Modèles et méthodes de traitement d'images pour l'analyse de la langue des signes

Frédérick Gianni — Christophe Collet — François Lefebvre

IRIT (UPS - CNRS UMR 5505)
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 9
{gianni,collet,lefebvre}@irit.fr

RÉSUMÉ. *Cet article s'intéresse aux méthodes de traitement informatique des vidéos en langue des signes (LS). Dans une première partie, nous présentons une méthode de suivi robuste. Elle détecte et suit les mains et le visage d'une personne en train de signer. Elle utilise un modèle de couleur de la peau et trois filtres à particules pour suivre les mains et la tête, avec des étapes de rééchantillonnage et de recuit simulé pour améliorer leur robustesse aux occultations et aux grandes variations de dynamique des gestes, fréquentes en LS. Les évaluations des filtres mettent en valeur les améliorations apportées par ces étapes. La seconde partie présente des modèles de la LS pour exploiter ces résultats. Un modèle d'espace de signation et un modèle de construction de cet espace permettent de rendre compte de la structure spatiale de l'énoncé signé. Nous exposons plusieurs pistes pour réaliser une segmentation automatique des signes. Nous concluons sur un ensemble d'applications et de perspectives rendues possibles par ces méthodes.*

ABSTRACT. *This paper focuses on methods applied for sign language video processing. In the first part, we present a robust traking method which detects and tracks the hands and face of a person performing Signs' language communication. The method uses a model of skin color and three particle filters to track the two hands and the face, with re-sampling and annealed update steps to increase their robustness to occultation and high acceleration variations of body parts that occure frequently in sign language (SL). Evaluations of the trackers show the improvement bringed by these enhancements. The second part presents SL models to process those results. A model of signing space allows us to represent the spatial structure of a signed sentence. We explain several ways which could be used to perform an automatic sign segmentation. To conclude, we expose a range of application perspectives enabled by such methods.*

MOTS-CLÉS : *langue des signes, espace de signation, annotation, suivi, filtre à particules.*

KEYWORDS: *Sign Languages, signing space, video annotation, tracking, particle filter.*

1. Introduction

Les recherches réalisées sur les langues des signes (LS) (Ong *et al.*, 2005), par des linguistes ou par des informaticiens, impliquent la réalisation d'un laborieux travail d'annotation de séquences vidéo, *à la main*, image par image. Ce travail était naguère réalisé à l'aide d'un traitement de texte ou d'un tableur (voire sur papier) en manipulant un magnétoscope. Il est aujourd'hui facilité par des outils informatiques d'annotation tel que ANVIL (Kipp, 2001) ou ELAN (Brugman *et al.*, 2002) qui intègrent et synchronisent le lecteur vidéo et l'annotation sous forme de partition. Malgré ces progrès, l'annotation nécessite toujours un expert humain pour interpréter les vidéos. Or, les critères d'annotation utilisés par l'expert sont en général visibles dans la séquence d'images et doivent pouvoir – dans une certaine mesure – donner lieu à un traitement automatique employant, par exemple, les techniques du traitement d'images et du suivi (Braffort *et al.*, 2004).

Les recherches en analyse automatique de la LS ont pour but d'élaborer des méthodes de traitement permettant l'interprétation des vidéos de LS. Cette approche implique que les résultats des traitements soient très robustes et donnent des valeurs précises. Le traitement d'images de LS présente de nombreuses difficultés : variabilités intra et interpersonnelles, informations occultées, vitesse des mouvements, comportement non linéaire, modélisation linguistique complexe et incomplète... La solution adoptée dans ces études consiste donc à réduire le champ d'application, d'un point de vue linguistique – lexique restreint, grammaire simplifiée, pas de contexte, pas de visée illustrative... – et les conditions de réalisation – peu de signeurs, tournage en studio, multiplication des caméras... – de manière à rendre possible l'étude de toute la chaîne de traitement depuis le traitement d'images jusqu'à la reconnaissance de signes ou d'énoncés en LS.

Étant confrontés à la nécessité de disposer de corpus vidéo annotés et donc au lourd travail d'annotation, nous cherchons tout d'abord à appliquer des méthodes d'analyse vidéo et de suivi pour faciliter cette tâche. Concrètement nous étudions des méthodes permettant de réaliser une annotation semi-automatique : l'outil propose une annotation que l'expert valide. Dans ce cadre-là, le résultat de la partie automatique de l'annotation n'est pas nécessairement juste à 100 %. Par exemple, pour la détection d'événements visuels – position de la main, mouvement du tronc, expression du visage... – il est possible d'avoir de fausses détections (le traitement détecte un événement alors que celui-ci n'a pas lieu), qui seront invalidées par l'expert. Les interventions étant très rares et faciles à réaliser, elles sont tolérables et ne dégradent pas le gain apporté par les outils de traitements automatiques. En revanche, les erreurs de détection (l'événement n'est pas détecté alors qu'il a lieu) doivent être évitées car cela impose à l'expert non seulement de valider les résultats mais aussi d'analyser l'ensemble du corpus, ce qui rend ce type d'outil inefficace. L'aspect interactif de l'outil d'annotation permet d'utiliser les connaissances de l'expert en amont du traitement – restriction de la zone de traitement ; initialisation des paramètres ou des modèles pour la détection, le suivi ou la reconnaissance – ou *a posteriori* – validation des résultats. Nous présentons une

méthode de traitement d'images permettant d'extraire des informations pertinentes d'une séquence vidéo : la détection et le suivi des mains et de la tête.

2. Détection et suivi de la tête et des mains

Dans les vidéos de LS que nous étudions les gestes sont effectués de manières naturelle, sans contrainte. Lors d'un énoncé en LS, les mouvements des mains peuvent être très rapides avec de brusques changements de direction. L'un des problèmes majeurs est de proposer des méthodes de suivi robustes pour ce type de mouvement. Le suivi de mouvements humains nécessite une détection précise de caractéristiques visuelles et une mise en correspondance d'image à image en utilisant les informations comme la position, la vitesse et/ou la couleur. Dans l'approche que nous proposons, la détection est réalisée par la couleur, et la mise en correspondance des caractéristiques est effectuée en utilisant des estimateurs statistiques par le biais de filtres à particules : un pour la tête et un pour chacune des mains. Puisque les filtres particulaires modélisent l'incertain, ils fournissent un cadre robuste pour le suivi des mains d'une personne communicant en LS. Nous présentons une méthode améliorée de suivi par filtrage particulaire. Nous détaillons la méthode utilisée pour modéliser et suivre les mains et la tête. Puis, dans une seconde partie, nous présentons des résultats de l'évaluation de la robustesse de cette méthode de suivi.

2.1. *Détection de la tête et des mains*

Nous utilisons la couleur « peau » comme indice pour détecter le visage et les mains d'une personne. C'est un attribut robuste si on le compare, par exemple, aux contours, étant données les variations géométriques d'un visage ou d'une main. Pour mettre en œuvre ce type de détection, il est tout d'abord nécessaire de choisir un espace de couleur, puis la représentation à utiliser pour modéliser la couleur, et enfin la manière d'exploiter le résultat produit par le détecteur. L'utilisation de cet attribut induit, bien entendu, certaines contraintes que nous détaillerons.

Le but d'un détecteur de couleur de peau est de construire une règle de décision pour faire la différence entre les pixels de couleur peau et les autres. On introduit habituellement une métrique pour mesurer la distance entre la couleur d'un pixel et la couleur de la peau. La métrique utilisée est définie par la méthode de modélisation de la couleur de la peau.

Nous pouvons distinguer trois types de méthodes pour modéliser la couleur : la modélisation **explicite**, **non paramétrique** et **paramétrique**.

2.1.1. *Modélisation explicite*

La modélisation explicite définit explicitement les limites d'un groupe de pixels dans un espace de couleur. Par exemple, (Kovac *et al.*, 2003) classe un pixel (R, G, B) comme un pixel de couleur peau si :

$R > 95$ et $G > 40$ et $B > 20$ et $max\{R, G, B\} - min\{R, G, B\} > 15$ et $|R - G| > 15$ et $R > G$ et $R > B$

2.1.2. *Modélisation non paramétrique*

La modélisation non paramétrique estime la distribution de couleurs recherchée à partir d'un ensemble d'apprentissage sans en dériver explicitement un modèle. Nous pouvons citer (Chen *et al.*, 1995) qui utilise des histogrammes pour segmenter les pixels de couleur peau. L'histogramme normalisé est utilisé en distribution de probabilité discrète. Un classifieur Bayésien peut être construit permettant de calculer la probabilité d'observer de la peau étant donné une couleur c. Pour calculer cette probabilité, la règle de Bayes est utilisée :

$$p(skin|c) = \frac{p(c|skin)p(skin)}{p(c|skin)p(skin) + p(c|\neg skin)p(\neg skin)}$$

$p(c|skin)$ et $p(c|\neg skin)$ sont calculées directement à partir des histogrammes peau et non peau. Les probabilités *a priori* $p(skin)$ et $p(\neg skin)$ peuvent également être calculées à partir de tous les échantillons de peau et non peau contenus dans l'ensemble d'apprentissage.

