, 2002.

[BH00] Matthew Brand and Aaron Hertzmann. Style machines. *Proceedings of ACM Siggraph*, pages 183–192, 2000.

[BHN04] The Duy Bui, Dirk Heylen, and Anton Nijholt. Combination of facial movements on a 3d talking head. *Proceedings of the Computer Graphics International (CGI)*, pages 284–291, 2004.

[BJR70] George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control.* Holden Day, San Francisco, 1970.

[BKC+08] L. Benedikt, V. Kajic, Darren Cosker, P.L. Rosin, and Dave Marshall. Facial dynamics in biometric identification. *Proc. of BMVC*, 2008.

[BM08] Emma Bould and Neil Morris. Role of motion signals in recognizing subtle facial expressions of emotion. *British Journal of Psychology*, 99(2):167–189, 2008.

[BN92] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. *ACM SIGGRAPH Computer Graphics*, 26(2):35–42, 1992.

[BN02] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2002.

[Boo89] F.L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[Bra03] Matthew Brand. Charting a manifold. *Advances in Neural Information Processing Systems*, 2003.

[BS95] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[BV99] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 2d faces. *Proceedings of ACM Siggraph*, pages 187–194, 1999.

[BVGP09] Ilya Baran, Dianel Vlasic, Eitan Grinspun, and Jovan Popovic. Semantic deformation transfer. *ACM Transactions on Graphics*, 28(3), 2009.

[BVS+96] M.S. Bartlett, Paul A. Viola, Terrence J. Sejnowski, Beatrice A. Golomb, Joseph C. Hager, and Paul Ekman. Classifying facial actions. *Advances in Neural Information Processing Systems*, 8:823–829, 1996.

[BW95] Armin Bruderlin and Lance Williams. Motion signal processing. *Proceedings of ACM Siggraph*, pages 97–104, 1995.

[BY97] Michael J. Black and Yaser Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[Byu07] Hae Won Byun. Online expression mapping for performance-driven facial animation. *Lecture Notes in Computer Science*, 4740:324–338, 2007.

[CB02] Erika S. Chuang and Christoph Bregler. Performance driven facial animation using blendshape interpolation. Technical report CS-TR-2002-02, Stanford University, 2002.

[CB05] Erika S. Chuang and Christoph Bregler. Mood swings: Expressive speech animation. *ACM Transactions on Graphics (TOG)*, 24(2):331–347, 2005.

[CBK+06] Cristobal Curio, Martin Breidt, Mario Kleiner, Quoc C. Vuong, Martin A. Giese, and Heinrich H. BÃijlthoff. Semantic 3d motion retargeting for facial animation. *Proceedings of the 3rd symposium on Applied perception in graphics and visualization*, pages 77–84, 2006.

[CC94] Trevor F. Cox and Michael A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, London, 1994.

[CCZB00] Diane Chi, Monica Costa, Liwei Zhao, and Norman Badler. The emote model for effort and shape. *Proceedings of ACM Siggraph*, pages 173–182, 2000.

[CDB02] Erika S. Chuang, Hrishikesh Deshpande, and Christoph Bregler. Facial expression space learning. *Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, pages 68–76, 2002.

[CES+00] Roddy Cowie, Douglas-Cowie Ellen, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc SchrÃűder. Feeltrace: An instrument for recording perceived emotion in real time. *Proc. of the ISCA Workshop on Speech and Emotion*, pages 19–24, 2000.

[CET98] T. F. Cootes, G.J. Edwards, and C. J. Taylor. Active appearance models. *European Conference on Computer Vision (ECCV)*, 2:484–498, 1998.

[CFP03] Yong Cao, Petros Faloutsos, and FrÃľdÃľric Pighin. Unsupervised learning for speech motion editing. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 225–231, 2003.

[CGH00] Ira Cohen, Ashutosh Garg, and Thomas S. Huang. Emotion recognition from facial expressions using multilevel hmm. *Neural Information Processing Systems (NIPS), Workshop on Affective Computing*, 2000.

[CH07] Jinxiang Chai and Jessica Hodgins. Constraint-based motion optimization using a statistical dynamic model. *Proceedings of ACM Siggraph*, 2007.

[Cho06] Danielle S. Chou. *Efficacy of Hammerstein Models in Capturing the Dynamics of Isometric Muscle Stimulated at Various Frequencies.* PhD thesis, MIT, 2006.

[CHT03] Ya Chang, Changbo Hu, and Matthew Turk. Manifold of facial expression. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 28–35, 2003.

[CHT04] Ya Chang, Changbo Hu, and Matthew Turk. Probabilistic expression analysis on manifolds. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:520–527, 2004.

[CK09] Yeongjae Cheon and Daijin Kim. Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, 42(7):1340–1350, 2009.

[CLK01] Byoungwon Choe, Hanook Lee, and Hyeong-Seok Ko. Performance-driven muscle-based facial animation. *Journal of Visualization and Computer Animation*, 12(2):67–79, 2001.

[CM90] Garrison Cottrell and Janet Metcalfe. Empath: face, emotion, and gender recognition using holons. *Proceedings of the conference on Advances in neural information processing systems*, 3:564–571, 1990.

[CM93] Michael M. Cohen and Dominic W. Massaro. Modeling coarticulation in synthetic visual speech. In Nadia Magnenat-thalmann and Daniel Thalmann, editors, *Models and Techniques in Computer Animation*. Springer-Verlag, Tokyo, 1993.

