Université Paul Sabatier – Toulouse III
D.E.A **Informatique de l'Image et du Langage** 2002-2003
Responsable du D.E.A : René CAUBET
Equipe **Traitement et Compréhension d'Images**

Analyse automatique des expressions du visage

Application à la Langue des Signes

Hugo MERCIER

Directeur de Recherche : Patrice DALLE
Responsable du stage : Patrice DALLE

Mots clefs : Expressions du visage, Langue des Signes.
Résumé : Le but de cette étude est l'analyse informatique des expressions du visage présentes dans un contexte de communication Homme-Homme. Les expressions du visage sont à prendre ici comme une composante d'un langage articulé. Elles interviennent dans les formes de communication co-verbale et en particulier dans la Langue des Signes, qui est ici l'application principale.
L'étude consiste dans un premier temps en un état de l'art des recherches dans le domaine de l'analyse des expressions du visage, d'abord en présentant le système humain d'analyse du visage et des expressions, en présentant les principales approches informatiques utilisées dans la littérature, puis en détaillant ces approches.
Des spécificités de la Langue des Signes, sont tirées les spécifications d'un système d'analyse des expressions. Le point central de cette étude est la définiton d'un formalisme de représentation des connaissances linguistiques.
Enfin de ce formalisme est tirée une architecture de système permettant l'analyse linguistique des expressions du visage présentes en Langue des Signes.

Keywords : Facial expressions, Sign Language.
Abstract : The goal of this work is the computer-aided analysis of facial expressions, in a context of Man to Man communication. Here, Facial expressions are considered as a component of an articulated language. It takes part in every kind of co-verbal communication and particularly in the Sign Language, which is here the main application.
The work consists in a state of the art in the domain of facial expressions analysis, first in introducing the human face and facial expressions processing system, in introducing the main approaches found in the literature, then in giving details about these approaches.
From specifities of the Sign Language are given specifications of a facial expressions analysis system. The main part of this work is the definition of a formalism to describe linguistics knowledges.
Finally, from this formalism, is given the structure of a system allowing a linguistic analysis of facial expressions found in the Sign Language.

Table des matières

Introduction

L'expression du visage est un des moyens le plus puissant, naturel et immédiat pour un humain de communiquer ses émotions ou intentions. Les premiers travaux sur le domaine sont dûs principalement à Darwin qui proposa une théorie évolutionniste sur les expressions du visage. Les expressions du visage forment un langage particulier dont l'étude est très importante dans le domaine des sciences comportementales, de la psychologie et de la Langue des Signes. Aujourd'hui, l'analyse assistée par ordinateur du visage et de ses expressions est un domaine émergent. Les applications sont nombreuses.

L'analyse et le codage des expressions du visage dans une séquence vidéo permet de réduire la quantité d'informations à transmettre. C'est en particulier un des objectifs du standard MPEG-4 ([49]) : une partie du protocole est lié à la reconnaissance du visage et de ses expressions. Le mouvement d'un personnage peut alors être reconstruit à partir d'un ensemble réduit de paramètres. Le processus d'analyse permet alors une compression des données.

En Interaction Homme-Machine, la mesure de la direction du regard de l'utilisateur pourrait être un moyen efficace d'effectuer certaines tâches dans une interface graphique (comme la sélection d'une fenêtre ou d'une zone de saisie). Les expressions du visage d'un utilisateur donnent aussi une indication de son état émotionnel. Une interface construite autour de ces nouvelles informations permettrait un meilleur retour d'informations de la part du logiciel (*feedback*). Les expressions du visage peuvent ainsi être vues comme une nouvelle modalité d'interaction.

D'une manière générale, les expressions du visage prennent une place importante dans le processus de communication humain et forment à elles seules un langage co-verbal. Les expressions du visage (au sens de mouvements musculaires) accompagnent le langage parlé, aussi bien en terme de mouvements physiques nécessaires à la parole (mouvements des lèvres), qu'en terme d'indicateur émotionnel accompagnant le langage parlé. Elles expriment une part non-négligeable du sens dans une communication orale. En particulier dans la Langue des Signes où les expressions du visage font partie intégrante du langage gestuel : elles peuvent être des unités grammaticales - la direction du regard peut servir de référence grammaticale (comme les pronoms en français) - ou représenter une intonation. *En LSF, les « intonations du visage » ont les mêmes fonctions que les intonations de la voix chez les entendants* ([28]).

Les chapitres 1 et 2 constituent un état de l'art sur l'analyse du visage. Le chapitre 1 définit les principaux concepts, présente les caractéristiques du système de vision humain liés à l'analyse du visage et introduit les principales approches utilisées pour l'analyse automatique. Le chapitre 2 détaille les principaux formalismes de description des expressions ainsi que les méthodes des différentes approches et souligne les spécificités d'une analyse des expressions pour la Langue des Signes.

Le chapitre 3 présente les problèmes liés à l'analyse informatique de la Langue des Signes, en détaillant quelques méthodes trouvées dans la littérature.

Le chapitre 4 présente un formalisme informatique pour la représentation des connaissances linguistiques et une architecture de système d'analyse des expressions.

Le chapitre 5 détaille la manière dont le système d'analyse fontionne.

Enfin, l'annexe A détaille les principales méthodes d'analyse automatique du visage.

1

Analyse

On présente dans ce chapitre un rappel des connaissances sur le domaine de l'analyse du visage, et des expressions du visage en particulier : quelles sont les particularités du visage humain et du système humain d'analyse du visage. Dans une deuxième partie, on présente les principales approches employées pour l'analyse *automatique* du visage. Ces approches seront développées dans le prochain chapitre.

1.1 Définitions

Expressions et émotions sont très liées et parfois confondues, c'est pour cette raison qu'on se tiendra aux définitions suivantes par la suite :

Mimique faciale : une mimique faciale est un état du visage composé par un ensemble de configurations des muscles faciaux. Le sourire est par exemple une mimique faciale composé d'un certain nombre d'activation des muscles faciaux (mouvements des muscles zygomatiques).

Emotion : l'émotion est un des générateurs des expressions faciales. L'émotion se traduit via de nombreux « canaux » comme la position du corps, la voix et les expressions faciales. Une émotion implique généralement une expression faciale correspondante (dont l'intensité peut être plus ou moins contrôlée selon les individus), mais l'inverse n'est pas vrai : il est possible de « mimer » une expression représentant une émotion sans pour autant ressentir cette émotion. Alors que les expressions dépendent des individus et des cultures, on distingue généralement un nombre limité d'émotions universellement reconnues.

Expression faciale : une expression faciale est une mimique faciale chargée de sens. Le sens peut être l'expression d'une émotion, un indice sémantique ou une intonation dans la Langue des Signes.

L'interprétation d'un ensemble de mouvements musculaires en expression est dépendante du contexte d'application. Dans le cas d'une application en interaction Homme-Machine où l'on désire connaître une indication sur l'état émotionnel d'un

individu, on cherchera à classifier les mesures en terme d'émotions. Pour une application en Langue des Signes, les mesures seront combinées pour contruire un sens, qui ne reflète pas forcément l'état émotionnel de l'individu.

1.2 Psychophysiologie du visage

Avant de s'intéresser à l'analyse automatique des expressions du visage, on doit connaître les particularités du visage : son anatomie et le lien entre le visage et les émotions.

1.2.1 Physiologie du visage

Le visage est une zone importante du corps humain qui possède une trentaine de muscles. L'électromyographie (EMG) est une technique permettant de mesurer l'activité musculaire au cours du temps. Cette technique a permis de déduire que l'activation musculaire, et en particulier l'activation des muscles du visage, peut généralement être découpée en trois phases :
– la phase d'attaque (« attack » ou « onset »), qui correspond à la période pendant laquelle l'activité du muscle passe de la valeur nulle à sa valeur maximale,
– la phase de soutien (« sustain » ou « apex »), qui correspond à la période pendant laquelle l'activité stagne à son maximum,
– la phase de relâchement (« relaxation » ou « offset »), qui correspond à la période pendant laquelle l'activité du muscle baisse jusqu'au niveau initial.

Indépendance des muscles

Il est à noter que les muscles de la zone supérieure du visage n'ont que peu d'influence sur les muscles de la zone inférieure et vice-versa ([21]). Il est donc possible de découper l'analyse en deux zones.

1.2.2 Psychologie du visage

Les expressions faciales peuvent former une indication sur l'état émotionnel d'un individu : ce sont les expressions dites « spontanées » en contraste avec les expressions qui peuvent être « forcées ». Les deux types d'expression sont générées par deux zones distinctes du cerveau ([12]).

On distingue par exemple deux types de sourire, dit « de Duchenne » ou non. Le sourire de Duchenne est un sourire sincère reflétant une émotion positive ; l'activation des muscles entourant les yeux accompagne en général ce sourire. Il semblerait

aussi que les mesures temporelles (onset, apex et offset) de l'activation des muscles soient différentes entre les deux types de sourire.

La distinction entre les deux types peut être mise en évidence si les différentes mesures sont suffisamment précises.

Aspect universel des émotions

Ekman et Friesen ([21]) ont établi qu'il existe un nombre limité d'expressions reconnues par tous, indépendamment de la culture. Ces expressions innées correspondent aux sept émotions suivantes : la **neutralité**, la **joie**, la **tristesse**, la **surprise**, la **peur**, la **colère** et le **dégoût**.

1.3 Système humain d'analyse du visage

On présente ici quelques informations relatives à l'analyse des visages chez l'humain. La *détection* du visage consiste à isoler dans quelle(s) zone(s) d'une image se trouve(nt) le(s) visage(s). La *reconnaissance* du visage consiste à retrouver à qui appartient un visage particulier.

Ces informations sont tirées principalement d'un état de l'art sur la reconnaissance des visages ([8], voir aussi [44]). Il est à noter que beaucoup de résultats dans ce domaine sont dus à l'étude d'une maladie neurologique : la *prosopagnosie*. Les prosopagnosiques ne *reconnaissent* pas ce qui fait l'identité d'un visage.

Le processus de reconnaissance du visage est un **processus dédié** chez l'humain. Trois indications principales permettent de vérifier cette proposition :

1. Les visages sont mieux mémorisés par les êtres humains que les autres objets,
2. Les prosopagnosiques n'identifient pas les visages. Bien qu'ils reconnaissent parfaitement les différents composants du visage (nez, bouche, yeux) et qu'ils sachent dire si un objet est un visage ou non, ils sont incapables de reconstituer ce qui forme l'individualité d'une personne. Ils reconnaissent ausi les expressions du visage (et les émotions sous-jacentes). Les prosopagnosiques reconnaissent les différences intra-individus, mais pas les différences inter-individus.
3. Le processus de reconnaissance des visages est un processus inné puisque les nouveaux-nés préfèrent suivre du regard des objets ressemblant à un visage plutôt que d'autres.

Le **mouvement** joue un rôle important dans la détection des visages familiers. Un visage familier est plus facile à détecter s'il est en mouvement que s'il est statique. Cependant, le mouvement n'apporte rien pour la détection des visages inconnus.

Les prosopagnosiques reconnaissent les expressions du visage, mais ont du mal à *identifier* ce visage. C'est donc que le processus de reconnaissance des expressions est un **processus parallèle** au processus de reconnaissance du visage.

On distingue deux types d'analyses du visage effectuées par le cerveau humain : l'une dite « globale » (*wholistic*) où le visage est traité comme un tout ([44]) et l'autre dite « par composantes » (*feature-based*) où le visage est vu comme un ensemble de composantes (yeux, nez, bouche, etc.).

Le processus de détection du visage consiste à isoler une zone qui « ressemble » à un visage « générique ». L'approche globale semble être la plus naturelle au problème de détection du visage, bien qu'il soit tout à fait possible de détecter un visage par une approche plus locale, en détectant le clignement des yeux par exemple.

D'un certain point de vue, le processus de détection du visage est antagoniste au processus d'identification. Détecter un visage utilise ce qui est commun à tous les visages, en ignorant donc les différences inter-individus. Identifier un visage particulier consiste au contraire à exploiter ces différences inter-individus pour la discrimination et donc la reconnaissance.

Le processus de reconnaissance des expressions, quant à lui, se base sur les différences intra-individus. *« facial expression identification requires finding something* common *across individuals, while face identification requires finding something* different »([44]).

Identifier un visage semble faire appel la plupart du temps à une analyse globale suivie d'une analyse par composantes pour affinement, bien que dans certains cas (quand certaines composantes sont « marquées » chez certains individus - grandes oreilles, nez tordu, etc.), une analyse globale suffise.

A l'inverse, reconnaître une expression fait appel généralement à une analyse locale (par composantes). En effet, il semblerait que le modèle utilisé par les humains pour reconnaitre une expression puisse se résumer à une indication sur la « forme » des composantes du visage. Ainsi un ensemble de points représentants la position de chaque composante du visage suffit pour qu'un humain reconnaisse une expression.

1.4 Système automatique d'analyse du visage

Quelque soit le type d'analyse (*détection* du visage, *reconnaissance* du visage, *reconnaissance* des expressions), il existe généralement deux approches principales pour traiter le problème : globale (*image-based*) ou par composantes (*feature-based*). On présente ici ces deux approches ainsi que les différentes approches existantes pour l'analyse de l'aspect *temporel* du visage et de ses mouvements. Les deux approches sont détaillées dans le chapitre suivant et les méthodes sont

développées en annexe.

Les différentes approches présentées ici sont généralement combinées. Bien que l'approche globale semble plus adaptée à la détection et la reconnaissance du visage et l'approche par composantes plus adaptée à l'analyse des expressions, les méthodes utilisées dans la pratique sont généralement une combinaison des deux approches.

1.4.1 Approche par composantes

L'approche par composantes – approche la plus couramment utilisée et la plus ancienne – consiste à considérer le visage comme un ensemble de composantes (yeux, nez, bouche, etc.).

Détecter un visage par cette approche consiste à identifier dans l'image les zones qui contiennent un ensemble de composantes faciales réparties spatialement d'une façon particulière.

Reconnaître un visage par cette approche consiste à *vérifier* la présence et les caractéristiques de certaines composantes précédemment mémorisées[1].

Reconnaître les expressions par cette approche, consiste généralement à détecter la présence[2] et les caractéristiques des composantes[3]. De plus, comme les expressions ont généralement une dimension temporelle significative, l'analyse est alors effectuée par des opérateurs qui intègrent la dynamique.

L'analyse du visage consiste donc à appliquer un ensemble d'opérateurs spécialisés. En tenant compte des connaissances sur ces différentes composantes, il est possible de construire des opérateurs spécialisés dans la détection et la mesure des caractéristiques d'une composante.

Les sourcils sont par exemple, quand ils sont présents, généralement plus foncés que le reste du visage et sont plus longs que larges. Ainsi, un opérateur basé sur le gradient de l'image peut être un opérateur adéquat pour les sourcils.

Les yeux, quand ils ne sont pas fermés, sont formés d'une zone sphérique (l'iris) entourée d'une zone très claire (la sclérotique). Un détecteur de zones claires peut former la base d'un opérateur efficace de détection des yeux.

1.4.2 Approche globale

L'approche globale (*image-based*), plus récente, consiste à considérer le visage comme un tout. Les méthodes de cette famille sont des méthodes de reconnaissance

[1] les opérateurs de *vérification* et de *détection* peuvent ne pas être les mêmes

[2] certaines composantes peuvent être cachées

[3] le nombre de composantes et leurs caractéristiques respectives sont dépendants du domaine d'application

des formes ou de mise en correspondance de modèles (*template-matching*).

La détection d'un visage par cette approche consiste alors à comparer une sous-image avec un modèle du visage. Le modèle du visage est généralement construit automatiquement à partir d'un ensemble de visages d'apprentissage (contrairement à l'approche par composantes ou le modèle du visage est généralement implicite et donné par des experts – voire par le concepteur du système d'analyse lui-même). Une mesure d'erreur entre le modèle et le visage observé permet d'avoir une idée de la « ressemblance ».

Il est possible que le modèle du visage ait été construit *a priori* , à partir de connaissances d'expert. On trouve généralement dans cette catégorie les modèles 3D du visage (Candide par exemple, voir plus loin).

La reconnaissance d'expressions par cette approche consiste par exemple à comparer un visage observé avec un modèle d'expression appris lui aussi à partir d'un certain nombre d'exemples.

Cependant, cette méthode est difficile à utiliser puisque le corpus d'apprentissage doit être très difversifié et doit représenter toutes les combinaisons possibles des mucles faciaux.

Reconnaître un visage par cette approche consiste à mesurer la différence entre le visage observé et chaque visage précédemment mémorisé. Le visage ayant le score de ressemblance le plus fort (à condittion que le score dépasse un certain seuil pour traiter le cas où aucun visage n'est reconnu) est alors le visage reconnu.

1.4.3 Aspect dynamique

Un aspect essentiel d'un système d'analyse automatique du visage est l'aspect temporel. Bien que les problèmes de détection et de reconnaissance des visages puissent se limiter à une analyse statique, la dimension temporelle semble très importante pour un système d'analyse des expressions[4] .

La mesure de la dynamique du visage et/ou de ses composantes permet généralement de guider (voire d'améliorer) l'analyse statique. Par exemple, si on sait détecter les mouvements du visage entre deux images, on pourra éviter l'étape de détection du visage[5]

[4] qu'il soit informatique ou humain. En effet, il est difficile de distinguer clairement une mimique faciale à partir d'une image fixe et ce, même pour un humain

[5] on peut aussi *vérifier* la prédiction de l'opérateur de suivi

2

Etat de l'art

On présente dans ce chapitre une vue d'ensemble des travaux précédemment entrepris dans le domaine de l'analyse du visage, tout d'abord avant, puis pendant l'ère informatique. Les travaux sur l'analyse manuelle des expressions du visage ont donné les premiers formalismes, qui ont été repris puis parfois étendus par les travaux sur l'analyse automatique des expressions du visage.

2.1 Détection du visage

Le problème de détection automatique des zones contenant un visage est un problème relativement bien couvert dans la littérature. On présente ici quelques uns de ces systèmes. On pourra se referer pour plus de détails aux états de l'art sur le sujet ([2] ou [7] par exemple).

Tous les systèmes de détection du visage consistent à appliquer une méthode de détection à des sous-fenêtres de l'image initiale. Cette recherche de sous-fenêtres doit être la plus exhaustive possible et est critique en temps de calcul. C'est pourquoi l'analyse consiste généralement en une première phase dont le but est de localiser les différentes régions de l'image susceptibles d'être des visages en appliquant un détecteur peu robuste[1]. Les méthodes de détection proprement dite (plus robustes) sont ensuite appliquées à ces régions.

Les techniques de localisation des zones de l'image candidates se font sur l'ensemble de l'image et sont généralement orientées bas-niveau : contours, couleur, mouvement et mesures.

2.2 Analyse humaine des expressions du visage

Les travaux fondamentaux dans le domaine de l'analyse des expressions du visage sont dus principalement à Charles Darwin, Guillaume Duchenne De Boulogne au

[1] dans le sens où peut se permettre de détecter plus de zones de visages qu'il n'y en a réellement, mais en aucun cas moins

XIXème siècle et plus récemment Paul Ekman.

Au XIXème siècle, Guillaume Duchenne de Boulogne est le premier à localiser individuellement les différents muscles faciaux par activation électrique[2]. Il est un des premiers à livrer à la communauté scientifique un ensemble de photographies montrant l'activation des différents muscles faciaux

FIG. 2.1 – Exemples de photographies de Duchenne de Boulogne

Charles Darwin, est le premier à traiter de l'universalité des expressions du visage et à proposer une théorie évolutionniste sur la formation des expressions. L'argument principal est que les expressions des enfants et des nouveaux-nés existent aussi chez les adultes. D'après lui, l'expression des émotions est un processus nécessaire à la survie. Ainsi, les expressions non-verbales sont aussi importante que les interactions verbales dans le processus de communication humain.

Paul Ekman, psychologue, s'intéresse à partir du milieu des années 1960 aux expressions et émotions humaines. Il met en évidence l'universalité de certaines émotions et développe un outil de codification des expressions du visage largement utilisé aujourd'hui. Il s'intéresse désormais à l'analyse des expressions de manière informatique.

[2]Il avait trouvé un cobaye idéal : une personne souffrant d'insensibilité musculaire au niveau du visage. C'est ce qui a permis d'utiliser l'électricité et de prendre le temps de « poser » pour les photographies de l'époque

2.3 Formalismes de description

La description des expressions du visage est un ensemble d'interprétations successives d'indices visuels. Un sens (dépendant du domaine d'application) est construit à partir de mesures de bas niveaux (présence ou non d'une composante, position éventuelle). Ces mesures sont combinées successivement spatiallement, puis de manière temporelle pour former le sens attendu (émotion sous-jacente par exemple). On introduit ici un vocabulaire nécessaire à la description.

Attribut facial : un attribut facial est une propriété élémentaire « centrée objet » caractérisant un visage. La position des yeux est un attribut facial. La présence de barbe est un autre attribut. Les attributs faciaux directement visibles sont dit de premier ordre. Les attributs qui ne peuvent être mesurés qu'à partir d'autres attributs de premier ordre, sont des attributs de second ordre et ainsi de suite ([44]). La mesure de distance entre les deux yeux est un attribut du second ordre, par exemple, puisqu'elle ne peut être calculée qu'à partir de la position respective des deux yeux qui est une mesure du premier ordre.

Certains attributs sont invariants par rapport au temps pour un même individu : la distance entre les deux yeux est un attribut statique, par exemple, composé cependant de deux attributs dynamiques (position des yeux qui change au cours du temps).

Indice visuel : un indice visuel est une propriété élémentaire « centrée observateur » du visage : c'est un attribut facial qui est observé et visible. Certains attributs ne sont pas visibles chez certaines personnes (barbe, moustache, sourcils) ; certains ne sont visibles qu'à certains moments (un oeil peut être caché lors d'une rotation de la tête par exemple).

Action faciale : une action faciale est un ensemble d'indices visuels intégrés de manière temporelle. Le relèvement des sourcils est par exemple une action faciale composée d'un ensemble de positions successives des sourcils. Une action faciale est généralement décrite par sa dynamique : le relèvement des sourcils consiste en une position actuelle des sourcils plus haute que sa position précédente. Les actions faciales sont généralement caractérisées par leur « profil temporel » : durée d'attaque, durée de maintien et durée de relâchement.