2.1.3. *Modélisation paramétrique*

La modélisation paramétrique permet une représentation plus compacte avec des possibilités de généralisation et d'interpolation des données d'apprentissage. Le modèle utilisé est généralement une loi gaussienne ou bien un mélange de lois gaussiennes. La couleur de peau est modélisée par une fonction de probabilités définie ainsi :

$$p(c|skin) = \frac{1}{2\pi^{\frac{d}{2}}|\Sigma_s|^{\frac{1}{2}}} \exp -\frac{1}{2}(c - \mu_s)^T \Sigma_s^{-1}(c - \mu_s)$$

avec c un vecteur de couleur de dimension d et μ_s, Σ_s les paramètres de la distribution (respectivement vecteur moyen et matrice de covariance). La probabilité $p(c|skin)$ peut être utilisée directement comme mesure de ressemblance entre la couleur c et la couleur de la peau. On utilise souvent la distance de Mahalanobis entre la couleur c et le vecteur moyen μ_s connaissant Σ_s.

$$\lambda_s = (c - \mu_s)^T \Sigma_s^{-1}(c - \mu_s)$$

Un modèle plus sophistiqué, capable de décrire des distributions de couleurs complexes, est le modèle de mélange de k lois gaussiennes. Ce modèle généralise le précédent. La fonction de distribution de probabilité est alors la suivante :

$$p(c|skin) = \Sigma_{i=1}^{k} \pi_i . p_i(c|skin)$$

Chacune de ces k distributions est pondérée par un poids π_i représentant les paramètres du mélange et obéissant à la contrainte de normalisation $\Sigma_{i=1}^{k} \pi_i = 1$. L'apprentissage du modèle est en général effectué à l'aide de la méthode itérative EM (Estimation Minimisation) qui suppose le nombre k de lois connu à l'avance.

2.1.4. *Observations*

Les méthodes de modélisation explicite sont trop rigides et les méthodes paramétriques produisent trop de faux négatifs à cause de leur capacité de généralisation. Aussi, dans la mise en œuvre des filtres particulaires utilisés pour le suivi (section 2.2), nous utilisons, comme observations, la méthode non paramétrique décrite en section 2.1.2. La règle de décision naturellement définie lors de l'utilisation de méthodes de détection probabiliste n'est ici pas définie explicitement. En effet, ces probabilités seront directement utilisées comme observation dans les filtres particulaires. Le nuage de particules se localisera alors sur les plus hautes probabilités.

2.2. *Filtrage particulaire*

Le problème du filtrage non linéaire consiste à estimer la loi conditionnelle d'un processus état indirectement lié à un processus observation dont on connaît une réalisation. Le filtre de Kalman permet de résoudre ce problème de façon exacte et rapide lorsque les dynamiques de l'état et de l'observation sont linéaires et gaussiennes. En dehors de ce cas, d'autres approches telles que les méthodes particulaires ont été développées. Fondées sur le principe de Monte-Carlo, les méthodes particulaires proposent une approximation faible de la loi conditionnelle recherchée en propageant un système de particules dans le temps. Elles permettent de s'affranchir des hypothèse de bruit gaussien et de linéarité du système, contraintes beaucoup trop fortes si l'on veut traiter des corpus naturels de langue des signes.

Les filtres particulaires permettent l'estimation de séquences de paramètres cachés $\mathbf{x}_t$ à partir des données observées $\mathbf{z}_t$. L'idée est d'approximer une distribution de probabilités par un ensemble d'échantillons pondérés : $\{(\mathbf{s}_t^{(0)}, \pi_t^{(0)}) \dots (\mathbf{s}_t^{(i)}, \pi_t^{(i)})\}$ avec $i = 0, \dots, n$ le nombre d'échantillons utilisés. Chaque échantillon s représente un état de l'objet poursuivi avec une probabilité π.

L'état est modélisé par $\mathbf{s}_t = [x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}]^T$, la position, la vitesse et l'accélération de l'échantillon $\mathbf{s}$ dans l'observation au temps t. Nous pensons qu'il faut prendre en compte ces trois niveaux pour rendre compte de la dynamique des gestes de la LS.

Trois états sont maintenus durant le suivi, un pour chaque partie du corps observée. Nous suivons la tête et les mains séparément, ces trois composantes étant fondamentales en LS. Chacune de ces zones est représentée par un ensemble d'échantillons.
Nous décrivons la méthode générique du filtrage particulaire puis les améliorations que nous avons apportées pour prendre en compte les particularités de la LS. Nous donnons en annexe le détail des algorithmes correspondants.
L'algorithme du filtre particulaire générique (annexe algorithme 1) procède en quatre étapes : « **Initialisation** », « **Prédiction** », « **Mise à jour** » et « **Reéchantillonnage** ».
Lors de l'**initialisation**, nous échantillonnons n fois à partir de la distribution de départ η_0. L'échantillonnage de $\mathbf{s}^{(i)}$ à partir d'une distribution μ, pour $i = 0, \ldots, n$ revient à simuler n échantillons aléatoires indépendants, c'est-à-dire les particules, à partir de μ. Nous obtenons ainsi n variables aléatoires $\mathbf{x}_0^{(i)}, 1 \leq i \leq n$ indépendantes selon η_0.
Ensuite, les valeurs des particules sont **prédites** pour la prochaine itération, en explorant l'espace d'état. Cette exploration s'effectue de manière indépendante selon une loi de transition $p(\mathbf{x}_t|\mathbf{x}_{t-1} = \mathbf{s}_{t-1}^{(i)})$ définissant la dynamique des particules. On obtient alors un état $\tilde{\mathbf{s}}_t$ issu de la prédiction.
À l'étape **mise à jour**, chaque particule prédite est pondérée par la vraisemblance $p(\mathbf{z}_t|\tilde{\mathbf{s}}_t^{(i)})$, qui est obtenue lors de l'observation.
Le **reéchantillonnage** peut être vu comme un cas spécifique d'une étape de sélection. Les particules sont choisies selon leur poids $\pi_t^{(i)}$. Cette étape crée d'autres particules au dépend des particules ayant un faible poids, qui sont supprimées.
Lors de l'implémentation de filtres à particules, il est nécessaire de pouvoir :

- simuler la loi initiale $p(\mathbf{x}_0) = \eta_0$
- simuler les lois de transition $p(\mathbf{x}_t|\mathbf{x}_{t-1} = \mathbf{s}_{t-1}^{(i)})$
- évaluer la fonction de vraisemblance $p(\mathbf{z}_t|\mathbf{x}_t^{(i)} = \tilde{\mathbf{s}}_t^{(i)})$

Nous utilisons une loi normale comme distribution de départ η_0. Lors de l'étape de prédiction, les échantillons sont propagés selon un modèle dynamique : un processus autorégressif du premier ordre : $\mathbf{x}_t = \mathbf{S}\mathbf{x}_{t-1} + \eta$, où η est une variable aléatoire gaussienne multivariée et $\mathbf{S}$ une matrice de transition. Nous utilisons la couleur de la peau comme paramètre d'observation, afin d'être robuste à la non-rigidité et au changement d'orientation des objets observés (les mains particulièrement). La densité d'observation $p(\mathbf{z}_t|\mathbf{x}_t)$ est modélisée par la distribution de la couleur de peau des pixels en utilisant la méthode décrite section 2.1.2 (fig.1).

Dans un filtre à particules, une étape de rééchantillonnage est requise pour éviter le problème de dégénérescence des échantillons, c'est-à-dire, pour éviter la situation où tous les poids d'importance sont proches de zéro sauf un. Nous utilisons le rééchantillonnage stratifié proposé par Kitagawa (Kitagawa, 1996) (annexe algorithme 2), car il est optimal en terme de variance sur la redistribution des poids de l'ensemble d'échantillons.

Cependant un tel filtre à particules ne permet pas de prendre en compte de grandes variations soudaines de dynamique. Les particules n'étant pondérées et diffusées qu'une seule fois, il est possible qu'elles ne soient plus positionnées sur la cible. Nous allons voir dans la prochaine section une amélioration de ce filtre permettant de résoudre un tel problème.