[CNH04] Naiwala P. Chandrasiri, Takeshi Naemura, and Hiroshi Harashima. Interactive analysis and synthesis of facial expressions based on personnal facial expression space. *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 105–110, 2004.

[CRRM07] Darren Cosker, Steven Roy, Paul L. Rosin, and David Marshall. Remapping animation parameters between multiple types of facial model. *Lecture Notes in Computer Science*, 4418:365–376, 2007.

[CSG+03] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen, and Thomas S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.

[CTFP05] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frederic Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302, 2005.

[CVTV05] Ya Chang, Marcelo Vieira, Matthew Turk, and Luiz Velho. Automatic 3d facial expression analysis in videos. *Analysis and Modelling of Faces and Gestures, Proceedings*, 3723:293–307, 2005.

[CXH03] Jinxiang Chai, Jing Xiao, and Jessica Hodgins. Vision-based control of 3d facial animation. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 193–206, 2003.

[Dar72] Charles Darwin. *The Expression of the Emotions in Man and Animals*. London, 1872.

[DBH+99] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[DBNN04] Zhigang Deng, Murtaza Bulut, Ulrich Neumann, and Shri Narayanan. Automatic dynamic expression synthesis for speech animation. *Proceed-*

ings of the IEEE International Conference on Computer Animation and Social Agents (CASA), pages 267–274, 2004.

[DBW+07] Yangzhou Du, Wenyuan Bi, Tao Wang, Yimin Zhang, and Haizhou Ai. Distributing expressional faces in 2-d emotional space. *Proceedings of the conference on Image and video retrieval*, pages 395–400, 2007.

[DCFN06] Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann. Animating blendshape faces by cross-mapping motion capture data. *Proceedings of the Symposium on Interactive 3D graphics and games*, pages 43–48, 2006.

[DiP89] Steve DiPaola. Implementation and use of a 3d parameterized facial modeling and animation system. *State of the Art in Facial Animation, SIGGRAPH 1989 Tutorials*, 22, 1989.

[DL02] Yangzhou Du and Xueyin Lin. Mapping emotional status to facial expressions. *Conference on Pattern Recognition (ICPR)*, 2:524–527, 2002.

[DL03] Yangzhou Du and Xueyin Lin. Emotional facial expression model building. *Pattern Recognition Letters*, 24(16):2923–2934, 2003.

[DM08] Zhigang Deng and Xiaohan Ma. Perceptually guided expressive facial animation. *Proc. of ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 67–76, 2008.

[DN07] Zhigang Deng and Jun-yong Noh. Computer facial animation: A survey. In Zhigang Deng and Ulrich Neumann, editors, *Data-Driven 3D Facial Animation*. Springer, London, 2007.

[dRPP+03] Fiorella de Rosis, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio, and Berardina de Carolis. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1-2):81–118, 2003.

[DRSV02] Doug DeCarlo, Corey Revilla, Matthew Stone, and Jennifer J. Venditti. Making discourse visible: Coding and animating conversational facial displays. *Proceedings of Computer Animation*, page 11, 2002.

[EBDP96] Irfan A. Essa, Sumit Basu, Trevor Darrell, and Alex P. Pentland. Modeling, tracking and interactive animation of faces and heads using input from video. *Proceedings of the Computer Animation*, page 68, 1996.

[Edw98] Kari Edwards. The face of time: Temporal cues in facial expressions of emotion. *Psychological Science*, 9(4):270–276, 1998.

[EF71] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[EF78a] Paul Ekman and W. Friesen. *Facial Action Coding System, Investigator's Guide*. Consulting Psychologists Press Inc., 1978.

[EF78b] Paul Ekman and W.V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[EG97] Peter Eisert and Bernd Girod. Facial expression analysis for model-based coding of video sequences. *Proceedings Picture Coding Symposium*, pages 33–38, 1997.

[EGP02] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. *Proceedings of ACM Siggraph*, pages 388–398, 2002.

[EHRW98] A. Eleftheriadis, C. Herpel, G. Rajan, and L. Ward. Mpeg-4 systems, text for iso/iec fcd 14496-1 systems. In *MPEG-4 SNHC*. 1998.

[Ekm82] Paul Ekman. *Emotion in the Human Face*. Cambridge University Press, New York, 1982.

[EKMt04] Arjan Egges, Sumedha Kshirsagar, and Nadia Magnenat-thalmann. Generic personality and emotion simulation for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):1–13, 2004.

[EL99] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. *Proceedings of the International Conference on Computer Vision*, page 1033, 1999.

[EL04] Ahmed ElGammal and Chan-Su Lee. Separating style and content on a nonlinear manifold. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2004.

[EL07] Ahmed ElGammal and Chan-Su Lee. Nonlinear manifold learning for dynamic shape and dynamic appearance. *Computer Vision and Image Understanding*, 106(1):31–46, 2007.

[EM04] Arjan Egges and Tom Molet. Personalised real-time idle motion synthesis. *Proceedings of the Computer Graphics and Applications*, pages 121–130, 2004.

[EP97] Irfan A. Essa and Alex P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[ESS00] Sheryl M. Ehrlich, Diane J. Schiano, and Kyle Sheridan. Communicating facial affect: it's not the realism, it's the motion. *Conference on Human Factors in Computing Systems*, pages 251–252, 2000.