Composante faciale : une composante faciale est une partie du visage. Le découpage en composantes est celui du langage naturel : les yeux, le nez, la bouche, les joues, les sourcils, la barbe, etc. Bien que certaines puissent être entièrement caractérisées par un ensemble d'attributs faciaux (les yeux peuvent être caractérisés par leur forme, leur couleur, la présence et la longueur des cils, etc.), d'autres ne sont que des mesures floues et sont difficiles à caractériser à partir d'indices visuels objectifs élémentaires. C'est le cas par exemple des joues dont les limites sont difficiles à fixer, même pour un observateur humain. Cependant, ces mesures sont importantes puisqu'un humain a plus de facilités à manipuler des données floues

que des données précises (« les joues sont gonflées »).

2.3.1 FACS

En 1978, Ekman et Friesen présentent un système de codification *manuelle* des expressions du visage ([22]). Leurs travaux d'observation leur permettent de décomposer tous les mouvements *visibles* du visage en terme de 46 *Actions Unitaires* (*Action Unit* , qui correspond aux *actions faciales* définies plus haut) qui décrivent les mouvements *élementaires* des muscles. N'importe quelle mimique observée peut donc être représentée sous le forme d'une combinaison d'*Actions Unitaires* . Ce système de codage est connu sous le nom de « Facial Action Coding System » (FACS).

FACS s'est imposé depuis comme un outil puissant de description des mimiques du visage, utilisé par de nombreux psychologues.

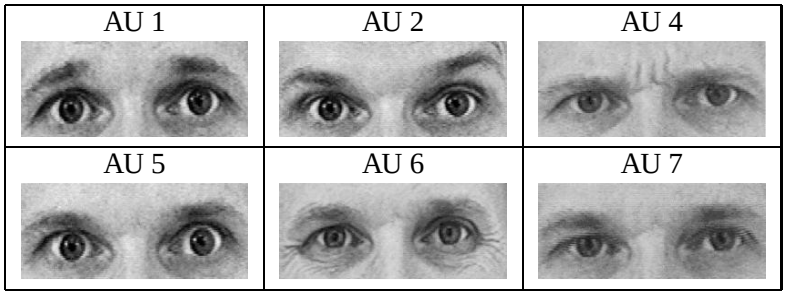


FIG. 2.2 – Exemples d'*Action Units* du haut du visage

2.3.2 FACS+

Bien que FACS soit un système de description bénéficiant d'une grande maturité (environ vingt années de développement), il souffre cependant de quelques inconvénients ([24]) :

Complexité : on estime qu'il faut 100 heures d'apprentissage pour en maîtriser les principaux concepts ([11]),

Difficulté de manipulation par une machine : FACS a d'abord été créé pour des psychologues, Certaines mesures restent floues et difficilement évaluables par une machine.

Manque de précision : les transitions entre deux « états » d'un muscle sont représentées de manière linéaire, ce qui est une approximation de la réalité. En particulier les mesures temporelles de l'activation des muscles faciaux (onset, apex et offset) ne sont pas mises en évidence.

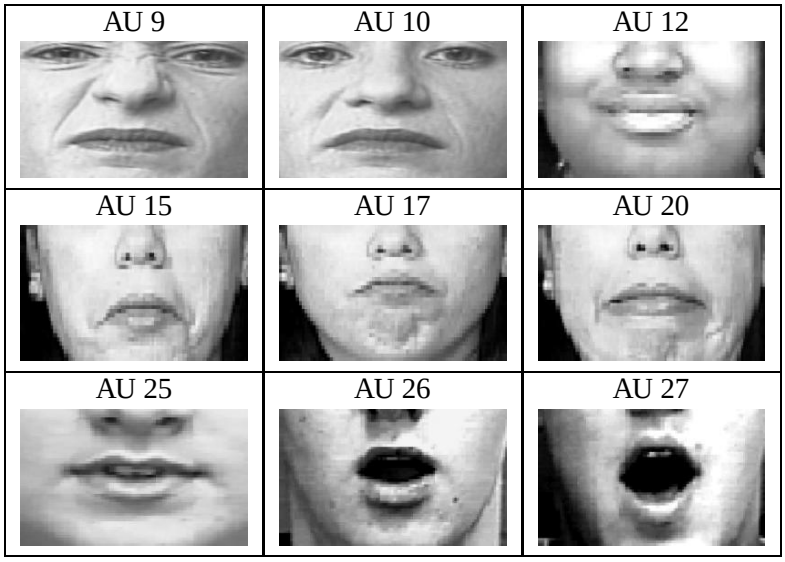


FIG. 2.3 – Exemples d'*Action Units* du bas du visage

Essa ([24]) dispose d'un modèle musculaire plus facilement manipulable par la machine (FACS+). Cependant, la principale critique ([11]) est sa difficulté d'interprétation par un humain.

2.3.3 MPEG4

La norme de codage vidéo MPEG-4 ([49]) dispose d'un modèle du visage humain développé par le groupe d'interêt « Face and Body AdHoc Group ». C'est un modèle 3D articulé.

Ce modèle est construit sur un ensemble d'attributs faciaux, appelés « Facial Feature Points » (FFP). Des mesures sur ces FFP sont effectuées pour former des unités de mesure (Facial Animation Parameter Units) qui servent à la description des mouvements musculaires (Facial Animation Parameters - équivalents des Actions Units d'Ekman).

Les *Facial Animation Parameter Units* (FAPU) permettent de définir des mouvements élémentaires du visage ayant un aspect naturel. En effet, il est difficile de définir les mouvements élémentaires des muscles de manière absolue : le déplacement absolu des muscles d'une personne à l'autre change, mais leur déplacement relatifs à certaines mesures pertinentes sont constantes. C'est ce qui permet d'animer des visages de manière réaliste et peut permettre de donner des expressions humaines à des personnages non-humains.

Comme exemples de FAPU, on peut citer *la largeur de la bouche, la distance de séparation entre la bouche et le nez, la distance de séparation entre les yeux et le*

nez, etc.

Par exemple, l'étirement du coin de la lèvre gauche (Facial Animation Parameter 6 `stretch_l_cornerlip`) est défini comme le déplacement vers la droite du coin de la lèvre gauche d'une distance égale à la longueur de la bouche. Les FAPUs sont donc des mesures qui permettent de décrire des mouvements élémentaires et donc des animations.

Cependant, les *Facial Animation Parameters* (FAP) de MPEG-4 ne représentent pas directement des mouvements du visage réalistes, contrairement à FACS. FACS décrit un ensemble de mouvements musculaires, alors que MPEG-4 décrit un ensemble de mouvements « visuels » qui ne sont pas forcément réalistes. Par exemple, l'Action Unit (AU) 26 de FACS (« Jaw Drop ») décrit le mouvement d'abaissement du menton ; cet abaissement est accompagné d'un abaissement de la lèvre inférieure. L'abaissement du menton de MPEG-4 (FAP 3 - `open_jaw`) ne décrit pas l'abaissement de la lèvre inférieure : la description n'est donc pas réaliste d'un point de vue musculaire.

On peut donc considérer que les FAPs de MPEG-4 sont des descriptions de plus bas niveau que les AUs de FACS.

FIG. 2.4 – exemple d'animation MPEG-4 - logiciel `wireface`([50]

2.3.4 Candide

Candide ([9]) est un modèle du visage. Il est composé d'un modèle en fil de fer représentant un visage « générique » et d'un ensemble de paramètres :

– **Paramètres de « forme »** (*Shape Units*) : ces paramètres permettent d'adapter le modèle générique à un individu particulier. Ils représentent les différences inter-individus et sont au nombre de 12 :
  1. Hauteur de la tête,
  2. Position verticale des sourcils,
  3. Position verticale des yeux,
  4. Largeur des yeux,
  5. Hauteur des yeux,
  6. Distance de séparation des yeux,
  7. Profondeur des joues,
  8. Profondeur du nez,
  9. Position verticale du nez,
  10. Degré de « courbure »du nez (s'il pointe vers le haut ou non),
  11. Position verticale de la bouche,
  12. Largeur de la bouche.

– **Paramètres d'animation** (Animation Units) : ces paramètres représentent les différences intra-individus, *i.e.* les différentes actions faciales. Ils sont composés d'un sous-ensemble de FACS et d'un sous-ensemble des FAPs de MPEG-4. Les FAPs sont définis par rapport à leur FAPU correspondant.

Ces paramètres, qu'ils soient d'animation ou de forme, sont représentés sous forme d'une liste de points du modèle de fil de fer à mettre à jour. Candide permet de voir clairement la différence entre les AUs de FACS et les FAPs de MPEG-4 : les AUs de FACS sont exprimées de manière absolue, à la différence des FAPs qui sont exprimés par rapport à des mesures du visage (les FAPUs).

On peut trouver une *applet* Java sur le site de l'auteur de Candide qui permet d'activer certaines AUs de FACS sur un modèle en fil de fer à l'adresse suivante : `http ://www.icg.isy.liu.se/candide/javacandide.html`

2.3.5 SignStream

Le *National Center for Sign Language and Gesture Resources* de l'université de Boston a défini une notation ([33]) permettant à des linguistes de coder les différents gestes de la Langue des Signes américaine. En particulier, un sous-ensemble du système de notation est dédié au codage des expressions du visage recontrées en Langue des Signes. Les différents mouvements du visage ont été choisis pour leur valeur linguistique.

Le logiciel `SignStream` du projet est un « éditeur de partitions » qui permet de décrire les différents gestes (du corps et du visage) effectués par le sujet observé au cours du temps.

Il apparaît clairement que la Langue des Signes est beaucoup plus riche en expressions que le langage co-verbal « classique ». En particulier certaines actions faciales considérées comme ayant peu d'importance pour les précédents formalismes, s'avèrent avoir une importance cruciale pour la langue des signes.

Ainsi, pour la langue des signes (américaine), il existe une différence entre un sourire où la langue est visible et un sourire où la langue n'est pas visible. La présence de la langue entre les dents constitue aussi un élément de sens.

Les précédents formalismes ne sont donc pas complètement adaptés à l'étude de la Langue des Signes. La langue n'est par exemple pas modélisée dans Candide et les mouvements de la joue par exemple ne sont que peu mis en avant.

2.4 Analyse automatique des expressions du visage

Les objectifs des systèmes d'analyse des expressions du visage sont multiples. On présente ici un ensemble de travaux sur l'analyse des expressions du visage. On pourra se référer aux états de l'art existants pour plus de détails ([36] [25]).

En **animation**, on cherche à animer des personnages virtuels qui doivent paraître le plus réaliste possible. On ajoute alors aux mouvements des muscles faciaux nécessaires à la parole, un ensemble de mouvements faciaux qui traduisent un état émotionnel. L'analyse de la formation des expressions est donc nécessaire aussi bien pour la description que pour la synthèse. On s'attarde surtout à reconstruire une expression qui semble réaliste d'un point de vue visuel et qui est porteuse d'un sens (d'une émotion), en accentuant éventuellement les expressions (pour des personnages caricaturés par exemple).

En **Interaction Homme-Machine**, on cherche à avoir une idée de l'état émotionnel de l'utilisateur pour la conception d'interfaces plus ergonomiques et présentant un meilleur retour d'informations (*feedback*). Bien que beaucoup de travaux dans ce domaine tentent de classer les expressions de l'utilisateur en émotions universelles, certains se focalisent sur des composantes particulières du visage qui servent à l'intéraction (suivi du mouvement des yeux pour la sélection par exemple). On ne cherche pas ici à avoir une description fine des expressions et des mouvements musculaires sous-jacents, mais plutôt à avoir une idée du mouvement de certaines composantes ou à avoir une idée d'un état émotionnel.

La **compression de données** s'intéresse à la description des expressions du visage. Le principe est de coder les expressions des visages présents dans une séquence vidéo et donc de réduire le nombre d'informations à transmettre. Encore une fois, on cherche à ce que la reconstruction à partir de ce codage soit la plus réaliste

possible d'un point de vue visuel.

Dans le domaine de la **linguistique**, des **sciences comportementales**, de la **psychologie** ou plus particulièrement de la **Langue des Signes**, on s'intéresse à une description détaillée des expressions en vue d'en fournir un sens. La classification en émotions universelles n'est généralement pas suffisante. Dans certains cas, les niveaux de détail d'analyse sont trés fins, puisque certains cherchent à détecter les expressions spontanées des expressions « forcées » (c'est la base d'un détecteur de mensonges basé sur les expressions).

On présente donc un ensemble de travaux dans le domaine de l'analyse des expressions du visage, classés par objectifs :
– Analyse des expressions pour la compression de données et l'animation réaliste,
– Analyse des expressions pour l'Interaction Homme-Machine,
– Analyse des expressions comme composantes langagières.

2.4.1 Compression et animation

En compression de données et en animation réaliste, on se limite à la détection et au suivi des composantes. L'étape d'interprétation de la dynamique des différentes composantes n'est géralement pas nécessaire.

Terzopoulos et Waters ([40]) ont cherché un moyen de transmettre suffisamment d'informations sur un visage et sa dynamique pour que la reconstruction soit suffisamment réaliste. Le but principal est d'obtenir un système de téléconférence où le visage et les expressions de l'intercoluteur sont transmis sur le réseau puis reconstruits. La principale contrainte est de limiter le nombre d'informations circulant sur le réseau tout en reconstituant un interlocuteur réaliste.
Les auteurs présentent tout d'abord le modèle utilisé pour la synthèse du visage et des expressions : c'est un modèle à trois dimensions contraint par un ensemble de muscles. Le tissu est modélisé par trois couches : le derme, l'épiderme et les muscles.
Ils présentent ensuite un système d'analyse automatique des expressions. Contrairement à l'étape de synthèse, l'analyse se fait sur un modèles à deux dimensions : plusieurs composantes sont suivies au cours de la vidéo par des *snakes* (sourcils, bouche, menton et rides nasolabiales). Le modèle à trois dimensions est ensuite animé par le mouvement des composantes précédemment analysé. Aucune interprétation des actions faciales n'est effectuée.

Ahlberg ([10]) se focalise aussi sur le problème de transmission du visage et de ses expressions (téléconférence par exemple). Il présente un système basé sur le concept d'*Active Appearance Model*. Le principe est d'adapter un modèle à une image. Ici, le modèle du visage utilisé est Candide. Différents paramètres de Candide (*Shape Units*) permettent d'adapter le modèle générique à un visage particulier.

Sur chaque image de la séquence, le modèle est adapté automatiquement (par une méthode d'optimisation) à la nouvelle configuration par des modifications des paramètres intra-individus (*Animation Units*).
L'avantage du modèle Candide est qu'il est déformé par un ensemble d'*Animation Units* qui sont une généralisation du concept d'*Action Unit* de FACS et de FAP de MPEG4. L'adaptation du modèle est guidé par les *Animation Units* et on obtient donc une description du mouvement en termes d'AUs ou de FAPs.

Oliver, Pentland, Bérard ([34]) présentent un système de suivi du visage et de la bouche en temps-réel. Le but est de suivre de manière robuste (rotation, translation, changement d'échelle) le visage et de reconnaître les différentes configurations de la bouche. Les applications principales sont l'animation d'avatars, le pilotage de caméra (pour le suivi du visage) et la compression vidéo. Dans ces applications, il n'est pas nécessaire d'effectuer une interprétation de plus haut niveau (actions faciales, émotions) des mouvements des composantes, puisque les mouvements sont les entités de base servant à la synthèse.

Goto, Kshirsagar et Magnenat-Thalmann ([27]) cherchent à animer un *clone* virtuel (une représentation virtuelle de l'utilisateur). Le système présenté est composé de deux modules : un module de construction du clone 3D à partir de deux vues (face et profil) et un module d'extraction des primitives de mouvements faciaux (FAPs de MPEG4).
Le premier module adapte un modèle à trois dimensions du visage aux deux vues du visage de l'utilisateur. Le modèle adapté et texturé forme le clone.
Le deuxième module suit un certain nombre de composantes (sourcils, yeux et bouche) et les codent sous forme de *Facial Animation Parameters* de MPEG4. Cette analyse permet alors l'animation du clone par un logiciel aux normes MPEG-4 pour l'animation faciale.

2.4.2 Interaction Homme-Machine

En Interaction Homme-Machine, on cherche généralement à interpréter les mesures de dynamiques des différentes composantes du visage en terme d'émotions. Il est rare qu'on ait besoin d'un niveau plus fin d'interprétation.

Black et Yacoob ([13]) présentent une approche pour le suivi et la reconnaissance des six expressions universelles. L'accent est mis sur la différence entre les mouvements du visage « rigides » (c'est à dire mouvements de la tête et rotation dans le plan) et les mouvements « non-rigides » (le mouvement des sourcils ou de la bouche n'est pas rigide). A partir d'une estimation de la dynamique de quelques composantes (yeux, sourcils et bouche uniquement), la décision (classement dans une des six classes d'émotions) est prise à partir d'un ensemble de règles.

Pantic et Rothkrantz ([37]) présentent un système de reconnaissance des expressions du visage pour une application en Interaction Homme-Machine. Les expres-

sions du visage sont considérés comme une nouvelle modalité d'interaction. Le but est, après analyse des différentes modalités, de déduire l'intention de l'utilisateur, dans un système multimodal. Le module d'analyse des expressions tente de les classer parmi les six expressions universelles.
L'analyse est effectuée par une approche par composantes. Les actions faciales sont traduites en *Action Units* de FACS. Les émotions sont détectées à partir des *Action Units* .
Le système se base sur deux vues du visage : de face et de profil. Un ensemble de points caractéristiques et de mesures entre ces différents points (angle, distance) permet, par un système de décision à base de règles, de déduire l'*Action Unit* correspondante. A partir des *Action Units* , le système déduit alors l'émotion correspondante.

Lyons, Akamatsu, Kamachi et Gyoba ([31]) tentent de classer les expressions du visage parmi les six classes universelles. Ils proposent d'utiliser des filtres de Gabor pour extraire l'information discriminante des expressions universelles. C'est la première étude de ce type. La motivation principale est que le filtrage par ondelettes de Gabor serait un processus présent dans le système de vision humain. Les auteurs démontrent qu'il est possible de construire un système de reconnaissance des expressions uniquement basé sur des images filtrées par des ondelettes de Gabor.

Zhang ([43]) compare alors l'approche par filtres de Gabor aux approches classiques basées sur des mesures de composantes. Le but est de mesurer la qualité de l'information véhiculée par ces deux approches pour la reconnaissance des expressions universelles. Les deux informations sont données en entrée à un perceptron.
Les résultats de la comparaison indiquent qu'une image filtrée par ondelettes de Gabor est beaucoup plus porteuse d'informations, pour la reconnaissance des expressions, que des mesures géométriques. Ces résultats ont aussi été obtenus par Cottrell ([44], voir plus loin), qui explique que les filtres de Gabor ne retiennent pas l'information d'*identité individuelle* (une information codant l'identité d'un individu est considéré comme du bruit pour un processus de reconnaissance des expressions).

2.4.3 Sciences comportementales et neuropsychologie

Dans le domaine des sciences comportementale, l'interprétation de la dynamique des différentes composantes du visage est essentielle et doit être la plus exhaustive et la plus précise possible.

Cohn, Zlochower, Lien, Kanade ([17]) ont développé un système d'analyse automatique du visage (*Automated Face Analysis* - AFA). Le but était de fournir un outil informatique permettant d'aider les psychologues. En effet, le codage manuel des différentes *Action Units* de FACS est estimé à environ 10 heures par minute de vidéo.

Les auteurs présentent donc un système d'analyse automatique basé sur le suivi de différentes composantes. La décision des différentes *Action Units* est prise par une analyse en fonctions discriminantes.
Le but de l'article est de montrer la validité d'un système automatique par rapport à un système manuel. Une étude comparative a permit de déduire que l'analyse par suivi de composantes obtient un score de correlation moyen de 91%, 88% et 81% (région des sourcils, des yeux et de la bouche respectivement) avec le codage manuel.
L'objectif est atteint puisque l'analyse passe d'une trentaine de secondes par image à 1 seconde par image. L'analyse n'est cependant pas entièrement automatisée, puisque les composantes sont détectées de manière manuelle. Néanmoins, le travail effectué manuellement est beaucoup plus réduit.

Tian, Kanade et Cohn ([30]) étendent le précédent système en basant la reconnaissance sur les composantes du visage « permanentes » (yeux, bouche, sourcils, etc.), mais aussi sur des composantes « temporaires » comme les rides d'expressions. 16 *Action Units* de FACS sont reconnues.
Chaque composante possède un détecteur particulier et les différentes mesures sur ces composantes (changements d'état, angles, distances) permettent de décider quelle *Action Unit* ou combinaison d'*Action Units* sont présentes. La décision est effectuée par un réseau de neurones.
Contrairement à la précédente version du système, l'analyse est complètement automatisée puisque le visage est détecté automatiquement (détecteur de visages de Kanade, basé sur des réseaux de neurones), les composantes sont suivies automatiquement et la décision est prise automatiquement.

Le système AFA est utilisé par quelques psychologues, en particulier Schmidt et Cohn ([39] [16]) qui s'en servent comme outil de validation. Un ensemble d'expériences est menée et plusieurs mesures sont effectuées : manuelle, informatique et par electromyographie. AFA permet ainsi de vérifier certaines observations.

Le but de Bartlett, Hager, Ekman et Sejnowski ([11]) est aussi de fournir un outil informatique d'aide aux psychologues et linguistes pour l'analyse des expressions du visage en codifiant automatiquement les expressions du visage en *Action Units* de FACS.
Les auteurs présentent une étude comparative de trois approches : une approche « holistique » (basée sur l'image), une approche par mesure de composantes et une approche par flux optique. La meilleure méthode semblerait être la méthode holistique, bien que la reconnaissance soit limitée à 6 *Action Units* et qu'il soit donc difficile de généraliser.
L'analyse n'est pas complètement automatique, puisque le visage est détecté par deux *clicks* de souris.

