Une vue d'ensemble de la reconnaissance de gestes*

Julien Thomet
Département d'informatique
Université de Fribourg
julien.thomet@unifr.ch

Résumé

Le désir de pouvoir interagir avec un ordinateur de manière intuitive et naturelle est grandissant. Dans cette optique, la recherche en reconnaissance de gestes propose de développer des systèmes capables de modéliser, d'analyser et de reconnaître les gestes de l'utilisateur. Nous donnons dans cet article une vue d'ensemble de la recherche actuelle tout d'abord en définissant un certain nombre de termes puis en présentant les problèmes principaux, les méthodes actuelles et des exemples d'applications en reconnaissance de geste. Nous concluons avec quelques réflexions sur les recherches futures.

1. INTRODUCTION

Aujourd'hui, le désir d'interactions avec des machines intelligentes est plus grand que jamais. Dans cette optique, la recheche en reconnaissance de gestes s'est attelée peu à peu au développement de systèmes capables de reconnaître les gestes de l'homme et de les interpréter afin d'enrichir l'expérience utilisateur. Hormis le besoin d'interfaces utilisateur intuitives/naturelles, des applications d'analyse et de surveillance peuvent profiter des méthodes que proposent les scientifiques dans ce domaine. La reconnaissance de gestes est une tâche complexe impliquant divers aspects dont la modélisation et l'analyse de mouvement, la reconnaissance de formes et les méthodes d'apprentissage, voire même des études psycholinguistiques [8].

Cet article présente un aperçu des problèmes, méthodes et applications relatifs au domaine de la recherche en reconnaissance de gestes. La première section (2) concerne la définition et la classification des gestes. La section 3 introduit les termes de « reconnaissance de gestes » et présente la structure générale d'un tel système. Puis sont présentés les applications principales (4) et quelques périphériques de capture (5). Finalement, la section 6 propose quelques visions du futur avant de conclure.

*Séminaire « Gesture recognition », semestre d'automne 2009, Université de Fribourg - `http://diuf.unifr.ch/diva/web/site/index.php/teaching-seminars/10-seminars/125-gesture-recognition`

2. DÉFINITION DU GESTE

Le geste peut à la fois être considéré d'un point vue physiologique (un geste est le résultat de contractions ou détractions musclaires visibles réflexes ou volontaires) ou du point de vue de l'interaction (un geste est une forme de communication non verbale). Dans cet article, nous considérons principalement l'aspect interactif du geste bien qu'ne partie des applications présentées se concentre sur certains mouvement du corps sans but interactif.

L'un des principaux problèmes dans le domaine de la recherche sur les gestes est le manque de termes communément admis pour décrire les interactions. On trouve par exemple dans la littérature pour désigner les gesticulations, les termes « *co-verbal gestures* », pantonimes [3] ou gestes naturels. Les gestes symboliques étant volontier désignés d'iconiques ou de *stroke gestures*. Plusieurs taxonomies sont proposées dans la littérature selon par exemple le type de geste, le domaine d'applications ou encore le type de technologies utilisées. Trois classifications sont présentées dans les paragraphes suivants.

On peut tout d'abord classer les gestes en fonction des parties du corps impliquées. On distingue généralement trois types de gestes :

- les gestes de la main et du bras : ils forment la principale catégorie de gestes interactifs. La main permet de réaliser des gestes précis et complexes. Les recherches autour de ces gestes concernent principalement la reconnaissance de positions de la main, l'interprétation du langage des signes et le développement d'interface homme-machine permettant la manipulation et l'interaction avec des données ou des éléments d'un environnement virtuel.
- les gestes de la tête et du visage : peu de gestes de la tête ont une signification spécifique ; l'orientation de la tête est quant à elle très utile pour la détection du champs de vision. Les recherches dans ce domaine s'intéressent à la reconnaissance faciale comme moyen d'authentification biométrique, comme soutien à d'autres système de reconnaissance tels que la reconnaissance de la parole ; l'analyse des gestes faciaux est également utiles pour la réalisation d'avatars virtuels réalistes ou encore pour décrypter les émotions à des fins marketing.

– les gestes impliquant tout le corps : les recherches dans ce domaine s'intéressent à tout le corps en interaction avec son environnement (analyse des gestes d'un danseur afin de générer de la musique idoine ; analyse des gestes d'un athlète pour améliorer ces performances).

On différencie également les gestes dynamiques des gestes statiques. Un geste statique, également appelé posture, concerne la configuration du corps ou d'une partie du corps à un moment fixe dans le temps alors que le geste dynamique désigne une succession continue de postures.

