HAL open science

# Analyse de vidéos en langue des signes : méthodes et stratégies

François Lefebvre-Albaret, Patrice Dalle

## To cite this version:

François Lefebvre-Albaret, Patrice Dalle. Analyse de vidéos en langue des signes : méthodes et stratégies. ORASIS'09 - Congrès des jeunes chercheurs en vision par ordinateur, 2009, Trégastel, France, France. inria-00404651

**HAL Id: inria-00404651**

**https://inria.hal.science/inria-00404651**

Submitted on 16 Jul 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Analyse de vidéos en langue des signes : méthodes et stratégies

# Computer analysis of sign language videos : methods and strategies

François Lefebvre-Albaret Patrice Dalle

IRIT - Université Paul Sabatier
118 Route de Narbonne, 31062 Toulouse cedex 9
lefebvre@irit.fr, dalle@irit.fr

## Résumé

*Les langues des signes (LS) sont les formes les plus abouties de communication gestuelle. Leur analyse automatique constitue donc un réel défi qui implique de prendre en compte leur organisation aux niveaux lexical et syntaxique. Cet article présente dans un premier temps les spécificités des langues des signes pour montrer la difficulté d'un traitement automatique de vidéos en LS. Nous montrons comment des contraintes propres aux LS peuvent être intégrées à un système de suivi des mains et de la tête dans les vidéos pour le rendre plus robuste et comment les résultats du suivi peuvent être intégrés dans un modèle de plus haut niveau rendant compte de l'organisation spatiale des LS.*

## Mots Clef

Langue des Signe, Suivi, Analyse.

## Abstract

*Sign Languages (SL) are the most complex and organised form of gesture communication. Automatic analysis of sign language videos constitutes a real challenge. In order to process SL utterances, we must take into account their organization at the lexical and the syntactic levels.*
*In this paper, we first present SL specificities in order to show the high complexity of a fully automatic processing of unconstrained sign utterances. We show how SL rules can be integrated in tracking algorithm. The prior knowledge about SL is used to combine less accurate measures in order to obtain more reliable results that can be integrated in a model taking into account SL spatial structure.*

## Keywords

Sign Languages, Tracking, Analysis

## 1 Introduction

Le but de cette étude est l'analyse de vidéos de langues des signes en mono-vision. Cet axe de recherche a été orienté par les observations suivantes :

- Il existe un besoin sociétal de faciliter la communication avec la communauté sourde signante. Le développement d'outils de traitement automatique des LS va dans ce sens.
- L'analyse de vidéos en LS constitue un réel défi en raison de la grande complexité des données à traiter.
- Cette analyse est désormais possible grâce aux progrès réalisés sur le matériel (dispositif d'acquisitions, puissance de calcul) et sur les algorithmes(détection, suivi, reconnaissance), et à la grande quantité de données disponibles.
- Les applications de cette recherche dépassent le cadre de l'analyse des LS et concernent différents domaines touchant à l'interaction gestuelle.

Cette recherche ouvre la porte à différentes applications :

- L'analyse automatique de phrases en LS (segmentation, caractérisation de signes),
- Les applications de type linguistique (indexation d'énoncés, création de dictionnaires),
- Les applications à l'enseignement de la langue des signes (didacticiels),
- La reconnaissance automatique de signes qui pourrait mener sur le long terme à des dispositifs de traduction automatique des LS vers les langues vocales,
- La production ou la validation de modèles permettant la génération d'énoncés par un signeur virtuel.

Toutes ces applications nécessitent une première phase de suivi et de segmentation des énoncés en LS.

## 2 Langues des signes et communication gestuelle

Notre étude se focalise sur des gestes communicationnels qui sont produits intentionnellement de manière à transmettre un message. Les LS constituent par conséquent les formes les plus élaborées de communications gestuelles et impliquent une grande variété de structures syntaxiques. Leur lexique permet de véhiculer des messages complexes à la manière de n'importe quelle langue vocale.

Bien que les langues des signes des différents pays possèdent de fortes variations au niveau lexical, leurs organisations grammaticales sont très proches ce qui explique que des sourds signants de différents pays arrivent à se

comprendrent rapidement. Quoi qu'il en soit, les règles énoncées dans la suite de cet article ainsi que les exemples ayant servi à la validation de nos algorithmes seront tous tirés de la Langue des Signes Française (LSF).
Les LS n'ayant pas de forme écrite, nous utilisons les vidéos des énoncés comme forme orale. Ceci ajoute des difficultés supplémentaires dues aux variations de contextes et de situations d'énonciation.
D'un autre côté, le fait que les LS soient hautement structurées et véhiculent un message intelligible peut aussi permettre d'accélérer l'analyse des vidéos si ces règles sont prises en compte durant les étapes de traitement.