Bartlett, Littlewort, Braathen, Sejnowski et Movellan ([12]) s'intéressent à l'analyse des expressions du visage dans un contexte plus générale. Le but est d'analyser les expressions « spontanées » plutôt que les expressions obtenues en laboratoires

qui sont la plupart du temps « forcées ». Le contexte n'étant pas contrôlable avec cette hypothèse, le système doit être capable de traiter avec des changements de pose du visage, en particulier des rotations hors du plan du visage. Les paramètres de rotation sont calculés par triangluation à partir de la position de 8 points de contrôles (placés manuellement).
Un banc de filtres de Gabor est appliqué sur chaque image. Ces données sont fournies en entrée à une *Support Vector Machine*. La décision sur l'ensemble de la vidéo (ajout de l'aspect temporel) est obtenue par un système à base de *HMM* (dont les entrées sont les sorties des différentes SVM de chaque image). La principale motivation est que les expressions spontanées et forcées se distinguent par la zone du cerveau qui les active. Une des hypothèses mise en avant par certains psychologues est alors qu'il est possible de distinguer *visuellement* les expressions spontanées des expressions forcées. En particulier, il semblerait que les expressions spontanées aient un profil temporel lègèrement différent. Malheureusement, les humains ont des difficultés à percevoir ces changements trop subtiles. L'idée est alors de savoir si un système automatique pourrait distinguer ces changements subtiles.

Cottrell, Dailey, Padgett et Adolphs ([44]) s'intéressent aux fonctions d'analyse du visage du cerveau humain. Le but est de savoir si toute l'analyse du visage effectuée par le cerveau humain est *holistique*. Un processus est holistique si le changement de configuration des parties change l'interprétation du tout. C'est la différence qu'on trouve entre le processus de reconnaissance d'un objet quelconque et la reconnaissance d'un visage. Les changements de configuration des parties d'un objet quelconque n'auront que peu d'influence sur la reconnaissance globale de l'objet. Par contre, reconnaître un visage semble être un processus holistique puisque le changement de configuration de ses parties (des composantes du visage par exemple) change l'identité que l'on peut lui associer.
Pour savoir si *tous* les processus d'analyse du visage du cerveau humain sont holistiques, les auteurs proposent d'utiliser différents processus d'analyse holistiques ou non, simulés par ordinateur, et d'en fournir les données en entrée à un réseau de neurones. Ils comparent ensuite les résultats pour la reconnaissance du visage et la reconnaissance des expressions.
Il semblerait alors que la reconnaissance du visage soit un processus holistique alors que la reconnaissance des expressions ne l'est pas.

2.5 Conclusion

Aucun système actuel ne permet de reconnaitre toutes les actions faciales définies dans les différents formalismes : aucun système n'est capable, en particulier, de reconnaître les 46 *Actions Unitaires* d'Ekman. Ceci reste un challenge intéressant et présente un intérêt pour le développement d'un outil d'analyse automatique dans le domaine des sciences comportementales, de la psychologie et de la linguistique.

Cependant, la question de savoir si l'ensemble des actions faciales est nécessaire à une compréhension aisée de la langue des signes reste une question ouverte.

Aucun système actuel ne permet de quantifier les actions faciales reconnues. Le problème n'est cependant pas trivial puisque l'intensité des expressions dépend des individus. Il peut être pourtant intéressant d'avoir une idée de l'intensité d'une expression relativement à l'individu. Les linguistes ne distinguent généralement que peu de degrés d'expressions (ouvert, demi-ouvert ou fermé par exemple) et on pourra se contenter d'une mesure floue.

En langue des signes, les expressions affichées sur le visage d'un locuteur sont bien souvent partiellement cachées par les gestes de la main, aussi bien parce que le signe nécessite que la main touche le visage ou parce que la main se trouve devant dans l'axe de la caméra. Bien que ces occultations partielles ne gènent pas la compréhension du signe (sinon le geste se serait adapté pour une meilleure compréhension) entre deux locuteurs, il est possible qu'elles gênent les opérateurs de traitements d'images qui ne sont généralement pas conçus pour traiter les occultations.

De plus, la position (légère rotation, mouvement) de la tête prend une part importante dans la définition du sens.

Le problème d'invariance aux mouvements « rigides » du visage est un problème qui est traité par certains chercheurs. Par contre, le problème d'invariance à l'occultation (partielle) du visage pour l'analyse des expressions est un problème qui n'a été que très peu traité. Lanitis, Taylor et Cootes ([29]) sont parmi les seuls à proposer une méthode générale prenant en compte le problème d'occultation.

Le problème d'invariance à la pose (mouvements rigides) du visage peut être traité, par exemple, en appliquant un prétraitement dont le but est de reconstruire la vue de face du visage. Une méthode efficace consiste à se servir de l'image du visage comme une texture d'un modèle 3D. L'image du visage vu de face est retrouvé par rotation du modèle.

Un problème important des expressions présentes en Langue des Signes est que certaines font intervenir des composantes dont la délimitation est floue (le gonflement des joues constitue un bon exemple). Il est possible que les opérateurs classiques de traitement d'images ne suffisent pas à détecter de manière robuste les changements d'état de ces zones, il est donc nécessaire de les « déduire ».

Une méthode efficace consiste à intégrer un certain nombre de contraintes (musculaires par exemple) dans le processus de reconnaissance. Les contraintes sont intégrées soit pendant la phase d'extraction des composantes (c'est alors généralement une approche globale où le modèle possède des contraintes musculaires [24]), soit pendant la phase de décision sous la forme d'un ensemble de règles.

Le regard a un rôle très important dans la Langue des Signes. Le clignement des yeux permet, généralement, de découper les phrases. L'orientation du regard est

très présente dans les transferts personnels. Il joue aussi un rôle important quand le locuteur regarde dans la direction de l'interlocuteur : cela signifie qu'il va avoir recours à un signe standard de la Langue des Signes.

Le système d'analyse des expressions doit donc décrire précisement les paramètres du regard (en fournissant par exemple un module de reconnaissance spécialisé).

On peut donc en conclure qu'un système automatique d'analyse des expressions du visage de la Langue des Signes doit :

1. pouvoir traiter un visage en **rotation** (dans le plan de l'image et hors plan),
2. pouvoir traiter un visage en partie **caché**,
3. pouvoir **déduire** certaines caractéristiques.

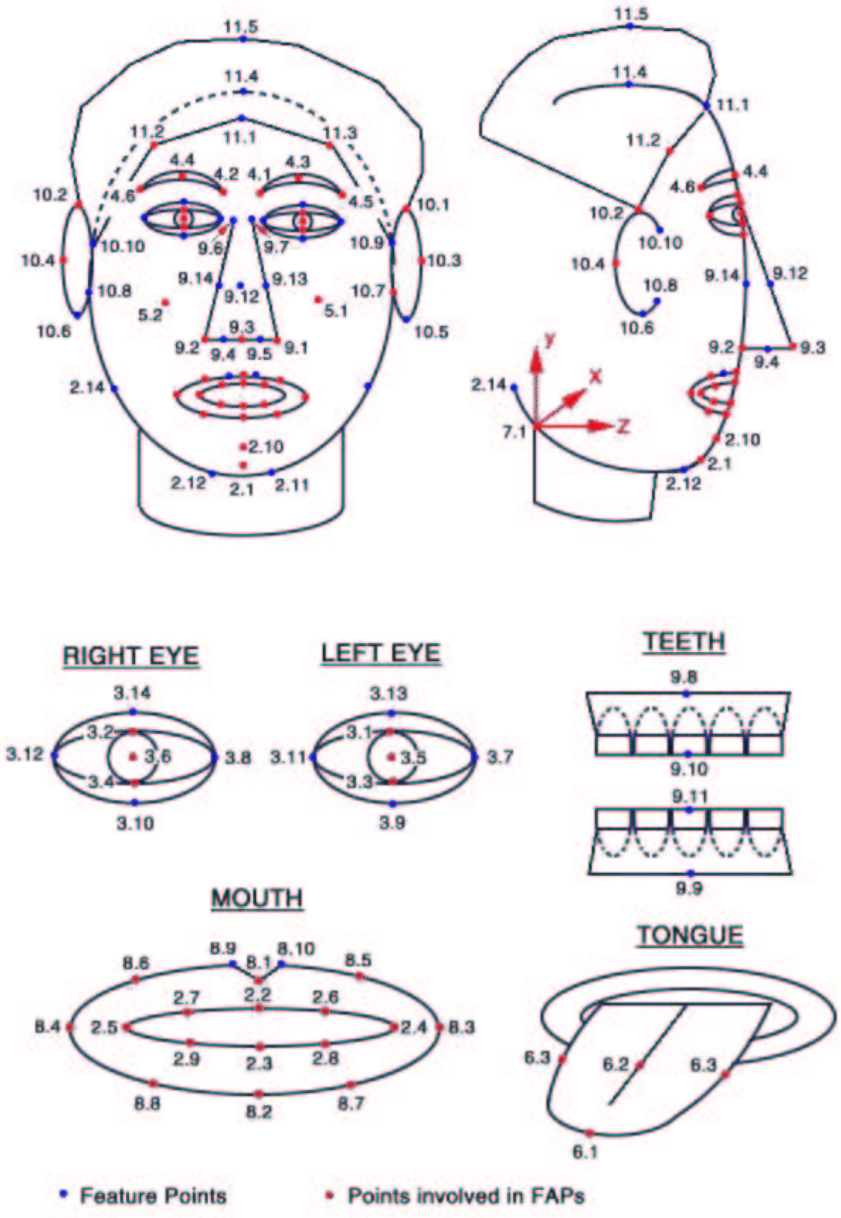


FIG. 2.5 – Modèle du visage MPEG-4 - Facial Definition Points

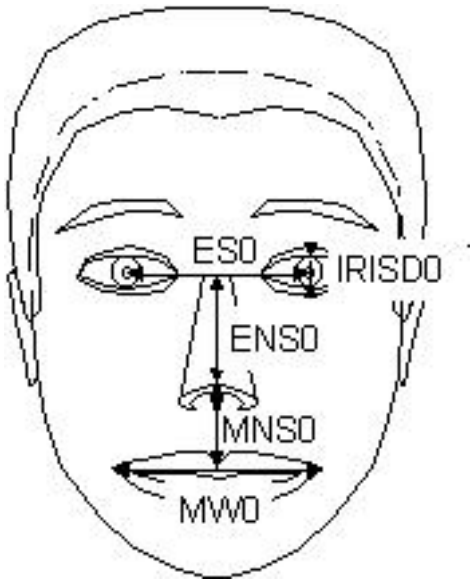


FIG. 2.6 – Modèle du visage MPEG-4 - Facial Animation Parameter Units

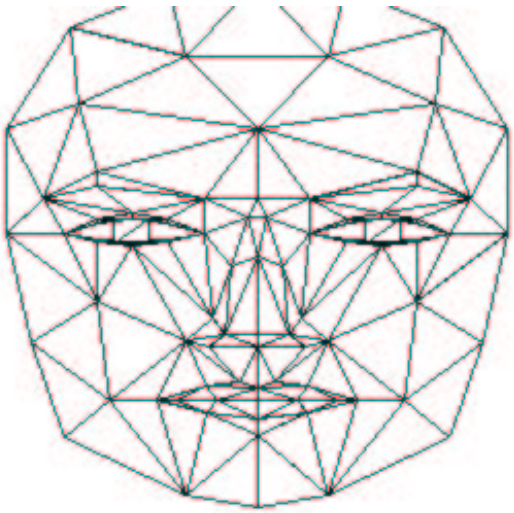


FIG. 2.7 – Version 3 du modèle Candide

3

Cahier des charges

On cherche à concevoir un système d'analyse des expressions, dans un contexte de communication humaine, et en particulier dans le contexte d'une communication en Langue des Signes.

Le système d'analyse des expressions a pour but de *décrire* pour *reconnaître* les expressions du visage d'un locuteur en Langue des Signes à partir d'une séquence vidéo.

L'analyse des expressions est un sous-système d'un système plus global *d'aide à la compréhension* de la Langue des Signes.

3.1 Particularités de la Langue des Signes

Le contexte particulier de la Langue des Signes amène de nombreuses contraintes quant à l'analyse des expressions du visage.

Les mouvements du corps et en particulier du visage (rotation, translation) sont très fréquents en Langue des Signes, puisqu'ils prennent part au sens du discours. Il est donc difficile d'étudier une locution en Langue des Signes sans prendre en compte les mouvements du visage. En particulier, ils sont très présentes dans les situtations de Transfert Personnel ou de Transfert Situationnel ; situations dans lesquelles on trouve beaucoup d'expressions du visage (voir plus loin).

Les mouvements des mains, constituant les éléments ayant le plus de sens, sont eux aussi très présents. Et de nombreux gestes des mains ont des intéractions avec le visage du locuteur. Ainsi, comme il a été dit, il arrive très souvent qu'une partie du visage soit en partie cachée par une main ; soit parce que la main touche réellement le visage du locuteur ou soit parce que la main se trouve dans l'axe de la caméra. Cette contrainte est donc à intégrer dans le processus d'analyse.

Les seules descriptions existantes des expressions du visage présentes en Langue des Signes ont été faites par des linguistes. Ces descriptions sont faites en langage « naturel » et sont donc difficile à exprimer de manière informatique.

Un système d'aide à la compréhension des expressions du visage se doit donc :

1. d'être **robuste aux changements de pose** du visage du locuteur,
2. d'être **robuste aux occultations partielles** du visage,
3. d'intégrer les **connaissances des linguistes**.

Les méthodes pour répondre aux deux premiers points sont décrites dans ce chapitre.

Le dernier point constitue la partie centrale de cette étude et est présenté dans le chapitre suivant.

3.2 Reconstruction 3D

Les mouvements du visage étant souvent présents dans les formes de communication faisant intervenir le visage, et en particulier dans la Langue des Signes, un processus d'analyse se doit de pouvoir détecter ces mouvements. D'abord pour pouvoir les quantifier, puisqu'ils sont chargés de sens, puis pour pouvoir mesurer les mouvements des composantes du visage, alors qu'il est en mouvement (en rotation par exemple).

Il existe deux approches principales pour traiter le problème des changements de pose du visage : la première adapte les opérateurs de détection en tenant compte des translations et rotations ; la deuxième considère l'étape de reconstruction 3D comme étape de prétraitement. Dans la deuxième approche, les opérateurs de détection sont spécialisés dans la détection à partir d'une image du visage vu de face. Le prétraitement consiste alors à estimer les paramètres de translation et de rotation du visage et à ramener le visage vers une vue de face. Ce prétraitement fait généralement intervenir un modèle 3D (même simplifié) du visage.

La première approche nécessite d'avoir des détecteurs très « souples » c'est à dire capables de s'adapter à de nombreuses situations. Malheureusement, il est très difficile de concevoir des opérateurs qui s'adaptent, par exemple, aux rotations du visage hors du plan de l'image.

La deuxième approche permet d'optimiser les détecteurs pour ne traiter que des images de visages vus de face. Le résultat du prétraitement, même si celui-ci utilise un modèle 3D, est généralement de fournir une image 2D du visage, vu de face, reconstituée. Les opérateurs de détection opéreront alors sur une image 2D classique.

Cependant, certaines actions faciales se distinguent difficilement sur une vue de face. C'est le cas par exemple de la projection des lèvres vers l'avant. Ce mouvement est par contre plus facilement visible à partir d'une vue de profil.

Il est alors peut-être nécessaire de considérer la vue de profil (voire d'autres vues, comme la vue de trois-quarts). Cela consiste alors à d'abord détecter quelle est

la « configuration » du visage la plus visible : le visage est-il plutôt de profil ou plutôt de face ?. L'analyse est alors adaptée aux différentes vues, et les opérateurs de détection changent.

3.2.1 Estimation des paramètres 3D

Le visage est une forme en trois dimensions. La partie observée (sur une séquence vidéo ou une image) a perdu une dimension. Le problème est alors de reconstituer cette troisième dimension.

Dans le cas général, la projection d'un objet à trois dimensions sur un plan s'effectue par le calcul suivant :

$$\begin{pmatrix} u \\ v \end{pmatrix} = A \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

où u et v représentent les coordonnées de la projection dans le repère image ; x, y, z représentent les coordonnées du point à projeter dans le repère scène et A représente la matrice de projection.

La matrice A est la multiplication de différentes matrices représentant les paramètres de translation, rotation et échelle du point à projeter. Par exemple, $A = RTE$ où R est la matrice de rotation, T la matrice de translation et E la matrice de changement d'échelle.

On considère dans le cas général les paramètres intrinsèques de la caméra comme la distance focale qui permet de modéliser la notion de perspective. Dans le cas des visages, on simplifie le problème en ignorant la perspective.

La rotation s'exprime en fonction de trois angles σ_x, σ_y et σ_z. On considère les translations parallèles aux trois axes t_x, t_y et t_z et le changement d'échelle selon les trois axes s_x, s_y et s_z.

Il est possible de simplifier le problème en notant que la translation selon l'axe des Z peut être vue comme un changement d'échelle global (sur tous les axes), puisqu'on ignore ici la déformation due à la perspective.

Le problème se réduit alors à estimer 6 paramètres $(\sigma_x, \sigma_y, \sigma_z, t_x, t_y, s)$.

Il est donc suffisant, théoriquement, de connaître 3 correspondances : connaître les coordonnées 3D d'origines et leurs projections respectives. Puisque l'équation matricielle de projection donne deux équations pour chaque correspondance de point. Etant donné qu'il y a 6 inconnues, il suffit de 6 équations (c'est à dire 3 points).

Cependant, en pratique, on ne connait que rarement la profondeur d'un ensemble

de points (même si cet ensemble se réduit à trois points).

Le principe est alors de *suivre* un ensemble de points de références : des points qui restent fixes par rapport au visage et ne sont pas déplacés par les expressions. Cet ensemble de points permet de contraindre davantage le système d'équations. En supposant qu'on connaisse la configuration de cet ensemble de points sur une image de référence (une image vue de face par exemple), les paramètres pourront être estimés sur les images suivantes.

Supposons que l'on détecte (de manière automatique ou assisté) trois points de référence sur un visage vu de face (typiquement les coins des deux yeux et un point sur le nez). Sur l'image suivante, à partir de la nouvelle position de ces trois points, il sera difficile de distinguer entre, par exemple, une rotation et un changement d'échelle suivi d'une translation.

On découpe alors généralement le problème en deux : estimation des paramètres de translation (X et Y) dans un premier temps, puis estimation des paramètres de rotation et de changement d'échelle. Le module de détection du visage permet, par exemple, d'avoir une estimation des paramètres de translations en X et Y.

3.2.2 Adaptation globale d'un modèle de visage

Il n'est pas toujours aisé de suivre un point d'une image à l'autre. Le suivi consiste généralement à mémoriser dans une image de référence la configuration du voisinage d'un point et de tenter de retrouver cette configuration dans une autre image.

Cependant, ces méthodes de suivi considèrent un certain nombre de contraintes : luminosité constante et aucune déformation, par exemple. Or, lors d'une rotation, la luminosité est rarement constante et les formes souvent déformées (surtout lors d'une rotation hors-plan).

Une solution est alors de mémoriser plus d'informations sur les points à suivre. Lanitis, Taylors et Cootes ([29]), présentent, par exemple, la notion de modèles d'apparences actifs, repris par la suite par Ahlberg ([10]).

Le principe est de se placer à un niveau global. On mémorise les informations d'un ensemble de points et non d'un point isolé : configuration spatiale et texture de la forme extraite (information d'apparence).

La méthode consiste, sur chaque image, à déformer ce modèle de visage pour qu'il corresponde le plus possible au visage observé. La mesure de correspondance est donnée par une décomposition en visages - propres de la forme extraite. La distance dans l'espace des visages donne la mesure de ressemblance. Le modèle se déforme alors itérativement pour augmenter la correspondance (*i.e.* minimiser la distance dans l'espace des visages). Cependant, pour garder une certaine cohérence, le modèle n'est pas déformé arbitrairement : il y a un ensemble réduit de déformations possibles.

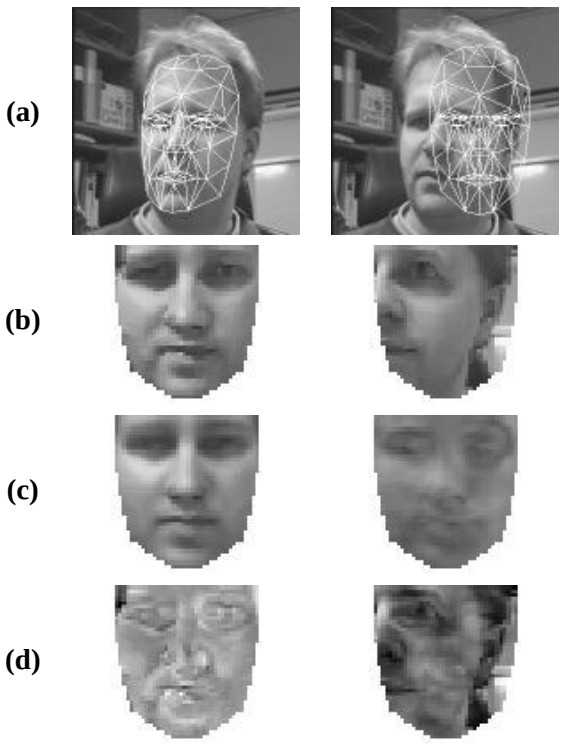


FIG. 3.1 – Méthode d'adaptation global du modèle Candide - **(a)** adaptation du modèle 3D. **(b)** texture extraite. **(c)** projection dans l'espace des visages-propres. **(d)** mesure de distance.

Cette méthode générale est applicable pour la reconstruction 3D. L'ensemble des déformations possibles du visage est alors l'ensemble des rotations / translations et changements d'échelle.

Plus généralement cette méthode peut servir aussi à la reconnaissance d'actions faciales si l'ensemble des déformations possibles contient l'ensemble des déformations dues aux actions faciales.

3.3 Occultation

Il est possible de concevoir un processus d'analyse du visage qui reste robuste, même si le visage est en partie caché, c'est à dire dont les détecteurs continuent à fournir des réponses satisfaisantes sur les parties du visage visibles, alors que certaines sont cachées.

Il est nécessaire que le système sache quelles parties du visage sont cachées. Pour ce faire, la méthode la plus sûre consiste à suivre la position des objets cachant le

visage. Dans le cas d'une communication en Langue des Signes, ces objets sont les mains. On ne traitera alors pas les parties cachées par d'autres objets (lunettes noires par exemple).

Un module de suivi de l'enveloppe externe des mains peut permettre de déduire quelles zones du visage sont cachées. Les mesures des détecteurs travaillant sur ces zones seront alors invalidées.

4

Formalisme et architecture

On présente dans ce chapitre les spécifications d'un système d'analyse des expressions du visage dans un contexte de communication en Langue des Signes, dans un but d'aide à la compréhension.

La conception du système est guidée par les connaissances des linguistes. Ces connaissances sont, dans un premier temps, traduites de manière informatique. La manière de représenter ces connaissances guide la manière de les utiliser.