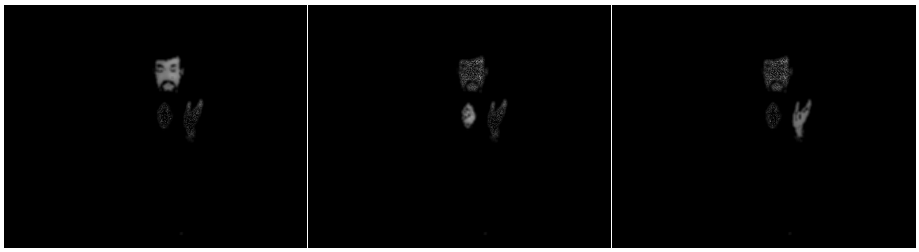
2.2.1. *Filtre particulaire avec une étape de recuit simulé*

Pour maintenir, durant le suivi, une bonne représentation de la probabilité *a posteriori*, il est possible d'itérer l'algorithme un certain nombre de fois. Mais ceci conduit à une surreprésentation de possibles maxima locaux. Ce phénomène est induit par l'application de la fonction de pondération à chaque itération. L'effet recuit proposé par (Deutscher *et al.*, 2000) et par (Gall *et al.*, 2006) avec une formulation plus générique fournit un moyen d'appliquer graduellement la fonction de pondération à l'ensemble d'échantillons (annexe algorithme 5).

L'étape de mise à jour recuite permet d'appliquer une fonction de pondération permettant de prendre en compte les particules de poids moyen et de diffuser dans leur alentours. Ceci a pour effet, lorsque nous itérons dans la boucle de recuit, de déplacer l'ensemble des particules en prenant en compte les particules situées aux limites des particules sélectionnées lors du rééchantillonnage. Le filtre peut alors se déplacer « plus rapidement » que le filtre particulaire simple.

2.2.2. *Suivi de multiples objets*

En LS, nous devons suivre trois objets (la tête et les deux mains). Si la tête est relativement stable, les deux mains produisent des mouvements très variés s'occultant ou occultant le visage. La présence de plusieurs objets à suivre engendre un problème d'association de données qui rend le suivi encore plus difficile à appliquer. Le suivi de multiples objets et les techniques d'association de données sont étudiés intensément dans (Bar-Shalom, 1987) et un large nombre de techniques statistiques d'association telles que les filtres d'association probabilistes, les filtres d'association de probabilités joints, les filtres de suivis à hypothèses multiples ont été développées. Elles utilisent en général une combinaison d'identifications de *« blobs »* et effectuent des hypothèses sur le mouvement de la cible. On peut penser éviter ces problèmes en identifiant les cibles à des *« blobs »* fusionnant et se divisant (Haritaoglu *et al.*, 1998). Or, une interprétation en *« blob »* ne maintient pas l'identité de la cible tout au long du suivi. C'est de plus peu évident à implanter pour des cibles qui ne sont pas aisément séparables. Nous optons pour un « principe d'exclusion » tel que celui fourni par MacCormick (MacCormick *et al.*, 2000). De cette manière nous ne tolérons pas que deux cibles puissent fusionner lorsque leurs vecteurs d'état deviennent similaires. Nous calculons dans un premier temps la densité postérieure pour chacun des filtres et l'utilisons pour pénaliser la mesure de chacun des autres filtres lors d'un deuxième calcul de cette densité (3).
L'effet peut être constaté figure 1, où pour chaque ensemble de particules les observations sont pénalisées par les particules maintenues par les autres filtres.



(1) (2) (3)

Figure 1. *Observations de la couleur de la peau pour la tête (1), la main droite (2) et la main gauche (3). Les niveaux de gris représentent les probabilités des pixels d'appartenir au modèle de couleur de peau appris. (1) Les probabilités pour les pixels des mains ont été pénalisées par la méthode de suivi multiple, (2) et (3) idem pour la tête et la main droite et gauche*

2.3. *Évaluation*

Nous avons évalué le suivi proposé sur une séquence vidéo d'une personne racontant une histoire en LSF – extrait du corpus LS-COLIN (Braffort *et al.*, 2001). Nous nous sommes intéressés à la robustesse de notre approche vis-à-vis des mouvements à hautes variations de dynamique et des occultations des parties du corps (fig.2). Le suivi multiple dans ce contexte est un challenge : grande similarité des cibles ; grandes variations de dynamique des gestes, cibles souvent occultées ; séquence très longue (plus de 3 000 images).

Le filtre à particules avec une étape de recuit simulé (*Annealed Particle Filter* **APF**) offre une meilleure robustesse que le filtre à particules simple (*Particle Filter* **PF**), face aux grandes variations de dynamique et également aux problèmes des maxima locaux (fig.3).

Pour évaluer le suivi image par image, nous étudions si un lien entre deux objets physiques détectés dans deux images consécutives est correctement calculé ou non. De cette manière nous pouvons comptabiliser les différentes erreurs de suivi pouvant être produites. La métrique utilise les informations de comparaison suivantes :

1) Les objets détectés au temps t et $t + 1$ identifient les mêmes données de référence en utilisant la distance euclidienne entre les centres de gravité des ensembles de particules aux temps t et $t + 1$ et un seuil.

2) Un lien existe entre les objets détectés aux temps t et $t + 1$ et un lien existe également dans les données de référence.

Si plusieurs liens entre les objets détectés identifient les mêmes données de référence, celui qui maximise le recouvrement avec les données est sélectionné comme lien cor-

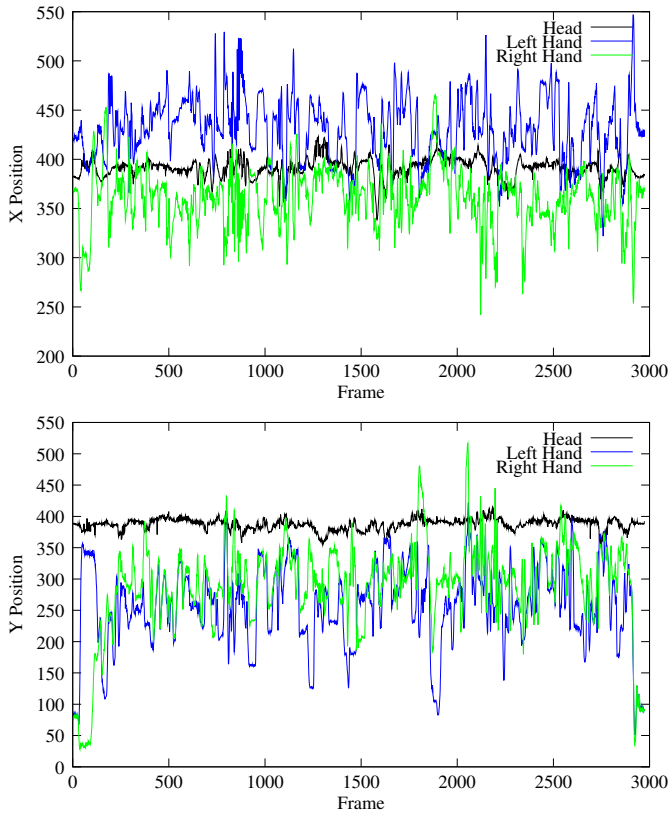


Figure 2. *Les trajectoires réelles des centres de gravité de la tête et des mains, coordonnées (x,y). Nous pouvons constater le faible mouvement de la tête et les grandes variation de dynamiques des deux mains. Des occultations de la tête par les mains peuvent également être distinguées ; par exemple aux environs de l'image 2 000, les trajectoires des mains en x et y coupent la trajectoire de la tête*

rect et ne sera pas utilisé pour les autres associations. En utilisant ces informations nous calculons quatre métriques (fig.4) :

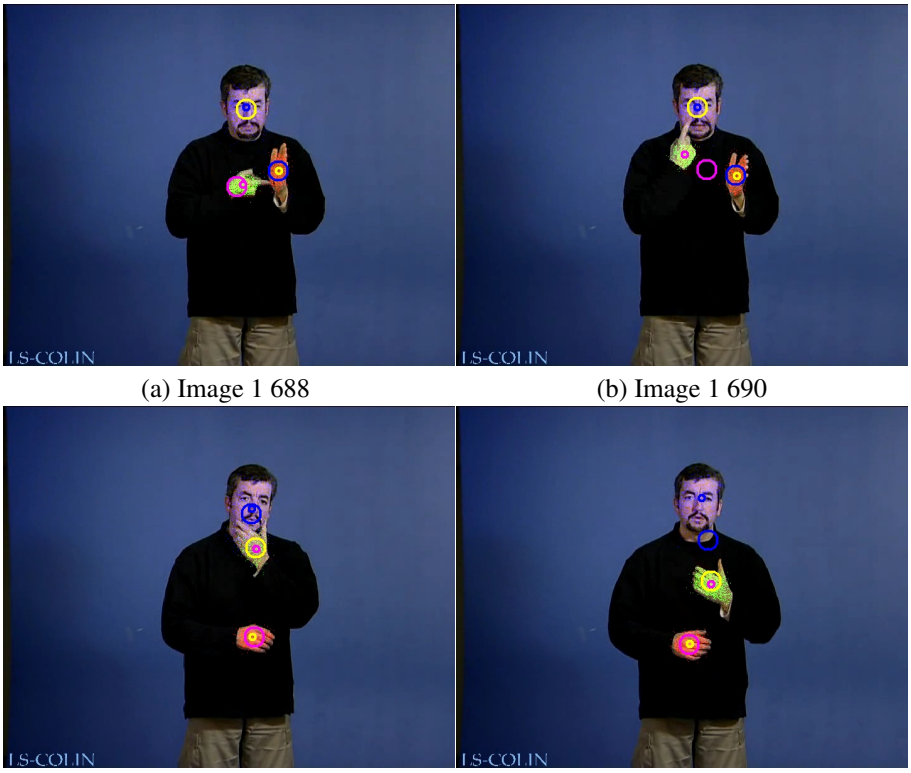
• **Suivi Bon** SB, un lien entre deux objets physiques identifie un lien des données de référence ;

• **Suivi Mauvais** SM, un lien entre deux objets physiques ne correspond à aucune données de référence ;

• **Suivi Raté 1** $SR1$, un lien présent dans les données de référence n'a pas été trouvé, rejet du cas (1) ;

• **Suivi Raté 2** $SR2$, un lien présent dans les données de référence n'a pas été trouvé, rejet du cas (1) et (2).