2.1 Styles de gestes

Dans [2], Karam *et al.* présentent cinq styles de gestes, principalement de la main et du bras, synthétisant les types d'interaction décrits à travers la littérature scientifique :

– les *gestes déictiques* sont des gestes de pointage permettant d'identifier un object ou son emplacement. Ces gestes sont typiquement utilisés dans des environnements virtuels. Ces gestes peuvent être considérés comme implicites dans d'autres formes de gestes (par exempe lorsque l'on pointe un objet à manipuler). Le premier exemple fut le « *Put that there* » de R. A. Bolt [1].
– les *gestes de manipulation* dont le but est de contrôler une entité en appliquant une relation étroite entre le mouvement du geste et l'entité qui est manipulée [2].
– les *gestes sémaphoriques* font partie de tout système gestuel basé sur un catalogue conventionnel de gestes statiques ou dynamiques (par exemple : le geste (statique) pour signifier « ok » ou le signe (dynamique) de la main pour dire « au revoir »).
– les *gesticulations* sont les gestes les plus naturels et désignés de *coverbal gestures* [2]. L'interprétation de ce type de gestes est le domaine de recherche en reconnaissance de geste le plus ambitieux car contrairement aux gestes sémaphoriques, leur signification ne peut pas être issus directement d'un *dictionnaire* de gestes et doivent être mis en relation avec d'autres modalités comme la parole.
– le *langage des signes* : les gestes utilisés dans le langage des signes sont souvent considérés indépendamment des autres types de gestes étant donné qu'ils sont basés sur des principes de linguistique et qu'ils permettent de combiner gestes et signes pour former des structures grammaticales utiles à la conversation.

Comme mentionné auparavant, aucune taxonomie n'est communément admise. Il est à préciser que la dernière classification présentée ne concerne pour ainsi dire que les gestes de la main et du bras.

3. LA RECONNAISSANCE DE GESTE

La reconnaissance de geste désigne l'ensemble des opérations permettant d'analyser une scène à savoir la capture des gestes (par exemple à l'aide d'une caméra ou d'un gant dotés de capteurs), la segmentation, l'évaluation des poses et l'interprétation à proprement parler. La sous-section suivante présente la structure générale d'un système de reconnaissance de geste décrite dans [4][5] ; bien que les auteurs décrivent un processus pour la capture et l'interprétation de gestes corporels amples (par exemple pour différencer un individu qui marche d'un individu qui court), leur analyse met en lumière des problèmatiques valables pour tout type de geste, y compris le gestes manuels et faciaux. Les méthodes décrites sont principalement *vision-based* ; l'utilisation de périphériques de captures tels que des gants sera présentée dans une section suivante.

3.1 Structure générale d'un système

La structure générale d'un système d'analyse des mouvements du corps humain se décompose, selon [4], en quatre processus plus ou moins indépendants et pas forcément présents dans tous les systèmes. Tout d'abord, tout système doit être *initialisé*, c'est-à-dire qu'un modèle adéquat du sujet doit être défini. Ensuite les mouvement du sujet sont suivis (*tracked*) impliquant typiquement une séparation du sujet de l'arrière plan et la recherche de correspondance entre des *frame*s successives. Puis la configuration du corps ou de parties du corps (*pose*) du sujet doit être *évaluée* avant de pouvoir, parfois à l'aide de paramètres supplémentaires, *reconnaître* les actions effectuées par le sujet. Les paragraphes suivants résument les quatre étapes susmentionnées décrites et illustrées dans [4][5].

3.1.1 Initialisation

L'initialisation consiste à s'assurer que le système commence à opérer avec une interprétation correcte de la scène actuelle. Parfois, l'initialisation désigne également le *preprocessing* des données. Pour simplifier, elle s'assure qu'un certain nombre d'hypothèses soient vérifiées au démarrage du système. Moeslund *et al.* [4] ont listé les principales hypothèses émises par les systèmes de capture de mouvements décrit dans les articles synthétisés dont voici un extrait :

– le sujet reste à l'intérieur d'une zone déterminée (*workspace*),
– la caméra ne bouge pas ou alors de manière constante,
– un seul individu est présent à la fois dans le *workspace*,
– le sujet fait toujours face à la caméra,
– il n'y a pas d'occlusion,
– les mouvements sont lents et continus,
– la lumière est constante,
– l'arrière plan est statique et uniforme,
– la posture de départ est connue,
– des marqueurs sont placés sur le sujet,
– les vêtements du sujet ont une couleur spécifique.

Lors de cette phase, un modèle humanoïde ayant une forme, une apparence, une structure cinématique et une position initiale proche de celle du sujet est définie. La plupart des systèmes partent du principe que la position initiale est connue comme position spéciale de départ ou est spécifiée manuellement. Certains systèmes utilisent un modèle générique qui est la moyenne de plusieurs individus (par exemple en utilisant une structure cinématique composée d'un nombre fixe d'articulations avec des degrés de liberté particuliers) tandis que d'autres mesurent le sujet courant et génère un modèle personnalisé (par exemple en adaptant un modèle moyen à la silouhette du sujet de face et de profil). La personnalisation complète d'un modèle reste une tâche ardue. Elle est quelque peu simplifiée pour la reconnaissance de geste de la main où l'anatomie quasiment complète de celle-ci peut être modélisée et pour laquelle un ou plusieurs modèles moyens suffisent à la plupart des interactions.