[FL03] B. Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[GBBB06] Oxana Govokhina, Gerard Bailly, Gaspard Breton, and Paul Bagshaw. Tda: A new trainable trajectory formation system for facial animation. *International Conference on Spoken Language Processing*, 2006.

[GCT+04] Lisa Gralewski, Neill Campbell, Barry Thomas, Colin Dalton, and David Gibson. Statistical synthesis of facial expressions for the portrayal of emotion. *Proceedings of the International Conference on Computer graphics and Interactive Techniques in Australasia and South East Asia*, pages 190–198, 2004.

[GGW+98] Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and FrÃľdÃľric Pighin. Making faces. *Proceedings of ACM Siggraph*, pages 55–66, 1998.

[GL06] Amandine Grizard and Christine Laetitia Lisetti. Generation of facial emotional expressions based on scherer psychological theory. *3rd Humaine Summer School*, 2006.

[Gle08] Michael Gleicher. Graph-based motion synthesis: an annotated bibliography. *SIGGRAPH Course*, 2008.

[GMHP04] Keith Grochow, Steven L. Martin, Aaron Hertzmann, and Zoran Popovic. Style-based inverse kinematics. *ACM Transactions on Graphics*, 23(3):522–531, 2004.

[HCFT04] Changbo Hu, Ya Chang, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*, 5:81–81, 2004.

[HCYZ05] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. *10th IEEE International Conference on Computer Vision (ICCV)*, 2:1208–1213, 2005.

[HN03] Xiaofei He and Partha Niyogi. Locality preserving projections. *Technical Report TR-2002-09, University of Chicago Computer Science*, 2003.

[HO00] A. HyvÃďrinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000.

[HW98] Christopher M. Harris and Daniel M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–784, 1998.

[Iza71] Carroll E. Izard. *The face of emotion.* Appleton-Century-Crofts, East Norwalk, CT, USA, 1971.

[Jol86] I. T. Jolliffe. *Principal component analysis.* Springer Series in Statistics. Springer, Berlin, 1986.

[Jos03] P. Joshi. Learning controls for blend shape based realistic facial animation. *Proceedings of the ACM Siggraph/Eurographics Symposium on Computer Animation*, 2003.

[JYL09] Sumit Jain, Yuting Ye, and C. Karen Liu. Optimization-based interactive motion synthesis. *ACM Transactions on Graphics*, 28(1):10, 2009.

[KCT00] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Proceeding of the 4th IEEE International Conference of on Automated Face and Gesture Recognition*, pages 46–53, 2000.

[KG03] Lucas Kovar and Michael Gleicher. Flexible automatic motion blending with registration curves. *Proceedings of the Symposium on Computer Animation*, pages 214–224, 2003.

[KG04] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. *Proceedings of ACM Siggraph*, pages 559–568, 2004.

[KGP02] Lucas Kovar, Michael Gleicher, and Frederic Pighin. Motion graphs. *Proceedings of ACM Siggraph*, 21(3):473–482, 2002.

[KH92] Hiroshi Kobayashi and Fumio Hara. Recognition of six basic facial expression and their strength by neural network. *International Workshop on Robot and Human Communication*, pages 381–386, 1992.

[KH93] Hiroshi Kobayashi and Fumio Hara. Dynamic recognition of basic facial expressions by discrete-time recurrent neural network. *International Joint Conference on Neural Networks*, 1:155–158, 1993.

[KHYS02] Kolja Kaehler, Joerg Haber, Hitoshi Yamauchi, and Hans-Peter Seidel. Head shop: Generating animated head models with anatomical structure. *Proceedings of ACM Siggraph*, pages 55–64, 2002.

[KMMtT91] Prem Kalra, Angelo Mangili, Nadia Magnenat-thalmann, and Daniel Thalmann. Smile: a multilayered facial animation system. *Proceedings of the IFIP Conference on Modeling in Computer Graphics*, pages 189–198, 1991.

[KMMTT92] Prem Kalra, Angelo Mangili, Nadia Magnenat-Thalmann, and Daniel Thalmann. Simulation of facial muscle actions based on rational free form deformations. *Computer Graphics Forum*, 11(3):59–69, 1992.

[KMt02] Sumedha Kshirsagar and Nadia Magnenat-thalmann. A multilayer personality model. *Proceedings of the international symposium on Smart graphics*, pages 107–115, 2002.

[KP07] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transaction on Image Processing*, 16(1):172–187, 2007.

[KPL98] C. Kouadio, P. Poulin, and P. Lachapelle. Real-time facial animation based upon a bank of 3d facial expressions. *Proceedings of Computer Animation*, pages 128–136, 1998.

[KS08] Taesoo Kwon and Sung Yong Shin. Motion modeling for on-line locomotion synthesis. *Proceedings of the ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 29–38, 2008.

[KSH+92] Masahiro Kato, Ikken So, Yoichi Hishinuma, Osamu Nakamura, and Toshi Minami. Description and synthesis of facial expression based on isodensity maps. *Proceedings of the International Conference on Visual computing*, pages 39–56, 1992.