4.1 Formalisme de représentation

Les connaissances sur les expressions du visage présentes en Langue des Signes sont décrites par les linguistes. Elles définissent la manière dont les expressions du visage prennent part à la construction d'un langage articulé tel que la Langue des Signes. Les expressions du visage interviennent à plusieurs niveaux langagiers : lexical, syntaxique et sémantique.

Puisque l'analyse informatique des expressions du visage se trouve être ici menée dans un contexte de communication humaine et en particulier dans le contexte d'une communication en Langue des Signes, il est nécessaire qu'elle prenne compte de l'aspect langagier, ce qui permettra de décoder des expressions qui sont, *a priori* , complexes ou coûteuses à analyser.

Une des approches préliminaires consiste alors à intégrer à un système informatique d'analyse, les différentes connaissances linguistiques établies par les spécialistes de la Langue des Signes (en particulier Christian Cuxac dans [19]).

Les descriptions des linguistes sont faites de manière informelle et relativement subjective. Par exemple, définir l'expression « fière » ou encore « sûr de soi » n'a de sens que si la description se réfère à un ensemble d'indices visuels objectifs, observables par chacun, expert ou non, comme l'activation de muscles faciaux ; ce qui n'est pas le cas pour un grand nombre des descriptions linguistiques.

Afin d'exprimer de manière informatique les différentes connaissances linguistiques, il est donc nécessaire de trouver un formalisme de représentation adéquat.

On propose ici un formalisme directement construit à partir des connaissances linguistes. La manière informatique de décrire les connaissances est proche de celle utilisée par les linguistes. Le formalisme est construit par simplification et généralisation des concepts utilisés par les linguistes.

Traduire des connaissances informelles et subjectives en des connaissances formelles et objectives ne se faisant pas sans perte, il est à noter que le formalisme informatique présenté ici n'est qu'une première approche et qu'il nécessite donc d'être validé et complété par les experts, *i.e.* les linguistes.

Après analyse exhaustive des différentes descriptions d'expressions présentes an Langue des Signes, on distingue un certain nombre d'éléments indépendants, entités de base du formalisme informatique que l'on présente ici.

4.1.1 Connaissances

Chaque expression de la Langue des Signes est décrite soit directement par un ensemble d'états musculaires, soit par composition (spatiale et/ou temporelle) d'autres expressions. On englobera ces entités (expression, émotion, état musculaire, etc.) sous le terme de **connaissance**.

On appelle **connaissances élémentaires** les connaissances directement observables (par exemple l'état des composantes : « les yeux sont ouverts, la bouche fermée, ... »). Elles seront généralement directement extraites par des opérateurs de Traitement d'Images.

On appelle **connaissances composées** les connaissances qui sont définies à partir de la composition d'autres connaissances (élémentaires ou non).

A chaque connaissance, sont associés un **nom** et un **état**. Par exemple, à la connaissance nommée `yeux-ouverts` est associé un **état** représentant ici un **degré** d'ouverture des yeux.

De manière générale, l'**état** d'une connaissance (élémentaire ou composée) représente toutes les informations nécessaires à la compréhension d'une connaissance, et il peut être formé de plusieurs **propriétés**.

A chaque connaissance est donc associé un ensemble de **propriétés** la caractérisant. Chaque propriété peut prendre un certain nombre de valeurs, défini par un **ensemble de définition** ou **domaine**.

Les connaissances ayant le même ensemble de propriétés corresponde généralement à un niveau d'analyse distinct. Par exemple, toutes les connaissances élémentaires partagent les mêmes propriétés (l'amplitude), toutes les actions faciales partagent les mêmes propriétés.

Connaissances élémentaires

La plupart des connaissances élémentaires correspondent directement avec l'état d'un muscle facial. L'**état** d'une connaissance élémentaire ne représente donc que l'amplitude d'activation de ce muscle.

L'ensemble minimal des connaissances élémentaires nécessaires à la description des expressions en Langue des Signes est donné en figure 4.1.

Certains muscles travaillent par deux. On les appelle ici les muscles antagonistes. Par exemple, il existe deux muscles pour le mouvement des sourcils : un muscle d'abaissement et un muscle de relèvement. De manière anatomique, ce sont bien deux muscles différents qui interviennent dans les mouvements des sourcils, mais un muscle ne peut être activé que quand son antagoniste ne l'est pas. C'est pourquoi, on considère ici une seule connaissance, même pour un couple de muscles antagonistes.

D'autre part, certains muscles n'ont qu'une activation isolée. C'est le cas, par exemple, du muscle d'ouverture de la bouche. Au repos, la bouche est fermée et l'activation du muscle ouvre la bouche.

On ne différenciera ici les états des muscles isolés des états des muscles antagonistes que par la position de la valeur nulle. La valeur nulle représente la valeur de repos du muscle ou du couple de muscles. Par exemple, pour le couple de muscles abaissement / relèvement des sourcils, la valeur nulle se trouve quand aucun des deux n'est activé. Pour le muscle d'ouverture de la bouche, la valeur nulle se trouve quand la bouche est fermée.

Ainsi, l'état d'un muscle isolé est représenté par une valeur positive ou nulle et l'état d'un couple de muscles est représenté par une valeur positive, négative ou nulle.

C'est ces valeurs positives ou négatives qui apparaissent sur la figure 4.1. Pour plus d'homogénéïté avec les descriptions linguistiques, des noms différents sont donnés pour les deux états d'un couple de muscles antagonistes.

En réalité, le système d'analyse ne donne qu'un seul nom pour chaque muscle ou couple de muscles. Par exemple, on parlera ici de `sourcil-gauche-froncé` et `sourcil-gauche-relevé` alors que ces deux connaissances n'en forment qu'une pouvant avoir une valeur positive, nulle ou négative : `froncement-sourcil-gauche`.

Valeur positive	**Valeur négative**	**Muscle**
sourcil-gauche-relevé	sourcil-gauche-froncé	relèvement-sourcil-gauche
sourcil-droit-relevé	sourcil-droit-froncé	relèvement-sourcil-droit
sourcil-intérieur- gauche-relevé	sourcil-intérieur-gauche-froncé	relèvement-sourcil-gauche-intérieur
sourcil-intérieur- droit-relevé	sourcil-intérieur-droit-froncé	relèvement-sourcil-droit-intérieur
yeux-plissés		plissement-yeux
yeux-fermés		fermeture-yeux
regard-à-gauche	regard-à-droite	regard-X
regard-en-haut	regard-en-bas	regard-Y
nez-plissé		plissement-nez
nez-pincé		pincement-nez
bouche-ouverte		ouverture-bouche
lèvre-supérieure-projetée		projection-lèvre-sup
lèvre-inférieure-projetée		projection-lèvre-inf
lèvre-gauche-étirée	lèvre-gauche-serrée	etirement-lèvre-gauche
lèvre-droite-étirée	lèvre-droite-serrée	étirement-lèvre-droite
coin-bouche-gauche-relevé	coin-bouche-gauche-abaissé	relèvement-coin-bouche-gauche
coin-bouche-droit-relevé	coin-bouche-droit-abaissé	relèvement-coin-bouche-droit
langue-visible		visibilité langue
dents-visible		visibilité dents
joue-gauche-gonflée	joue-gauche-creusée	gonflement-joue-gauche
joue-droite-gonflée	joue-droite-creusée	gonflement-joue-droite
langue-saillante-joue-gauche	langue-saillante-joue-droite	saillance-langue
mâchoires-contractées		contraction-mâchoires
visage-avancé	visage-reculé	translation-Z
visage-relevé	visage-abaissé	rotation-X
visage-tourné-gauche	visage-tourné-droite	rotation-Y
visage-incliné-gauche	visage-incliné-droite	rotation-Z

FIG. 4.1 – Connaissances élémentaires

4.1.2 Connaissances composées

A partir des connaissances élémentaires qui viennent d'être définies, il est possible de définir toutes les autres connaissances, par composition. Il existe plusieurs types de compositions :
– les **redéfinitions** qui correspondent au renommage d'un ensemble de connaissances,
– les **compositions langagières** qui permettent de définir l'ensemble des expressions présentes en Langue des Signes.
– les **actions faciales** qui expriment l'évolution dans le temps d'un état musculaire, selon un certain profil (simple, hochement, tremblement),

Quelque soit le type de composition, elle peut être définie par un **ensemble de valeurs** combinées entres elles par un certain nombre d'**opérateurs de composition**. L'ensemble des valeurs de chaque connaissance intervenant dans la définition de la composition est désigné par un ensemble de **sélecteurs**.

Opérateurs de composition

Les compositions de type **redéfinitions** et **compositions langagières** partagent les mêmes opérateurs de compositions. Ces opérateurs sont les suivants :

– un opérateur de **conjonction** (représenté ici par le symbole ','') qui indique que deux faits ont lieu **en même temps**,
– un opérateur de **disjonction** (représenté ici par le symbole '||') qui indique l'alternative,
– un opérateur de **négation** (représenté ici apr le symbole '-') qui indique qu'un fait n'a pas lieu,
– un opérateur de **séquencement** (représenté ici par le symbole '+') qui indique qu'un fait a lieu **après** l'autre,
– un opérateur de **répétition** (représenté ici par le symbole '*') qui indique qu'un fait a lieu **plusieurs fois successivement**,
– un opérateur d'**éventualité** (représenté ici par le symbole ' ?') qui indique qu'un fait a **éventuellement** lieu,
– les parenthèses qui permettent d'exprimer les priorités entres opérateurs.

Les **actions faciales** quant à elles ont des opérateurs spécialisés. Ce sont des opérateurs unaires qui permettent d'exprimer le type d'évolution temporelle d'un état musculaire.

Une **action faciale** représente l'évolution d'un muscle de l'état nul vers l'état nul (ou proche de l'état nul), en passant par un certain nombre d'états non nuls.

Dans l'ensemble des descriptions, on distingue trois types d'actions faciales :
– l'action faciale **simple** (notée ici 'af-simple()') qui consiste à passer de l'état nul à l'état nul par un seul état non-nul,

– l'action faciale de **tremblement** (notée ici 'af-tremblement()') qui consiste à passer de l'état nul à l'état nul par un certain nombre d'activations de de relâchements musculaires,
– l'action faciale de **hochement** (notée ici 'af-hochement()') qui consiste à passer de l'état nul à l'état nul en passant dans les négatifs, puis dans les positifs.

Le profil temporel de ces trois types d'actions faciales est représenté en figure 4.2

FIG. 4.2 – Profil temporel des actions faciales. (a) simple (b) tremblement (c) hochement

Chaque action faciale est définie par rapport à une connaissance élémentaire. Par exemple, la connaissance `tremblement- joues` est définie par rapport à `gonflement-joues`. Pour distinguer les deux sens d'évolutions d'un couple de muscles antagonistes, est aussi associé à une action faciale, un **signe**. Une action faciale au signe positif a un profil temporel tel que présenté en figure 4.2. Une action faciale au signe négatif a un profil temporel symétrique. Ainsi, une action faciale de **hochement** négative commence par une évolution dans le domaine des négatifs, puis dans le domaine des positifs.

A chaque connaissance élémentaire est associée son action faciale **simple** dont le nom est préfixé par **af-**.

Sélecteurs

Il est possible de préciser la valeur des connaissances qui font partie de la définiton d'une composition à l'aide de **sélecteurs**.

Un **sélecteur** permet de sélectionner un sous-ensemble des valeurs possibles d'une connaissance et peut être vue comme la sélection d'un sous-ensemble de valeurs dans une base de données. Sans précisions, une connaissance utilisée comme composante indique que n'importe quelle valeur de la composante est valable. C'est ce mécanisme qui permet de traduire les qualificatifs utilisés dans les descriptions linguistiques : « yeux **légèrement** ouverts » par exemple.

Par exemple, `joues-gonflées` indique que la règle s'applique si les joues sont gonflées (quelque soit « l'amplitude » de gonflement). Par contre, `joues-gonflées (amplitude = légèrement)` indique que la règle ne s'applique que quand les joues sont *légèrement* gonflées. Ou encore `regard-en-haut (durée = longue)` indique que la règle ne s'applique que si le visage observé regarde en haut pendant une longue période.

Un sélecteur est donc une expression logique permettant de ne garder que les valeurs qui vérifient cette expression. L'expression logique se refère aux **propriétés** de chaque connaissance. Par exemple dans le sélecteur `(amplitude = légèrement)` on se refère à un ensemble de valeurs de la propriété **amplitude** de la connaissance.

Il est à noter que les valeurs des propriétés sont généralement données ici par des qualificatifs flous (*légèrement, long, grand, ...*) afin de garder une cohérence avec les descriptions linguistiques d'origine. Ces qualificatifs flous seront discrétisés au moment de l'analyse : `durée = longue` pourra par exemple se traduire par `durée >= 3`.

L'ensemble des propriétés des différents types de connaissances est donné en figure 4.3. 0+ indique un domaine à valeurs positives uniquement. -0+ indique un domaine à valeurs positives ou négatives. Un domaine peut aussi être constitué d'une union de plusieurs valeurs symboliques (*simple, tremblement, hochement*).

On remarquera alors que la propriété de *durée* n'existe pas sur les connaissanes *statiques*. Cependant, il est quand même possible de s'y référer. Sélectionner une durée sur une connaissance statique est équivalent à composer plusieurs fois cette connaissance avec l'opérateur de *séquencement*. Par exemple, `A ( durée >= 3 )` est équivalent à `A + A + A + A?`.

Type de connaissance	**Propriété**	**Domaine**
statique	position (numéro de l'image)	0+
	amplitude	0+
action faciale	position	0+
	type	simple | tremblement | hochement
	amplitude (pic)	-0+
	durée	0+
expression	position	0+
	amplitude (max)	0+
	durée	0+

FIG. 4.3 – Propriétés des différents types de connaissances

A partir des sélecteurs et des opérateurs, il est donc possible de traduire les connaissances linguistiques de manière formelle.

La figure 4.4 représente l'ensemble des compositions par redéfinition.
L'ensemble des actions faciales du corpus est donné en figure 4.5 (seuls quelques exemples d'actions faciales simples sont donnés).
La définition des expressions du visage présentes en Langue des Signes (d'après les descriptions de Cuxac [19]) est donnée en figure 4.7.
Enfin, la définition des émotions (d'après Cuxac [19] et Pantic [37]) est donnée en figure 4.6

Nom	Composition
sourcils-relevés	sourcil-gauche-relevé, sourcil-droit-relevé
sourcils-froncés	sourcil-gauche-froncé, sourcil-droit-froncé
sourcils-intérieurs-relevés	sourcil-intérieur-gauche-relevé, sourcil-droit-relevé
sourcils-intérieurs-froncés	sourcil-intérieur-gauche-froncé, sourcil-droit-froncé
lèvres-projetées	lèvre-inférieure-projetée, lèvre-supérieure-projetée
lèvres-étirées	lèvre-gauche-étirée, lèvre-droite-étirée
lèvres-serrées	lèvre-gauche-serrée, lèvre-droite-serrée
coins-bouches-relevés	coin-bouche-gauche-relevé, coin-bouche-droit-relevé
coins-bouches-abaissés	coin-bouche-gauche-abaissé, coin-bouche-droit-abaissé
joues-gonflées	joue-gauche-gonflée, joue-droite-gonflée
joues-creusées	joue-gauche-creusée, joue-droite-creusée
visage-incliné	visage-incliné-gauche || visage-incliné-droite
sourire-en-coin	(lèvre-gauche-étirée, -(lèvre-droite-étirée), coin-bouche-gauche-relevé?) || (lèvre-droite-étirée, -(lèvre-gauche-étirée), coin-bouche-droit-relevé?)
souffle-air	bouche-ouverte, joues-gonflées
lèvres-arrondies	bouche-ouverte, lèvres projetées

FIG. 4.4 – Redéfinition de connaissances

Action faciale	**Définition**
tremblement-joues	`af-tremblement [gonflement-joues, +]`
visage-oui-gd	`af-hochement [rotation-X, +]`
visage-oui-dg	`af-hochement [rotation-X, -]`
visage-oui	`visage-oui-dg || visage-oui-gd`
visage-non-hb	`af-hochement [rotation-Y, +]`
visage-non-bh	`af-hochement [rotation-Y, -]`
visage-non	`visage-non-hb || visage-non-bh`
af-ouverture-bouche	`af-simple [ouverture-bouche, +]`
af-relèvement-sourcils	`af-simple [relèvement-sourcils, +]`
af-froncement-sourcils	`af-simple [relèvement-sourcils, -]`

FIG. 4.5 – Définitions des actions faciales

Nom	**Définition**
émotion-joie	`lèvres-étirées, coins-bouches-relevées, yeux-plissés, bouche-ouverte?`
émotion-tristesse	`sourcils-intérieurs-relevés, yeux-plissés, coins-bouches-abaissés`
émotion-colère	`yeux-plissés, sourcils-abaissés, ( ( lèvres-serrées, - lèvres-projetées ) || lèvre-inférieure-projetée )`
émotion-dégoût	`nez-pincé, lèvres-ouvertes?`
émotion-peur	`sourcils-intérieurs-relevés, yeux-plissés, bouche-ouverte`
émotion-surprise	`sourcils-relevés, bouche-ouverte(grandement)`
émotion-douleur	`sourcils-froncés, nez-pincé`

FIG. 4.6 – Traduction des émotions

Nom	**Définition**	**Texte original**
grande quantité de	`(joues-gonflées, yeux-plissés) + souffle-air(durée=longue)`	*gonflement des joues, plissement des yeux, souffle d'air en continu*
gros	`(joues-gonflées, bouche-fermée) + (joues-neutres, bouche-ouverte)`	*gonflement des joues, souffle d'air retenu puis très brève explosion d'air*
flasque	`langue-visible, tremblement-joues`	*tremblement des joues, langue légèrement sortie entre les dents, souffle d'air*
spongieux	`flasque*`	*idem, mais répété*
vite	`lèvres-arrondies + (présence-langue, joues-gonflées, (sourcils-relevés)?)`	*souffle d'air, lèvres arrondies, arrête net de l'émission d'air au moyen d'un mouvement rapide de la langue aboutissant à une obstruction de la colonne d'air (plissement du front fréquent, mais non obligatoire*
minuscule	`yeux-plissés, nez-froncé, sourcils-froncés, lèvres-serrées, lèvres-projetées`	*plissement des yeux, froncement du nez et des sourcils, lèvres serrées et projetées vers l'avant*
minuscule et fin	`minuscule, souffle-air`	*idem, accompagné d'un souffle d'air*
petit	`yeux-plissés, lèvres-projetées(legèrement), (visage-incliné(légèrement)?)`	*plissement des yeux moindre, moue avec les lèvres, visage en général légèrement incliné*
normal	`lèvres-projetées(légèrement)`	*légère moue des lèvres, sans le plissement des yeux*
maigre	`joues-creusées, nez-pincé, lèvres-projetées`	*joues creusées, nez pincé, lèvres arrondies projetées*
fort	`joues-gonflées(légèrement), sourcils-relevés, mâchoires-serrées`	*joues légèrement gonflées, front plissé, air dur, mâchoires serrées*
pointu, piquant	`sourcils-relevés, nez-plissé, bouche-ouverte, émotion-douleur`	*front et nez plissés, aspiration d'air, grimace de douleur*
exhaustivité	`sourcils-relevés, tremblement-lèvres`	*Elle se réalise en plissant le front et en fronçant légèrement les sourcils en même temps qu'un souffle d'air fait trembler les lèvres.*

FIG. 4.7 – Traduction des connaissances linguistiques (1)

Nom	Définition	Texte original
beau	`normal, sourcils-relevés`	*« normal » combiné à un relèvement des sourcils, expression « fière », regard neutre*
normalité	`joues-creusées, lèvres-projetées`	*en creusant un peu les joues et en projetant en légère lippe arrondie les lèvres*
conditionnel	`visage-incliné, visage-reculé(légèrement), regard-en-haut, sourcils-relevés`	*inclinaison et léger mouvement de recul du visage vers l'arrière ; le regard, désinvesti, est dirigé vers le haut, les sourcils sont relevés*
hypothèse mentale	`clignement-yeux + regard-en-haut`	*clignement très bref des yeux suivi d'une fuite du regard ultra-rapide vers le haut*
croire à tort	`regard-en-haut(durée=grande)`	*un regard qui part vers le haut pendant toute la durée de l'énoncé*
détrimental actif	`(langue-saillante-joue-gauche + langue-saillante-joue-droite) || (langue-saillante-joue-droite + langue-saillante-joue-gauche)`	*la langue format saillance contre la joue droite (pour un locuteur droitier) et qui, très rapidement vient buter contre la joue gauche*
détrimental passif	`langue-visible`	*langue sortie visible, sans mouvement*
impératif	`sourcils-froncés, visage-avancé`	*en fronçant les sourcils, avec un mouvement du visage en direction de l'interlocuteur*
impératif négatif	`sourcils-froncés, visage-avancé, visage-non`	*la même expression, accompagnée d'un « non » de la tête*
volitif	`mâchoires-serrées(fortement)`	*en serrant fortement les mâchoires, l'expression (regard) est plus ou moins intense*
incitatif	`visage-incliné, lèvres-arrondies, yeux-plissés`	*mouvement d'inclinaison du visage accompagné d'une légère moue (avancée des lèvres faiblement arrondies) et d'un plissement des yeux*
réprobatif	`incitatif, sourcils-froncés, visage-non`	*proche de la précédente, mais les sourcils sont froncés en même temps que le locuteur fait « non » de la tête*
ironique	`visage-incliné, sourire-en-coin`	*bref mouvement d'inclinaison du visage, proche de l'interrogation, regard plutôt vague, sourire en coin*

FIG. 4.8 – Traduction des connaissances linguistiques (2)

Nom	**Définition**	**Texte original**
dubitatif	`sourcils-relevés, visage-relevé, lèvres-projetées`	*les sourcils sont relevés, petit mouvement du visage vers le haut, puis redescendant très faiblement, moue accentuée*
assertif	`visage-oui*`	*expression sérieuse, hochement(s) de tête*
assertif négatif	`visage-non`	
capacitif	`sourcils-relevés(légèrement), lèvre-inférieure-projetée`	*les sourcils sont légèrement relevés, expression du visage « sûre de soi », la lèvre inférieure est projetée vers l'avant*
problématisation	`froncement-sourcils`	*le froncement des sourcils est associé à une problématisation*
concessif	`relèvement-sourcils`	*un relèvement marqué des sourcils est associé à un changement thématique*
interrogatif	`sourcils-relevés, visage-reculé, visage-relevé`	*les sourcils sont relevés, le front légèrement plissé, le visage se porte vers l'arrière, le menton fortement relevé*
négatif	`visage-non, sourcils-froncés`	*« non » de la tête + froncement des sourcils*
interro-négatif	`visage-relevé, visage-non, (sourcil-gauche-froncé, sourcils-droit-relevé) || (sourcil-gauche-relevé, sourcil-droit-froncé)`	*menton et visage relevés, combinaison du relèvement et du froncement des sourcils, « non ... non » de la tête*
duratif	`tremblement-lèvres(légèrement), souffle-air`	*léger tremblement des lèvres avec un souffle d'air*
continu	`souffle-air(léger)`	*léger souffle d'air*
ponctuel	`(joues-gonflées, bouche-fermée) + (joues-neutres, bouche-ouverte)`	*très brève explosion d'air faite par les lèvres*
résultatif	`lèvres-serrées + lèvres-étirées`	*en serrant et en rétractant les lèvres*

FIG. 4.9 – Traduction des connaissances linguistiques (3)

4.1.3 Représentation interne

De manière interne, les connaissances sont représentées sous forme d'un ensemble de structures de données, qui découlent de l'analyse.