1. Bolt, R. A. 1980. Put-that-there : Voice and gesture at the graphics interface. In Proceedings of the 7th annual conference on Computer graphics and interactive techniques. ACM Press, 262-270

Les obstacles à l'initialisation des modèles et à la détection de la position initiale impliquent que peu de systèmes possèdent une phase d'initialisation entièrement automatique. Durant les sept dernières années, un nombre non négligeable de recherches ont été menées dans le but d'automatiser l'initialisation de la forme du modèle à partir de plusieurs vues d'une image. Pour l'initialisation de la structure cinématique du modèle, plusieurs approches ont été étudiées utilisant entre autre des techniques d'apprentissage et des modèles anthropométriques. Seulement un nombre restreint de recherches se sont intéressées aux changements dans l'apparence d'une personne en mouvement.

3.1.2 *Tracking*

Par *tracking* on désigne la mise en relation du sujet à travers les *frame*s. Le *tracking* peut être considéré comme une tâche indépendante, en tant que préparation pour l'évaluation de la position du sujet, ou en tant que préparation des données pour la reconnaissance. Si le processus de *tracking* prépare les données pour l'évaluation de la position du sujet son but est d'extraire les informations spécifiques de l'image, de bas niveau, comme les contours, ou de haut niveau, comme les mains et la têtes. Si par contre ce processus sert à préparer les données pour la reconnaissance, la tâche sera de transformer les données de manière appropriée (pour pouvoir par exemple être utilisées par un classificateur).

Trois aspects communs peuvent être identifiés pour les processus de *tracking* : la segmentation du sujet du reste de l'image, la transformation des images afin de réduire la quantité d'information ou pour convenir à un algorithme particulier et la définition de la méthode permettant de suivre le sujet de *frame* en *frame*.

Tout d'abord l'algorithme débutera par la segmentation des objets d'intérêts (corps, main, tête ...) du reste de l'image (*figure-ground segmentation*). La première solution est d'analyser des données temporelles (des *frame*s successives) et de les comparer (point par point ou par caractéristiques) pour détecter les différences, c'est-à-dire principalement les objets qui ont bougés ou les objets parasites. On considère généralement un arrière-plan statique (peu parasité) et un unique objet mouvant, le sujet. Une autre méthode consiste à considérer des flux, autrement dit des mouvements cohérents de points ou de caractéristiques entre les *frame*s.

Alternativement aux données temporelles, des données spatiales peuvent être utilisées. On distingue deux approches : le seuillage (*thresholding*) et les approches statistiques. La première émet généralement des hypothèses environmentales telles qu'un arrère-plan monochrome ou des vêtements unicolores pour le sujet. Une autre approche très populaire est l'utilisation de marqueurs (passifs ou actifs) facilement segmentables par seuillage. Une approche alternative est l'utilisation d'une caméra infrarouge révélant uniquement les objets « chauds » de la scène, également segmentable aisément.

L'utilisation du seuillage dépend du nombre d'hypothèses émises sur l'apparence du sujet et de la scène et est donc très efficace en environnement contrôlé. Bien que certaines applications évolueront toujours dans de tels environnements, nombre d'applications requièrent des méthodes adaptives comme les approches statistiques.

Les approches statistiques utilisent les caractéristiques de pixels individuels ou de groupes de pixels pour extraire les formes de l'arrière-plan. Les caractéristiques usuelles sont les couleurs et le contours. Des méthodes s'inspirent de la soustraction de l'arrière-plan : une séquence d'images de la scène vide est enregistrée et la variation et la moyenne des intensités ou des couleurs de chaque pixel sont calculées. Dans l'image courante chaque pixel est comparé aux statistiques de l'arrière-plan et classé comme appertant ou non à ce dernier. D'autres méthodes utilisent des contours (statiques ou dynamiques) représentant le sujet ou une partie du sujet.

La seconde étape du *tracking* concerne la représentation des données utiles. Les données de l'étape de segmentation peuvent être directement interprétées pour représenter les zones d'intérêts comme des points (suffisant pour un système utilisant des marqueurs), des boîtes (*box*), des silhouettes, des *blob*s (groupes d'objets partageant les mêmes caractéristiques), des caractéristiques, des contours ....

La dernière étape du processus de *tracking* consiste à trouver des objets similaires (ou correspondances) dans des *frame*s consécutives. Les difficultés de cettes tâche sont proportionnelles à la complexité de la scène et des objets tracés qui dépend des degrés de liberté des objets et de leur représentation. L'analyse des correspondances est souvent réalisée grâce aux prédictions. Ces dernières consistes à délimiter des zones d'intérêts (et ainsi de réduire la taille des données à traiter) en fonction des objets précédemment détectés et de connaissances de haut niveau sur l'état des objets (position, apparence ...) qui sont comparées avec les informations relatives aux objets courants. Des modèles de vélocité, d'accélération ou de mouvements (marche, course ...) peuvent être utilisés. Une méthode habituellement utitlisée pour la prédiction est le filtre de Kalman qui permet d'estimer les incertitudes de la prédiction.