[KVMK99] Kostas Karpouzis, George Votsis, George Moschovitis, and Stefanos Kollias. Emotion recognition using feature extraction and 3-d models. *Proceedings of IMACS International Multiconference on Circuits and Systems Communications and Computers*, pages 5371–5376, 1999.

[LAAB02] Craig Latta, Nancy Alvarado, Sam S. Adams, and Steve Burbeck. An exressive system for endowing robots or animated characters with affective facial displays. *Proceedings of AECSI - AISB*, 2002.

[LAKG98] Michael Lyons, Shigeru Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. *IEEE International Conference on Automatic Face and Gesture Recognition*, page 200, 1998.

[Las87] John Lasseter. Principles of traditional animation applied to 3d computer animation. *ACM Siggraph Computer Graphics*, 21(4):35–44, 1987.

[LBA99] Michael Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[LBF+06] Gwen Littlewort, M.S. Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[LCR+02] Jehee Lee, Jin-Xiang Chai, Paul Reitsma, Jessica Hodgins, and Nancy Pollard. Interactive control of avatars animated with human motion data. *Proceedings of ACM Siggraph*, pages 491–500, 2002.

[LE05] Chan-Su Lee and Ahmed Elgammal. Facial expression analysis using nonlinear decomposable generative models. *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 17–31, 2005.

[LE06] Chan-Su Lee and Ahmed Elgammal. Human motion synthesis by motion manifold learning and motion primitive segmentation. *IV Conference of Articulated Motion and Deformable Objects (AMDO)*, pages 464–473, 2006.

[LEM06] Chan-Su Lee, Ahmed Elgammal, and Dimitris Metaxas. Synthesis and control of high resolution facial expressions for visual interactions. *IEEE International Conference on Multimedia and Expo (ICME)*, pages 65–68, 2006.

[LHP05] C. Karen Liu, Aaron Hertzmann, and Zoran Popovic. Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics*, 24(3):1071–1081, 2005.

[LK00] Jennifer S. Lerner and Dacher Keltner. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, 14(4):473–493, 2000.

[LKCL98] J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li. Automated facial expression recognition based on facs action units. *Proceedings of the International Conference on Face and Gesture Recognition*, page 390, 1998.

[LLX+01] Lin Liang, Ce Liu, Ying-Qing Xu, Baining Guo, and Heung-Yeung Shum. Real-time texture synthesis by patch-based sampling. *ACM Transactions on Graphics*, 20(3):127–150, 2001.

[LRF93] H. Li, P. Roivainen, and R. Forcheimer. 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, 1993.

[LSZ01] Zicheng Liu, Ying Shan, and Zhengyou Zhang. Expressive expression mapping with ratio images. *Proceedings of ACM Siggraph*, pages 271–276, 2001.

[LSZ09] John C. Langford, Ruslan Salakhutdinov, and Tong Zhang. Learning nonlinear dynamic models. *Proceedings of the International Conference on Machine Learning*, 382:593–600, 2009.

[LTC97] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[LTW95] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Realistic modeling for facial animation. *Proceedings of ACM Siggraph*, pages 55–62, 1995.

[LTZ09] Xiao-min Liu, Hua-chun Tan, and Yu-jin Zhang. New research advances in facial expression recognition. *International Conference on Machine Learning and Cybernetics*, 2009.

[LWS02] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: A two-level statistical model for character motion synthesis. *Proceedings of ACM Siggraph*, pages 465–472, 2002.

[LZK09] Manfred Lau, Bar-Joseph Ziv, and James Kuffner. Modeling spatial and temporal variation in motion data. *ACM Transactions on Graphics (SIGGRAPH ASIA 2009)*, 29(5):171, 2009.

[Mas91] K. Mase. Recognition of facial expression from optical flow. *IEICE transactions*, 74(10):3473–3483, 1991.

[MCC09] Jianyuan Min, Yen-Lin Chen, and Jinxiang Chai. Interactive generation of human animation with deformable motion models. *ACM Transactions on Graphics*, 29(1):9, 2009.

[MD05] Hugo Mercier and Patrice Dalle. Face analysis: identity vs. expressions. *International Society for Gesture Studies Symposium*, 2005.

[Meh68] Albert Mehrabian. Communication without words. *Psychology Today*, 2(4):53–56, 1968.

[Mor70] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[MR74] A. Mehrabian and J.A. Russell. *An Approach to Environmental Psychology*. MIT Press, Cambridge, MA, USA, 1974.

[MRK06] Lori Malatesta, Amaryllis Raouzaiou, and Stefanos Kollias. Mpeg-4 facial expression synthesis based on appraisal theory. *3rd IFIP conference in Artificial Intelligence Applications and Innovations (AIAI)*, 2006.

[MTPT88] Nadia Magnenat-Thalmann, E. Primeau, and Daniel Thalmann. Abstract muscle action procedures for human face animation. *Visual Computer*, 3(5):290–7, 1988.

[MVBV06] Ives Macedo, Emilio Vital Brazil, and Luiz Velho. Expression transfer between photographs through multilinear aam's. *SIBGRAPI, Brazilian Symposium on Computer Graphics and Image Processing*, pages 239–246, 2006.

[NHS88] Monique Nahas, Herve Huitric, and Michel Saintourens. Animation of a b-spline figure. *The Visual computer*, 3(5):272–276, 1988.