Nous différencions les « décrochages » spécifiés par les **Suivi Raté 1** et **Suivi Raté 2** car ils ne résultent pas des mêmes erreurs. Le **Suivi Raté 1** traduit



(a) Image 1 688 (b) Image 1 690

(c) Image 116 (d) Image 125

Figure 3. *Le filtre à particules recuit illustré par les petits cercles est plus robuste face aux grandes variations de dynamique (a,b) et aux maxima locaux (c,d) que le filtre à particules simple illustré par les grands cercles*

une erreur dans le cadre du suivi d'objets multiples tandis que le **Suivi Raté 2** sera dû à la dynamique du système. Dans le cadre de l'analyse de LS, l'erreur **Suivi Raté 1** sera pénalisante car une interprétation basée sur les trajectoires données par le suivi sera erronée (la main prenant la place de la tête et vice versa). En revanche, l'erreur **Suivi Raté 2** est moins pénalisante, en effet les filtres ne sont pas en situation de décrochage très souvent et, lorsqu'ils sont positionnés sur une région du fond de scène, ils restent cependant sur une trajectoire plausible du mouvement ; ils sont juste « en retard » par rapport à la position réelle de la cible.

À partir de ces métriques nous calculons des valeurs de :

$$\text{Précision } \frac{SB}{SB+SM}, \quad \text{sensibilité 1 } \frac{SB}{SB+SR1} \quad \text{sensibilité 2 } \frac{SB}{SB+SR2}$$

$$\text{sensibilité globale } \frac{SB}{SB+SR1+SR2}$$

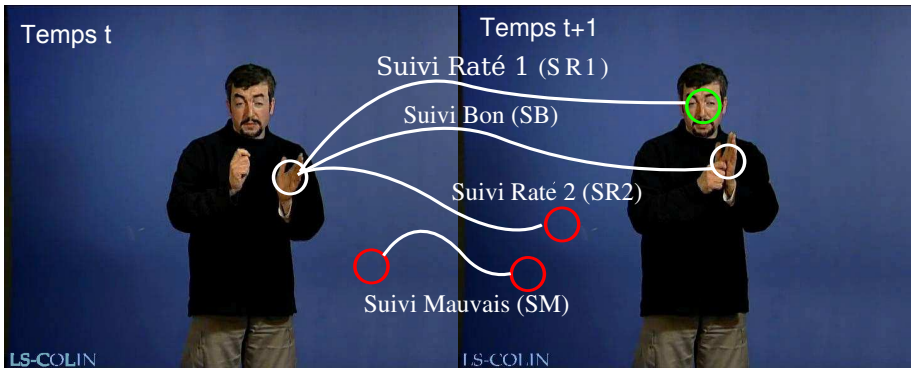


Figure 4. *Métriques utilisées pour l'évaluation de la qualité du suivi image par image*

Nous avons effectué des évaluations avec différents nombres de particules : 2 000, 3 000 et 6 000 pour étudier le comportement des filtres. Le nombre optimal de particules est autour de 3 000 pour les mains et 6 000 pour la tête. Ces chiffres correspondent à la taille en pixels des régions suivies (tableau 1). Le nombre de particules variera suivant la taille des régions suivies dans l'image.

L'erreur en position a été calculée à partir de la distance euclidienne entre la position détectée et la vérité terrain. Les pics d'erreurs sont causés par des situations où les mains et la tête sont proches les unes des autres ou lorsqu'elles s'occultent et s'éloignent suivant une accélération (fig. 5). Cependant les filtres particulaires ne ratent pas leur cible longtemps, ils les retrouvent dans les images qui suivent. Cette capacité de la méthode à se « réinitialiser » automatiquement est particulièrement importante dans le cas d'analyse de corpus en LS qui sont toujours assez longs. Avec un seuil de 50 pixels de distance, signifiant qu'à partir de cette distance la cible est considérée comme ratée par le filtre, nous comptons 224 erreurs pour la tête (7,4 %), 25 et 79 pour chaque main (0,8 % et 1,7 %) sur 3 000 images.

3. Conclusion et perspectives sur le suivi

Nous avons présenté une procédure effectuant un suivi visuel de la tête et des mains, objets très similaires, en utilisant des filtres à particules. Le filtre à particules a été proposé avec une étape de recuit simulé, dans le but d'améliorer la robustesse vis-à-vis des maxima locaux et des grandes variations de dynamique. L'évaluation quantitative des résultats de l'application de cette procédure sur des images réelles a montré l'amélioration apportée par cette méthode comparée à l'original. Les résultats sont prometteurs et peuvent être améliorés sur plusieurs points :

– un meilleur cadre de suivi multicible pourrait être proposé pour réduire les erreurs d'identification des cibles. En effet, le suivi n'utilise pas de connaissances spéci-

Nombre de particules		Tête	Main droite	Main gauche
2 000	P	0,99963	0,99925	0,99961
	S1	0,94091	0,98491	0,91875
	S2	0,95833	0,86624	0,97810
	SG	0,90393	0,85489	0,90024
3 000	P	1	0,99862	1
	S1	0,93890	0,97428	0,99323
	S2	0,98892	0,99242	0,99256
	SG	0,92912	0,96708	0,98589
6 000	P	1	0,99929	0,99860
	S1	0,95024	0,96801	0,95929
	S2	0,97775	0,98647	0,98985
	SG	0,93013	0,95532	0,94995

Tableau 1. *Précision et sensibilité du suivi avec différents nombres de particules*

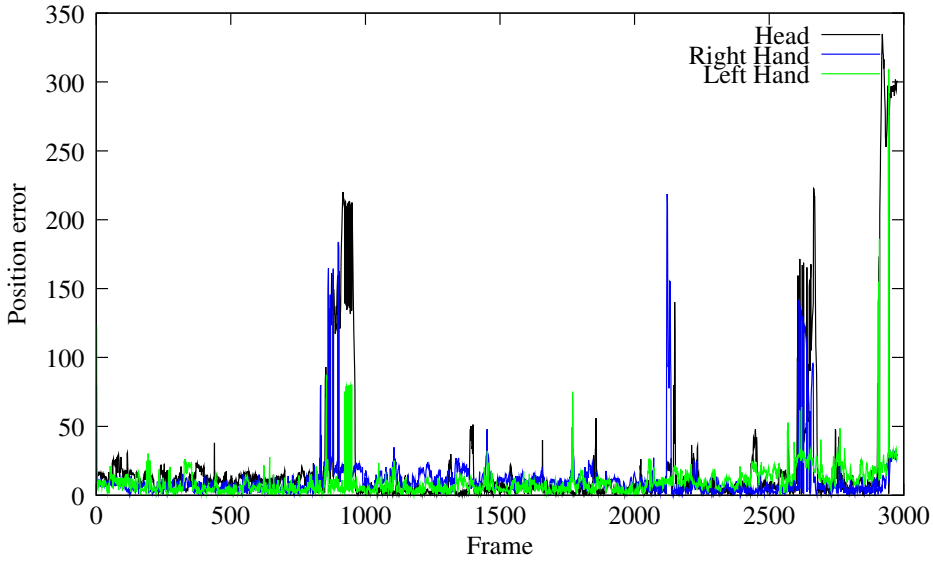


Figure 5. *Erreur en position pour la tête, la main droite et la main gauche*

fiques sur les cibles telles que le fait que la tête bouge beaucoup moins que les mains ce qui est le cas de l'observation d'un discours en LS, ni la notion de main droite/main gauche ou main dominante/main dominée ;

– l'optimisation du nombre de particules par cible en fonction de la dynamique du mouvement permettrait d'accélérer le processus pour aller vers une exécution en temps réel ;

– l'exploitation de connaissances de plus haut niveau, issues d'une interprétation, permettrait de corriger certaines erreurs de suivi et de prédire des localisations ou des mouvements. Par exemple, l'interprétation de l'espace de signation permet d'anticiper l'utilisation de loci – donc mouvement d'une main vers l'un des loci – ou de désignations – mouvement de tête, épaule ou signe en direction d'un des loci. Ces informations peuvent provenir d'une passe d'interprétation de l'expert *via* un éditeur d'annotation tel que AnColin (Braffort *et al.*, 2004) – éditeur de partitions – ou VIES (Lenseigne *et al.*, 2006) – éditeur de l'espace de signation.