Un autre aspect du traçage apparaît lors de l'utilisation de plusieurs dispositifs de capture. Certains systèmes utilisent plus de caméras que nécessaires et doivent choisir quelle(s) image(s) utiliser à chaque instant.

La recherche autour de la segmentation *figure-ground* a passablement avancé ces dernières années, motivées par l'essor des applications de surveillance. Les méthodes de segmentation doivent être adaptives en environnement réel. Certaines avancées ont été réalisée notamment dans la détection de l'arrière-plan en analysant plusieurs heures de vidéos. Le récent intérêt pour les scènes naturelles à aussi contribué à faire avancer les méthodes pour la correspondance temporelle, spécialement pour le problème d'occlusion (grâce notamment aux méthodes probabilistes).

3.1.3 *Estimation de pose*

L'estimation de pose est le processus d'identification de la configuration du corps humain et/ou des membres pris individuellement dans une scène donnée. Cette estimation peut être une étape de *postprocessing* dans l'algorithme de traçage ou être une partie active de ce dernier. Certains système n'utilisent qu'une estimation grossière (centre de masse du corps, information sur la tête et les mains du sujet) ou à l'inverse une estimation précise de la position, de l'orientation, de la largeur de chacun des membres du sujet. Étant

donné la complexité d'une estimation précise, généralement un seul sujet ou quelques parties du corps sont considérés.

Un apsect commun à l'estimation de pose est l'utilisation d'un modèle humain. Habituellement, un modèle géométrique du corps humain est appliqué, mais d'autres modèles, comme des modèles de mouvement, peuvent aussi être appliqués. Le concept général concernant l'utilisation d'un modèle humain est d'exploiter le fait que le système est dédié à l'analyse du corps humain et donc peut intégrer des connaissances sur les humains dans son traitement. Dans [4], Moeslund *et al.* distinguent trois classes d'estimation de pose : *model-free*, *indirect model use* et *direct model use*. Ces classes sont brièvement décrite ci-dessous.

L'estimation de pose sans modèle n'utilise aucun modèle *a priori*. Les représentations de la pose sont des points, des formes simples ou des *stick-figure*s. La pose du sujet peut être représentée par un ensemble de points, représentation largement utilisée lorsque le sujet et muni de marqueurs. Sans marqueurs, les mains et la tête peuvent être estimées représenté par seulement trois points. Cette représentation compacte suffit à bon nombre d'applications. Le sujet peut également être représenté par de simples *boundary boxes*. Cette représentation est généralement une représentation intermédiaire durant le traitement contrairement à des formes plus *human-like* comme des ellipses qui peuvent former la représentation finale.

La représentation en *stick-figure* contient des informations plus précises sur la structure du sujet et est une représentation populaire lorsqu'on s'intéresse par exemple à la démarche du sujet.

Les méthodes d'estimation de pose avec utilisation indirecte de modèle utilisent un modèle *a priori* lors de l'estimation. Elles utilisent le modèle comme une référence ou table de correspondance (*look-up table*) de laquelle peuvent être extraites des informations utiles à l'interprétation des données mesurées. Divers types de modèles et de niveaux de détails sont utilisés. Le niveau de détail peut aller de la taille du sujet à toutes les informations concernant la structure et la dynamique du sujet. Un exemple simple de modèle comprend les proportions des divers membres du corps humain. À la limite de l'utilisation directe d'un modèle, on peut utiliser un modèle afin de s'assurer que les poses prédites par le *tracking* sont réalistes.

Par utilisation directe de modèle on entend l'utilisation d'un modèle *a priori* comme modèle représentant le sujet observé. Ce modèle est continuellement mis à jour par les observations. Donc le modèle fournit n'importe quelle information désirée sur la pose en tout temps. Moeslund *et al.* rapport dans [4] que 40% des articles étudiés utilisent un modèle de cette manière. Les modèles utilisés sont très détaillés, existent explicitement à l'intérieur du programme et sont utilisés de façon intensive dans la phase de traitement. Les avantages d'un tel modèle sont la capacité à traiter les occlusions et la facilité avec laquelle peuvent être introduites des contraintes cinématiques dans le système.

Un modèle humain est représenté par un nombre d'articulations et de bâtons (les os) les reliant. Les os et la chair les recouvrant peuvent être représentés de manière diverses suivant le niveau de détails nécessaire. Plus le modèle est précis et complexe, meilleurs sont les résultats, au prix de traitements et d'entraînements plus importants.