[NN98] Jun-yong Noh and Ulrich Neumann. A survey of facial modeling and animation techniques. Technical Report Technical Report 99-705, University of Southern California, 1998.

[NN01] Jun-yong Noh and Ulrich Neumann. Expression cloning. *Proceedings of SIGGRAPH 2001*, pages 21–28, 2001.

[NPCP09] Mohammad Ali Nazari, Pascal Perrier, Matthieu Chabanas, and Yohan Payan. Simulation of muscle-based orofacial movement dynamics using a muscle activation dependent varying constitutive law. *International Workshop on Dynamic Modeling of the Oral, Pharyngeal and Laryngeal Complex for Biomedical Applications*, 2009.

[NPP+08] Mohammad Ali Nazari, Yohan Payan, Pascal Perrier, Matthieu Chabanas, and Claudio Lobos. A continuous biomechanical model of the face: A study of muscle coordination for speech lip gestures. *International Seminar on Speech Production*, 2008.

[OCC88] Andrew Orthony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotion.* Cambridge University Press, Cambridge, UK, 1988.

[OO98] T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using hmm. *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 442–447, 1998.

[OPS08] Magalie Ochs, Catherine Pelachaud, and David Sadek. An empathic virtual dialog agent to improve human-machine interaction. *Autonomous Agent and Multi-Agent Systems (AAMAS)*, pages 89–96, 2008.

[PA96] Penio S. Penev and Joseph J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[Pan09] Maja Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society B*, 364:3505–3513, 2009.

[Par72] Frederick I. Parke. Computer generated animation of faces. *Proceedings of the ACM annual conference SESSION: SIGGRAPH - Computer graphics*, 1:451–457, 1972.

[Par74] Frederick I. Parke. *A Parametric Model for Human Faces.* PhD thesis, University of Utah, 1974.

[PB81] Stephen M. Platt and Norman I. Badler. Animating facial expressions. *ACM SIGGRAPH Computer Graphics*, 15(3):245–252, 1981.

[PB02] Katherine Pullen and Christoph Bregler. Motion capture assisted animation: texturing and synthesis. *ACM Transactions on Graphics (TOG)*, 21(3):501–508, 2002.

[PB03] Catherine Pelachaud and Massimo Bilvi. Computational model of believable conversational agents. In M.-P. Huget, editor, *Communications in Multiagent Systems*, pages 300–317. Springer, Berlin, 2003.

[PB07] Maja Pantic and M.S. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, 2007.

[PBMD07] Julien Peyras, Adrien Bartoli, Hugo Mercier, and Patrice Dalle. Segmented aams improve person-independent face fitting. *British Machine Vision Conference*, 2007.

[PBS96] Catherine Pelachaud, Norman I. Badler, and Mark Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.

[PC97] Curtis Padgett and Garrison Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge, MA, 1997.

[PCNS05] Bongcheol Park, Heejin Chung, Tomoyuki Nishita, and Sung Yong Shin. A feature-based approach to facial expression cloning. *Computer Animation and Virtual Worlds*, 16(3-4):291–303, 2005.

[Per95] Ken Perlin. Real time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):5–15, 1995.

[PF03] Igor S. Pandzic and Robert Forchheimer. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley and Sons, Inc., New York, NY, USA, 2003.

[PG10] E.K. Patterson and A. Gaweda. Toward using dynamics of facial expressions and gestures for person identification. *Proceedings of Computational Intelligence*, 2010.

[PHL+98] FrÃľdÃľric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. Synthesizing realistic facial expressions from photographs. *Proceedings of ACM Siggraph*, pages 75–84, 1998.

[PKC+03] Hyewon Pyun, Yejin Kim, Wonseok Chae, Hyung Woo Kang, and Sung Yong Shin. An example-based approach for facial expression cloning. *Proceedings of the SIGGRAPH/Eurographics symposium on Computer animation*, pages 167–176, 2003.

[PL06] M. Paleari and Christine Laetitia Lisetti. Psychologically grounded avatars expressions. *Emotion and Computing workshop at Annual German Conference on Artificial Intelligence*, 2006.

[Ple03] Robert Pless. Image spaces and video trajectories: Using isomap to explore video sequences. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2:1433, 2003.

[Plu80] Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harper and Row, New York, USA, 1980.

[PP97] Kris Popat and Rosalind W. Picard. Cluster-based probability model and its application to image and texture processing. *IEEE Transactions on Image Processing*, 6(2):268–284, 1997.

[PP01] Stefano Pasquariello and Catherine Pelachaud. Greta: A simple facial animation engine. *6th Online World Conference on Soft Computing in Industrial Appications, Session on Soft Computing for Intelligent 3D Agents*, 2001.

[PP02] Catherine Pelachaud and Isabella Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 13(5):301–312, 2002.

[PP06] Maja Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man and Cybernetics*, 36(2):433–449, 2006.

[PR04] Maja Pantic and Leon J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man and Cybernetics*, 34(3):1449–1461, 2004.

[PRM00] Vladimir Pavlovic, James M. Rehg, and John Maccormick. Learning switching linear models of human motion. *Advances in Neural Information Processing Systems*, 13:981–987, 2000.

[PSS02] Sang Park, Hyun Joon Shin, and Sung Yong Shin. On-line locomotion generation based on motion blending. *Proceedings of the Symposium on Computer Animation*, pages 105–111, 2002.