Pour résumer :
- Une connaissance possède un *nom* et un *état*.
- L'*état* d'une connaissance est décrit par un ensemble de *propriétés*, d'un certain *type* (entier, chaine, symbole) et définies sur un certain *domaine*.
- Une connaissance est associée à une fonction qui permet de la *vérifier* et/ou une fonction qui permet de l'*extraire*.
- Chaque fonction est soit un opérateur de Traitement d'Images, soit une *règle de composition*.
- Une *règle de composition* permet de composer plusieurs connaissances pour exprimer une *action faciale*, une *redéfiniton* ou une *expression*.
- Les compositions autre que la définition des actions faciales sont formées de plusieurs connaissances combinées par un ensemble d'*opérateurs de composition*.
- L'état d'une connaissance utilisée dans la définition d'une autre peut être spécifié par un *sélecteur*.
- Un *sélecteur* sélectionne un *sous-ensemble* d'une *propriété* d'une connaissance.

Toutes ces règles sont résumées par un schéma UML (voir figure 4.10).

4.1.4 Représentation externe

Toutes les connaissances sont traitées par le système à partir de leur représentation interne. Le stockage externe (en dehors du système) des connaissances est du même type que le stockage interne.

On a choisi ici une représentation externe des connaissances basée sur XML. Ce langage de haut niveau possède plusieurs avantages qui ont motivé ce choix, en particulier le fait que XML soit un langage de description générique normalisé et donc associé à un ensemble d'outils de traitements largement disponibles.

La structure d'un document XML peut être validée par une description de sa structure, à l'aide d'un fichier DTD (Document Type Definition). De nombreux outils permettent, à partir d'un document DTD, de valider la structure d'un document XML.

La structure du document XML est directement guidée par la structure du formalisme interne adopté précédemment.

```
<?xml version="1.0"?>
<!DOCTYPE knowledges [

  <!ELEMENT knowledges
```

```
          (knowledge-definition*)>

  <!ELEMENT knowledge-definition
          (function)*>
  <!ATTLIST knowledge-definition name CDATA #REQUIRED>

  <!ELEMENT function
          (composition-rule |
           image-operator |
           facial-action |
           EMPTY)>

  <!ATTLIST function type
          (extract | check | check-or-extract)
           #REQUIRED>

  <!ELEMENT image-operator
          (EMPTY)>
  <!ATTLIST image-operator name CDATA #REQUIRED>

  <!ELEMENT facial-action
          (EMPTY)>

  <!ATTLIST facial-action
           type
          (simple | tremblement | hochement)
          #REQUIRED>
  <!ATTLIST facial-action reference CDATA #REQUIRED>
  <!ATTLIST facial-action sign (+ | -) #REQUIRED>

  <!ELEMENT composition-rule
          (composition | selector)>

  <!ELEMENT composition
          (composition | selector)*>
  <!ATTLIST composition operator CDATA #REQUIRED>

  <!ELEMENT selector
          (selection)?>
  <!ATTLIST selector name CDATA #REQUIRED>

  <!ELEMENT selection
          (EMPTY)>
  <!ATTLIST selection property CDATA #REQUIRED>
```

```
  <!ATTLIST selection expression CDATA #REQUIRED>
]>
```

Cette définition permet de décrire les différentes connaissances introduites précédemment. Dans l'exemple suivant, on trouve la définition de plusieurs connaissances : un exemple de définition d'action faciale (`af-visage-non`), un exemple de définition d'expression modale (`interro-négatif`), et un exemple de définition constituée uniquement d'une fonction de vérification (`joues-gonflées`).

```
<knowledges>

<!-- Hochement du visage de gauche à droite -->

<knowledge-definition name="af-visage-non-gd">
    <function type="check-or-extract">
        <facial-action reference="rotation-Y"
                       type="hochement"
                       sign="+"/>
    </function>
</knowledge-definition>

<!-- Hochement du visage de droite à gauche -->

<knowledge-definition name="af-visage-non-dg">
    <function type="check-or-extract">
        <facial-action reference="rotation-Y"
                       type="hochement"
                       sign="+"/>
    </function>
</knowledge-definition>

<!-- Hochement du visage -->

<knowledge-definition name="af-visage-non">
    <function type="check-or-extract">
        <composition-rule>
            <composition operator="disjonction">
                <selector name="af-visage-non-gd"/>
                <selector name="af-visage-non-dg"/>
            </composition>
        </composition-rule>
    </function>
</konwledge-definition>
```

```
<!-- Expression modale interro-négatif -->

<knowledge-definition name="interro-négatif">
    <function type="check-or-extract">
        <composition-rule>
            <composition operator="conjonction">
                <selector name="visage-relevé"/>
                <selector name="af-visage-non"/>
                <composition operator="disjonction">
                    <composition operator="conjonction">
                        <selector name="sourcil-gauche-froncé"/>
                        <selector name="sourcil-droit-relevé"/>
                    </composition>
                    <composition operator="conjonction">
                        <selector name="sourcil-gauche-relevé"/>
                        <selector name="sourcil-droit-froncé"/>
                    </composition>
                </composition>
            </composition>
        </composition-rule>
    </function>
</konwledge-definition>

<!-- expression de la surprise -->

<knowledge-definition name="émotion-surprise">
    <function type="check-or-extract">
        <composition-rule>
            <composition operator="conjonction">
                <selector name="sourcils-relevés"/>
                <selector name="bouche-ouverte">
                    <selection property="amplitude"
                               expression="grand"/>
                </selector>
            </composition>
        </composition-rule>
    </function>
</knowledge-definition>

<!-- joues gonflées ? -->

<knowledge-definition name="joues-gonflées">
    <function type="check">
        <image-operator name="correlation-joues"/>
```

```
    </function>
</knowledge-definition>

</knowledges>
```

4.2 Architecture du système d'analyse

La section précédente présentait un formalisme de représentation des connaissances des expressions de la Langue des Signes. Ce formalisme est constitué d'un ensemble de **faits**, appelés *connaissances élémentaires*, qui correspondent à l'état musculaire de certaines composantes du visage, et d'un ensemble de **définitions** qui permettent de traduire les *connaissances composées*.

Les **définitions** des connaissances composées sont représentées, implicitement, au moyen de **règles** : un nouveau fait (partie gauche de la règle) est défini par un ensemble d'autres faits (partie droite de la règle) assemblés par des **opérateurs de compositions**.

Le formalisme est adapté au mécanisme de déduction de connaissances. Les connaissances de « haut niveau » sont composées à partir d'autres de « bas niveau », les connaissances de plus bas niveau étant les connaissances représentant l'état musculaire des composantes faciales.

Les *règles* permettent de guider l'analyse. Pour chaque règle, si les connaissances de la partie droite sont présentes, on peut déduire la nouvelle connaissance de la partie gauche.

4.2.1 Analyse ascendante

Les connaissances élémentaires sont extraites à l'aide d'opérateurs de Traitement d'Images. Dans ce type d'analyse, il est nécessaire que chaque connaissance élémentaire puisse être extraite par un opérateur adéquat.

Or, il est difficile d'extraire toutes les connaissances élémentaires à partir d'opérateurs de Traitement d'Images.

Un module d'analyse des lèvres, comme celui présenté par Delmas ([20], voir A.2), permet d'extraire les connaissances `ouverture-bouche`, `étirement-lèvres`. Les connaissances `langue-visible` et `dents-visible` peuvent être déduites de leur profil de couleur, une fois les lèvres identifiées.

Les connaissances sur l'état des sourcils (`relèvement-sourcil-gauche` et `relèvement-sourcil-droit` peuvent être facilement déduites par une analyse du gradient de la partie haute du visage.

Les connaissances sur les mouvements « rigides » du visage (translation et rota-

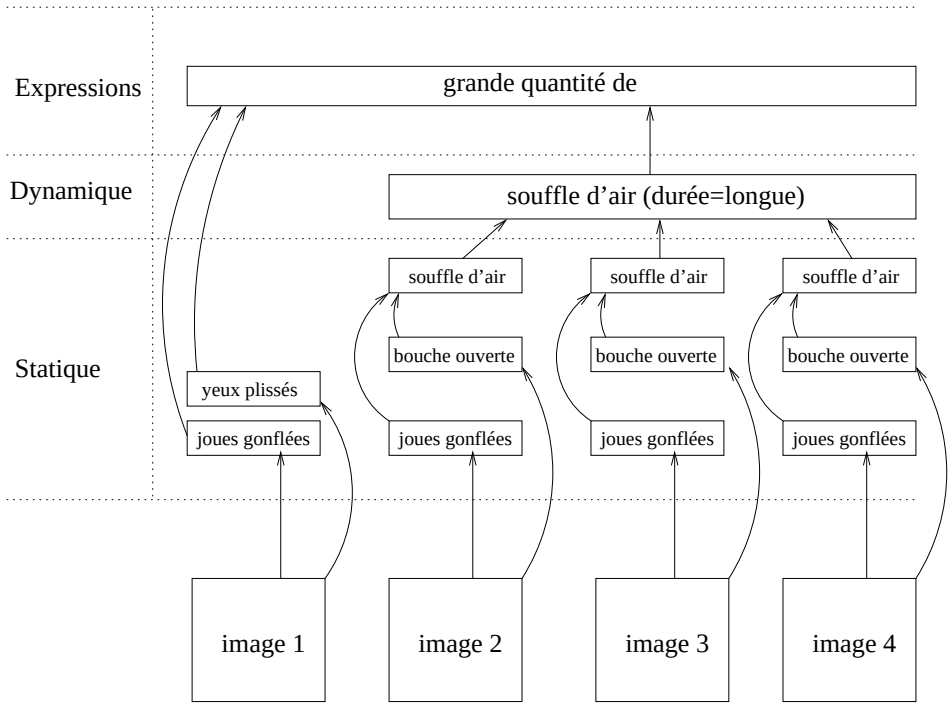
tion) peuvent être extraites à partir d'un module de reconstruction des paramètres 3D du visage.

Cependant, certaines connaissances sont difficiles à extraire directement. C'est le cas par exemple des connaissances sur l'état des joues (gonflement ou saillance de la langue), des yeux (plissement, regard), des mâchoires (contraction), des lèvres (projection). Soit parce que la définition de la séquence vidéo risque de ne pas être assez élevée pour différencier un état sur une image (regard, contraction des mâchoires), soit parce que certaines connaissances ne sont pas visibles à partir d'une vue de face (projection des lèvres), soit parce qu'il n'existe pas d'opérateur de Traitement d'Images adéquat.

Il peut être par contre plus facile d'avoir une « idée » de l'état de certaines composantes. Par exemple, il est possible de savoir que les joues, sur une image, ne sont pas dans le même état que sur une autre image, sans savoir exactement dans quel état elles se trouvent. On peut, par exemple, utiliser des mesures de corrélations (décomposition en sous-espaces propres par exemple) pour savoir si une zone de l'image a changé d'état par rapport à une configuration initiale (neutre). On distingue alors l'état neutre des autres, plutôt que de distinguer **tous** les états possibles.

Dans ce type d'analyse (ascendante), les connaissances sont composées à partir de connaissances de bas niveau. Du fait de l'aspect temporel de certaines connaissances, il existe bien des cas où aucune connaissance ne peut être directement déduites d'une image isolée. Ainsi, les décisions sont prises au dernier moment, quand on a extrait suffisamment de connaissances pour pouvoir distinguer entre les différentes compositions possibles, et l'analyse ascendante implique de stocker beaucoup de connaissances avant la décision.

Dans certains cas, il est plus facile d'utiliser des opérateurs de **vérification** que des opérateurs d'**extraction**. C'est le cas par exemple pour l'état des joues, où il est plus facile de *vérifier* qu'une joue est gonflée plutôt que d'*extraire* son état.



4.2.2 Analyse descendante

L'analyse descendante part de connaissances de haut niveau et les *vérifie* par plusieurs mécanismes : *prédiction* et/ou *suivi*. L'intérêt de l'analyse descendante est qu'elle traite des connaissances de plus haut niveau, qui sont généralement proches de la compréhension humaine. De plus, étant située à un plus haut niveau, elle est généralement moins coûteuse qu'une analyse ascendante classique.

Si certaines informations sont connues à un instant donné de l'analyse, il est parfois possible de *prédire* les connaissances sur les prochaines images. Ces prédictions sont alors à *vérifier*. L'hypothèse est alors que les opérateurs de *vérification* sont moins coûteux que les opérateurs d'*extraction*.

De même, il est possible de tirer profit des connaissances déjà accumulées à un instant de l'analyse pour *suivre* leur évolution, plutôt que d'appeler des opérateurs d'extraction qui procèdent sans informations préalables. Il est généralement plus facile de trouver le prochain état d'un état donné, plutôt que d'extraire un état sans connaissances préalables.

Par exemple, connaissant la position de quelques points de références des sourcils sur une image donnée, il est moins coûteux de *suivre* la position de ces points sur l'image suivante plutôt que d'*extraire* leur nouvelle position.

On a donc tout intérêt à mener l'analyse le plus possible de manière descendante. Malheureusement, il est imposible de mener une analyse exclusivement descen-

dante. D'abord parce qu'elle tire profit d'un certain nombre d'informations qu'il a bien fallu extraire d'une manière ou d'une autre (par une analyse ascendante) et ensuite parce que les mécanismes de vérification ne sont pas toujours suffisants.

Lorsque la *prédiction* échoue, il est nécessaire de faire un retour arrière et de traiter une autre possibilité. Si aucune possibilité n'est valable, on manque d'informations pour conclure, il est donc nécessaire de mener à nouveau une analyse ascendante pour extraire de nouvelles connaissances.

Les opérateurs de *suivi* sont peu coûteux, mais peuvent donner des résultats faussés. Par exemple, suivre la position d'un point de référence d'une image à l'autre par mesure de corrélation, est difficile si, par exemple, la luminosité change beaucoup sur la prochaine image. Il est alors nécessaire de mener à nouveau une analyse ascendante.

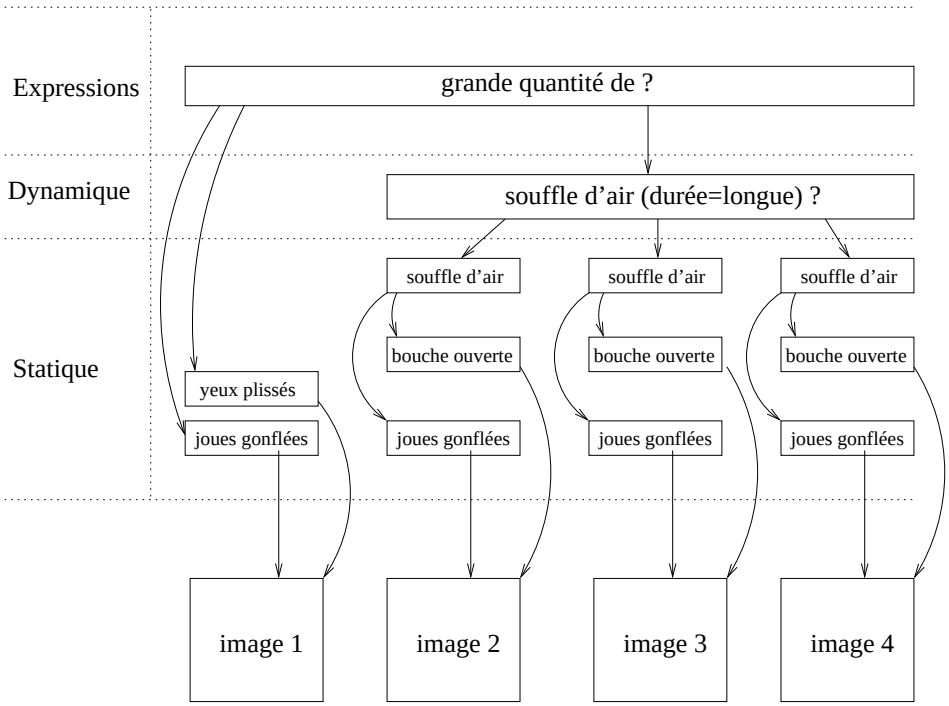
Prédiction

Il est possible de prédire une connaissance future, du à l'aspect temporel de certaines connaissances. Par exemple, dans le cas de la vérification de l'expression *hypothèse mentale*, décrite par la succession de `clignement-yeux` et `regard-en-haut`, si `clignement-yeux` a été vérifiée sur une image, on peut *prédire* la présence de `regard-en-haut` sur les images suivantes.

Ainsi, dans ce type d'analyse, il est possible de « sauter » l'analyse de certaines images, ou de ne traiter que partiellement certaines images et de simplement vérifier la prédiction. En cas d'erreur, le principe est de faire un « retour arrière » puisque la prédiction n'est plus vrai. On gagnerait ainsi en temps de calcul dans le cas où la prédiction serait vérifiée.

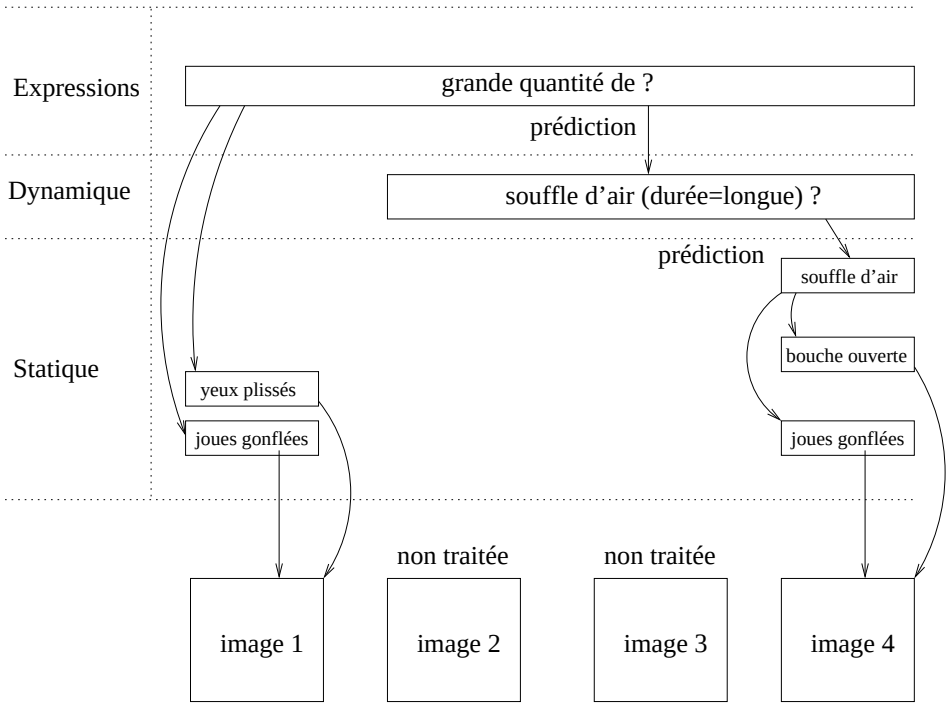
On se limitera ici à la prédiction d'une image à l'autre. Le mécanisme de « retour arrière » n'étant nécessaire qu'en cas de prédiction sur plus d'une image, on ne traitera donc pas le cas des retours arrières.

Par exemple, soit une analyse descendante sans prédiction, qui correspond à une chaînage arrière (partir des connaissances de haut niveau) d'une analyse ascendante (à chaînage avant, en partant des connaissances de bas niveau) :



Le mécanisme de prédiction entre en jeu pour la vérification de la connaissance *grande quantité de* qui est définie par `(joues-gonflées, yeux-plissés) + souffle-air (durée = longue)`. En effet, en sachant que la première partie temporelle de la règle est vérifiée (`joues-gonflées, yeux-plissés`), il est possible de *prédire* les prochaines connaissances.

Prédire consiste donc à traiter en priorité un certain nombre de vérification pour les prochaines images.



Le mécanisme de *prédiction* peut être pris en compte à un niveau encore plus haut, en ayant connaissance de l'agencement temporel des expressions dans une locution en Langue des Signes.

Par exemple, Christian Cuxac indique, à propos de l'expression modale « capacitif » qu'elle est souvent accompagnée de l'assertif ou de l'assertif négatif. Ainsi, en prenant en compte ces informations, il est possible de mettre en priorité la vérification des expressions « assertif » et « assertif négatif » lorsque l'on a détecté l'expression « capacitif ».

Malheureusement ce type d'informations sur la statistique des expressions du visage en Langue des Signes est relativement peu connu est donc peu décrit. C'est pourquoi, on ne le traitera pas ici.

4.2.3 Analyse bi-directionnelle

L'analyse bi-directionnelle consiste à combiner les deux précédentes : ascendante et descendante.

Comme indiqué précédemment, l'analyse descendante n'étant efficace que si elle se base sur un ensemble de connaissances, le système est « amorcé » par une première analyse ascendante qui extrait un certain nombre de connaissances (celles que l'on sait extraire ou qu'il est plus facile d'extraire). A partir de ces premières connaissances, on peut lancer une analyse descendante, qui vérifiera les différentes connaissances incomplètes.

La manière de traiter entre analyse ascendante et descendante représente le point central du système, le choix judicieux entre les deux types d'analyse accroisant l'efficacité du système.