# 3 Stratégies de traitement

Nous ne réalisons pas pour l'instant d'analyse en temps réel de vidéos en LS. Notre étude se base par conséquent exclusivement sur des énoncés filmés. Malgré tout, la taille des corpus vidéos à traiter nécessite l'emploi de méthodes rapides et robustes pour effectuer automatiquement la détection et le suivi (même si certaines interventions humaines peuvent ponctuellement être ajoutées dans des phases de pré- ou post- traitement). Il est donc nécessaire d'intégrer un modèle de LS dans une ou plusieurs étapes de traitement des vidéos.
La première stratégie de traitement consiste à sélectionner des méthodes élaborées et à les adapter aux caractéristiques des LS. La deuxième consiste à intégrer des connaissances sur le fonctionnement des LS pour rendre les résultats plus fiables et plus robustes à partir d'une combinaison de mesures moins précises. Ces deux approches de traitement d'une vidéo seront illustrées dans le cadre du suivi de la position des mains dans une séquence vidéo mono-vue. Nous discuterons enfin de l'intégration de ces mesures dans un modèle de plus haut niveau permettant une représentation de la structure spatio-temporelle des énoncés en LS.

# 4 Adaptation de méthodes existantes à l'analyse du geste

La connaissance de données de bas niveau comme les trajectoires des mains dans les images est une nécessité pour la plupart des traitements opérés sur un énoncé en LS. La partie suivante traite de ce problème de suivi de la tête et des mains. Comme nous allons le montrer, les mesures fournies par les algorithme de suivi traditionnels sont difficiles à utiliser seules en raison de fréquentes occultations qui apparaissent en LS.

## 4.1 Suivi de la tête et des mains, état de l'art

Un grand nombre d'algorithmes dédiés à la détection et au suivi de la tête et des mains ont été développés ces dernières années. La couleur de peau est souvent utilisée pour détecter les zones de peau (seule, ou en combinaison avec le mouvement). Ces modèles de peau sont soit explicites [13], soit appris à partir d'images de peau [1] et [20]. Il est ainsi possible de déterminer la probabilité qu'a un pixel de la vidéo d'appartenir à la peau du signeur par le biais d'une modélisation de la distribution de la couleur de la peau sous forme d'un mélange de gaussienne ou d'un histogramme [23], [21]. Ensuite, d'autres procédés doivent être appliqués pour déterminer si les blobs de peau appartiennent ou non aux mains ou à la tête du signeur. D'autres caractéristiques peuvent aussi être utilisées pour identifier la tête dans une vidéo. La méthode la plus connue a été proposée dans [22]. Elle consiste à combiner des filtres élémentaires grâce à une technique de boosting [6]. Les résultats de ces méthodes sont extrêmement fiables tant que le visage n'est pas occulté et ne subit pas de rotations. Le même procédé peut être utilisé pour détecter d'autres parties du corps dans une vidéo. Par exemple [17] l'utilise pour détecter le tronc et les mains dans une vidéo. D'autres techniques reposent sur la caractérisation de la forme des mains ou de la tête. Parmi elles, on citera la caractérisations de blobs de peau par des moments de Hu utilisés par [11], les machines à vecteur support [4] ainsi que la corrélation avec un modèle de forme de la partie du corps à suivre [19]. Toutes ces techniques d'identifications montrent de faibles résultats en présence d'occultations. Or cette situation est très fréquente en LS puisque de nombreux signes impliquent un contact des mains avec d'autres articulateurs du corps. Un algorithme de suivi qui ne prendrait pas en compte cette spécificité ne serait pas adapté au traitement automatique de la LS. A cela s'ajoute le fait que les LS impliquent des mouvements hautement non linéaires et de grandes amplitudes. De telles caractéristiques rendent difficile l'application de prédicteurs de Kalman et diminuent la robustesse des approches adaptées aux mouvements de faible amplitude.