Un modèle est concrètement représenté par un *state space* où chaque axe représente un degré de liberté d'une articulation du modèle. Une pose du sujet correspond à un point dans le *state space* alors qu'il correspond à plusieurs points dans l'image. L'approche générale pour mettre en relation les données de l'image des données de la pose est connue sous le nom de *analysis-by-synthesis* et utilisée à la manière d'un processus de type *predict-match-update*. L'idée est de prédire la pose du modèle correspondant aux images suivantes dans la séquence. Le modèle prédit est ensuite synthétisé à un certain niveau d'abstraction pour pouvoir être comparé avec les données de l'image. On compare ensuite les données réelles avec les données synthétisées pour déterminer le niveau de similitude. On réitère le processus pour plusieurs prédictions de modèle jusqu'à trouver la meilleure prédiction.

Évidemment, ce *state space* décrit un très grand nombre de poses possibles du modèle qui n'est pas raisonnable pour la comparaison avec les données réelles. C'est pourquoi des contraintes sont introduites pour réduire ce *state space*. Par exemple, l'introduction de contraintes cinématiques du système moteur humain permet de réduire considérablement la plage de valeurs (par exemple le coude ne peut former que des angles entre 0° et 140°). Le fait que deux corps humain ne peuvent pas se passer au travers introduit également des contraintes. Une autre approche pour réduire le nombre de poses modélisées possibles est de considérer des *pattern* de mouvement (particulièrement cycliques comme la marche ou la course).

3.1.4 Reconnaissance

La dernière étape d'un système de reconnaissance de geste est la reconnaissance à proprement parler. On peut voir la reconnaissance comme une sorte de *postprocessing*. Le but de ce processus est habituellement de classifier les movements capturés en différents types d'actions. Les actions sont normalement simples telles que marcher, courrir, prendre un objet, ou plus complexes comme par exemple l'étude de pas de danse.

On peut distinguer deux types de reconnaissance : la reconnaissance statique ou dynamique, c'est-à-dire respectivement basée sur une ou plusieurs *frame*s.

La reconnaissance statique utilise les données spatiales, une *frame* après l'autre. Les approches statiques comparent des information pré-enregistrées avec l'image courante. Ces informations peuvent être des *template*s, des *template*s transformés, des silhouettes normalisées ou des postures [2]. Le but est généralement de reconnaître diverses postures comme le pointage, le fait d'être assis ou debout ou des postures spéciales définies pour utiliser une interface.

Les approches dynamiques utilisent des caractéristiques temporelles dans la tâche de reconnaissance. Les données utili-

2. voir [4], section 6.1 pour les références des exemples mentionnés

sées peuvent être de bas ou de haut niveau. La reconnaissance bas niveau est basée sur des données spatio-temporelles sans trop de traitements. Le but est habituellement de reconnaitre un individu dans uns scène qui marche ou non. Les méthodes de plus haut niveau sont basées sur les données de l'estimation de pose. Ces méthodes vont de la corrélation et de la correspondance de silhouette (*silhouette matching*) aux *Hidden Markov Model*s (*HMM*s) et réseaux de neurones artificiels. L'objectif est de reconnaître des actions comme marcher, porter des objets, ôter ou placer des objets, pointer, des gestes pour le contrôle, se tenir debout versus marcher, marcher versus jogger, marcher versus courrir et classifier divers exercices d'aérobique ou de pas de danses [3].

Un travail intéressant de Bregler [4] représente les données de mouvement par des *moveme*s, similaires aux phonèmes dans la reconnaissance de la parole. Cela rend possible la composition d'activités complexes à partir de *moveme*s simples. Wren *et al.* [5] utilise également une représentation symbolique de haut niveau à l'aide d'un alphabet de comportement (*behavior alphabet*) et modélise chaque comportement en utilisant un *HMM*. L'alphabet est utilisé pour classifier différents types d'actions dans un jeu de réalité virtuelle et pour distinguer les style de jeu de différents sujets. Ces dernières méthodes s'approchent des domaines de recherche de la sémantique et de l'intelligence artificielle.

Les différentes sous-sections précédentes ont présenté les tâches principales d'un système de reconnaissance de geste. La plupart des exemples concernaient plutôt les mouvements du corps dans son ensemble ou du bras et de la main. Avant de présenter les principaux domaines d'application de la reconnaissance de gestes, la sous-section suivante est dédiée à la reconnaissance faciale.

3.2 Reconnaissance des gestes faciaux

Les humains peuvent aisément détecter et identifier des visages dans une scène. La robustesse avec laquelle nous le faisons est déconcertante vu le nombre de changements inhérents au stimulus visuels dus [3] :
- aux conditions de visualisation (telles que les variations de luminosité),
- aux expressions du visage,
- à l'âge,
- au genre,
- à l'occlusion,
- aux « distractions » comme les lunettes, la coupe de cheveux et autres déguisements.