Il existe plusieurs manières de guider ce choix. Il peut, par exemple, être guidé par un module *superviseur* qui choisit (selon une heuristique) quel type d'analyse lancer et quand. Il est possible aussi que chaque type d'analyse « passe la main » à l'autre type quand l'analyse est bloquée.

4.2.4 Niveaux langagiers

La Langue des Signes, comme tout langage articulé possède un certain nombre de « niveaux langagiers ». Les expressions du visage appartiennent elles aussi à plusieurs niveaux du langage.

Les expressions du visage en Langue des Signes ont plusieurs rôles :

1. un rôle **lexical** puisqu'elles permettent de définir des « signes » (*i.e.* de mots du langage), seules (qualifiants) ou accompagnées d'autres indices (comme la position du corps et/ou la configuration de la main). Par exemple, l'expression du visage permet de distinguer entre les signes « content » et « mal au coeur », où la configuration, l'orientation, l'emplacement et le mouvement de la main sont les mêmes. Accompagnée de l'expression de « joie », cette configuration définie le signe « content ». Accompagnée de l'expression « crispée », elle définie le signe « mal au coeur ». Le quantifiant « grande quantité de », quant à lui, est défini uniquement grâce aux expressions du visage.
2. un rôle **syntaxique** puisqu'elles permettent, à elles seules de définir la plupart des modes du discours (conditionnel, intérrogatif, assertif, etc.),
3. un rôle **sémantique** puisque le regard permet (généralement avec un geste du corps) d'initier une situation de Transfert Personnel. Dans cette situation, les émotions que reflètent le locuteur sont celles ressenties par le personnage joué.

4.2.5 Niveaux d'analyse

L'analyse d'une langue est généralement guidée par ces niveaux langagiers. Dans le cas d'un langage informatique, le découpage est clair et tous les traducteurs / interpréteurs / compilateurs travaillent à trois niveaux : lexical, syntaxique et sémantique. Les niveaux langagiers sont des niveaux d'abstraction du langage.

Pour l'analyse d'un langage informatique, le niveau lexical est le niveau le plus bas. Les données brutes à traiter sont les caractères. Les caractères sont composés pour former des lexèmes.

Pour l'analyse informatique d'un langage articulé, le niveau lexical n'est pas le plus bas. La composition des différents indices en éléments lexicaux (*i.e.* les états musculaires en expression) n'est pas immédiate.

C'est pourquoi le niveau lexical langagier pour l'analyse des expressions est constitué de différents niveaux d'analyse. Ces niveaux d'analyse sont des niveaux de composition. Chaque connaissance d'un niveau est composée de connaissances des niveaux inférieurs.

Les niveaux d'analyse sont ainsi différents des niveaux langagiers, puisqu'on peut avoir des connaissances d'un niveau d'analyse peu elevé interprété par un niveau langagier haut. Par exemple, une information statique sur le regard, ne nécessitant que peu de compositions peut être interprété en terme de haut niveau langagier (passage en situation de Transfert Personnel par exemple).

On distingue ici essentiellement trois niveaux d'analyse : le niveau statique où les connaissances sont extraites directement des informations brutes (de la séquence vidéo), le niveau dynamique qui compose temporellement les connaissances du niveau statique et le niveau des expressions qui compose les connaissances des niveaux précédents.

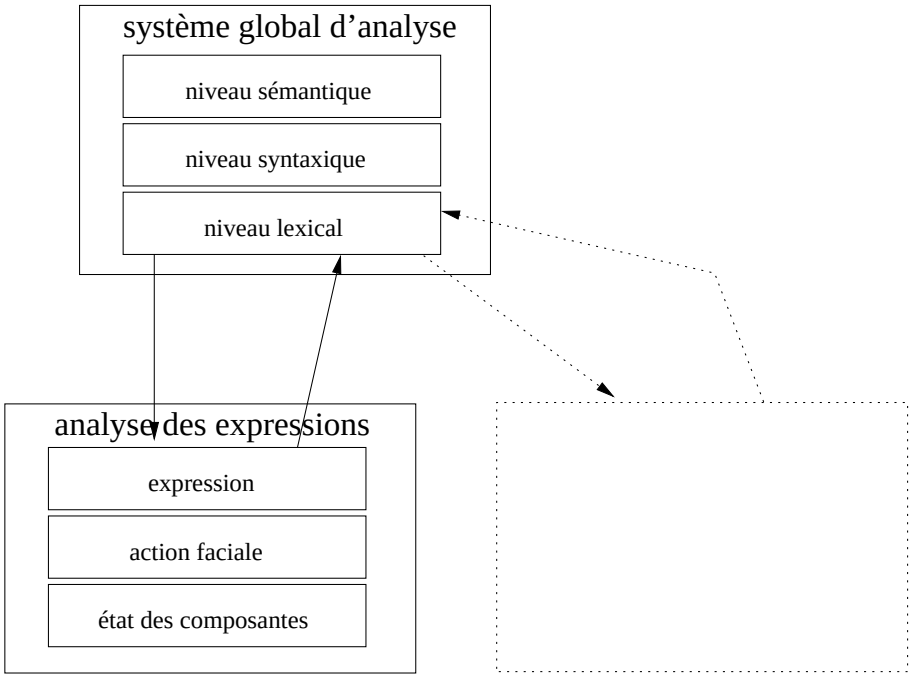
Le niveau statique correspond aux attributs du visage directement observables. Les connaissances extraites sont généralement des connaissances sur l'état actuel (*i.e.* sur l'image courante) d'une certaine composante du visage, ou d'un certain muscle. Les connaissances sont extraites isolément sur chaque image.

Le niveau dynamique correspond à l'intégration temporelle des connaissances du niveau statique. Les connaissances statiques sont intégrées temporellement soit pour former une nouvelle connaissance sur l'évolution simple d'un muscle, soit pour former une action faciale, symbolisant un ensemble d'évolutions.

Le niveau des expressions compose les connaissances des deux niveaux précédents. Les compositions sont modélisées par les règles décrites précédemment.

Les niveaux langagiers (niveau lexical, syntaxique et sémantique) ne peuvent que difficilement aider à la compréhension s'ils sont uniquement composés d'informations sur les expressions du visage.

Le système d'analyse des expressions du visage présentes en Langue des Signes ne constitue donc qu'un module d'un système plus global de compréhension, qui intègre les connaissances des différents modules (expressions du visage, mouvements du corps, configurations de la main).


système global d'analyse
niveau sémantique
niveau syntaxique
niveau lexical
analyse des expressions
expression
action faciale
état des composantes

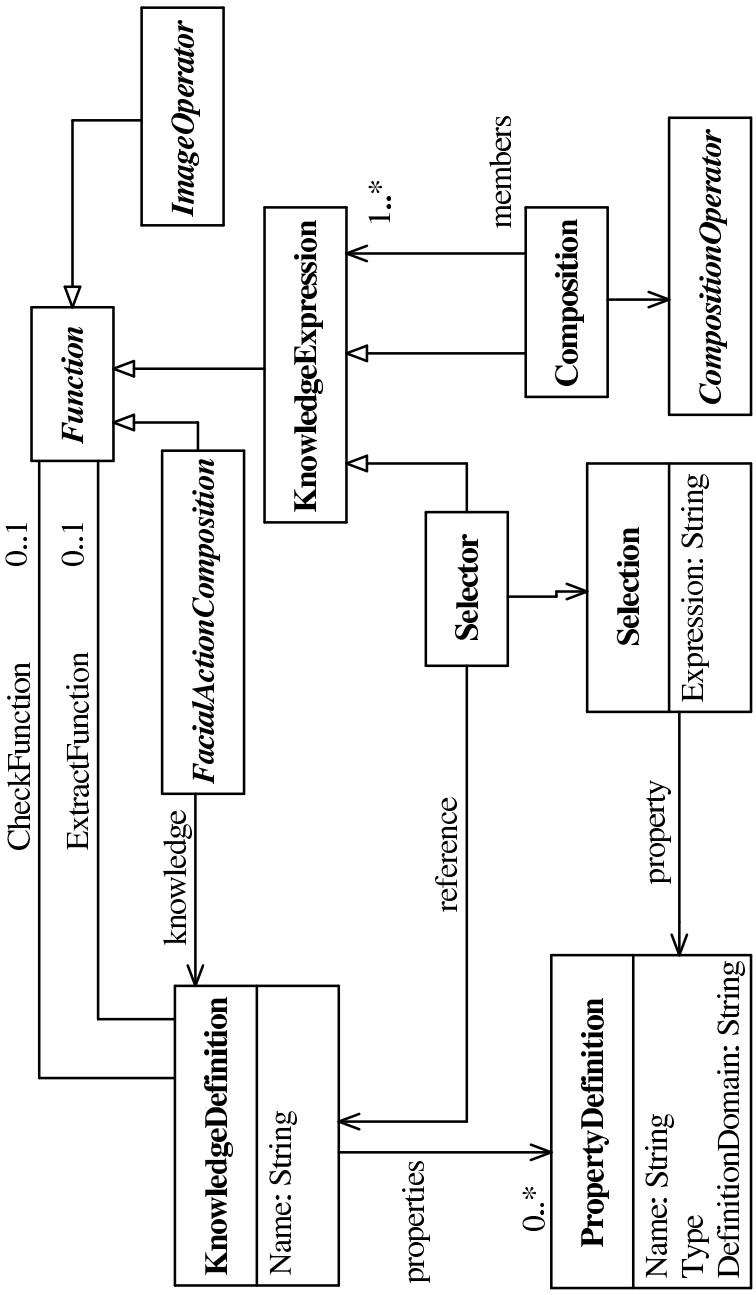


FIG. 4.10 – Description UML du formalisme de représentation interne des connaissances

5

Conception

On présente dans ce chapitre les détails algorithmiques d'un système à analyse bi-directionnelle des expressions du visage.

Comme présenté au chapitre précédent, le système mène une analyse bi-directionnelle à l'aide d'un ensemble d'opérateurs (extraction, prédiction, suivi, vérification) en vue d'une *interprétation* linguistique d'une locution en Langue des Signes.

On présente dans ce chapitre la manière dont vont être exploités les connaissances et les différents opérateurs par un système informatique d'analyse.

5.1 Mécanisme d'extraction

Le mécanisme d'extraction est le mécanisme classique de l'analyse ascendante. A partir des données brutes (les images de la séquence vidéo), les connaissances élémentaires sont extraites à l'aide d'opérateurs de Traitement d'Images. Si ces connaissances peuvent se composer à l'aide de règles de composition, elles le sont.

On propose ici un premiere liste d'opérateurs d'extraction, qu'il faudra valider par la suite.

5.1.1 Connaissances sur les composantes faciales

Pour les connaissances liées à la bouche, à savoir l'ouverture de la bouche, le degré d'étirement des lèvres, le degré de relèvement des commissures, la visibilité des dents et de la langue, on utilisera l'opérateur d'extraction de Delmas, détaillé en annexe (voir en A.2).

Pour les connaissance liées aux yeux et au regard, à savoir le degré de fermeture des paupières et la direction du regard, on utilisera l'opérateur d'extraction présenté par Christophe Collet (voir en A.2).

Les connaissances sur la configuration des sourcils seront extraites à l'aide d'un opérateur basé sur le gradient de l'image. Quatre points de référence seront détectés

sur les sourcils et permettent de déterminer le degré de relèvement / abaissement des sourcils extérieurs et intérieurs.

5.1.2 Prétraitements

Les connaissances sur les translations et rotations du visage seront extraites par une méthode de reconstruction 3D avec adaptation de modèle 3D (en ne prenant en compte que le haut du visage, *i.e.* sans la partie du menton dont les mouvements introduisent des rotations hors plan) du type Candide.

Les zones susceptibles d'être un visage seront extraites par analyse de la couleur, en utilisant un modèle de couleur de la peau. Les pixels appartenant à la peau seront regroupés par une analyse en composantes connexes pour former (au maximum) trois formes, dont une plus grande : le visage et les deux mains du locuteur. La position de l'enveloppe externe des mains pourra ainsi servir à traiter le problème d'occultation.

La position des boîtes englobantes des différentes composantes sera extraite avec une technique basée sur les projections verticales et horizontales du visage, dans la même idée que l'opérateur présenté par Pantic (voir A.1).

Toutes les autres connaissances élémentaires (gonflement des joues, plissement des yeux, projection des lèvres, etc.) n'ont pas d'opérateur d'extraction correspondant, mais des opérateurs de vérification, qui seront, *a priori* , des mesures de corrélation.

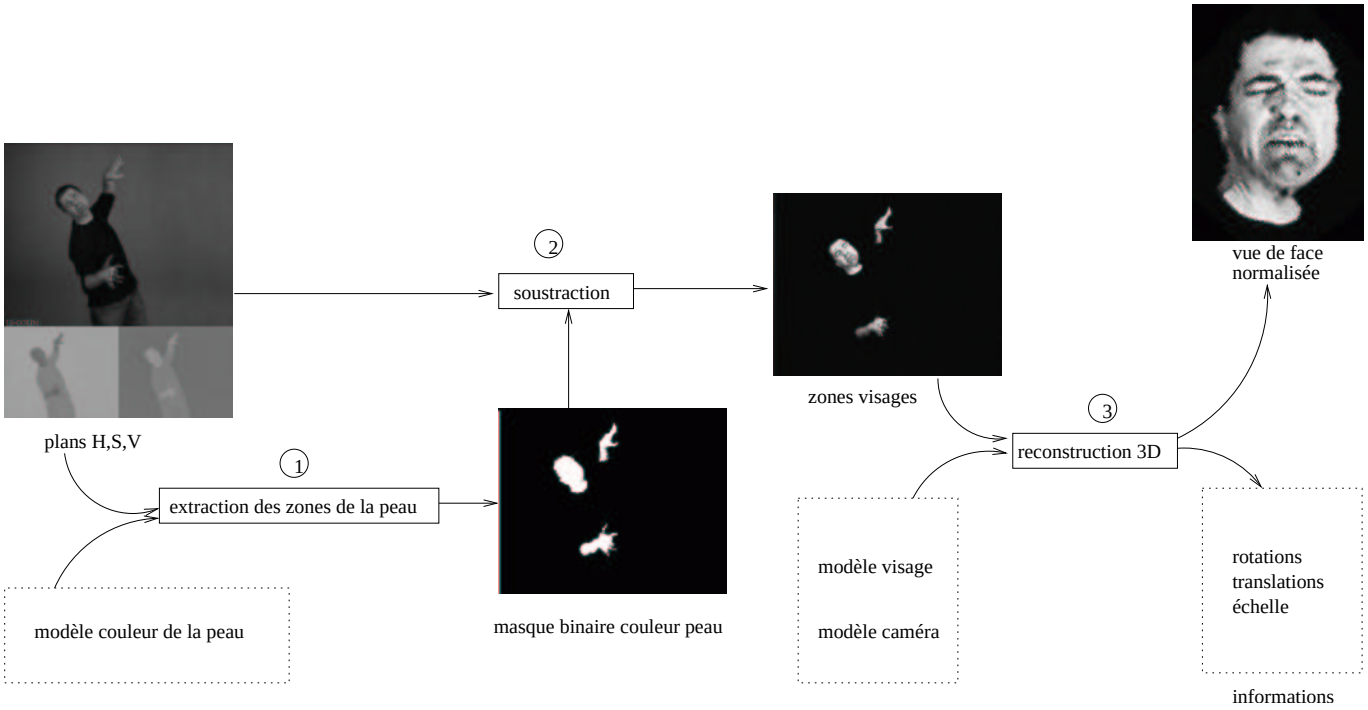


FIG. 5.1 – Chaîne des opérateurs de prétraitement

L'extraction par règles de composition est menée de manière classique, au même titre que la déduction de nouveaux faits dans un système expert. Chaque règle est parcourue et si tous les faits de la partie droite sont présents dans la base des connaissances, la connaissance de la partie gauche est ajoutée à la base. Puis, les règles sont à nouveau parcourues jusqu'à ce qu'aucun ajout n'ait été fait à la base. L'extraction par règles de compositions peut donc être vue comme une déduction par chaînage avant dans un système expert.

5.2 Mécanisme de vérification

Une connaissance quelconque peut être vérifiée à tout moment. Les opérateurs de *vérification* peuvent être les mêmes que les opérateurs d'extraction, auquel cas, le résultat de l'extraction est comparée au résultat attendu. Mais l'intérêt est d'avoir des opérateurs de vérification moins coûteux que les opérateurs d'extraction.

Les opérateurs de Traitement d'Images adaptés au mécanisme de vérification sont les opérateurs de mesures par corrélation, puisque la vérification consiste à comparer l'observation à un modèle hypothètique.

La vérification par règles de compositions peut être vue comme une déduction par chaînage arrière dans un système expert. La règle dont la partie gauche est la connaissance à vérifier est choisie et si toutes les connaissances de la partie droite sont présentes dans la base, la règle est vérifiée. Si une connaissance de la partie droite est elle-même définie par une règle de composition, cette règle est à son tour développée.

5.3 Mécanisme de prédiction

Le mécanisme de prédiction est le point central de l'analyse bi-directionnelle. C'est ce mécanisme qui permet de gagner en temps de calcul par rapport à une analyse classique.

La prédiction peut intervenir à tous les niveaux. Par exemple, à bas niveau, il est possible de déduire la position de la boîte englobante de l'oeil gauche à partir de la position de la boîte englobante de l'oeil droit. On peut donc ainsi prédire puis vérifier cette position.

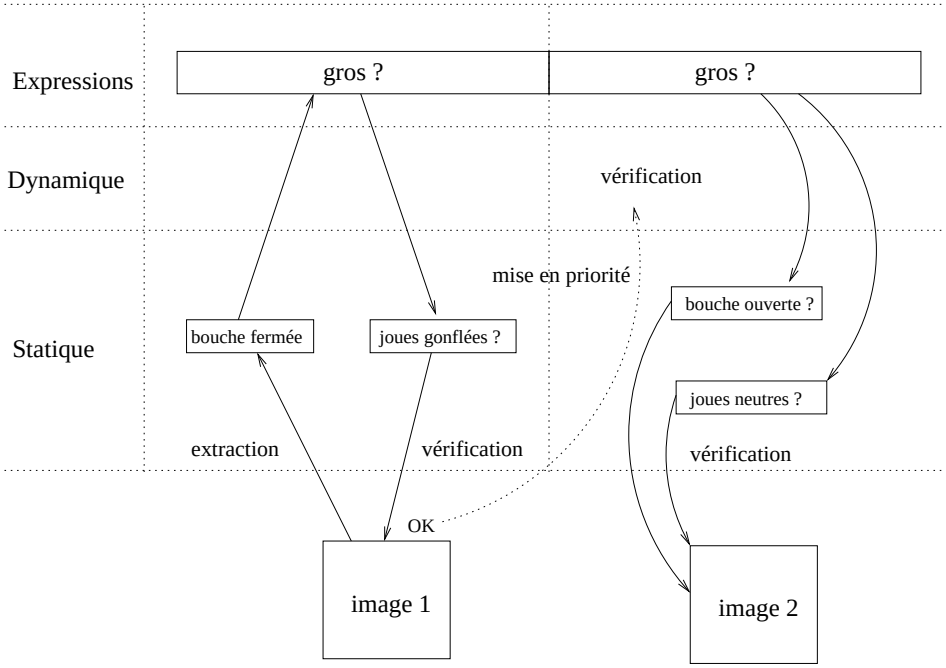
A haut niveau, comme il a été dit, il est possible de prédire les prochains états musculaires du visage si on a extrait un certain nombre d'informations sur une image donnée, en ayant une règle qui indique le lien temporel entre les deux images.

Le mécanisme de prédiction consiste donc à traiter certains opérateurs **en priorité** par rapport à d'autres. C'est pourquoi on propose ici d'utiliser une **file des appels d'opérateurs**.

Par exemple, avec une règle du type `Z := A + B`, si `A` a été validée sur l'image courante, le premier opérateur à appeler pour la prochaine image sera la fonction de vérification de la règle `Z := B`. Ainsi, cette règle sera traitée en priorité.

Lors d'une analyse ascendante, la file est remplie par les opérateurs d'extraction de connaissances, puis par les règles de composition, dans un ordre arbitraire.

Par exemple, pour l'expression *gros* définie par `(joues-gonflées, bouche-fermée) + (joues-neutres, bouche-ouverte)`, et en supposant que les connaissances sur les joues ne peuvent être que vérifiées, on a, par exemple :



5.3.1 Algorithme

Comme dit précédemment, le point crucial d'une analyse bi-directionnelle est sa capacité à basculer entre analyse ascendante et analyse descendante. Idéalement, les deux processus travaillent en parallèle et l'analyse ascendante est traitée en priorité quand il manque de l'information à l'analyse descendante.

On propose ici un premier algorithme d'analyse bi-directionnelle qui utilise un mécanisme de *prédiction* d'une image sur l'autre à l'aide d'une *file d'appel*. Si une règle a été *partiellement* vérifiée sur une image, la partie temporelle restante sera traitée en priorité sur les images suivantes.

La *vérification* est traitée de la manière suivante : si une part non négligeable des connaissances de la partie droite d'une règle est déjà présente dans la base, on pourra se contenter de *vérifier* celles qui ne s'y trouvent pas.

Soient K un ensemble représentant la base des connaissances, F et F2 représentant une file d'appel. On a alors l'algorithme suivant :

```
K := 0;
F := 0;
Pour chaque image i faire
    ajouter à F les fonctions d'extraction
    Tant que F non vide faire
        f := F.défiler();
        Si (f est un opérateur)
            appeler f et ajouter à K
                   les nouvelles connaissances;
        Fin Si
        Si (f est une règle)
            Validation(f, r);
            Si (r.état == VALIDATION-REPORTEE)
                F2.enfiler(r.reste);
                f.marque := i;
            Fin Si
            Si (r.état == VALIDATION-PARTIELLE)
                F.enfiler(r.manque);
            Fin Si
            Si (r.état == VALIDE)
                k := r.connaissance;
                Si (f.marque != 0)
                    k.durée := i - f.marque;
                Fin Si
                Ajouter r.connaissance à K;
            Fin Si
        Fin Si
    Fin Tant Que
    échanger F et F2;
Fin Pour
```

La validation des règles est menée par la fonction `Validation` qui prend en paramètre une règle. Le résultat de cette fonction possède plusieurs champs. Le champ `état` indique si la règle est entièrement validée (`VALIDE`), si la validation est reportée aux images suivantes par un mécanisme de prédiction (`VALIDATION-REPORTEE`) ou si la validation n'est que partielle, auquel cas on utilise des opérateurs de vérification (`VALIDATION-PARTIELLE`).