4. Espace de signation

Pour exploiter les résultats fournis par les outils d'analyse du signal vidéo (détection, suivi...), il est nécessaire de développer des modèles permettant de décrire de manière satisfaisante le contenu des énoncés signés (Lenseigne *et al.*, 2005). En raison des spécificités de la LSF, il est nécessaire de repenser les outils traditionnellement utilisés dans le domaine du traitement automatique des langues pour arriver à modéliser correctement les relations spatiales structurant la phrase signée. Après avoir mis en avant les éléments distinguant la langue des signes des langues orales, nous aborderons les contraintes qu'ils induisent sur les modèles appliqués à la langue des signes. Nous exposerons ensuite le modèle d'espace de signation développé dans notre laboratoire. Nous conclurons par une réflexion sur les améliorations envisageables de ce modèle et ses perspectives d'application.

4.1. *Spécificités de la langue des signes*

La conception des modèles à appliquer à la LSF est extrêmement délicate en raison des spécificités qui font l'originalité de cette langue.

Le fait de passer par le canal visuel induit une prise en compte de plusieurs paramètres qui contribuent à l'élaboration du message. Si on en dénombre traditionnellement cinq (configuration, mouvement, orientation et placement des mains, expression du visage), il est important de souligner que c'est le corps tout entier qui participe à la réalisation de la phrase signée. Ainsi, un transfert personnel (Sallandre, 2001) modifiera à la fois la posture générale du signeur, la façon de réaliser ses signes et son expression du visage. De même, le regard est utilisé pour distribuer les rôles ou pour pertiniser des emplacements de l'espace de signation. Si les premières approches descriptives des langues signées se focalisaient presque exclusivement sur les aspects phonologiques des signes manuels (Stokoe., 1960), les modèles actuels prennent de plus en plus en compte les autres composantes non manuelles que nous venons d'évoquer.

Une autre spécificité concerne la structure de la LS fondée sur l'iconicité. Comme l'explique C. Cuxac (Cuxac, 2003), la langue des signes est basée sur l'iconicité. L'organisation iconique des langue des signes rejaillit inévitablement sur sa structure syntaxique. En plus d'une organisation temporelle de la phrase signée, il faut également respecter le placement des différentes entités discursives dans l'espace. Ainsi, chaque concept pourra se voir assigner une position dans l'espace et la simple « réactivation » de cette position permettra d'y faire référence. Ceci entraîne une grande variabilité des signes exécutés dont plusieurs paramètres peuvent être modifiés en fonction du contexte de leur utilisation. Ces flexions sont particulièrement importantes dans le cas des verbes directionnels. Ainsi, le verbe *donner* pourra se signer de plusieurs manières suivant l'actant et l'objet impliqué (figure 6).

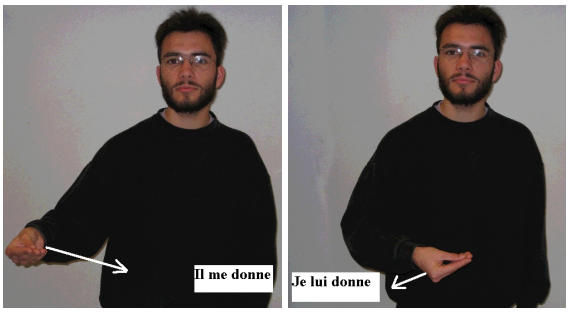


Figure 6. *Verbes directionnels*

Cette variabilité s'applique également pour des signes « à fort potentiel iconique » dont l'orientation peut également être modifiée pour être en correspondance avec la position réelle de l'objet désigné (figure 7).

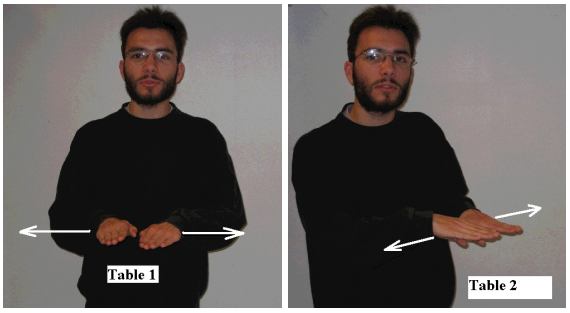


Figure 7. *Iconicité*

D'autres paramètres tels que l'expression du visage, la vitesse de réalisation du signe ou la fluidité des gestes rentrent également en ligne de compte dans les mo-

dulations des signes, mais les flexions spatiales présentent un intérêt particulier dans le cadre du développement de notre modèle de l'espace de signation. Comme le révèlent les études sur les flexions des signes directionnels, il est impossible d'énoncer une règle prédictive qui puisse s'appliquer à tous les signes (Huenerfauth, 2006). Cependant, l'analyse de la directionnalité des signes en tenant compte des entités déjà placées dans l'espace peut permettre de reconstituer la syntaxe de la phrase signée.

Le troisième point spécifique à la LS et découlant également de l'utilisation du canal gestuel est la possibilité d'exprimer simultanément plusieurs informations. Ce cas de parallélisme d'informations est particulièrement fréquent dans le cas de transferts personnels où on peut noter une dissociation entre les signes non manuels (posture, expression du visage) qui qualifient l'état de la personne représentée par le transfert, et les concepts représentés par les signes manuels. On observe aussi fréquemment le cas où les deux mains servent à représenter deux concepts différents.

Si on se place dans le cadre d'une modélisation du contenu d'une phrase en LS, il est indispensable de concevoir un modèle qui prenne en compte à la fois l'exploitation des différents paramètres de la langue des signes, de la spatialisation des signes et de la simultanéité de plusieurs informations. Ceci permet d'expliquer la raison pour laquelle les outils utilisés traditionnellement pour la modélisation des langues orales sont souvent inadéquats pour une modélisation des LS ou demandent une certaine adaptation.

4.2. *Une approche différente des langues orales*

Nous dressons dans les lignes qui suivent quelques points à surmonter pour l'adaptation des outils existants du traitement automatique des langues ou pour la création de nouveaux modèles. La spatialisation de la LS se superpose à la nature linéaire d'un énoncé signé analogue à celle des langues orales. Cela impose de prendre en compte l'agencement des signes dans l'espace en plus de leur succession temporelle. Comme le soulignent plusieurs études (Voisin, 2006; Lejeune, 2004), c'est souvent l'organisation spatiale des concepts qui conditionnera l'ordre de réalisation des signes d'un énoncé signé. On citera à titre d'exemples certaines règles qui consistent à énoncer les concepts du global au détail, du contenant au contenu, ou encore, du plus stable au plus mobile. Il semble donc évident que certaines informations de nature spatiale, peu prises en compte dans le domaine du traitement automatique des langues orales, doivent être rajoutées pour permettre de rendre compte correctement de l'organisation de la phrase signée.

En plus de la position spatiale des concepts dans l'espace, il sera également nécessaire de tenir compte de leur type : temporel, spatial, actant... Il faudra également considérer l'arité des actions c'est-à-dire le nombre des concepts mis en relation par ces actions.

Des modèles développés ces dernières années pour les LS s'orientent vers une représentation spatiale de la phrase signée et viennent en complément des modèles

utilisés dans le cadre du traitement automatique des langues. On citera par exemple le projet Visicast (Elliott *et al.*, 2000) qui allie à une description de la phrase de type « *Discours Representation Structure* » des information sur la spatialisation des signes. Plus récemment, des modélisations de l'espace de signation ont été appliquées à la description des phrases signées impliquant l'utilisation de proformes comme *« la voiture se gare entre le chat et la maison »* (Huenerfauth, 2006). L'intérêt de telles études est qu'elles permettent de modéliser très clairement le processus de spatialisation des concepts pour aboutir à une phrase signée.

Le modèle d'espace de signation proposé par l'IRIT (Lenseigne, 2004) se place également dans cette perspective d'une modélisation spatiale et temporelle d'une phrase signée.

4.3. *Propoposition de modèles*

Ce modèle permet de rendre compte étape par étape du processus de placement des différents concepts dans l'espace de signation (Dalle, 2006). Nous présentons les spécificités de ce modèle ainsi que ses perspectives d'applications.

4.3.1. *Cadre de conception du modèle d'espace de signation*

Dans un premier temps, (Lenseigne, 2004) choisit de concevoir une représentation du modèle de (Cuxac, 2003) applicable à des phrases simples portant sur l'interrogation d'une base de données renseignant sur les films de cinéma (réalisateur, horaires, lieu…). Ce choix permet à la fois de limiter les types de requêtes possibles et de préserver une grande diversité dans les types de concepts étudiés (lieux, dates, personnes, objets…). Cette modélisation ne prétend donc pas à une représentation complète de la langue des signes.