L'objectif de la détection de visage est d'efficacement identifier et localiser des visages humains indépendamment de leur position, leur taille, leur orientation, leur pose et de l'éclairage. Les applications dans ce domaines sont multiples : identification criminelle, surveillance, vérification d'utilisation de carte de crédit, télécommunication, télévision haute définition, médecine, requêtes multimédia, interface homme-machine ou transmission d'informations faciales sur faible bande passante. Il y a deux approches principales dans la reconnaissance automatique de visage [3] :
- approche analytique : des modèles mathématiques flexibles sont développés prenant en compte les déformations du visage et les changements de luminosité. Des caractéristiques discrètes locales, comme les iris, sont extraites afin de localiser et d'identifier les visages. La position de ces caractéristiques par rapport aux autres détermine l'emplacement global du visage. Des méthodes statistiques telles que les *HMM*s peuvent être appliquées à ses mesures.
- approche holistique : cela implique l'utilisation de templates en niveau de gris pour une reconnaissance globale. Un *feature vector* est utilisé pour représenter le template de tout le visage. Cette approche inclut entre autre les réseaux de neurones artificiels.

Les variations entre deux images de visage peuvent être de deux types :
- interpersonel : différence entre deux visage de deux individus différents. Cette catégorie correspond à la reconnaissance faciale.
- intrapersonel : changements dans l'apparance de la même personne dus à différentes expressions du visage ou a des variations de luminosité. Cette catégorie correspond à la reconnaissance d'expressions faciales.

Les méthodes concrètes pour la reconnaissance faciale et d'expressions faciales gravitent autour des modèles de Markov cachés, l'utilisation de modèles de contours ou de filtre de Gabor, l'interprétation de *FACS* [6] ou encore l'emploi d'approches connexionnistes (réseaux de neurones artificiels) [3].

Comme mentionné précédemment, plusieurs applications possibles de la reconnaissance faciale ont besoin de pouvoir identifier de façon unique les individus. La question est de savoir à quel point un visage humain est unique et de savoir si un algorithme pourrait approximer les performances humaine avec un taux d'erreur inférieur à 1%.

4. APPLICATIONS

Au fil de cet article, nous avons évoqué quelques exemples d'applications de la reconnaissance de gestes. On dénombre trois catégories d'applications :
- les applications de surveillance,
- les applications de contrôle (interfaces pour les jeux vidéos, pour les environnements de réalité virtuelle ou plus généralement interfaces homme-machine),
- les applications d'analyse (diagnostiques médicaux, optimisations des performances d'un athlète, annotation automatique de documents, compression vidéo . . .)

Trois applications principales sont birèvement présentée dans les sections suivantes.

4.1 Interaction homme-machine

À l'heure actuelle, les interfaces de communication les plus utilisés de l'homme vers la machine sont toujours le clavier et la souris. Pour que des systèmes intelligents soient capables d'interpréter efficacement et précisément les gestes

3. voir [4], section 6.2 pour les références des exemples mentionnés

4. C. Bregler, Learning and recognizing human dyamics in video sequences, in *Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997.*

5. C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, Pfinder : real-time tracking of human body, *Trans. Pattern Anal. Mach. Intelligence* **19**(7), 1997, 780-785.

6. *Facial Action Coding System*

complexes de l'homme et permettre une interaction naturelle entre l'homme et la machine (et remplace peut-être le clavier et la souris), beaucoup de problèmes restent à résoudre. Les applications qui profiteraient (ou profitent déjà) d'intefaces gestuelles sont par exemple les applications permettant la visualisation de (grands volumes de) données, la navigation et le contrôle en environnement virtuel et la conception assisté par ordinateur (CAO). Une interface gestuelle en CAO permet de manipuler directement des objets virtuels ou des outils virtuels nécessaires à la réalisation des objets ; l'utilisateur se contente ainsi de reproduire les gestes qu'il ferait naturellement pour manipuler réellement les outils et les objets. L'un des avantage est l'apprentissage facilité pour l'utilisation des outils informatiques. Concernant la quête de l'immersion totale en réalité virtuelle, elle passe forcément par une manipulation et une navigation naturelle en environnement virtuel. Quant à la visualisation de données, on commence à voir les limites de l'interaction classique clavier-souris qui permet d'interagir avec un seul objet à la fois. Évidemment, il faut que les système logiciels sous-jacents soient capables de gérer ces nouvelles interactions pour que leur succès soit garantit. La sous-section suivante présente une interaction très particulière entre l'homme et la machine et souvent utilisé comme application modèle en reconnaissance de gestes.

4.2 Langages des signes

L'une des applications évidentes lorsqu'on pense à la reconnaissance de gestes est la reconnaissance du langage des signes. De nombreuses recherches ont été menées autour de l'interprétation de tels langages.

Le langage des signes consiste généralement en trois composants principaux [3] :
– un alphabet dactylologique (*finger-spelling*),
– un vocabulaire de signes (*word-level sign vocabulary*),
– des caractéristiques non manuelles.

L'alphabet dactylologique est utilisé pour épeler des mots lettre par lettre (généralement des noms propres ou des mots exclus du vocabulaire de signes). Le vocabulaire de signes est formé de signe représentant des mots et est utilisé majoritairement pour la communication. Les caractéristiques non manuelles consistent en expressions faciales, à la position de la langue, de la bouche et du corps.