[PWWH86] A. Pearce, B. Wyvill, G. Wyvill, and D. Hill. Speech and expression: a computer solution to face animation. *Proceedings on Graphics Interface*, pages 136–140, 1986.

[PZ07] Yuru Pei and Hongbin Zha. Stylized synthesis of facial speech motions. *Computer Animation and Virtual Worlds*, 18(4-5):517–526, 2007.

[RCB98] C. Rose, M. F. Cohen, and Bobby Bodenheimer. Verbs and adverbs: multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, 18(5):32–40, 1998.

[RGM06] Jane Reilly, John Ghent, and John McDonald. Investigating the dynamics of facial expression. *Advances in Visual Computing*, 4292:334–343, 2006.

[RM77] J.A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 1977.

[RNtH03] Zsofia Ruttkay, Han Noot, and Paul ten Hagen. Emotion disc and emotion squares: tools to explore the facial expression space. *Computer Graphics Forum*, 22(1):49–53, 2003.

[RP06] Mauricio Radovan and Laurette Pretorius. Facial animation in a nutshell: past, present and future. *Proceedings conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, pages 71–79, 2006.

[RPD01] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. *Applied Regression Analysis: A Research Tool*. Springer, New York, 2001.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[RSH02] Sam T. Roweis, Lawrence K. Saul, and Geoffrey E. Hinton. Global coordination of local linear models. *Advances in Neural Information Processing Systems*, 14, 2002.

[Rus80] J.A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[Ryd87] M. Rydfalk. Candide: A parameterized face. Technical Report S-581 83, Linkoping University, Dept. of Electrical Engineering, 1987.

[SALGS07] Abdul Sattar, Yasser Aidarous, Sylvain Le Gallou, and Renaud Seguier. Face alignment by 2.5d active appearance model optimized by simplex. *International Conference on Computer Vision Systems*, 2007.

[SBS02] Hedvig Sidenbladh, Michael J. Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. *Proceedings of the Conference on Computer Vision*, pages 784–800, 2002.

[SC01] K. L. Schmidt and J. F. Cohn. Dynamics of facial expression: Normative characteristics and individual differences, 2001.

[SC03] K. L. Schmidt and Jeffrey F. Cohn. Signal characteristics of spontaneous facial expressions: automatic movement in solitary and social smiles. *Biological Psychology*, 65(1):49–66, 2003.

[SCCH09] Takaaki Shiratori, Brooke Coley, RakiÃľ Cham, and Jessica K. Hodgins. Simulating balance recovery responses to trips based on biomechanical principles. *Proceedings of the Symposium on Computer Animation*, pages 37–46, 2009.

[Sch41] Harold Schlosberg. A scale for the judgement of facial expressions. *Journal of Experimental Psychology*, 29(6):497–510, 1941.

[Sch52] H. Schlosberg. The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology*, 44(4):229–237, 1952.

[Sch84] Klaus R. Scherer. On the nature and function of emotion: A component process approach. In Klaus R. Scherer and Paul Ekman, editors, *Approaches to Emotion*, pages 293–318. Lawrence Erlbaum Associates, Hillsdale, New Jersey, London, 1984.

[Sch87] Klaus R. Scherer. Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communications*, 1:1–98, 1987.

[SCR09] Robin J.S. Sloan, Malcolm Cook, and Brian Robinson. Considerations for believable emotional facial expression animation. *International Conference on Visualization*, 2009.

[SDF06] Klaus R. Scherer, Elise Dan, and Anders Flykt. What determines a feeling's position in affective space? a case for appraisal. *Cognition and Emotion*, 20(1):92–113, 2006.

[SDT+07] Mingli Song, Zhao Dong, Christian Theobalt, Huiqiong Wang, Zicheng Liu, and Hans-Peter Seidel. Generic framework for efficient 2d and 3d facial expression analogy. *IEEE Transaction on Multimedia*, 9(7):1384–1385, 2007.

[SGM05a] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Appearance manifold of facial expressions. *Computer Vision in Human-Computer Interaction (ICCV workshop on HCI)*, 3766:221–230, 2005.

[SGM05b] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Robust facial expression recognition using local binary patterns. *International Conference on Image Processing*, 2:370–373, 2005.

[SGM06] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. A comprehensive empirical study on linear subspace methods for facial expression analysis. *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, page 153, 2006.

[SK08] Jaewon Sung and Daijin Kim. Pose-robust facial expression recognition using view-based 2d + 3d aam. *IEEE Transaction on Systems, Man and Cybernetics*, 38(4):852–866, 2008.

[SL09] Hyun Joon Shin and Yunjin Lee. Expression synthesis and transfer in parameter spaces. *Computer Graphics Forum*, 28(7):1829–1835, 2009.

[SNF05] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Transactions on Graphics (TOG)*, 24(3):417–425, 2005.

[SP86] Thomas W. Sederberg and Scott R. Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH Computer Graphics*, 20(4):151–160, 1986.

[SR03] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4:119–155, 2003.

[SSB09] Nicolas Stoiber, Renaud Seguier, and Gaspard Breton. Automatic design of a control interface for a synthetic face. *Proceedings of the conference on Intelligent User Interfaces*, pages 207–216, 2009.