## 4.2 Suivi de la tête basé sur des filtres particulaires

Nous avons repris et adapté une méthode développée par [8] basée sur les filtres particulaires pour surmonter ce problème d'occultation et fournir un suivi à la fois fiable et robuste des mains et de la tête dans une vidéo. Cette méthode fait appel à une étape de recuit rendue nécessaire par les mouvements de grande amplitude observés dans le cadre des LS. Trois filtres particulaires (correspondant respectivement aux deux mains et à la tête) sont utilisés pour le suivi. Les filtres particulaires sont constitués de nuages pondérés $\{(s_0(t), \pi_0(t))...(s_N(t), \pi_N(t))\}$ de N particules. Chaque particule a un état (s) et un poids ($\pi$). Ici, l'état est constitué du vecteur $s(t) = [x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}]^t$ où $(x, y), (\dot{x}, \dot{y})$ et $(\ddot{x}, \ddot{y})$ représentent respectivement la position, la vitesse et l'accélération de la particule à l'instant t. La méthode utilise le filtre particulaire décrit dans [7] basé sur la détection de la couleur de peau. Cette densité d'observation sera notée $p(z(t)|x(t))$ probabilité conditionnelle d'observer une couleur en fonction de la zone (peau / non-peau) où se trouve le pixel. Une étape de ré-échantillionage est ajoutée après chaque phase de recuit. Elle permet d'éviter la dégénérescence de l'algorithme

(sans ré-échantillonnage, seulement quelques particules auraient un poids significatif à la fin de l'algorithme).

Le processus de recuit est itéré M fois et l'algorithme général peut être décrit de la manière suivante :

**Initialisation** A l'initialisation de l'algorithme, attribuer le même poids à chaque particule :
$Pour\ i = 0..N \quad \pi_i(t) = 1/N$

**Etape 1** Effectuer une normalisation des poids :
$A = \sum_{i=0..N} \pi_i(t)$
$Pour\ i = 0..N \quad \pi_i(t) = \pi_i(t)/A$

**Etape 2** Effectuer un rééchantillonage stratifié du nuage grâce à l'algorithme proposé par [12] explicité dans [5] section 5.3.1 (La probabilité de survie d'une particule s'accroît avec son poids. Les particules de poids important ($\pi_i(t) > 1/N$) peuvent aussi être dupliquées) :
$Pour\ i = 0..N \quad x_i(t) \leftarrow x_j(t)$

**Etape 3** L'état suivant de la particule est calculé grâce à un produit de son vecteur d'état actuel par la matrice de transition $S$. Un mouvement aléatoire gaussien multivarié $P(t)$ est ajouté à ce nouvel état. L'amplitude de ce mouvement additionnel est diminuée au fur et à mesure des cycles de recuits :
$s_i^m(t) = S.s_i^{m-1}(t) + P(m)$

**Etape 4** On attribue à chaque particule un nouveau poids en fonction de la densité d'observation. Le coefficient $\beta$ est positif et inférieur à un ; il croît à chaque itération.
$Pour\ i = 0..N \ \ \pi_i^m(t) = p(z(t)|x_i^m(t))^\beta$

FIGURE 1 – Soustraction des nuages particulaires correspondant aux mains droite et gauche avant le suivi de la tête

A l'issue du processus de recuit, la position des mains est déterminée comme étant le centre de gravité des nuages de particules.
Pour éviter que deux nuages suivent la même partie du corps (par exemple, que deux filtres particulaires suivent la même main), l'algorithme est appliqué à deux reprises sur chaque image. Le suivi est d'abord appliqué sur l'image originale ($I_0$). Ensuite, les nuages sont soustraits de $I_0$. Par exemple, les nuages correspondant aux mains droites et gauches seront soustraits de $I_0$ avant de localiser la tête (fig 1). Grâce à cette méthode de pénalisation, les deux nuages ne suivront pas la même partie du corps mais seront tout de même autorisés à se superposer en cas d'occultation. La validation de cet algorithme a été effectuée sur une vidéo de 3000 images pour évaluer le nombre d'images où les mains et la tête n'étaient pas suivies par les nuages. Les cibles sont ratées dans 7,4% de la vidéo pour la tête et 1,7% seulement pour les mains. Les critères d'évaluation sont détaillés dans [9].
Le nombre de particules de chaque filtre est proportionnel au nombre de pixels de peau de la partie du corps qu'il suit. Le temps de traitement par image augmente donc avec la résolution de la vidéo. Le modèle de mouvement utilisé est de type marche aléatoire car l'observation des mouvements de la LSF révèle qu'il n'est pas possible de prédire la position qu'aura la main dans l'image suivante (grande non-linéarité du mouvement).
Cette méthode fournit d'excellent résultats qui ne peuvent cependant pas être utilisés directement dans le cadre d'une analyse de vidéo en LS. En effet, si le nuage de la tête se distingue de celui des mains par son nombre de particules qui lui permet de couvrir une plus grande de surface, les deux nuages utilisés pour suivre les deux mains sont identiques, ce qui explique que les fréquentes occultations engendrent des inversions entre les mains droite et gauche. Il est donc nécessaire d'ajouter d'autres informations pour *désambigüiser* les positions des mains.