Le modèle représente la manière dont les concepts s'agencent dans l'espace et dans le temps mais ne représente pas la façon dont la phrase a été signée. Ainsi, on notera par exemple l'activation d'un emplacement de l'espace de signation sans indiquer la manière dont cette activation a été effectuée (pointage déictique, regard, proforme, signe…).

Les types de concepts modélisés sont assez généraux pour permettre la modélisation d'un grand nombre de phrases. Toutefois, le modèle n'inclut pas encore la modélisation de transferts personnels, ni toutes les structures de grande iconicité qui forment à elles seules un problème à part entière. Une fonctionnalité intéressante du modèle est qu'il permet de modéliser également une conversation entre deux personnes. Un locuteur pourra ainsi faire référence à un concept placé par l'autre locuteur grâce à la modélisation d'un espace de signation partagé.

La partie suivante décrit les lignes directrices du modèle développé et un exemple concret permettra de montrer pas à pas la construction de l'espace de signation

4.3.2. *Présentation des constituants du modèle*

L'**espace de signation** représente l'espace situé devant le signeur où sont placés les différents signes qui constituent la phrase signée. Dans l'implémentation actuelle, cet espace de signation est partitionné en volumes de la forme de pavés dans lesquels seront placées les entités.

Chaque pavé de l'espace de signation est appelé **emplacement** et contiendra une ou plusieurs entités. Un emplacement est dit pertinisé s'il contient au moins une entité.

Une phrase signée peut être décrite sous la forme d'une **transcription**. Il s'agit d'un découpage de la phrase signée en intervalles temporels. Chaque **intervalle** contient la description de l'état de l'espace de signation dans l'intervalle de temps considéré.

Une **entité** (figure 8) ne correspond pas toujours à un signe mais désigne plutôt un concept (lieu, objet, personne…). Le signe faisant référence à l'entité est parfois exécuté dans l'espace neutre. Le concept est alors relié à son emplacement par pointage ou simplement par le contexte de la phrase dans le cas de qualificatifs.

À chaque entité est associé un **référent** qui indique le type de l'entité ainsi que les autres entités qu'elle met en relation. Par exemple, dans la phrase *« je te vois »*, l'entité *« vois »* aura un référent de type « action » qui fera référence aux entités *« moi »* et *« toi »*.

Un référent peut avoir trois types de fonctions : **actantielle**, **locative** ou **temporelle**.

Il existe plusieurs types de référents selon les entités :

• **action** : entité à laquelle il est possible de faire référence sous forme de référence actancielle ;

• **date** : entité à laquelle il est possible de faire référence sous forme de référence temporelle ;

• **lieu** : entité à laquelle il est possible de faire référence sous forme de référence locative ;

• **objet** : entité susceptible de changer d'emplacement au cours d'un énoncé.

Pour permettre une plus grande précision du modèle, il a été nécessaire de subdiviser ces différents types.

Ainsi, le type objet peut être spécialisé en **animé**, lui même spécialisé dans le type **personne** qui est susceptible de servir à un transfert personnel.

Une **action autoréflexive** est une action où le locuteur est le seul protagoniste (ex. : *penser*).

Pour finir, notons l'ajout du type **inconnu** qui permet de désigner une entité dont l'appartenance à l'une des catégories précédemment citées ne peut pas encore être décidée.

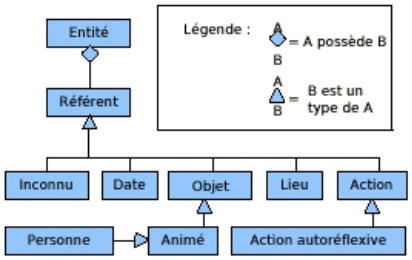


Figure 8. *Représentation des différents types d'entités*

4.3.3. *Exemple de modélisation d'un espace de signation*

À titre d'exemple, la figure 10 montre pas à pas la construction de la phrase *« Il va au cinéma à Toulouse »* (en LSF : [Toulouse]LSF [Cinéma] LSF [Lui] LSF [il y va] LSF). Le contenu de l'espace de signation du dernier intervalle est détaillé pour mettre en évidence les différentes entités et références qui le composent.

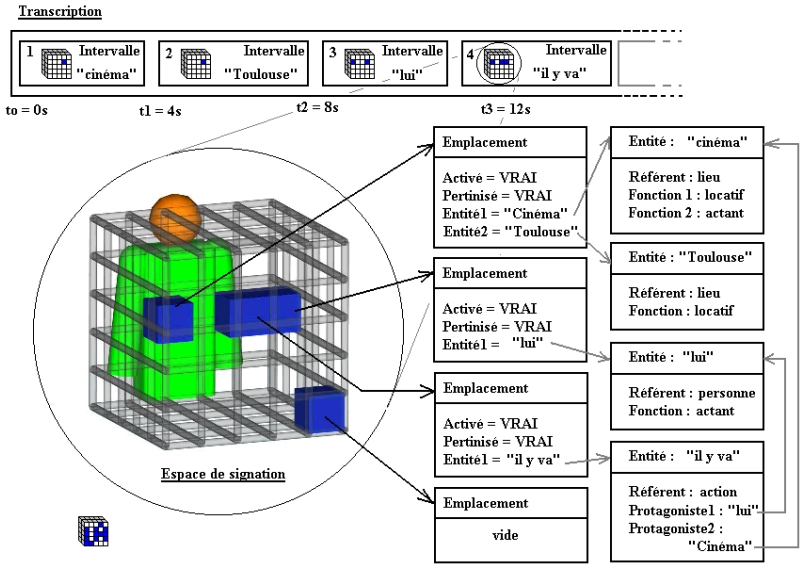


Figure 9. *Modélisation de l'énoncé : « Il va au cinéma à Toulouse »*

4.3.4. *Module de visualisation*

Il est également intéressant de visualiser la modélisation de l'espace de signation obtenu en fin de l'analyse de la phrase signée. Les outils disponibles ne proposent qu'une description textuelle affichée sous forme de partition. Ils permettent de rendre compte de la simultanéité des informations, mais rendent la spatialisation difficile à appréhender.

Nous avons donc développé un module de visualisation de l'espace de signation dans lequel les différentes entités sont représentées dans un espace tridimensionnel avec un symbole différent selon les types de référents : « X » pour une date, « ll » pour une personne, « O » pour un lieu, « □ » pour un objet et « → » pour une relation (action, ...). L'illustration (figure 10) représente le schéma 3D de l'espace de signation obtenu pour la phrase : *« Quel est le réalisateur du film qui passe le jeudi 26 février à 9 h 30 au cinéma Utopia de Toulouse ? »* (en LSF : *« à Toulouse (A), au cinéma Utopia (B), le film qui passe (C), jeudi 26 février à 9 h 30 (D), la personne (E), qui a fait le film (F), qui-est-ce ? (G) »*).

Cette représentation peut être construite de manière interactive, l'utilisateur ayant la possibilité de faire pivoter l'espace pour mieux l'observer, et d'ajouter un second locuteur pour les situations de dialogue.

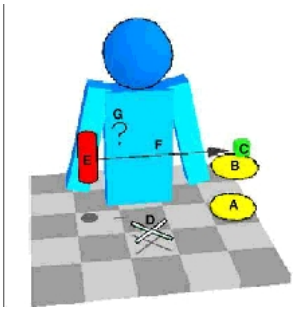


Figure 10. *Illustration du module de visualisation de l'espace de signation*

4.3.5. *Vers un modèle d'énoncé signé*

Afin d'exploiter ce modèle, il est nécessaire de modéliser sa construction par le locuteur. Ce modèle servira à analyser les constituants corporels intervenant dans la construction de l'énoncé et notamment dans la segmentation.

Le modèle de phrase signée devra également inclure des contraintes sur les dépendances des composantes manuelles et non manuelles et donc sur leur ordre d'apparition. On trouve dans le travail de (Lenseigne *et al.*, 2004) à ce sujet quelques pistes intéressantes qui permettent de mettre en correspondance le type de signe observé avec les réalisations successives du signeur. Il est par exemple possible de se servir d'un enchaînement d'observations sur la direction du regard et le placement des

mains pour émettre une hypothèse sur la nature de la séquence. Les observations effectuées par (Cuxac, 2003) peuvent être exploitées dans ce but. Pour illustrer cette notion, on pourra faire référence à l'exemple d'une « action à destination implicite » tiré de (Lenseigne *et al.*, 2004) p. 116 :

• la main dominante prend la configuration figurant l'agent (proforme), le regard installe l'agent ;

• la main dominante effectue le mouvement pendant que le regard anticipe sur sa destination ;

• à l'emplacement d'arrivée, il y a création implicite d'une entité matérialisant la destination.