[SSM98] Bernhard Schoelkopf, Alexander Smola, and Klaus-Robert Mueller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[SSSE00] Arno Schoedl, Richard Szeliski, David H. Salesin, and Irfan A. Essa. Video textures. *Proceedings of ACM Siggraph*, pages 489–498, 2000.

[SYH07] Kenji Suzuki, Hiroshi Yamada, and Shuji Hashimoto. A similarity-based neural network for facial expression analysis. *Pattern Recognition Letters*, 28(9):1104–1111, 2007.

[Tan06] Emmanuel Tanguy. *Emotions: the Art of Communication Applied to Virtual Actors*. PhD thesis, University of Bath, 2006.

[TBTHF03] J. Teran, S. Blemker, V. Ng Thow Hing, and Ronald Fedkiw. Finite volume methods for the simulation of skeletal muscle. *Proceedings of the Symposium on Computer Animation*, pages 68–74, 2003.

[TdSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[THJ07] Pohsiang Tsai, Tom Hintz, and Tony Jan. Facial behavior as behavior biometric? an empirical study. *Conference on Systems, Man and Cybernetics*, pages 3917–3922, 2007.

[THR06] Graham W. Taylor, Geoffrey E. Hinton, and Sam Roweis. Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems*, 19, 2006.

[TJ81] Frank Thomas and Ollie Johnston. *The Illusion of Life: Disney Animation*. Abbeville Press, Disney Editions, New York, 1981.

[TKC01] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Recognizing action units for facial expression analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[TLJ07] Yan Tong, Wenhui Liao, and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[TLP07] Adrien Treuille, Yongjoon Lee, and Zoran Popovic. Near-optimal character animation with continuous control. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 26:7, 2007.

[TMC07] Barry-John Theobald, Iain A. Matthews, and Jeffrey F. Cohn. Real-time expression cloning using appearance models. *Proceedings conference on Multimodal interfaces*, pages 134–139, 2007.

[TR03] Yee Whye Teh and Sam T. Roweis. Automatic alignment of local representations. *Advances in Neural Information Processing Systems*, 2003.

[TRK+02] Nicolas Tsapatsoulis, Amaryllis Raouzaiou, Stefanos Kollias, Roddy Cowie, and Douglas-Cowie Ellen. Emotion recognition and synthesis based on mpeg-4 faps. In Igor S. Pandzic and Robert Forchheimer, editors, *MPEG-4 Facial Animation - The standard, implementations and applications*, pages 141–167. John Wiley and Sons, Hillsdale, NJ, USA, 2002.

[TW90] Demetri Terzopoulos and Keith Waters. Physically-based facial modeling, analysis, and animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[Val07] M.F. Valstar. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. *Workshop on Human Computer Interaction. ICCV*, 2007.

[VBPP05] Dianel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. *ACM Transactions on Graphics (TOG)*, 24(3):426–433, 2005.

[vdMPvdH07] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: A comparative review. *Published online*, 2007.

[VGS+06] Vinoba Vinayagamoorthy, Marco Gillies, Anthony Steed, Emmanuel Tanguy, Xueni Pan, Celine Loscos, and Mel Slater. Building expression into virtual characters. State of the art report, Eurographics, 2006.

[VT02] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. *European Conference on Computer Vision*, pages 447–460, 2002.

[VY92] Marie-Luce Viaud and Hussein Yahia. Facial animation with wrinkles. *Eurographics Worshop on Animation and Simulation*, 1992.

[Wat87] Keith Waters. A muscle model for animation three-dimensional facial expression. *Proceedings of ACM Siggraph*, 21:17–24, 1987.

[WFH08] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008.

[WH97] D.J. Wiley and J.K. Hahn. Interpolation synthesis for articulated figure motion. *IEEE Computer Graphics and Applications*, 17(6):39–45, 1997.

[Whi89] C.M. Whissel. The dictionary of affect in language. In Robert Plutchik and Henry Kellerman, editors, *Emotion: Theory, Research, and Experience*, pages 113–131. Academic Press, New York, 1989.

[WHL+04] Yang Wang, Xiaolei Huang, Chan-Su Lee, Song Zhang, Zhiguo Li, Dimitris Samaras, Dimitris Metaxas, Ahmed Elgammal, and Peisen Huang. High resolution acquisition, learning and transfer of dynamic 3-d facial expressions. *Computer Graphics Forum*, 23:III: 677–686, 2004.

[Wil90] Lance Williams. Performance-driven facial animation. *Proceedings of ACM Siggraph*, pages 235–242, 1990.

[WK88] Andrew Witkin and Michael Kass. Spacetime constraints. *Proceedings of ACM Siggraph*, pages 159–168, 1988.

[WKSS00] Thomas Wehrle, Susanne Kaiser, Susanne Schmidt, and Klaus R. Scherer. Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology*, 78(1):105–119, 2000.

[WS04] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:988–995, 2004.

[YD94] Yaser Yacoob and Larry S. Davis. Computing spatio-temporal representations of human faces. *Computer Vision and Pattern Recognition*, pages 70–75, 1994.

[YD96] Yaser Yacoob and Larry S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[YL10] Yuting Ye and C. Karen Liu. Synthesis of responsive motion using a dynamic model. *Computer graphics Forum (Proceedings of Eurographics)*, 29(2), 2010.

[YSZ09] Lihua You, Richard Southern, and Jian J. Zhang. Adaptive physics-inspired facial animation. *Motion in Games International Workshop*, pages 207–218, 2009.