# 5 Intégration d'un modèle de LS dans une méthode d'analyse

Les modèles utilisés dans le cadre du traitement des LS peuvent être classés entre plusieurs domaines :
- Si on travaille au niveau du signe, un modèle phonologique peut suffire. Une telle approche sera illustrée dans le §5.1.
- Au contraire, l'analyse d'un énoncé complet en LS implique une structure spatio-temporelle, qui doit être utilisée pour exploiter des mesures de bas niveau ou prédire de futures observations (cette approche descendante §5.2 peut aussi permettre des traitements plus rapides).

## 5.1 Désambigüisation des mains droite et gauche

Comme il est mentionné dans le §4.2, les filtres particulaires permettent de déterminer les coordonnées des deux mains $(H_1, H_2)$ sans pour autant indiquer de quelle main (droite ou gauche) il s'agit. Certaines études résolvent ce problème en effectuant un suivi du tronc, des mains et de la tête et utilisent un modèle de postures corporelles pour effectuer la désambigüisation [16][17]. Cette approche tend à considérer un croisement des deux mains comme une pose non naturelle ce qui est problématique étant donné que les mains peuvent être croisées jusqu'à 10,1% [1] du temps dans

1. Mesures effectuées sur des traductions de brèves en langue des signes d'une durée moyenne de 30 secondes fournies par la société Web-

un énoncé standard. Pour obtenir des résultats plus fiables, nous avons effectué une détermination grossière des positions des coudes $(E_r, E_l)$ basée sur l'apparence de leur silhouette. Ce type de modèle est utilisable car la plupart des signes sont effectués dans l'espace neutre (devant le signeur) et que le signeur fait face à la caméra, ce qui donne aux coudes une apparence peu variable.

1. La désambigüisation des mains sera opérée grâce à des estimateurs bayésiens. Nous nommerons I l'évènement "H2 représente la main droite" et $p(I|z)$ la probabilité de I conditionnellement à l'observation z. Cette probabilité peut être déterminée conformément à la règle de Bayes.

   $$p(I|z) = p(z|I)p(I)/P(z)$$

   Plusieurs mesures mènent à différentes probabilités conditionnelles :
   La **comparaison des abscisses des deux mains**,

   $$P_1(t) = p(I/(x_{rh}(t) - x_{lh}(t)))$$

   La **comparaison des ordonnées des deux mains**,

   $$P_2(t) = p(I/(y_{rh}(t) - y_{lh}(t)))$$

   Ce dernier critère peut être utilisé **dans un contexte de langue signée** car la main dominante [2] est souvent plus haute que la main dominée (85,4%[1] du temps dans notre vidéo test)
   La **comparaison des distances main-coude**,

   $$P_3(t) = P(I/(H_r, E_r, H_l, E_l))$$

   Comme il est impossible de quantifier précisément les relations de dépendance entre les trois mesures ($P_1$,$P_2$ et $P_3$), on les fusionne en utilisant une approche de classificateur bayésien naïf (le résultat de cette fusion nommé $B(t)$ ne peut donc pas être interprété comme une probabilité).

   $$B(t)(H_r, H_l, E_r, E_l) = P_1(t).P_2(t).P_3(t)$$

   On calcul $B(t)$ pour les hypothèses I et $\bar{\text{I}}$ et le plus petit $B(t)$ correspondra à l'hypothèse la plus vraisemblable. Une telle approche permet une désambigüisation des mains avec un taux de succès de 95,2%[1].
2. Pour améliorer la fiabilité de notre algorithme, nous essayons également d'optimiser la trajectoire des deux mains de manière à ce que le déplacement total des mains droite et gauche soit minimal. Ce critère d'économie peut s'écrire :

   $$argmin(\sum_{t=1..T} dis_R(t) + dis_L(t))$$

   où $dis_R(t)$ et $dis_L(t)$ représentent les déplacements des mains droites et gauche et T représente la durée de la vidéo.
3. Les meilleurs résultats sont obtenus en cherchant à minimiser l'expression suivante tenant compte des critères locaux et globaux :

   $$argmin(\sum_{t=1..T} [dis_R(t) + dis_L(t) + \alpha.ln(B(t)])$$

   (Le paramètre $\alpha$ est déterminé empiriquement). Un algorithme de programmation dynamique est utilisé pour minimiser cette expression.