Ces contraintes sur la position des signes et la synchronisation avec le regard ont été exprimées à l'aide d'un formalisme logique mais n'ont pas encore été exploitées en traitement d'images.

Dans le traitement automatique d'une vidéo en langue des signes, cette implémentation devra notamment mettre en œuvre une segmentation temporelle réalisée en combinant plusieurs indices :

– la comparaison des vitesses 2D ou 3D des mains dominante et dominée permet de détecter les signes admettant une symétrie planaire, centrale ou un mouvement identique des deux mains. Une rupture de symétrie peut être l'indication d'un changement de signe ;

– l'autocorrélation des positions des mains pendant le signe permet aussi de détecter les signes à répétition (ex. : [TRAVAILLER]LSF) (Chateau *et al.*, 2004) ;

– il est également possible d'utiliser comme indice les points de singularité (contacts, changement de vitesse brusque) qui caractérisent souvent les signes (Braffort, 1996) ;

– un changement de position de main dominée entre deux positions stationnaires indique également souvent un changement de signe ;

– même s'il est actuellement difficile d'estimer avec précision la configuration d'une main, des indicateurs sur sa forme permettent de détecter des changements de configuration qui arrivent fréquemment lors des changements de signe ;

– on peut également envisager une détection automatique de primitives de mouvement caractéristiques telles que les mouvements circulaires que l'on retrouve souvent dans les signes standard.

La plupart de ces pistes ont été testées séparément mais il est nécessaire de les combiner pour améliorer les résultats de segmentations temporelles.

Le problème de la segmentation des signes non standard est encore bien plus délicat. Il est d'ailleurs fréquent que les experts en langue des signes soient en désaccord sur la segmentation de gestes relevant de la grande iconicité (voir (Risler, 2005)).

4.4. *Utilisation des modèles*

Ces modèles vont être instanciés à partir des observations effectuées sur la vidéo. Ils vont également servir à piloter l'analyse de la vidéo.

Nous disposons actuellement d'outils d'analyse vidéo appliqués au suivi de l'expression du visage (Mercier, 2007) et au suivi des épaules d'un signeur (Gianni *et al.*, 2007) qui viennent s'ajouter au module de détection et de suivi robuste des mains que nous avons décrit dans la première partie.

Ces observations doivent être intégrées dans nos modèles d'énoncés signés et d'espace de signation afin de permettre de reconstituer la structure exacte de la phrase. Ces modèles n'interviendront pas systématiquement de manière constructive à chaque étape de l'analyse, mais ils pourront également être utilisés à des fins déductives ou prédictives. Ce dernier point est particulièrement intéressant dans le cadre de l'optimisation d'un traitement d'image car il permet d'orienter l'analyse de la séquence vidéo pour en extraire rapidement les indices les plus pertinents à la reconstruction du sens de l'énoncé.

D'autres méthodes, moins explicites, permettent également l'exploitation des données issues de l'analyse de la vidéo. C'est notamment le cas des « Modèles de Markov Cachés » (MMC) qui ont déjà montré leur efficacité dans le domaine de la reconnaissance de la parole et dont le principe est le suivant : l'opérateur fournit dans un premier temps un corpus d'apprentissage au programme qui, à partir des observations issues de l'analyse d'image et d'une caractérisation manuelle des signes, établit les probabilités de transitions entre les états et les probabilités d'observation, et organise ces états dans la structure du MMC. Dans un second temps, l'ordinateur exploitera ce modèle pour indiquer automatiquement les caractéristiques des signes.

Cette approche a déjà été utilisée avec succès dans (Braffort, 1996) dans le cadre d'une segmentation automatique des signes. Cependant, en reconnaissance des LS, Vogler a montré l'intérêt d'une représentation des signes de plus bas niveau (phonémique) pour éviter d'avoir à reconstruire un MMC par signe standard (Vogler, 2003).

Les approches par formulation explicite d'un modèle et/ou par MMC pourront évidemment être combinées pour interpréter les données vidéo.

4.5. *Perspectives d'amélioration*

Les applications des modèles que nous avons décrits sont nombreuses. Des projets sont en cours dans les domaines de l'analyse et de la synthèse de la langue des signes (pilotage d'un signeur virtuel) ainsi que dans celui de la pédagogie (enseignement de la LS).

Il est maintenant nécessaire de compléter ces modèles pour rendre compte de façon plus exhaustive des différentes spécificités de la langue des signes.

L'élargissement du modèle concernera en premier lieu la modélisation des transferts. Ce point est particulièrement délicat car il s'agira alors de tenir compte des composantes de grande iconicité. De plus, comme l'a montré Huenerfauth, la prise en compte des transferts personnels entraînera inévitablement une rotation dans l'espace de signation due au changement de point de vue. Il sera donc nécessaire de compléter notre modèle en fonction de cette contrainte.

En second lieu, il nous faudra modéliser également la « désactivation » des emplacements de l'espace de signation. En effet, dans le cadre d'énoncés plus importants mettant en scène un grand nombre d'entités, on ne fait référence qu'aux entités récemment placées, et il est nécessaire de les réactiver après un certain temps si on veut de nouveau y faire référence. Ce mécanisme d'oubli couplé à un mécanisme de rappel permet d'éviter la saturation de l'espace de signation par un grand nombre d'entités qui engendrerait une ambiguïté sur l'interprétation des énoncés.

À ces deux améliorations, s'ajoute l'utilisation de notre modèle à des fins prédictives et déductives pour pouvoir orienter les points d'analyse à effectuer sur l'image et atteindre ainsi une pertinence et une vitesse d'analyse plus élevée dans la phase de traitement d'image. On sait, en effet, que le mécanisme d'interprétation d'une image, comme celui d'un message, utilise une voie ascendante à partir des données, mais aussi une voie descendante à partir des modèles et des connaissances *a priori*.

5. Conclusion

L'alimentation de notre modèle par les données issues de l'analyse vidéo n'en est qu'à ses débuts, mais les travaux menés ouvrent déjà la voie à des applications diversifiées dans les domaines de la synthèse, de l'analyse et de la pédagogie pour l'enseignement de la langue des signes.

Comme nous l'avons vu, nous disposons actuellement d'un processus robuste de suivi des mains testé dans le cadre de l'analyse de la LS. Ce dispositif permet de prendre en compte les phénomènes d'occultations fréquents caractéristiques des vidéos de productions signées. Il reste maintenant à faire le pont entre ces observations et la construction de l'espace de signation. Il s'agit de quantifier les différentes observations et d'en extraire les informations sémantiques pertinentes et nécessaires à l'alimentation du modèle et à l'interprétation des phrases signées.

Le développement des modèles d'énoncés signés contribuera également à une meilleure compréhension de la structure iconique de la langue des signes.

6. Bibliographie

Bar-Shalom Y., *Tracking and data association*, Academic Press Professional, Inc., San Diego, CA, USA, 1987.

Braffort A., Reconnaissance et compréhension de gestes, application à la langue des signes, PhD thesis, Université Paris-XI Orsay, 1996.

Braffort A., Choisier A., Collet C., Cuxac C., Dalle P., al., « Projet LS-COLIN. Quel outil de notation pour quelle analyse de la LS ? », *Actes des 3^{e} Journées d'études « Recherches sur les Langues des Signes » , R-LSF'01*, Toulouse, France, 23–24 November, 2001.

Braffort A., Choisier A., Collet C., Dalle P., Gianni F., Lenseigne B., Segouat J., « Toward an annotation software for video of Sign Language, including image processing tools and signing space modelling », *Proc. of 4th International Conference on Language Resources and Evaluation - LREC 2004*, vol. 1, Lisbon, Portugal, p. 201-203, 26–28 May, 2004.

Brugman H., Wittenburg P., Levinson S. C., Kita S., « Multimodal annotations in gesture and sign language studies », *In Proc of the 3rd International Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association, Paris., Las Palmas, Canary Islands, p. 176-182, 29–31 May, 2002.

Chateau T., Curie F., Marc R., « Reconnaissance de gestes par la vision monoculaire temps réel », *Acte de l'atelier acquisition du geste humain par vision artificiel conjointement au colloque RFIA2004*, 2004.

Chen Q., Wu H., Yachida M., « Face detection by fuzzy pattern matching », *ICCV '95 : Proceedings of the Fifth International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, p. 591, 1995.

Cuxac C., « Phonétique de la LSF : une formalisation problématique », *Journées "la linguistique de la LSF : recherches actuelles"*, 2003.

Dalle P., « High level models for sign language analysis by a vision system », *Workshop on the Representation and Processing of Sign Language :Lexicographic Matters and Didactic Scenarios*, Evaluations and Language resources Distribution Agency, p. 17-20, 2006.

Deutscher J., Blake A., I. R., « Articulated Body Motion Capture by Annealed Particle Filtering », *Computer Vision and Pattern Recognition*, 2000.

