Segmentation semi-automatique de corpus vidéo en Langue des Signes

Matilde Gonzalez, Christophe Collet

Segmentation semi-automatique de corpus vidéo en Langue des Signes

M. Gonzalez C. Collet

IRIT (UPS - CNRS UMR 5505) Université Paul Sabatier
118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9
gonzalez@irit.fr

Résumé

De nombreuses études sont en cours afin de développer des méthodes de traitement automatique de langues des signes. Plusieurs approches nécessitent de grandes quantités de données annotées pour l'apprentissage des systèmes de reconnaissance. Nos travaux s'occupent de l'annotation semi-automatique afin de permettre de collecter les données. Nous proposons une méthode de suivi de composantes corporelles, de segmentation de la main pendant occultation et de segmentation des gestes à l'aide des caractéristiques de mouvement et de forme de la main.

Mots Clef

Langue des signes, annotation, corpus.

Abstract

Many researches focus on the study of automatic sign language recognition. Many of them need a large amount of data to train the recognition systems. Our work address the annotation of sign language video corpus in order to collect training data. We propose a robust tracking algorithm for hands and head, a method to segment hand during occlusion and an approach to segment gestures using motion and hand shape features.

Keywords

Sign language, annotation, corpora.

1 Introduction

La langue des signes (LS) est une langue gestuelle développée par les sourds pour communiquer. Un énoncé en LS consiste en une séquence de signes réalisés par les mains, accompagnés d'expressions du visage et de mouvements du haut du corps, permettant de transmettre des informations en parallèles dans le discours. Même si les signes sont définis dans des dictionnaires, on trouve une très grande variabilité liée au contexte lors de leur réalisation. De plus, les signes sont souvent séparés par des mouvements de co-articulation (aussi appelé *'transition'*). Cette extrême variabilité et l'effet de co-articulation représentent un problème important dans les recherches en traitement automatique de la LS. Il est donc nécessaire d'avoir de nombreuses vidéos annotées en LS, si l'on veut étudier cette langue et utiliser des méthodes d'apprentissage automatique. Les annotations de vidéo en LS sont réalisées manuellement par des linguistes ou experts en LS, ce qui est source d'erreur, non reproductible et extrèmement chronophage. De plus, la qualité des annotations dépend des connaissance en LS de l'annotateur. L'association de l'expertise de l'annotateur à des traitements automatiques facilite cette tâche et représente un gain de temps et de robustesse. Cet article présente une méthode permettant de segmenter semi-automatiquement des énoncés en LS, sans utiliser d'apprentissage automatique. Plus précisement, nous cherchons à détecter les limites de début et fin de signes. Cette méthode de segmentation nécessite plusieurs traitements de bas niveau afin d'extraire les caractéristiques de mouvement et de forme de la main. D'abord nous proposons une méthode de suivi des composantes corporelles robuste aux occultations. Ensuite, un algorithme de segmentation des mains est développé afin d'extraire la région des mains même quand elles se trouvent devant le visage. Puis, les caractéristiques de mouvement sont utilisées pour réaliser une première segmentation qui est par la suite améliorée grâce à l'utilisation de caractéristiques de forme. En effet,celles-ci permettent de supprimer les limites de segmentation détectées en milieu des signes. Cet article est structuré comme suit. La section 2 présente une synthèse des méthodes de suivi des composantes corporelles, de segmentation de la main devant le visage et d'annotation automatique appliquées à la LS. La section 3 détaille notre approche pour le suivi de composantes corporelles et la section 4 décrit notre méthode de segmentation des mains. Nous montrons ensuite dans la section 5 l'extraction de caractéristiques de mouvement et de forme afin de segmenter la séquence vidéo. Des résultats expérimentaux sont ensuite présentés en section 6. Enfin, en section 7, une conclusion rappelle les principaux résultats obtenus et évoque quelques perspectives de recherche.

2 Etat de l'art

Un des problèmes majeurs dans les systèmes de reconnaissance de la LS concerne les méthodes de suivi des composantes corporelles et de segmentation à cause du grand nombre d'occultations entre les mains et le visage et de la similarité de couleur entre ces membres.