Les combinaisons d'indices locaux et globaux aboutissent à de meilleurs résultats car nous obtenons un taux d'appariement correct de 98,7%[1]. Ceci montre que l'inclusion de contraintes propres aux LS comme la notion de signation dans l'espace neutre (utile pour la détection des coudes) ou la notion de main dominante (qui permet la notion de comparaison de hauteur entre les deux mains) peuvent améliorer sensiblement les résultats fournis par les techniques de suivi traditionnel malgré le fait que le suivi des coudes soit peu précis.
Nous avons déjà utilisé les résultats du suivi pour opérer une segmentation assistée d'énoncés en LS[14]. L'étude du mouvement peut donc être pertinente au niveau du signe. Comme nous allons le voir, le suivi peut également être intégré à des modèles de plus haut niveau s'attachant à décrire la structure syntaxique des LS.

## 5.2 Intégration au niveau d'un énoncé en LS

Comme l'a montré notre méthode de suivi, la combinaison de différentes informations peut conduire à des estimations plus robustes (ceci est illustré dans notre cas par le suivi des coudes). D'autres mesures telles que la direction du regard, la position des épaules peuvent également être prises en compte pour augmenter la précision des mesures ou aboutir à un niveau d'interprétation de plus haut niveau des énoncés en LS. Ces deux derniers paramètres possédent à la fois un rôle prosodique et syntaxique et ils impliquent par conséquent l'utilisation d'un modèle d'énoncé LS basé sur l'espace de signation.

En conséquence du canal visuo-gestuel utilisé dans le cadre d'une communication en LS, la syntaxe d'un énoncé doit être représentée à la fois dans le temps et l'espace. Nous avons proposé un modèle d'espace de signation [15], [3] où chaque entité est définie par son rôle (temporel, locative, actanciel) et sa position spatiale. Il est important de remarquer que tout système d'analyse de geste communicationnel devra inclure un tel modèle. Le modèle d'espace de signation, décrit dans un formalisme UML (cf fig 2), permet de représenter l'espace de signation à un instant précis. Chaque emplacement tridimentionnel situé devant le signeur peut recevoir une ou plusieurs entités impliquées dans un énoncé.

Il existe plusieurs catégories d'entités :
- Les **dates** sont utilisées pour désigner des références temporelles absolues ou relatives.

sourd (www.websourd.org)

2. La main dominante est la main utilisée le plus fréquemment dans les signes impliquant la mobilisation d'une seule main

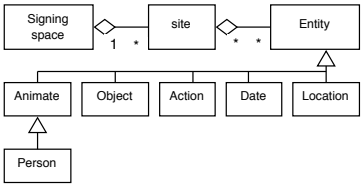


FIGURE 2 – Représentation dans le formalisme UML du modèle d'espace de signation

- Les **lieux** concernent des entités qui peuvent jouer le rôle de références spatiales.
- Les **animés** peuvent réaliser une action ou être associés à une référence spatiale.
- Les **personnes** sont des types particuliers d'animés car le signeur peut les utiliser pour effectuer des transferts de rôle [2]. (le signeur prend alors le rôle de la personne qu'il joue).
- Les **actions** sont des relations impliquant au plus trois entités.
- Les **objets** désignent toute entité qui n'appartient pas à une des catégories mentionnée ci-dessus, ils peuvent être situés et impliqués dans une action.

Un énoncé peut être modélisé comme une succession d'espaces de signation. Aux cours de l'énonciation, les entités peuvent être soit déplacées, soit ajoutées.

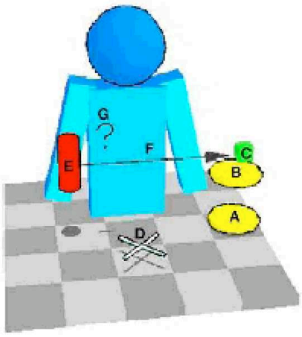


FIGURE 3 – Représentation de la structure spatiale de la requête : Quel (G) est le réalisateur (E) du(F) film(C) qui passe le jeudi 26 février à 9H30(D) au cinéma Utopia(B) de Toulouse ?(A). Dans l'énoncé en LSF, les entités sont créées dans l'ordre A .. G