La résultat de `Validation` contient la règle temporaire à valider sur les images suivantes en cas de validation reportée (champ `reste`). Il contient la liste des connaissances à *vérifier* en cas de validation partielle (champ `manque`). Il contient

l'instanciation de la nouvelle connaissance en cas de validation totale (champ `connaissance`).

Lors d'une validation reportée, la règle originale est marquée de la position à partir de laquelle la validation a été reportée. Lors de la validation totale d'une règle, si elle a été marquée, la durée de la nouvelle connaissance correspond à la durée entre la position actuelle et la position marquée lors du report de la validation.

5.4 Ajout de nouvelles connaissances

L'ajout d'une nouvelle connaissance peut se faire parce qu'une connaissance à été vérifiée ou extraite par un opérateur de Traitement d'Images. Dans ce cas, les opérateurs renvoient directement une connaissance instanciée. C'est le cas pour les connaissances élémentaires (éventuellement aussi pour quelques connaissances composées ayant un opérateur spécialisé).

Les connaissances composées sont déduites à partir de règles de composition. Instancier la connaissance de la partie gauche d'une règle, *i.e.* donner une valeur à chacune de ses propriétés, se fait de manière différente selon le type de connaissance.

Les connaissances définies comme des **redéfinitions** sont du même « type » (*i.e.* partagent les mêmes propriétés) que les connaissances qui les composent. Par exemple, `souffle-air` est une connaissance statique, ayant comme propriétés une *position* et une *amplitude*, de même que les connaissances qui la composent (`bouche-ouverte` et `joues-gonflées`).

Quelle amplitude donner à `souffle-air` quand `bouche-ouverte` et `joues-gonflées` sont présentes dans la base ? L'amplitude de la nouvelle connaissance sera composée à partir des amplitudes des connaissances composantes. On peut ici utiliser plusieurs opérateurs : faire la moyenne des amplitudes, le minimum, le maximum, ou autre. On choisira généralement d'utiliser le maximum.

La propriété d'*amplitude* des **actions faciales** correspond au maximum d'ampltiude de la connaissance élémentaire qui les compose. Par exemple, l'amplitude de l'action faciale `af-tremblement-joues` correspond au maximum d'amplitude de chaque gonflement de joue qui compose l'action faciale.

De même, l'amplitude d'une expression correspond au maximum d'amplitude des connaissances qui la compose.

La propriété de *durée* pour les actions faciales et les expressions est calculée comme la somme des durées de chaque connaissance qui les compose.

6

Conclusion

Le but de cette étude était de mettre en évidence les spécificités de l'analyse des expressions du visage dans un contexte de communication Homme-Homme, et en particulier dans le contexte de la Langue des Signes, riche en expressions.

La Langue des Signes apporte son lot de problèmes quant à l'analyse informatique des expressions : problème de changement de pose, d'occultation et de représentation des connaissances linguistiques. C'est ce dernier point, la représentation des connaissances, qui a été traité ici.

Le formalisme informatique présenté a été choisi parce que proche du formalisme utilisé par les linguistes. La difficulté réside dans le fait que les descriptions existantes sont informelles et relativement subjectives. Un formalisme informatique permet de léver les différentes ambiguités d'interprétation en se référant à des indices objectifs : l'état des muscles faciaux.

De ce formalisme informatique est proposé l'architecture d'un système permettant l'analyse informatique des expressions d'un locuteur en Langue des Signes. L'accent est mis sur le mécanisme d'analyse et en particulier sur le mécanisme de prédiction et vérification qui permettent de placer le système à un niveau de compréhension proche de l'humain et de décoder les différentes expressions à moindre coût.

La prochaine étape consiste à valider ou invalider ce formalisme dans un système d'analyse sur des cas réels, avec la participation de linguistes ; d'abord en conditions « maitrisées », en mettant de côté les problèmes de pose et d'occultation ; puis en condition réelle, en traitant ces problèmes.

Quelques opérateurs de Traitement d'Images ont été présentés dans l'état de l'art. Il reste à choisir un ensemble d'opérateurs, existants ou à créer, qui permettront d'extraire suffisamment de connaissances des séquences vidéo, et compatibles avec le mécanisme d'analyse bi-directionnelle présenté.

Enfin, à terme, il reste à étudier la manière dont intéragit un tel système d'analyse des expressions avec un système global d'aide à la compréhension de la Langue des Signes.

Annexe A

Méthodes d'analyse du visage

Ce chapitre présente quelques méthodes utiliées pour l'analyse automatique du visage. La première section se focalise sur les méthodes de détection des zones candidates du visage. C'est sur ces zones que l'on appliquera ensuite une méthode de détection du visage proprement dite.

Les deux sections suivantes présentent des méthodes d'analyse du visage valables aussi bien pour la détection du visage que pour l'analyse des expressions. Les méthodes sont classées en deux approches distinctes : l'approche image et l'approche par composantes.

A.1 Détection des zones candidates du visage

Détecter un visage implique généralement de tester chaque région et sous-région de l'image en vérifiant si elle contient un visage. C'est ce balayage qui est le plus critique en temps de calcul. C'est pourquoi la détection du visage s'effectue en deux étapes : recherche de zones susceptibles de contenir un visage par des méthodes rapides, mais assez peu précises et recherche dans ces zones particulières la présence de visages par une analyse plus fine.

Trouver les zones candidates du visage consiste à trouver des régions de l'image ayant une certaine propriété partagée généralement par un visage. Les propriétés généralement utilisées sont la couleur de la peau, la forme et le mouvement.

Modèle de couleur de la peau

Une des méthodes de détection de zones candidates consiste à détecter les zones de la peau. Cette méthode nécessite donc une modélisation de la couleur de la peau.

La méthode la plus simple ([1]) consiste à représenter les couleurs dans le modèle YCrCb, puisque les différences de couleur de peau entre les différents individus (et ce, quelque soit le type de pigmentation de la peau) sont dûes à une différence de

luminance plutôt qu'une différence de chrominance. Les pixels ayant des valeurs de *Cr* et *Cb* comprises dans un certain intervalle ([RCr1, RCr2] et [RCb1, RCb2] respectivement) sont sélectionnés comme pixels faisant partie de la peau.

Avantages : Cette méthode est rapide à exécuter. La décision est effectuée directement sur chaque pixel.

Inconvénients : Le nombre de fausses alarmes est assez grand, d'abord parce que l'unique information de couleur n'est parfois pas assez discriminante et parce que le modèle est construit *a priori* .

Deux autres méthodes couramment utilisées consistent à construire le modèle de la peau à partir d'exemples.

La classification de la peau basée sur un histogramme consiste à modéliser l'histogramme type de la peau à partir d'exemple. Les pixels appartenant à la peau sont étiquetés sur chaque image. A chaque étiquetage, le probabilité que cette couleur fasse partie de la peau est incrémentée. On construit ainsi un histogramme (qui peut être vu comme une distribution de probabilité) de la couleur de la peau. Après apprentissage, un pixel est considéré comme faisant partie de la classe « peau » si la valeur de l'histogramme modèle de la couleur considérée dépasse un certain seuil. Il est possible d'améliorer cette méthode en prenant en compte une modélisation de la classe « non-peau ». Après apprentissage, un pixel appartient à la classe « peau » si le rapport entre la valeur de l'histogramme « peau » et la valeur de l'histogramme « non-peau » dépasse un certain seuil.

La classification d'un pixel dans une des deux classes (« peau » ou « non-peau ») peut être vue comme une mesure de probabilité qu'un pixel appartienne à l'une ou l'autre classe. On modélise généralement la distribution de probabilités qu'un pixel appartienne à la peau comme un mélange de lois normales multi-variées (généralement de dimension 3 pour la couleur). Lors de la phase d'apprentissage, chaque nouvelle image permet de préciser le nombre et les paramètres de chaque gaussienne (moyenne et variance). Un algorithme classique pour l'adapation d'un mélange de gaussiennes donné est l'algorithme *Expectation Maximization* ([48]).

Jones et Rehg ([3]) ont mené une étude comparative des deux méthodes. Ils ont construit les deux modèles à partir d'un ensemble d'images séléctionnées de manière automatique sur le Web (6822 photos au total). Leurs tests montrent que la modélisation par histogramme est plus efficace et plus rapide (à construire et à évaluer) que la modélisation par mélanges de lois gaussiennes.

Leur meilleur histogramme (taille 32) possède une aire de 0.942 sur la courbe de ROC (mesure statistique indiquant le rapport entre réussite et « fausses alarmes »), alors qu'un modèle idéal a une aire de 1. Ce détecteur permet de classer correctement à 80% avec 8.5% de fausses alarmes ou à 90% avec 14.2% de fausses alarmes (le taux de réussite par rapport au taux de fausses alarmes est paramétrable).

Avantages : Comme pour la méthode précédente, puisque la décision est effectuée

au niveau du pixel, l'exécution est rapide. Après l'apprentissage, le modèle peut être réutilisé à volonté.

Inconvénients : Le nombre de fausses alarmes, bien que moins elevé que la méthode précédente, reste relativement élevé. De plus, le modèle nécessite un grand nombre d'exemples et est donc difficile à construire.

Composantes connexes Décider si un pixel appartient ou non à la peau ne suffit généralement pas. Avec une telle décision, un grand nombre de pixels isolés de l'arrière plan est selectionné. Le but est alors de ne retenir que les ensembles de pixels constituant une forme.

L'algorithme le plus souvent utilisé consiste à ne garder des pixels précédemment sélectionnés que s'ils sont entourés d'autres pixels sélectionnés. Cette méthode permet de supprimer les pixels isolés de l'arrière-plan.

Si on cherche à connaître l'enveloppe du visage, il est aussi possible de faire suivre cette composition par une transformation par morphologie mathématique. Ce qui permet, par exemple, de combler les « trous » de la forme.

Analyse de la forme

Une approche pour détecter les zones candidates consiste à sélectionner les objets ayant une forme proche de celle du visage.

La forme la plus simple du visage est l'ellipse (si le visage est vu de face). La détection par critère de forme consiste généralement à trouver les contours de l'image et à vérifier qu'ils sont organisées géométriquement selon un certain modèle (oval, ellipse). Le problème est alors de trouver la position optimale de l'ellipse dans l'image. On cherche donc à maximiser une certaine mesure d'« adéquation » entre l'ellipse et l'image. On peut, par exemple, utiliser la somme des gradients entourant l'ellipse comme mesure d'adéquation. Un des problèmes est que cette méthode est peu efficace lorsque l'arrière-plan est relativement complexe. On peut ensuite ajouter une mesure de corrélation sur l'histogramme de la partie interne de l'ellipse entre la position initiale et la position courante.

Pour savoir si une région de l'image a une certaine forme, on mesure généralement différentes caractéristiques.

Une des caractéristiques souvent utilisée est la projection. On projette verticalement et horizontalement les pixels de la forme. Les deux profils obtenus pourront être comparés à la forme désirée.

Pantic ([37]) utilise cette méthode pour la détection du visage et de ses composantes. Les pics du profil vertical donnent les positions de la frontière entre cheveux et front, des yeux, des narines, de la bouche et de la frontière entre menton et cou. La ligne horizontale des yeux est le maximum local du deuxième pic.

A partir de cette ligne, la ligne verticale coupant le visage en deux et passant par la zone entre les deux yeux est donnée par le minimum des différences de contraste sur la ligne horizontale des yeux extraite précédemment.

Le visage ainsi découpé en quatre zones permet de définir les zones des yeux, de la bouche et du nez. La zone de l'oeil gauche est définie initialement comme ayant les mêmes dimensions que le coin supérieur gauche du visage et étant coupé en deux verticalement par la ligne horizontale traversant les yeux. Cette zone est réduite en ne gardant que les maxima locaux. La même analyse est faite pour l'oeil droit, la bouche et le nez.

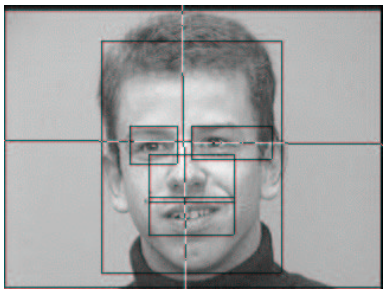


FIG. A.1 – Découpage du visage par analyse du profil horizontal et vertical

Avantages : Le principal avantage de cette méthode de découpage du visage en boites englobantes est sa rapidité d'exécution.
Un autre avantage est que les résultats sont indépendants de la taille du visage observé.

Inconvénients : L'inconvénient majeur est que cette méthode est très contrainte. Les décisions étant prises par rapport à des critères de luminosité, les résultats sont grandement faussés si la luminosité n'est pas idéale. Une information de luminosité bruitée peut aussi beaucoup fausser la décision. Cette méthode nécessite donc une phase de prétraitement.
De plus, cette méthode suppose que le visage soit vu de face.

Analyse du mouvement

Si l'objet d'étude est le visage d'une personne filmé par une caméra fixe, l'arrière-plan est généralement fixe. Ainsi, détecter les zones de l'image en mouvement revient à détecter les zones de l'image susceptibles d'être un visage.

Une simple analyse par différences d'images permet de sélectionner les zones en mouvements.

A.2 Analyse par extraction des composantes

La première approche pour l'analyse du visage consiste à considérer le visage comme un ensemble de composantes ayant une certaine configuration spatiale. L'analyse par cette approche consiste alors à détecter les composantes (en utilisant des détecteurs spécialisés) et à les réorganiser spatialement pour les faire correspondre à un modèle du visage.

La section précédente présentait des méthodes de sélection de zones candidates du visage. Ces méthodes peuvent aussi servir de pré-analyse pour la détection de composantes. Par exemple, modéliser la couleur des lèvres, qui est un cas particulier du modèle de la peau, donne un ensemble de régions des lèvres candidates.

On présente dans cette section des méthodes spécialisées dans le détection et la mesure des différentes composantes du visage. Ces méthodes sont généralement une combinaison judicieuse des méthodes générales.

Sourcils

Les sourcils, s'ils sont présents, sont généralement plus foncés que le reste du visage. Un détecteur de contours peut permettre la détection des sourcils.

Yeux

La zone des yeux est composée de plusieurs composantes intéressantes pour l'analyse des expressions : les paupières, les yeux (iris, blanc), les muscles entourant les yeux qui entrent en jeu dans l'action faciale « plissement des yeux » et les rides sur le côté.

paupières Les paupières ont la couleur de la peau. Détecter une zone ayant la couleur de la peau à l'endroit des yeux indique la présence des paupières.

Le clignement d'yeux est un processus naturel qui intervient relativement fréquemment. Il est possible de détecter les yeux en se basant sur le principe qu'ils clignent avec une certaine fréquence et tous les deux en même temps.

iris, blanc des yeux Les yeux possède un profil horizontal de couleur très spécifique : clair, foncé et clair. Une analyse du gradient ou des histogrammes verticaux / horizontaux peut permettre de détecter les caractéristiques des yeux.

Le forme ronde de l'oeil peut aussi servir à la détection, par exemple, par une transformée de Hough.

Une première méthode basée sur la luminosité consiste à dire que l'iris est la partie la plus sombre de la zone de l'oeil. On peut ainsi faire varier un seuil de binarisation jusqu'à obtenir deux formes distinctes. En se basant sur des critères morphométriques, on peut donc détecter la position des yeux.

Christophe Collet ([18]) propose une méthode pour détecter les yeux (et leur configuration). L'analyse est basée sur le gradient. Les yeux ayant un profil de luminosité horizontal spécifique, il tente de « reconnaître » ce profil dans différentes zones de l'image. La zone des yeux est celle ayant le plus de correspondances.

Bouche

La zone de la bouche est une des zones du visage la plus mobile. La mâchoire peut ainsi être abaissée ou décalée à gauche ou à droite. Les éléments les plus importants de cette zone sont les lèvres. Les autres éléments sont les dents, la langue et les rides naso-labiales.

Les techniques de contours actifs sont souvent utilisées pour l'extraction de la forme de la bouche. Le principal problème des contours actifs est leur difficulté d'initialisation. Delmas ([20]) propose une méthode de détection des commissures de la bouche pour initialiser un *snake*. L'idée est que sur une image de la partie inférieure du visage, les zones les plus sombres de l'image correspondent aux commissures de la bouche, qu'elle soit ouverte ou fermée.

Le principe est alors de determiner, pour chaque colonne de l'image, le minimum de luminance. Afin de tenir compte de l'aspect symétrique et du centrage horizontal de la bouche, on introduit une fonction de pondération (semblable à une gaussienne) favorisant les minima proches du centre de l'image plutôt que ceux situés sur les bords, *a priori* en dehors de la bouche. On construit alors un vecteur d'accumulation V_{roi}, somme des projections pondérées des minima précédemment détectés. La composante la plus forte de ce vecteur donne alors la position verticale de la bouche. Pour trouver les commissures, on effectue alors un chaînage des minima de luminance.

Une fois les deux lèvres repérées, il est plus facile de différencier la configuration de l'intérieure de la bouche : la présence des dents ou de la langue peut être détecté par l'analyse de la couleur, puisque les deux ont des couleurs très différentes.

Une autre technique consiste à suivre la position du menton. Etant donné que le mouvement du menton entraîne le mouvement de la lèvre inférieure, on peut déduire la position de la lèvre inférieure de la position du menton.

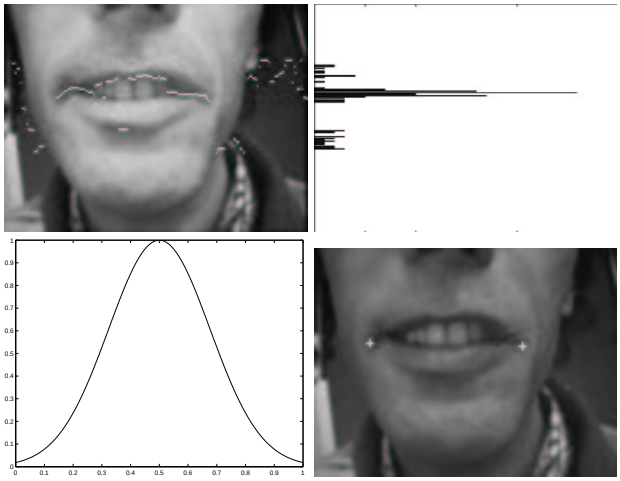


FIG. A.2 – Détection des commissures de la bouche. En haut à gauche : minima de luminance par colonne. En bas à gauche : fonction de pondération. En haut à droite : vecteur Vroi. En bas à droite : commissures

Rides

En plus de la configuration des différentes composantes (yeux, sourcils, bouche, etc.) du visage, il est important de pouvoir caractériser les « rides d'expression ». Les rides correspondent à au relâchement de la peau. Ce relâchement peut être du à l'« usure » de la peau ou à l'activation d'un muscle. Il est donc nécessaire, non seulement de *détecter* la présence ou l'absence de rides, mais aussi de les *quantifier* ceci pour distinguer les rides « permanentes » de celles engendrées par les muscles.

Les rides les plus importantes du visage sont les rides du haut du nez (froncement du nez), du front (relèvement des sourcils), du coin des yeux (plissement des yeux) et les rides nasio-labiales qui interviennent généralement lors du sourire.

Les rides apparaissent sur les images sous forme d'une forte différence de luminosité. Utiliser un détecteur de contours (détectant aussi son orientation) sur des zones pertinentes de l'image (front, nez, coin des yeux, ...) permet d'extraire l'information sur les rides.

A.2.1 Evaluation

L'avantage de ces méthodes est qu'elles sont généralement très simple à metter en oeuvre.

Malheureusement, puisque les modèles sous-jacents sont construits *a priori* , ces méthodes sont généralement spécifiques à une configuration donnée (visage vu de face, luminosité constante, etc.). Il devient difficile de les généraliser (rotation du

visage, etc.). L'approche couramment utilisée consiste à préceder ces méthodes de phases de prétraitement (reconstruction 3D pour les rotations, égalisation d'histogramme pour la luminosté, etc.) pour que l'observation soit présentée conformément aux hypothèses d'application des méthodes.

A.3 Mise en correspondance de modèles

La deuxième approche pour l'analyse du visage consiste à voir le visage comme un tout. L'analyse consiste alors à mesurer la ressemblance du visage observé à un modèle (connu ou appris). Les méthodes de cette approche font généralement appel à des méthodes de mise en correspondance de modèles.

L'interêt de ces méthodes est qu'elles peuvent être appliquées de manière plus locale. Ainsi, la plupart des méthodes présentées dans cette section, valable pour le visage, sont généralement aussi valable pour n'importe quel objet et plus particulièrement pour les composantes du visage.

A.3.1 Analyse en composantes principales

On présente ici l'analyse en composantes principales et plus particulièrement la décomposition en « visages-propres » (*eigen-faces*) ([6]).

A partir d'un ensemble d'images de visages caractéristiques (généralement de taille fixée – nécessitant éventuellement un redimensionnement de l'image – et en niveaux de gris), considérées comme un vecteur (en concaténant chaque ligne de l'image) :

$$E = \{\Gamma_1, \Gamma_2, ..., \Gamma_n\}$$

on construit un visage « moyen » :

$$\bar{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i$$

On construit ensuite la matrice suivante :

$$X = [\Gamma_1 - \bar{\Gamma} \quad \Gamma_2 - \bar{\Gamma} \quad ... \quad \Gamma_n - \bar{\Gamma}]$$

La matrice de covariance est X multipliée par sa transposée :

$$L = XX^t$$

Pour trouver les valeurs propres, on résoud l'équation $det(L - \lambda I) = O$ où I est la matrice identité.

On a $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$

Les vecteurs propres sont alors :

$$\lambda_i v_i = C v_i$$

La nouvelle image Φ est projetée dans l'espace des visages propres : $P(\Phi) = [v_1 \quad v_2 \quad ... \quad v_k]^t(\Phi - \bar{\Gamma})$, où $k \leq n$

Turk et Pentland ([6]) utilisent cette méthode pour la reconnaissance des visages. Les résultats de reconnaissance sont très bons pour un visage faisant partie de la base d'apprentissage (même en présence d'une forte occultation). La reconnaissance d'un nouveau visage (*i.e.* , ne faisant pas partie de la base d'apprentissage) nécessite la construction d'une base d'apprentissage la plus représentative possible.