2.1 Suivi des Composantes Corporelles

Dans la littérature, les méthodes de suivi sont principalement basées soit sur des mesures de coût entre l'image et un motif [1], soit sur des modèles dynamiques qui estiment la fonction de densité de probabilité a posteriori du système. Dans le cas de systèmes non-linaires ou non-Gaussiens, le filtrage particulaire est très populaire [2]. Il nécessite un modèle d'observation qui en général tient compte de la couleur ou des contours [3, 4, 5]. L'inconvénient des approches ne considérant que la couleur, est le fait de représenter plusieurs objets avec le même modèle, e.g. un blob de peau peut modéliser les mains et le visage. Dans ce cas d'autres traitements sont nécessaires afin d'identifier chaque cible. De plus les occultations entre les objets de la même couleur sont difficilement gérables car l'information spatiale est ignorée. Les techniques de suivi basées contours prennent en considération cette information spatiale. Cependant elles ne sont pas souhaitables pour suivre des objets extrêmement déformables comme les mains et sont sensibles aux occultations.

Les occultations entre les mains et la tête sont généralement traitées avec l'association des données dans l'image et des caractéristiques globales ou locales [3, 6, 5]. Gianni *et al.* [3] ont proposé une approche basée sur le filtrage particulaire et la couleur. Ils considèrent que chaque objet ciblé est un nuage des points et utilisent le principe d'exclusion [7] pour associer et interpréter les données. Ceci évite que les filtres convergent vers le même objet et permet de gérer implicitement les occultations. Cependant la position des objets pendant l'occultation n'est connue avec précision. Lefebvre [5] a présenté une approche qui utilise un modèle anatomique en plus de la couleur. La reconnaissance du torse et des coudes permet d'estimer la position des autres membres en partitionnant l'espace de recherche. Cette technique est très rapide en temps de calcul mais est source d'erreurs car les objets deviennent dépendants entre eux. Tanibata et Shimada [6] ont introduit un algorithme de suivi utilisant des caractéristiques locales et basé sur le recalage de motifs pour gérer les occultations de la tête et des mains. Toutefois la forme de la main peut changer pendant l'occultation et ne correspondra plus au motif préalablement enregistré.

2.2 Segmentation de la main devant le visage

L'étude de la segmentation de la main constitue un aspect important dans la LS car les mains transmettent la majeure partie des informations. Des recherches antérieures proposent des techniques de segmentation où la main est le seul objet dans la scène [8] ou encore la seule région de peau [9, 10]. Ces approches ne considèrent pas les occultations potentielles entre objets de la même couleur comme c'est le cas des mains et de la tête. D'autre méthodes basées sur des contours actifs [11, 12] ou sur le recalage de motifs [6] ne donnent des résultats satisfaisants que si la forme change peu, or en LS ce n'est pas le cas. Dans [13] est présenté une méthode pour résoudre le problème d'occultation de la main devant le visage avec le concept de champ de force de l'image. Cette méthode permet de retrouver grossièrement la main mais ne permet pas de la segmenter.

2.3 Segmentation Automatique des Signes

Actuellement plusieurs recherches s'intéressent au problème de l'analyse automatique de la LS [14], plus particulièrement de sa reconnaissance [15, 16, 17]. Dans [18] les données d'apprentissage sont des signes isolés réalisés plusieurs fois par un ou plusieurs signeurs. La réalisation des signes est dépendante du contexte et, dans le cas des signes isolés, la co-articulation n'est pas prise en compte. L'annotation de corpus vidéo est nécessaire pour collecter des données à partir d'énoncés en LS. Il existe plusieurs logiciels d'annotation pour assister l'annotateur dans cette tâche, e.g. Ancolin [19], Elan [20]. Dreuw et Ney [21] proposent une approche pour générer automatiquement l'annotation des corpus en gloses, c.a.d. segmentation et identification des signes. Ils utilisent un système de reconnaissance pour identifier les gloses. Bien que cette approche produise de l'annotation, elle ne résout pas le problème de collection de données car le système de reconnaissance nécessite des données pour l'apprentissage qui sont, en général, manuellement annotées. Yang *et al.* [22] proposent d'annoter automatiquement les données d'apprentissage mais ne considèrent que les caractéristiques de bas niveau comme la position et la segmentation des mains. Nayak *et al.* [23] ont proposé une méthode qui permet d'extraire automatiquement un signe à l'aide de plusieurs occurrences du signe dans la vidéo. Ils considèrent la forme et la position relative des mains par rapport au corps à l'aide de représentations multi-dimensionnelles. Pour la plupart des signes ces caractéristiques varient énormément selon le contexte cantonnant cette approche à quelques exemples typiques. Lefebvre et Dalle [24] ont présenté une méthode utilisant des caractéristiques de bas niveau afin de segmenter semi-automatiquement les signes. Ils ne considèrent que le mouvement dans le but d'identifier plusieurs types de symétries. Or plusieurs signes sont composés de plusieurs séquences avec différents types de symétrie, ces signes seront sur-segmentés.