Notre modèle d'espace de signation a déjà été utilisé pour modéliser des requêtes simples adressées à une base de donnée (cf fig 3). Nous l'avons aussi étendu pour qu'il puisse rendre compte du partage de l'espace de signation entre deux interlocuteurs parlant du même concept, représenté par la même entité (dans une situation de dialogue). Même si la création de la représentation de la structure des énoncés est encore manuelle, les données issues du suivi des mains dans la vidéo peuvent déjà être utilisées pour situer automatiquement les entités dans l'espace de signation. La position des entités dans l'espace de signation correspond aux positions des mains ou au centre de gravité des deux mains (dans le cas où le signe ferait intervenir un contact avec une partie du corps du signeur, une autre technique de placement comme le pointage déictique, le pointage visuel, l'emploi d'un proforme ou la rotation du buste permet de préciser l'emplacement de l'entité).
Un modèle complémentaire, nommé ACTES (Action de Transformation de l'Espace de Signation) permet de modéliser séquentiellement les différentes actions à effectuer pour créer ou modifier la position d'une entité. On citera à titre exemple l'enchaînement d'actions nécessaires pour localiser une entité déjà créée dans l'espace de signation à l'aide d'un proforme [3] :

1. La main dominante prend la configuration caractéristique de l'entité.
2. Le regard vise la localisation finale de l'entité
3. La main dominante se déplace jusqu'à l'emplacement final de l'entité

Des séquences coordonnées de comportements comme celle-ci peuvent être observées pour différents évènements modifiant l'espace de signation et sont décrits dans [2].
Le modèle ACTES peut être utilisé dans un but prédictif et permettre une approche descendante lors d'une analyse automatique de la vidéo qui peut augmenter la vitesse des traitements. Il est ainsi possible d'émettre des hypothèses sur les indices de la vidéo qui n'ont pas encore été suivis. Par exemple, si l'analyse des épaules et des mains ainsi que l'orientation du visage indiquent la création d'une nouvelle entité, on pourra plus aisément prédire la direction du regard (très difficile à mesurer dans des vidéos basse définition). Or la vérification d'un indice prédit fait généralement appel à des opérateurs moins coûteux que sa détection. Il est ainsi possible de décider uniquement de vérifier certains indices qui permettent de confirmer ou d'infirmer les différentes hypothèses.

# 6 Critique du modèle proposé

La chaîne de traitement que nous venons d'exposer fait plus ou moins explicitement référence à des hypothèses sur la structure sous-jacentes de vidéos traitées. Il est important de les lister afin d'anticiper les difficultés à surmonter pour analyser des vidéos différentes par leur mode d'acquisition, leur contexte d'énonciation ainsi que par leur contenu. Nous nous proposons donc d'énumérer nos différentes hypothèses en dégageant les contraintes éventuelles qu'elles induisent sur le type d'énoncé que le modèle permet de traiter.

## 6.1 Hypothèses effectuées au niveau du filtre particulaire

**Suivi basé sur la couleur de peau** : La méthode de suivi de la tête et des mains par couleur de peau s'avère per-

3. Un proforme est une configuration manuelle utilisée pour représenter et situer ou la forme d'une entité

tinente dans la plupart des méthodes de traitement des vidéos en LS [18]. Elle implique par contre de prévoir la possibilité d'initialiser l'algorithme avec un modèle de peau adapté au signeur. Elle nécessite également que le fond et les vêtements du signeur soient de couleur différente de celle de la peau.

**Forme des mains et de la tête** : Dans la version actuelle de notre filtre particulaire, aucune hypothèse n'est émise quant à la forme de la tête et des mains. Cela permet à un nuage d'être réparti entre plusieurs blobs de peau ce qui est en contradiction avec la connexité physique des blobs de peau. Cependant, ceci n'affecte pas les performances de l'algorithme de suivi. La position des mains est estimée comme étant le centre des nuages de particules de suivi. Cette deuxième hypothèse n'est valable que dans le cas où le signeur porte des manches longues.

**Gestion des occultations** : Le principe de pénalisation utilisé dans le cadre du filtre particulaire autorise le recouvrement partiel ou total de plusieurs blobs de peau. Ceci est indispensable car près de 37,9% [4] des images vidéos LS peuvent contenir des occultations.

**Blobs de peau visibles** Même si l'algorithme prend en compte les cas d'occultation les plus fréquents (main-main, tête-main) en autorisant les nuages à se recouvrir, il ne traite pas le cas où une main serait cachée par d'autres membres ou sortirait du cadre. Ce dernier cas survient très fréquemment dans les vidéos en LS lorsque la prise de vue ne permet pas de filmer l'intégralité du haut du corps du signeur.