L'intérêt du langage des signes en reconnaissance de gestes est sa structure relativement précise, permettant la définition de règles contextuelles et grammaticales strictes pouvant être appliquées pour faciliter la reconnaissance. Cependant, il n'y a pas de forcément de frontières claires avec les signes personnels et la reconnaissance de langages signés reste très difficile [8]. Ce domaine de recherche est parallèle à la reconnaissance de la parole sachant que les deux sont des processus variant dans le temps montrant des variations statistiques rendant l'utilisation de *HMM*s un choix approprié pour la modélisation des processus. Les signes isolés peuvent être facilement extraits (grâce à la présence de silences entre les signes) et présentés individuellement à un *HMM* entraîné. La reconnaissance continue de signes est plus ardue. Dans ce cas les modèles de Markov cachés offrent un avantage indéniable en étant capable de segmenter des flux de signes automatiquement avec l'algorithme de Viterbi [8].

La coarticulation [7] est l'un des problèmes difficiles dans la reconnaissance de gestes continus. L'une des approches développée consiste à considérer des « phonèmes » (aussi appelés *viseme* dans [3]) pour modéliser les mouvements entre les signes. Une approche étendue consiste à ne plus considérer le signe comme unité basique d'un langage signé, mais des phonèmes et d'entraîner les *HMM*s à les reconnaître. Puisque le nombre de phonèmes est limité, il est possible d'utiliser de modèles de Markov cachés pour reconnaître de larges vocabulaires.

La phase d'initialisation d'un système de reconnaissance de langage des signes est également problèmatique. Des marqueurs ou de gants monochromes sont souvent utilisés pour faciliter la capture. Reste que lorsqu'un système est initialisé, les gestes qui suivent doivent souvent tous faire partie du langage afin de ne pas bruiter le flux.

4.3 Systèmes de surveillance

Dans [4] et [5] les auteurs présentent le besoin de systèmes de surveillance comme la principale raison de l'avancée dans le domaine de la capture et la reconnaissance de mouvements corporels. Ces systèmes ont pour but d'analyser les comportements humains et de détecter de manière automatique des comportement hors norme comme le fait de commettre un crime, par exemple un vol de voiture. Les systèmes de surveillance automatiques doivent répondre à plusieurs défis importants de la reconnaissance de gestes ; initialisation automatique, reconnaissance de lieu, segmentation de sujets dans une foule ... Des applications dans ce domaine pourraient par exemple compter les individus d'une foule, analyser la congestion ou les comportements dans une file d'attente ou identifier des individus.

5. PÉRIPHÉRIQUES DE CAPTURE

La majorité des exemples des articles considérés concerne la reconnaissance de gestes par vision par ordinateur. Les périphériques de capture habituels de ces systèmes sont des caméras. En fonction du type de système, de sa capacité à supporter l'occlusion, du nombre de degrés de liberté considéré ... on utilisera une ou plusieurs caméras. L'utilisation de plusieurs périphériques visuels complique le *tracking* comme mentionné précédemment.

L'alternative à la vision par ordinateur est l'utilisation de périphériques dédiés à la capture de mouvement. Le périphérique alternatif par excellence est le gant. Il fut très populaire dans les années 90 quand la puissance de calcul des ordinateurs n'était pas nécessaire pour avoir de bonnes performances en *vision based recognition* [7]. Les principaux inconvénients du gant sont le caractère intrusif du périphérique (bien que depuis les années 90 ils aient énormément évolués du point de vue ergonomique) et son prix.

L'utilisation d'un gant ne résout évidemment pas tous les problèmes par rapport aux méthodes visuelles. Les principaux avantages concernent le *tracking* qui est facilité puisqu'aucune segmentation de la main ou du corps n'est nécessaire. Une partie de l'initialisation du système peut égale-

7. transition entre les signes

ment être facilitée ; la récupération de la posture initiale est automatique et on considère uniquement les gestes lorsque le gant est utilisé (contrairement à une méthode visuelle où le moment du début de capture doit généralement être spécifié). Toutes les autres difficultés liées à la modélisation et à l'interprétation des gestes restent entières.

Bon nombre d'applications utilisant un gant peuvent être réalisées en vision par ordinateur, principalement pour la manipulation de données et le contrôle d'environnements virtuels. D'autres applications utilisant des gants à retour de force sont *a contrario* quasiment irremplaçables. L'un des domaines d'application les plus prometteurs pour la reconnaissance geste reste le domaine des jeux vidéos. Les contrôleurs de jeux vidéos actuels intègrent quasiment tous un système de *feedback* (par exemple vibration de la manette). Il est certain que des systèmes utilisant des gants à retour de force seront plus appréciés par les joueurs cherchant un haut niveau de réalisme. Un autre domaine nécessitant des périphériques haptiques est le domaine de la réhabilitation moteur. Des systèmes utilisant des gants à retour de force ont été développés afin d'analyser les capacités motrices de patients et planifier des traitements adéquats [1]. Avec de tels périphériques on atteint la limite entre la reconnaissance de geste et les interfaces tactiles.