Afin de résoudre certains problèmes émergents de l'état de l'art nous proposons un algorithme de suivi robuste aux occultations [25], une méthode d'extraction de la main pendant les occultations [26] et une méthode de segmentation automatique des signes qui exploite le caractéristiques de mouvement et de forme de la main.

3 Suivi des composantes corporelles

Dans cette section nous présentons notre algorithme de suivi des main et de la tête basé sur le filtrage particulaire. Le filtre Bayésien estime la fonction de densité de probabilité a posteriori de l'état actuel $\mathbf{x}_t$ conditionné aux observations $\mathbf{z}_{1:t} = \mathbf{z}_1 \dots \mathbf{z}_t$ avec $\mathbf{z}_t$ le vecteur d'observations. La fonction de densité de probabilité $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$ pour un processus Markovien de premier ordre est obtenu en deux

étapes. D'une part l'étape de prédiction,

$$p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \cdot p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1}) dx_{t-1}, \quad (1)$$

estime la distribution a priori pour $t+1$ comme la convolution entre la distribution a posteriori $p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1})$ et la distribution de probabilité de transition $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, c.a.d le modèle dynamique du système. D'autre part l'étape de mise à jour,

$$p(\mathbf{x}_t \mid \mathbf{z}_{1:t}) = k \cdot p(\mathbf{z}_t \mid \mathbf{x}_t) \cdot p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1}), \quad (2)$$

calcule la densité de probabilité a posteriori en utilisant la probabilité des observations $p(\mathbf{z}_t \mid \mathbf{x}_t)$ et la distribution temporelle a priori, $p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1})$, sur $\mathbf{x}_t$ connaissant les observations passées.

3.1 Les filtres particulaires

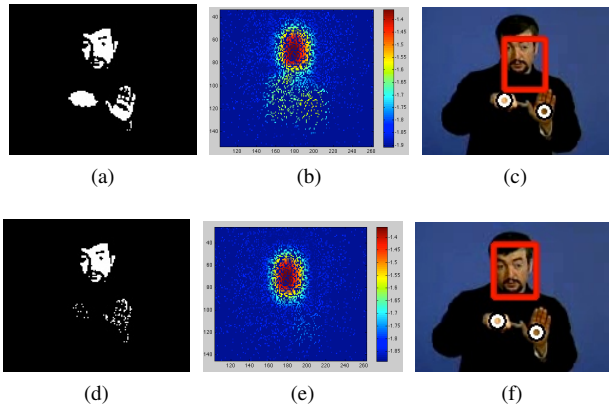
Les filtres particulaires [2] sont une bonne solution pour le suivi de mouvements stochastiques. Ils estiment successivement l'état $\mathbf{x}_t$ du système grâce à l'implémentation d'un filtre récursif Bayésien par simulation de Monte Carlo. La densité de probabilité a posteriori $p(\mathbf{x}_t/\mathbf{z}_t)$ de l'état actuel $\mathbf{x}_t$ est approximée par une série de particules pondérées, $\{\mathbf{s}_t^n, \pi_t^n\}_{n=1}^N$. Les filtres particulaires maintiennent de multiples hypothèses, c.a.d chaque particule est un état hypothétique de l'objet, pondérées par la probabilité des échantillons $\pi_t^n \propto p(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{s}_t^n)$. Le poids des particules correspond à l'observation générée par l'état hypothétique et reflète la pertinence de chaque particule, e.g. la couleur de la peau. L'algorithme des filtres particulaires est composé des étapes suivantes :

1. **Échantillonnage** des N particules de la collection $\{\mathbf{s}_{t-1}^n, \pi_{t-1}^n\}_{n=1}^N$ vers $\{s'^n_t, \frac{1}{N}\}_{n=1}^N$. Les particules sont sélectionnées en fonction de leur poids. Les particules de poids élevé sont dupliquées alors que les particules de poids faible sont supprimées.
2. **Propagation** des particules à l'aide du modèle dynamique du système $s_t^n \sim p(\mathbf{x}_t \mid \mathbf{x}_{t-1} = \mathbf{s}'^n_{t-1})$ afin d'obtenir $\{\mathbf{s}_t^n, \frac{1}{N}\}_{n=1}^N$.
3. **Pondération** de la nouvelle collection de particules avec les observations $\mathbf{z}_t$ avec $\pi_t^n \propto p(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{s}_t^n)$ et normalisation pour obtenir $\sum_{n=1}^N \pi_t^n = 1$.