Pour que cette méthode soit robuste aux changements de pose et/ou d'illumination, il faut intégrer dans la base d'apprentissage des visages ayant différentes poses et/ou illuminations. La construction de la base d'apprentissage devient alors difficile.

Avantages : Robustesse à l'occulation pour des visages de la base d'apprentissage.

Inconvénients : La construction du corpus d'apprentissage est critique.
De plus, la taille de la matrice d'apprentissage est fixée et les images à décomposer doivent être éventuellement redimensionnées.

Cette méthode de décomposition, généralement utilisée pour le problème de *reconnaissance de visage*, peut être utilisée pour analyse certaines caractéristiques du visage. En effet, si la base d'apprentissage contient suffisamment de différences en pose, illumination et/ou expression, les principaux « modes » de la décomposition peuvent correspondre à ces différences. Ainsi, Taylor et Cootes ont observés que faire varier un visage autour du premier axe, changeait son illumination ; le deuxième axe correspondrait aux différences de pilosité (barbre et sourcils) et le troisième axe au sourire.

Le concept de visage-propre peut être étendu au concept de composantes-propres (*eigen-features*). Plutôt que d'effectuer une décomposition en valeurs propres de l'image du visage, on effectue la décomposition sur les images des différentes composantes.

Taylor et Cootes ([29]) ont introduit le concept d'*Active Shape Model* et d'*Active Appearance Model* qui consiste à modéliser le visage en prenant en compte à la fois les informations de forme et les informations d'apparence. Un ensemble de points de contrôle est placé manuellement sur un ensemble de visages d'apprentissage. De ces points, on déduit un arrangement spatial et on mémorise l'information de couleur (ou de niveaux de gris) de cette forme. En effectuant une analyse en composantes principales sur les données d'apprentissage (aussi bien sur les informations de forme que d'apparence), on peut ainsi recomposer un visage.

Avantages : Robustesse à l'occulation, même sur des visages non-connus (*i.e.* hors de la base d'apprentissage).

Inconvénients : La construction du modèle est très longue : l'extraction intiale de la forme des visages est effectuée manuellement.

Ahlberg ([10]) utilise une méthode inspirée des *Active Shape Models* à la différence près qu'il utilise un modèle de visage « générique » et non construit à partir des données (Candide). Ainsi, le système d'Ahlberg possède une bonne robustesse à l'occultation sans que le modèle soit difficile à construire.

A.3.2 Apprentissage par réseaux de neurones

Un réseau de neurones peut être vu comme une fonction ayant un certain nombre d'entrées et un certain nombre de sorties. Le principe de l'apprentissage est de donner en entrée au réseau un certain nombre d'exemples et de fixer la sortie à la valeur désirée. Une méthode d'apprentissage permet alors au neurone de s'adapter au mieux pour qu'il affiche la même sortie quand on lui donnera des donnés *proches* des données d'apprentissage. L'un des avantages des réseaux de neurones est leur robustese au bruit.

Cottrell et Padgett ([44]) ont cherché des méthodes d'analyse automatique du visage les plus proches possibles de la réalité biologique. Ainsi, un réseau de neurones dit « auto-supervisé » , c'est à dire dont la couche d'entrée et la couche de sortie sont identiques et égales à l'image d'un visage, effectue une analyse en composantes principales. Le nombre de composantes principales est donné par le nombre de neurones de la couche cachée. Chaque neurone de la couche cachée correspond aux valeurs propres de la décomposition.

Rowley, Baluja et Kanade ([5]) ont construit un réseau de neurones qui, à partir d'une image prétraitée de 20x20 pixels indique s'il s'agit d'un visage ou non. Le prétraitement consiste à égaliser l'histogramme. L'image est balayée en fenêtres de 20x20. Pour détecter les visages de différentes tailles, une analyse multi-résolutions est effectuée. L'extension du système consistait à ajouter un réseau de neurones indiquant le degré de rotation d'un visage. Ainsi, le système est capable de détecter des visages ayant subi des rotations dans le plan et de différentes échelles.

Les réseaux de neurones peuvent aussi servir pour la reconnaissance des expressions faciales. Par exemple, la reconnaissance des configurations de la bouche peut être détectée par un réseau de neurones. Le réseau aura alors été entraîné sur un ensemble d'images de bouches ayant des configurations différentes ([37]).

Avantages : Les réseaux de neurones sont généralement utilisés pour leur faible sensibilité au bruit et leur capacité d'apprentissage.

Inconvénients : Malheureusement, les réseaux de neurones, sont souvent difficile à construire. Leur structure (nombre de couches cachées pour les perceptrons

par exemple) influe beaucoup sur les résultats et il n'existe pas de méthode pour déterminer automatiquement cette structure.
La phase d'apprentissage est difficile à mener puisque les exemples doivent être correctement choisis (en nombre et configuration).

A.3.3 Modèle statistique

L'approche statistique consiste à trouver quelle est la probabilité qu'un échantillon observé fasse partie du modèle. Le modèle à construire est la distribution de probabilité.

La distribution de probabilité peut être apprise à partir des données. On considère généralement que la distribution de probabilité est un mélange de lois gaussiennes dont le nombre et les paramètres (variance et moyenne) doivent être estimés. L'algorithme utilisé dans la plupart des cas pour cet apprentissage est l'algorithme d'*Expectation Maximization* ([48]).

Les modèles statistiques peuvent servir à modéliser le visage, ses composantes ou certaines de ses caractéristiques. Au plus bas niveau, une distribution de probabilité qu'un pixel fasse partie du visage peut être donnée par une information sur la couleur. Cette information seule n'étant généralement pas assez discriminante, on peut ajouter des informations spatiales (comme la notion de *blob* dans [34]).

A.3.4 Evaluation

Les avantages principaux de ces méthodes sont qu'elles sont génériques et peuvent donc s'adapter à beaucoup de problèmes. Ces méthodes sont généralement moins sensibles au bruit que les méthodes classiques, puisqu'il existe un modèle sous-jacent de comparaison.

Par contre, la construction du modèle (*i.e.* l'apprentissage) est souvent long et nécessite un corpus « intelligent » (*i.e.* adapté au problème).

Du point de vue de l'analyse des expressions du visage, ces méthodes souffrent d'un autre inconvénient : elles ne donnent que des « configurations » et non des mesures. Elles peuvent être vues comme des méthodes de classification, *i.e.* des méthodes qui indiquent que l'observation se trouve dans un ensemble préétabli de configurations.

Les méthodes de mise en correspondance de modèles sont mal adaptées au problème de quantification : il est, par exemple, difficile de concevoir un réseau de neurones qui indique quelle est le degré d'ouverture, en pixels, de la bouche ; il est plus facile de construire un réseau de neurones qui décide si la bouche est ouverte ou fermée.

Ces méthodes sont donc bien adaptées à l'extraction d'informations sur des com-

posantes dont certains états sont difficile à quantifier par des opérateurs classiques (le gonflement de la joue par exemple) ou à l'extraction d'informations sur des composantes n'ayant qu'un nombre restreint d'états.

A.4 Autres méthodes.

La décompostion des méthodes en deux approches (par composantes et globale) est relativement naíve. Il reste un certain nombre de méthodes qui exploitent les avantages des deux approches présentées.

D'une manière générale, toutes les méthodes globales, peuvent être appliquées de manière locale sur des composantes du visage en particulier, puisqu'elles ont été conçues dans un but de généricité.

A.4.1 Composantes propres

La technique des composantes-propres est une technique voisine de la décomposition en visages-propres. Plutôt que de décomposer la globalité du visage en sous-espaces, on décompose ici une certaine composante du visage. On parle alors de décomposition en yeux-propres, bouches-propres, etc.

Cette méthode peut être utilisée pour extraire des informations qu'il est difficile d'extraire avec des opérateurs classiques. Il est par exemple difficile d'extraire le degré de gonflement des joues à partir d'informations de bas niveau (couleur, gradient, etc.). Par contre, la décomposition en joues-propres permet de distinguer les différentes états de la joue : gonflée, neutre et creusée par exemple.

L'inconvénient de cette méthode est le même que pour la méthode de décomposition en visages-propres, à savoir que le modèle est très long à construire.

A.4.2 Contours actifs

Les méthodes de contours actifs sont des méthodes qui peuvent être utilisées aussi bien au niveau global que local (bien que pour le visage, elles soient plus souvent utilisées de manière locale, pour une composante particulière). Elles se basent sur les contours de l'image.

Le but est de faire évoluer une forme. Cette forme possède un certain nombre de contraintes. Les courbes sont définies comme suit pour les modèles continus : ([26]).

$$C = \{v(s,t) = (x(s,t), y(s,t)); s \in [a,b] \text{ et } t \in [0,T]\}$$

où a et b sont les exterémités du contour. Le déplacement de la courbe est effectuée de sorte à minimiser l'énergie de la courbe. Le calcul de l'énergie est découpé en trois :

1. L'**énergie interne** permet de définir la rigidité, la longueur, la courbure et l'élasticité de la courbe.

$$E_{interne}(C) = \int_a^b \alpha(r) \left| \frac{\delta v(s)}{\delta s} \right| + \int_q^b \beta(r) \left| \frac{\delta^2 v(s)}{\delta s^2} \right|^2$$

   la dérivée première agit sur la rigidité et la dérivée seconde sur l'élasticité. Les coefficients $\alpha(s)$ et $\beta(s)$ permettent de pondérer ces grandeurs.

2. L'**énergie externe** permet de définir des points d'attraction ou de répulsion qui influencent le déplacement de la courbe.

3. L'**énergie image**, calculée à partir du gradient de l'image :

$$E_{image}(C) = -\int_a^b |\nabla^* I(v(s))|^2 \, ds$$

   où $\nabla^* I(v(s))$ représente le gradient de l'image au voisinage de la courbe $v(s)$.

Les snakes sont très utilisés pour l'extraction et le suivi de composantes. Certains l'utilisent pour extraire et suivre les yeux, d'autres pour la bouche ([37]).

Avantages : L'avantage des contours actifs est qu'ils peuvent « modéliser » des formes relativement complexes, généralement « arrondies » . Ces formes sont très présentes dans le corps humain et particulièrement sur le visage. Il est possible de mettre en correspondance le l'élasticité et la rigidité du contour actif et des muscles du visage qu'il « suit ».

Inconvénients : L'inconvénient majeur des contours actifs est l'initialisation. La méthode des contours actifs étant basée sur le gradient de l'image (une information locale), le choix de la position et des paramètres initiaux est cruciale pour avoir de bons résultats. Le contour est donc intialisé par une autre méthode (découpage en boîte englobante par exemple).

A.5 Estimation de la dynamique

Une fois les composants du visage extraits, la mesure de leur dynamique permet de caractériser les expressions.

La méthode dite du « flux optique » permet une estimation du mouvement sous certaines contraintes (généralement contrainte d'illumination constante). Il existe

de nombreuses méthodes de flux optique (Horn et Schunck [45], Simoncelli [51], Lucas et Kanade [47]). Leur application est globale ou locale.

Une méthode de flux optique appliquée à l'ensemble de l'image permet une très bonne estimation des mouvements (même légers) et est utilisée généralement pour ajouter des contraintes à un modèle pré-établi (dans [24] par exemple où l'estimation du mouvement global permet de préciser le modèle musculaire). Cependant, l'application globale requiert une grande complexité de calcul.

Appliquer la méthode du flux optique à des regions locales de l'image permet de limiter les temps de calculs tout en obtenant une estimation correcte de la dynamique. De plus, les vecteurs vitesse peuvent servir à la prédiction de la future position d'une composante ([13]).

Les méthodes par différence d'images consistent à mesurer la différence entre l'image courante et une image particulière, une image « neutre » pour le problème considéré. Par exemple, sous l'hypothèse que la première image d'une séquence représente un visage affichant l'expression « neutre », les mouvements des différentes composantes pourront être mis en évidence sur chaque image suivante par différence avec la première image.

Bibliographie

Détection du visage

[1] Douglas CHAI and King N. NGAN. « Locating Facial Region of a Head-and-Shoulders Color Image ». In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition (FG'98)*, pages 124–129, Nara, Japan, April 1998.

[2] Erik HJELMAS and Boon Kee LOW. « Face Detection : A Survey ». *Computer Vision and Image Understanding*, 83(3) :236–274, 2001.

[3] Michael J. JONES and James M. REHG. « Statistical Color Models with Application to Skin Detection ». In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 1274, Fort Collins, Colorado, June 1999.

[4] Alexandre LEMIEUX and Marc PARIZEAU. « Experiments on Eigenfaces Robustness ». In *International Conference on Pattern Recognition (ICPR)*, volume 1, page 10421, Québec, 2002.

[5] Henry A. ROWLEY, Shumeet BALUJA and Takeo KANADE. « Rotation Invariant Neural Network-Based Face Detection ». In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, page 38, Santa Barbara, CA, June 1998.

[6] Matthew TURK and Alex PENTLAND. « Eigenfaces for recognition ». *Journal of Cognitive Neuroscience*, 3(1) :71–86, Winter 1991.

[7] Ming-Hsuan YANG, David J. KRIEGMAN and Narenda AHUJA. « Detecting Faces in Images : A Survey ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1) :34–58, January 2002.

[8] ZHAO, R. CHELLAPPA, A. ROSENFELD and P. PHILLIPS. « Face recognition : A literature survey ». Technical Report CAR-TR-948, Center for Automation Research, University of Maryland, 2000.

Analyse des expressions

[9] Jörgen AHLBERG. « CANDIDE-3 - un updated parameterised face ». Technical Report, Dept. of Electrical Engineering, Linköping University, Sweden, January 2001.

[10] Jórgen AHLBERG. « Real-Time Facial Feature Tracking using an Active Model with Fast Image Warping ». In *International Workshop on Very Low Bitrate Video (VLBV)*, pages 39–43, Athens, Greece, 2001.

[11] Marian Stewart BARTLETT, Joseph C. HAGER, Paul EKMAN and Terence J. SEJNOWSKI. « Measuring Facial Expressions by Computer Image Analysis ». *Psychophysiology*, 36 :253–263, 1999.

[12] Marian Stewart BARTLETT, Gwen LITTLEWORT, Bjorn BRAATHEN, Evan SMITH and Javier R. MOVELLAN. « An Approach to Automatic Analysis of Spontaneous Facial Expressions ». In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)*, page 360, Washington, D.C, 2002.

[13] Michael J. BLACK and Yaser YACOOB. « Tracking and Recognizing Facial Expressions in Image Sequences, Using Local Parameterized Models of Image Motion ». Technical Report CS-TR-3401, University of Maryland, January 1995.

[14] N. Badler C. PELACHAUD and M. VIAUD. « Final Report to NSF of the Standards of Facial Animation Workshop (http ://hms.upenn.edu/pelachaud/workshop_face/workshop_face.html) ».

[15] Ira COHEN, Nicu SEBE, Larry CHEN, Ashutosh GARG and Thomas S. HUANG. « Facial Expression Recognition from Video Sequences : Temporal and Static Modelling ». *Computer Vision and Image Understanding, in Special Issue on Face Recognition, a paraître*, 2003.

[16] Jeffrey F. COHN, Karen SCHMIDT, Ralph GROSS and Paul EKMAN. « Individual Differences in Facial Expression : Stability over Time, Relation to Self-Reported Emotion, and Ability to Inform Person Identification ». In *IEEE International Conference on Multimodal Interfaces (ICMI)*, page 48, Pittsburgh, USA, 2002.

[17] Jeffrey F. COHN, A. J. ZLOCHOWER, J. LIEN and Takeo KANADE. « Automated Face Analysis by Feature Point Tracking has High Concurrent Validity with Manual Faces Coding ». *Psychophysiology*, 36 :35–43, 1999.

[18] Christophe COLLET. « CapRe : un système de capture du regard dans un contexte d'Intéraction Homme-Machine », 1998. Thèse de doctorat, LIMSI.

[19] Christian CUXAC. *La Langue des Signes Française - Les voies de l'iconicité*. Ophrys, 2000.

[20] Patrice DELMAS. « Extraction des contours de lèvres d'un visage parlant par contours actifs - Application à la communication multimodale. », Avril 2000. Thèse de doctorat, INPG, Grenoble.

[21] Paul EKMAN and W. V. FRIESEN. *Unmasking the Face*. New Jersey : Pentice Hall, 1975.

[22] Paul EKMAN and W. V. FRIESEN. *Facial Action Coding System (FACS) : Manual*. Palo Alto : Consulting Psychologists Press, 1978.

[23] Irfan ESSA. « Visual Coding and Tracking of Speech Related Facial Motion ». Technical Report GIT-GVU-TR-01-16, Georgia Institute of Technology, 2001.

[24] Irfan ESSA and Alex PENTLAND. « Coding, Analysis, Interpretation, and Recognition of Facial Expressions ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :757–763, 1997.

[25] Beat FASEL and Juergen LUETTIN. « Automatic Facial Expression Analysis : A Survey ». *Pattern Recognition*, 36(1) :259–275, 2003.

[26] Sébastien GALVAGNO. « Suivi du geste de la Langue des Signes ». DEA IIL 1998-1999, Université Paul Sabatier, Toulouse.

[27] Taro GOTO, Sumedha KSHIRSAGAR and Nadia MAGNENAT-THALMANN. « Real Time Facial Tracking and Speech Acquisition for Cloned Head ». In *International Workshop on Very Low Bitrate Video Coding (VLBV01)*, Athenes, Grece, 2001.

[28] Guillemette JAUSIONS. « Analyse de traitement de l'image des expressions du visage d'un locuteur en LSF ». DEA 2IL 2001-2002, Université Paul Sabatier, Toulouse.

[29] Andreas LANITIS, Chris J. TAYLOR and Timothy F. COOTES. « Automatic Interpretation and Coding of Face Images Using Flexible Models ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :743–756, July 1997.

[30] Ying li TIAN, Takeo KANADE and Jeffrey F. COHN. « Recognizing Action Units for Facial Expression Analysis ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2) :97–115, 2001.

[31] Michael LYONS, Shigeru AKAMATSU, Miyuki KAMACHI and Jiro GYOBA. « Coding Facial Expressions with Gabor Wavelets ». In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, Nara, Japan, april 1998.

[32] B. MOGHADDAM and Alex PENTLAND. « Face Recognition using View-based and Modular Eigenspaces ». In *Automatic Systems for the Identification and Inspection of Humans, (SPIE'94)*, volume 2257, San Diego, 1994.

[33] Carol NEIDLE. « SignStream[tm] Annotation : Conventions used for the American Sign Language Linguistic Research Project ». Technical Report 11, Boston University, August 2002.

[34] Nuria OLIVER, Alex PENTLAND and Françcois BÉRARD. « LAFTER : a real-time face and lips tracker with facial expression recognition ». *Pattern Recognition*, 33 :1369–1382, 2000.

[35] T. OTSUKA and J. OHYA. « Spotting segments displaying facial expression from image sequences using HMM ». In *Proceedings of the Third IEEE*

International Conference on Automatic Face and Gesture Recognition, pages 442–447, Nara, Japan, 1998.

[36] Maja PANTIC and Leon J. M. ROTHKRANTZ. « Automatic Analysis of Facial Expressions : The State of the Art ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1424–1445, 2000.

[37] Maja PANTIC and Leon J. M. ROTHKRANTZ. « Expert system for automatic analysis of facial expressions ». *Image and Vision Computing Journal*, 18(11) :881–905, 2000.

[38] Maja PANTIC, Milan TOMC and Leon J. M. ROTHKRANTZ. « A hybrid approach to mouth features detection ». In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2001.

[39] K. L. SCHMIDT and Jeffrey F. COHN. « Dynamics of facial expression : Normative characteristics and Individual differences ». In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 728–731, Tokyo, Japan, August 2001.

[40] D. TERZOPOULOS and K. WATERS. « Analysis and Synthesis of Facial Image Sequences using Physical and Anatomical Models ». In *Proc. Third International Conf. on Computer Vision (ICCV'90)*, pages 727–732, Osaka, Japan, 1990.

[41] Ying-Li TIAN, Takeo KANADE and Jeffrey F. COHN. « Recognizing Lower Face Action Units for Facial Expression Analysis ». In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 484 – 490, Nara, Japon, March 2000.

[42] Alper YILMAZ, Khurram SHAFIQUE and Mubarak SHAH. « Estimation of Rigid and Non-Rigid Facial Motion Using Anatomical Face Model ». In *International Conference on Pattern Recognition (ICPR)*, volume 1, page 10377, Québec, 2002.

[43] Zhengyou ZHANG. « Feature-Based Facial Expression Recognition : Experiments With a Multi-Layer Perceptron ». Technical Report 3354, Institut National de Recherche en Informatique et en Automatique, Février 1998.

Divers

[44] Garrison W. COTTRELL, Mathhew N. DAILEY, Curtis PADGETT and Ralph ADOLPHS. « Is All Face Processing Holistic ? The view from UCSD ». *Chapitre à paraitre dans Computational, Geometric, and Process Perspectives on Facial Cognition : Contexts and Challenges (M. Wenger and J. Townsend)*, 2000.

[45] B. HORN and B. SCHUNCK. « Determining Optical Flow ». *Artificial Intelligence*, 17(1) :185–203, 1981.

[46] Thanarat HORPRASERT, Yaser YACOOB and Larry S. DAVIS. « Computing 3-D Head Orientation from a Monocular Image Sequence ». In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, page 242, Killington, Vermont, USA, October 1996.

[47] B. LUCAS and T. KANADE. « An Iterative Image Registration Technique with an Application to Stereo Vision ». In *Proceedings of the Joint Conference on Artificial Intelligence*, pages 674–680, Vancouver, Canada, 1981.

[48] Thomas MINKA. « Expectation-Maximization as lower bound maximization », 1998. http ://www.stat.cmu.edu/ minka/papers/em.html.

[49] MPEG Working Group on VISUAL. « International Standard on Coding of Audio-Visual Objects, Part 2 (Visual) », 2001. ISO/IEC 14496-2 :2001.

[50] MPEG Working Group on VISUAL. « Reference software for MPEG-4 », 2002. ISO/IEC 14496-5 :2001/Amd 1 :2002.

[51] E. SIMONCELLI. « *Distributed Representation and Analysis of Visual Motion* ». PhD thesis, MIT, 1993.

[52] Y. YACOOB and L. S. DAVIS. « Computing Spatio-Temporal Representations of Human Faces ». In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 70–75, Seatle, Washington, 1994.