Hormis les gants, d'autres périphériques ont été conçus pour la reconnaissance de geste comme le contrôleur de la console de jeux Nintendo Wii utilisé pour la détection des gestes des bras [6], ou des *bodysuit* garnis de capteurs permettant, avec les avantages d'un gant pour les gestes manuels, de capturer les mouvements de tout le corps.

6. ET APRÈS ?

Aujourd'hui les outils les plus utilisés en reconnaissance de gestes sont les *HMM*s, les filtres particulaires et l'algorithme condensation, les *final state machine*s (*FSM*s) et les réseaux de neurones artificiels. Les *HMM*s étant lourds en calculs, dans [3] les auteurs proposent une hybridation entre les *HMM*s et les *FSM*s possiblement plus fiable. Le besoin grandissant de méthodes de requête pour des bases de données picturales, de méthodes d'annotations ou de compression offre de bonnes perspectives d'avenir pour la reconnaissance de geste. Toujours dans [3], Mitra *et al.* mettent en avant l'utilisation de *fuzzy sets* et de *rought sets* comme cadre à la reconnaissance d'expressions faciales afin de déterminer les émotions, sachant que les émotions humaines ne sont jamais pures.

La recherche en reconnaissance de geste avec l'utilisation de gant reste également très active et il va de soit que les avancés technologiques en informatique, en prériphérique sensoriels, en matériel et en techniques de traitement/classification feront des gants des périphériques moins chers, plus puissants, polyvalents et peut-être plus ubiquitaires.

Dipietro *et al.* précisent, dans leur conclusion de [1], que le domaine du logiciel a un rôle très important ; des logiciels sous-jacents intuitifs et très bien intégrés facilitent l'adoption des nouvelles technologies par le public. Finalement, dans l'optique de pouvoir évaluer et comparer les différentes méthodes de reconnaissance, des bases de données de test sont encore à enrichir, principalement pour la reconnaissance de gestes impliquant tout le corps, sachant que plusieurs bases de données existent déjà pour la reconnaissance faciale [8].

7. CONCLUSION

Depuis toujours, l'homme rêve de machines intelligentes, capables de dialoguer avec lui, de le comprendre, et de répondre aux questions qu'il s'est toujours posé. Il rêve de pouvoir interagir avec une elles comme avec ses semblables. Dans cette quête du Graal interactif, la recherche en reconnaissance de gestes joue un rôle important tout comme celle en reconnaissance de la parole. Dans cet article nous avons présenté diverses facettes de la reconnaissance de gestes. Dans la première partie, nous avons tenté de définir certains termes et avons évoqué le problème de taxonomie communément admise inexistante pour ce domaine de recherche. Puis la structure générale d'un système de reconnaissance de geste fut présentée avant de s'intéresser aux applications possibles et à la reconnaissance de geste via des gants.

La littérature étudiée a permis de dégager les principaux défis de la reconnaissance de geste, de présenter les outils et méthodes utilisées et d'entrevoir quelques directions futures. Le principal problème de la reconnaissance de geste reste la segmentation et la traçage d'objets d'intérêts dans une scène réelle (problème quasiment inexistant en utilisant des gants) et l'interprétation des gestes dans un flux d'information généralement très bruité. Des approches multimodales, bien que relativement complexes, pourraient rendre certains système plus robustes en ajoutant de la redondance nécessaire à une meilleure interprétation des gestes.

Il va sans dire que les fantasmes de l'homme actuel concernant sa relation avec les machines sont loin de devenir réalité, mais offrent d'innombrables opportunités pour le recherche en reconnaissance de gestes.

8. REFERENCES

[1] Laura Dipietro, Angelo M. Sabatini, and Paolo Dario. A survey of glove-based systems and their applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(4) :461–482, 2008.

[2] Maria Karam and m. c. schraefel. A taxonomy of gestures in human computer interactions, 2005.

[3] Sushmita Mitra and Tinku Acharya. Gesture recognition : A survey. *IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS - PART C*, 37(3) :311–324, 2007.

[4] Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Comput. Vis. Image Underst.*, 81(3) :231–268, 2001.

[5] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2) :90–126, 2006.

[6] Thomas Schlömer, Benjamin Poppinga, Niels Henze, and Susanne Boll. Gesture recognition with a wii controller. In *TEI '08 : Proceedings of the 2nd international conference on Tangible and embedded interaction*, pages 11–14, New York, NY, USA, 2008. ACM.

8. `http://www.face-rec.org/databases/`

[7] David J. Sturman and David Zeltzer. A survey of glove-based input. *IEEE Comput. Graph. Appl.*, 14(1) :30–39, 1994.

[8] Ying Wu, Thomas S. Huang, and N. Mathews. Vision-based gesture recognition : A review. In *Lecture Notes in Computer Science*, pages 103–115. Springer.