Le poids des particules est utilisé pour estimer l'état actuel du système grâce à la relation suivante :

$$E[\mathbf{x}_t] = \sum_{n=1}^{N} \pi_t^n \mathbf{s}_t^n. \quad (3)$$

Le modèle d'observation des particules peut être composé de plusieurs caractéristiques visuelles. Dans notre cas, les mains et la tête sont caractérisées par leur forme et leur couleur. La différence de dynamique du mouvement et de forme entre la main et la tête nous permet de choisir le modèle d'objet adapté. Pour la tête nous proposons un modèle



(a) (b) (c)

(d) (e) (f)

FIGURE 1 – (a) et (d) carte de probabilité de la peau, (b) et (c) poids des particules de la tête et le résultat avec et sans pénalisation des particules respectivement.

rectangulaire où l'état $x_t^{head} = \{x, y\}$ représente les coordonnées du centre du rectangle. Pour les main nous utilisons un modèle de nuages de points où chaque particule est un pixel dont l'état représente la position, la vitesse et l'accélération, $x_t^{hand} = \{x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}\}$.

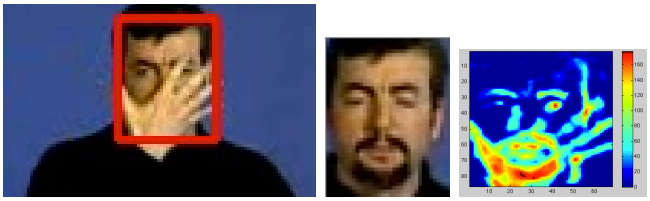
3.2 Le suivi de multiples objets

Le modèle d'observation basé sur la couleur de la peau rend difficile le suivi de plusieurs objets à cause de la superposition d'objets de même couleur. En effet quand un objet en occulte un autre ou se trouve proche, le poids des particules est influencé par des pixels de la couleur de la peau n'appartenant pas à la cible. Par exemple la figure 1 montre la carte de probabilité de la peau $\mathbf{S}_k$ (a), la carte de poids des particules (b) et le résultat de suivi (c) quand la main s'approche du visage. Nous remarquons que le poids des particules de la tête est perturbé par les pixels de la main. L'estimation de la position de la tête se trouve décalée vers la main. Pour y remédier nous utilisons le principe d'exclusion [7] qui affirme que l'observation pour une particule ne peut appartenir qu'à un filtre. Cependant quand les objets se recouvrent partiellement ou complètement, les observations peuvent appartenir à plusieurs filtres. Appelons f_j le filtre particulaire associé à la cible j tel que

$$f_j(\mathbf{S}_k) = \sum_{n=1}^{N} \pi_{t,j}^n \mathbf{s}_{t,j}^n, \quad (4)$$

où $\mathbf{S}_k$ représente la carte de probabilité de la peau utilisée pour calculer le poids des particules associées à la cible j. L'utilisation d'une carte de probabilité $\mathbf{S}_k^j$ adaptée augmente la robustesse du système car les valeurs les plus élevées représentent la probabilité pour un pixel de peau d'appartenir à la cible j. Pour ce faire $\mathbf{S}_k$ est pénalisée à l'aide des particules des autres filtres pour obtenir $\mathbf{S}_k^j$. Appelons $g(\mathbf{S}_k, j)$ la fonction de pénalisation de la carte de probabilité $\mathbf{S}_k$,

$$g(\mathbf{S}_k, j) = \mathbf{W}(\mathbf{s}_{t,j}^n) \cdot \mathbf{S}_k(\mathbf{s}_{t,j}^n), \quad (5)$$



(a) (b) (c)

FIGURE 2 – (a) Résultat du suivi, (b) image avant occultation et (c) coefficients de pénalisation de la tête.