Elliott R., Glauert J., Kennaway J., Marshall I., « The Development of Language Processing Support for the ViSiCAST Project », *4th International ACM SIGCAPH Conference on Assistive Technologies*, 2000.

Gall J., *« Generalised Annealed Particle Filter - Mathematical Framework, Algorithms and Applications »*, Master's thesis, University of Mannheim, 2005.

Gall J., Potthoff J., Schnoerr C., Rosenhahn B., , Seidel H. P., Interacting and Annealing Particle Filters : Mathematics and a Recipe for Applications, Technical Report n° MPI-I-2006-4-009, Max-Planck Institute, 2006.

Gianni F., Collet C., Dalle. P., « Robust tracking for processing of videos of communication's gestures », *International Workshop on Gesture in Human-Computer Interaction and Simulation*, 2007.

Haritaoglu I., Harwood D., Davis L., « Ghost : a human body part labelling system using silhouette », *Proc. of IEEE Conf. on Pattern Recognition*, n° 1, p. 77-82, 1998.

Huenerfauth M., Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation, PhD thesis, University of Pennsylvania, 2006.

Kipp M., « ANVIL – A Generic Annotation Tool for Multimodal Dialogue », *In Proc. of the 7th European Conference on Speech Communication and Technology, Eurospeech*, Aalborg, Scandinavia, p. 1367-1370, September, 2001.

Kitagawa G., « Monte Carlo filter and smoother for non-Gaussian nonlinear state space models », *Journal of Computational and Graphical Statistics*, vol. 5, n° 1, p. 1-25, 1996.

Kovac J., Peer P., Solina F., « Human skin color clustering for face detection », *EUROCON 2003 Computer as a Tool*, vol. 2, p. 144-148, 2003.

Lejeune F., Analyse sémantico-cognitive des énoncés en Langue des Signes Française pour une génération automatique de séquences gestuelles, PhD thesis, Université d'Orsay, 2004.

Lenseigne B., Intégration de connaissances linguistiques dans un système de vision, application à l'étude de la Langue des Signes, PhD thesis, Université Paul Sabatier, 2004.

Lenseigne B., Dalle P., « A Model for Sign Language Grammar », *2nd Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 256-260, 2005.

Lenseigne B., Dalle P., « Using Signing Space as a Representation for Sign Language Processing », *in* S. Gibet, N. Courty, J.-F. Kamp (eds), *Gesture in Human-Computer Interaction and Simulation : 6^{th} International Gesture Workshop, GW 2005, Revised Selected Papers*, n° 3881 in *Lecture Notes in Computer Science*, Springer-Verlag, Berder Island, France, p. 25-36, 18–20 May2005, 2006.

Lenseigne B., Gianni F., Dalle P., « A new gesture representation for sign language analysis », *Workshop on the representation and processing of sign language : from sign writing to image processing*, p. 109-110, 2004.

MacCormick J., Blake A., « A Probabilistic Exclusion Principle for Tracking Multiple Objects », *Int. Journal on Computer Vision*, 2000.

Mercier H., Modélisation et suivi des déformations faciales : Applications à la description des expressions du visage dans le contexte de la langue des signes, PhD thesis, Université Paul Sabatier, 2007.

Ong S., Ranganath S., « Automatic sign language analysis : A survey and the future beyond lexical meaning », *PAMI*, vol. 27, n° 6, p. 873-891, June, 2005.

Risler A., « Construction / déconstruction de l'espace de signation », *12ème Conférence Internationale sur le Traitement Automatique du Langage Naturel*, 2005.

Sallandre M., Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité, PhD thesis, Université Paris 8, 2001.

Stokoe. W. C., « Sign Language Structure », *Studies in Linguistics. Occasional Papers n° 8*, 1960.

Vogler P., ASL recognition : reducing the complexity of the task with phonem based modelling and Parallel Hidden Markov Models, PhD thesis, University of Pennslylvania, 2003.

Voisin E., « Flexion et ordre des Signes en Langue des Signes française », *Actes de l'Atelier des Doctorants en Linguistique*, 2006.

ANNEXE

Algorithme 1 : Filtre particulaire générique (Gall, 2005)

1. **Initialisation** $t \leftarrow 0$

Pour $i = 0, \ldots, n$, échantillonner $\mathbf{s}_0^{(i)}$ à partir de η_0 la distribution initiale

2. **Prédiction**

Pour $i = 0, \ldots, n$ échantillonner $\tilde{\mathbf{s}}_t^{(i)}$ à partir de $p(\mathbf{x}_t | \mathbf{x}_{t-1} = \mathbf{s}_{t-1}^{(i)})$

3. **Mise à jour**

Pour $i = 0, \ldots, n$, $\pi_t^{(i)} \leftarrow p(\mathbf{z}_t | \mathbf{x}_t = \tilde{\mathbf{s}}_t^{(i)})$

Pour $i = 0, \ldots, n$, $\pi_t^{(i)} \leftarrow \frac{\pi_t^{(i)}}{\sum_{j=1}^{n} \pi_t^j}$

4. **Rééchantilonnage**

Pour $i = 0, \ldots, n$, $\mathbf{s}_t^{(i)} \leftarrow \tilde{\mathbf{s}}_t^{(j)}$ avec la vraisemblance $\pi_t^{(j)}$

$t \leftarrow t + 1$ et retourner à l'étape **2**.

Algorithme 2 : Rééchantillonnage stratifié (Kitagawa, 1996)

ENTRÉES: sommes cumulées de poids $S_1 = \pi_k^{(1)}$

$i = 2, \ldots, n$, $S_i = S_{i-1} + \pi_k^{(i)}$

$t = S_n/n$, $u_j = \mathcal{U}[0, t]$

pour $j = 0, \ldots, n$, **faire**

  tantque $i < n$ et $u_j > S_i$ **faire**

    $i = i + 1$

  fin tantque

  $\tilde{\mathbf{s}}_k^{(j)} = \mathbf{s}_k^{(i)}$

  $u_j = u_j + t$

fin pour

Algorithme 3 : Filtre particulaire recuit (Gall *et al.*, 2006)

1. **Initialisation** $t \leftarrow 0$
Pour $i = 0, \ldots, n$, échantillonner $\mathbf{s}_0^{(i)}$ à partir de η_0
2. **Prédiction**
Pour $i = 0, \ldots, n$ échantillonner $\tilde{\mathbf{s}}_t^{(i)}$ à partir de $p(\mathbf{x}_t | \mathbf{x}_{t-1} = \mathbf{s}_{t-1}^{(i)})$
3. **Mise à jour (recuit)**
Pour m=$M, \ldots, 1$

Pour $i = 0, \ldots, n$, $\pi_{t,m}^{(i)} \leftarrow p(\mathbf{z}_t | \tilde{\mathbf{x}}_{t,m} = \tilde{\mathbf{s}}_{t,m}^{(i)})^{\beta_m}$

Pour $i = 0, \ldots, n$, $\pi_{t,m}^{(i)} \leftarrow \frac{\pi_{t,m}^{(i)}}{\sum_{j=1}^{n} \pi_{t,m}^{(j)}}$

Pour $i = 0, \ldots, n$, $\mathbf{x}_{t,m}^{(i)} \leftarrow \tilde{\mathbf{x}}_{t,m}^{(j)}$ avec la vraisemblance $\pi_{t,m}^{(j)}$

Pour $i = 0, \ldots, n$, $\tilde{\mathbf{x}}_{t,m-1}^{(i)} = \tilde{\mathbf{s}}_{t,m-1}^{(i)} = \tilde{\mathbf{s}}_{t,m}^{(i)} + \mathbf{B}_m$

Pour $i = 0, \ldots, n$ $\pi_{t,0}^{(i)} \leftarrow p(\mathbf{z}_t | \mathbf{x}_t = \tilde{\mathbf{s}}_t^{(i)})$
Pour $i = 0, \ldots, n$ $\pi_{t,0}^{(i)} \leftarrow \frac{\pi_{t,0}^{(i)}}{\sum_{j=1}^{n} \pi_{t,0}^{(j)}}$
4. **Rééchantillonnage**
Pour $i = 0, \ldots, n$, $\mathbf{s}_{t+1}^{(i)} \leftarrow \tilde{\mathbf{s}}_{t+1}^{(j)}$ avec la vraisemblance $\pi_{t+1}^{(j)}$
$t \leftarrow t + 1$ et retour à l'étape 2.

Calcul de la densité :

Pour $m \in M$

Pour $n \in \{M - m\}$

Si $(s_{t,m}^{(i)} == s_{t,n}^{(j)})$

Alors $p(z_{t,m} | \tilde{s}_{t,m}^{(i)}) \simeq p(z_{t,m} | \tilde{s}_{t,m}^{(i)}) - \lambda p(s_{t,n}^{(j)} | z_{t,n})$

où M est l'ensemble des filtres (celui de la tête et les deux des mains) et $\lambda = 0,8$. Cette valeur a été fixée de manière empirique.