**Mouvements fortement non linéaires** Les grandes non-linéarités sont une caractéristique de tout moyen d'expression gestuel. C'est d'ailleurs un des paramètres qui permet de donner du réalisme et du sens au mouvement [10].

## 6.2 Hypothèses effectuées au niveau de la désambigüisation

**Apparence peu variable des coudes** Cette hypothèse de faible variabilité de l'apparence des coudes a pu être vérifiée sur les différentes vidéos qui nous ont permis de valider notre algorithme de désambigüisation. Toutefois, il n'est possible de l'utiliser qu'à condition que le signeur ait une orientation constante face à la caméra. Ceci induit des contraintes sur le cadrage de la vidéo ainsi que sur la structure de l'énoncé qui doit contenir un nombre limité de transferts personnels [5]. Le suivi des coudes induit également des contraintes sur l'apparence vestimentaire qui doit permettre de localiser les coudes dans la silhouette du signeur.

4. Mesure obtenue sur une traduction de brève fournie par la société Websourd ; 37,9% représente la part des images où au moins un contact main/main ou main/tête est observé dans l'espace 2D image

5. Transfert personnel : structure syntaxique dans laquelle le signeur prend le rôle d'un des actants de l'énoncé

**Mouvement continu** L'hypothèse de continuité du mouvement est justifiée du point de vue physique.

**Main dominante au dessus de la main dominée** Le fait de prendre en compte la notion de main dominante est justifié du point de vue statistique dans le cas d'une analyse de vidéo LS (cf §5.2). Ceci nécessite toutefois de prévoir une variation de l'algorithme suivant la latéralité du signeur (droitier ou gaucher) à l'aide d'une initialisation automatique ou assistée.

**Mains décroisées la majorité du temps** Cet indice est également vérifié statistiquement .

**Distance main-coude** Le fait que la distance main coude puisse être utilisée pour désambigüiser les deux mains se justifie d'un point de vue physique mais ne permet que rarement de conclure avec certitude sur la latéralité de chaque main identifiée dans l'image, du fait de la projection.

## 6.3 Contraintes induites sur les vidéos analysables

A l'aide des différentes hypothèses de travail que nous venons de dégager, il est possible de caractériser les vidéos que notre méthode de suivi sera à même de traiter. Elles devront être conformes à plusieurs catégories de contraintes :

**Contraintes vestimentaires** : Le signeur devra porter un vêtement à manche longue de couleur différente de celle du fond et de la peau.

**Contrainte de cadrage** : Le signeur devra être filmé en plan américain et les mains du signeur devront constamment rester à l'intérieur du cadre.

**Contrainte sur la structure des énoncés** : La fréquence des transferts personnels et tout autre structure grammaticale impliquant une rotation hors plan importante du buste du signeur devra être limitée.

Nos travaux actuels visent à relâcher ces deux dernières contraintes.

L'analyse que nous venons de présenter est faite du point de vue du traitement d'image pour évaluer les limites d'application de notre approche et les contraintes qu'il faudrait lever. On pourrait également la faire d'un point de vue de linguiste pour expliciter le modèle de LS utilisé dans l'état actuel de notre système. Ceci permettrait de cibler les cas d'usage d'un tel système, de préciser le cahier des charges des fonctionalités à ajouter, mais aussi de valider le modèle linguistique sous-jacent.

# 7 Conclusion

En raison de la complexité des LS, il est nécessaire de tenir compte de manière consciente et explicite des contraintes sur la structure des énoncés signés. Il est aussi important de faire le lien entre ces contraintes et la possibilité d'appliquer ou non des algorithmes en situation réelle pour aboutir à un compromis entre sur-spécialisation des algorithmes et fiabilité des résultats obtenus. Pour aboutir à un

tel compromis, des connaissances sur la LS étudiée doivent être intégrées de l'analyse bas niveau à l'interprétation des énoncés. Nous avons montré dans cet article, à partir du problème de suivi des mains, comment la combinaison d'opérateurs bas niveau donnant parfois des résultats peu fiables peut aboutir à des données utilisables pour une analyse de la langue. Nous avons également montré comment de telles observations peuvent être utilisées dans le cadre d'une analyse de plus haut niveau et intégrées dans un modèle d'espace de signation.
Nous espérons pouvoir à terme utiliser ces données de haut niveau afin d'affiner la précision du suivi et coupler l'approche ascendante précédemment décrite à une approche descendante.