[Zaj89] F.E. Zajac. Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control. *Critical reviews in biomedical engineering*, 17(4):359–411, 1989.

[ZG05] Lukasz Zalewski and Shaogang Gong. 2d statistical models of facial expressions for realistic 3d avatar animation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:217–222, 2005.

[Zha99] Zhengyou Zhang. Feature-based facial expression recognition: Sensitivity analysis and experiments with a multi-layer perceptron. *International Journal of Pattern Recognition and Artificial Intelligence*, 13(6):893–911, 1999.

[ZJ03] Yongmian Zhang and Qiang Ji. Facial expression understanding in image sequences using dynamic and active visual information fusion. *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2:1297, 2003.

[ZL05] Chuan Zhou and Xueyin Lin. Facial expressional image synthesis controlled by emotional parameters. *Pattern Recognition Letters*, 26(16):2611–2627, 2005.

[ZLGS03] Qingshan Zhang, Zicheng Liu, Baining Guo, and Harry Shum. Geometry-driven photorealistic facial expression synthesis. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 177–186, 2003.

[ZNKS09] Liming Zhao, Aline Normoyle, Sanjeev Khanna, and Alla Safonova. Automatic construction of a minimum size motion graph. *Proceedings of the Symposium on Computer Animation*, pages 27–35, 2009.

[ZSCS04] Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz. Spacetime faces: High resolution capture for modeling and animation. *Proceedings of ACM Siggraph*, 23(3):548–558, 2004.

[ZWMC07] Shen Zhang, Zhiyong Wu, Helen M. Meng, and Lianhong Cai. Facial expression synthesis using pad emotional parameters for a chinese expressive avatar. *Proceedings of the Conference on Affective Computing and Intelligent Interaction*, 4738:24–35, 2007.

[ZZL05] Qijun Zhao, David Zhang, and Hongtao Lu. Supervised lle in ica space for facial expression recognition. *International Conference on Neural Networks and Brain*, 3:1970–1975, 2005.

VU :

Le Directeur de Thèse

VU :

Le Responsable de l'Ecole Doctorale

VU pour autorisation de soutenance

Rennes, le

Le Président de l'Université de Rennes1

Guy CATHELINEAU

VU après soutenance pour autorisation de publication :

Le Président de Jury,

Abstract

In all computer-graphics applications, one stimulating task has been the integration of believable virtual characters. Above all other features of a character, its face is arguably the most important one since it concentrates the most essential channels of human communication. The road toward more realistic virtual characters inevitably goes through a better understanding and reproduction of natural facial expressiveness.

In this work we focus on emotional facial expressions, which we believe represent the most interesting type of non-verbal facial communication. We propose an animation framework that learns practical characteristics of emotional facial expressions from human faces, and uses these characteristics to generate realistic facial animations for synthetic characters.
Our main contributions are:
- A method that automatically extracts a meaningful representation space for expressive facial deformations from the processing of actual data. This representation can then be used as an interface to intuitively manipulate facial expressions on any virtual character.
- An animation system, based on a collection of motion models, which explicitly handles the dynamic aspect of natural facial expressions. The motion models learn the dynamic signature of expressions from data, and reproduce this natural signature when generating new facial movements.

The obtained animation framework can ultimately synthesize realistic and adaptive facial animations in real-time interactive applications, such as video games or conversational agents. In Addition to its efficiency, the system can easily be associated to higher-level notions of human emotions; this makes facial animation more intuitive to non-expert users, and to affective computing applications that usually work at the semantic level.

Résumé

Dans les mondes virtuels, une des tâches les plus complexes est l'intégration de personnages virtuels réalistes et le visage est souvent considéré comme l'élément le plus important car il concentre les canaux de communications humains les plus essentiels. La création de personnages virtuels convaincants passe ainsi par une meilleure compréhension et une meilleure reproduction de l'expressivité faciale naturelle.

Dans ces travaux, nous nous concentrons sur les expressions faciales émotionnelles, qui selon nous représente le plus intéressant aspect de la communication non-verbale. Nous proposons une approche qui apprend les caractéristiques des expressions faciales directement sur des visages humains, et utilise cette connaissance pour générer des animations faciales réalistes pour des visages virtuels.
Nos contributions sont les suivantes:
- Une méthode capable d'extraire de données brutes un espace simple et pertinent pour la représentation des expressions faciales émotionnelles. Cet espace de représentation peut ensuite être utilisé pour la manipulation intuitive des expressions sur les visages de n'importe quel personnage virtuel.
- Un système d'animation, basé sur une collection de modèles de mouvement, qui pilote l'aspect dynamique de l'expressivité faciale. Les modèles de mouvement apprennent la signature dynamique des expressions naturelles à partir de données, et reproduisent cette signature lors de la synthèse de nouvelles animations.

Le système global d'animation issu des ces travaux est capable de générer des animations faciales réalistes et adaptatives pour des applications temps-réel telles que les jeux vidéos ou les agents conversationnels. En plus de ses performances, le système peut être associé aux notions plus abstraites d'émotions humaines. Ceci rend le processus d'animation faciale plus intuitif, en particulier pour les utilisateurs non-experts et les applications d''affective computing' qui travaillent généralement à un niveau sémantique.