## Remerciements

Cette étude est financée par la société Websourd et la région Midi-Pyrénées dans le cadre du projet SESCA (Système pour l'Enseignement de la langue des Signes et la Communication par Avatar) du Pôle de REcherche Signe TOlosan (PRESTO).

## Références

[1] Chen, Q.,Wu, H.,Yachida, M., 1995, *Face detection by fuzzy pattern matching*, dans International Conference on Computer Vision 1995, p. 591

[2] Cuxac, C., 2000, *LSF, les voies de l'iconicité*, Orphys ed., Paris

[3] Dalle, P., 2006, *High level models for sign language analysis by a vision system*, dans International Conference on Language Resources and Evaluation 2006, 17–20

[4] Evgeniou, T., Pontil, M., Papageorgiou C., Poggio, T., 2000, *Image representations for object detection using kernel classifiers*, dans Asian Conference on Computer Vision 2000, p. 687–692

[5] Fearnhead, P.,1998, *Sequential Monte Carlo methods in filtter theory*, PhD thesis, University of Oxford

[6] Freund, Y., Schapire, R.E., 1996, *Experiments with a new boosting algorithm*, in proc. International Conference on Machine Learning 1996, 148–156

[7] Gall, J.,Potthoff, J., Schnörr, C., Rosenhahn, B., Seidel, H.P., 2007, *interacting and annealing particle filter*, Journal of Mathematical Imaging and Vision, **28**, 1, 1–18

[8] Gianni, F., Collet, C., Dalle, P., 2007, *Robust tracking for processing of videos of communication's gestures*, Gesture Workshop 2007, 34–37

[9] Gianni, F., Collet, C., Lefebvre-Albaret, F. , à paraître, *Modèles et méthodes de traitement d'images pour l'analyse de la langue des signes*, revue Traitement Automatique des Langues.

[10] M. Gleicher and N. Ferrier,*Evaluating Video-Based Motion Capture*,Proceedings of Computer Animation,p. 75-80(2002)

[11] Hruz, M., Campr, P.,2008, *Semi-automatic Annotation of Sign Language Corpora*,dans International Conference on Language Resources and Evaluation 2008, W25, 78–81

[12] Kitagawa, G., 1996,*Monte Carlo filter and smoother for non-Gaussian nonlinear state space models*, Journal of Computational and Graphical Statistics, **5(1)**, 1–25

[13] Kovac, J., Peer, P., Solina, F., 2003,*Human skin color clustering for face detection*, dans International Conference on "Computer as a Tool", **2**, 144–148

[14] Lefebvre-Albaret, F., Dalle, P., 2008, *Une approche de segmentation de la Langue des Signes Française*, dans Traitement Automatique des Langues Naturelles 2008, Avignon

[15] Lenseigne, B., Dalle, P., 2005, *Using Signing Space as a Representation for Sign Language Processing*, dans Gesture Workshop 2005, 25–36

[16] Micilotta, A.S., Bowden, R., 2004, *View-based Location and tracking of body parts for Visual interaction*, in proc. British Machine Vision Association 2004, 849–858

[17] Micilotta, A.S., Ong, E.J., Bowden, R., 2005, *Detection and tracking of humans by probabilistic body parts assembly*, in proc. British Machine Vision Association 2005, **1**, 429–438

[18] Ong, S.C.W., Ranganath, S., 2005, *Automatic sign language analysis : a survey and the future beyond lexical meaning*, PAMI, **27**, 6, 873–89

[19] Roberts, T.J., Mckenna, S.J., Ricketts, L.W., 2004, *Human pose estimation using learnt probabilistic region similarities and partial configurations*, dans European Conference on Computer Vision 2004, 291–303

[20] Sigal L., Sclaroff, S., 2000, *Estimation and prediction of evolving colour distribution under varying illumination*,dans Conference on Computer Vision and Pattern Recognition 2000, **2**, 152–159

[21] Terrillon, J.C., Shirazi, M.N., Fukamachi H., Akamatsu, S., 2000, *Comparative Performance of Different skin Chrominance models and Chrominance spaces for the automatic detection of human faces in colour images.*, dans International Conference on Automatic Face and Gesture Recognition 2000, 54–61

[22] Viola P., Jones, M.J., 2004, *robust real time face detection*, International Journal of Computer Vision, **57**, 137–154

[23] Yang M.H., Ahuja, N., 1999, *Gaussian mixture model for human skin colour and its application in image and video database*,dans SPIE, Visualization in Biomedical Computing, **3656**, 458–466