avec $\mathbf{W}$ une matrice positive de valeurs comprises entre 0 et 1. Le poids des particules $\pi^n_{t,j}$ est maintenant calculé à partir des observations obtenues à partir de la nouvelle carte de probabilité de la peau,

$$\mathbf{S}^j_k = \prod_{i=0}^{M} g(\mathbf{S}_k, i) \quad i \neq j \tag{6}$$

où M représente le nombre total de cibles.
La figure 1 montre la carte de probabilité de la peau (d), associée à la tête, $\mathbf{S}^{head}_k$. Cette fois quand le poids des particules est calculé, les pixels des mains sont négligés. Le poids des particules se retrouve concentré au niveau de la tête (Fig. 1e) dont l'estimation de la position (Fig. 1f) est moins influencée par la proximité d'autres objets.
La pénalisation de la carte de probabilité de la peau nécessite le calcul de la matrice de pénalisation $\mathbf{W}$. Quand le modèle de l'objet est un nuage de points la matrice de pénalisation peut être définie comme la matrice unitaire U multipliée par une constante c, tel que :

$$W(\mathbf{s}^n_{t,j}) = c * U(\mathbf{s}^n_{t,j}) \quad avec \;\; c \,\epsilon [0,1] \tag{7}$$

Ceci est pertinent car chaque particule a la même probabilité d'appartenir à la cible j. Ce n'est plus valable dans le cas où les particules ont une forme plus complexe comme dans le cas des particules de la tête à forme rectangulaire. Dans ce cas une particule contient un groupement de pixels où la probabilité pour chaque pixel d'appartenir à la cible peut varier en fonction des autres objets.
Pour calculer la matrice de pénalisation en cas de superposition d'objets, nous utilisons la différence de luminosité entre les images de la tête avant et pendant l'occultation. L'image de la tête avant occultation T est mise à jour en tenant compte de la distance entre la main et la tête à l'aide d'un seuil calculé automatiquement en fonction de la taille de la tête. La figure 2 (a) montre l'image de la tête pendant l'occultation H, l'image enregistrée avant occultation T (b) et la matrice de pénalisation pour la tête (c). Les coefficients les plus élevés se trouvent dans la région de la main. L'algorithme de suivi proposé (Algo 1) est composé d'une séquence de détections et pénalisations à l'aide des filtres particulaires. La tête est d'abord détectée et pénalisée à l'aide de son image avant occultation puis les mains sont détectées et pénalisées pour corriger leur influence sur la tête et enfin la tête et une main sont pénalisées afin d'éviter leur influence sur la détection de l'autre main.

Algorithme 1 : Algorithme de suivi des composantes corporelles basé sur les filtres particulaires

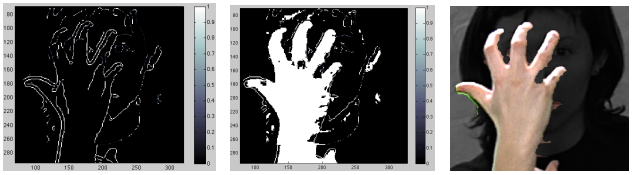
Data : Image k à traiter
Result : Position de la tête et des mains respectivement ; h',h'_1 et h'_2
1. Calcul de la carte de probabilité de la peau $\mathbf{S}_k$
2. Estimation de la position de la tête $h = f_{tte}(\mathbf{S}_k)$
3. Pénalisation de la tête $\mathbf{S}^{mains}_k = g(\mathbf{S}_k, h)$
4. Estimation de la position des mains
   (a) $h_1 = f_{main_1}(\mathbf{S}^{mains}_k)$
   (b) $h_2 = f_{main_2}(\mathbf{S}^{mains}_k)$
5. Pénalisation des mains à partir des particules, $\mathbf{S}^{tte}_k = \prod_{i=1}^{2} g(\mathbf{S}_k, H_i)$
6. Estimation de la position de la tête à partir de la nouvelle carte de probabilité $h' = f_{tte}(\mathbf{S}^{tte}_k)$
7. Pénalisation des particules
   (a) pour $mains$, $\mathbf{S}^{mains}_k = g(\mathbf{S}_k, h)$
   (b) pour $main_1$, $\mathbf{S}^{h1}_k = g(\mathbf{S}^{mains}_k, h_2)$
   (c) pour $main_2$, $\mathbf{S}^{h2}_k = g(\mathbf{S}^{mains}_k, h_1)$
8. Estimation des positions des mains $h'_1 = f_{main_1}(\mathbf{S}^{h1}_k)$ et $h'_2 = f_{main_2}(\mathbf{S}^{h2}_k)$

4 Segmentation de la main devant le visage

La segmentation des mains est une tâche difficile, principalement quand la main se trouve devant le visage. En effet il s'avère laborieux de dissocier les pixels de la main de ceux de la tête. Certaines informations complémentaires peuvent être utiles pour la classification des pixels. Nous proposons, ici, de combiner les caractéristiques des contours et de couleur. En classifiant les contours comme appartenant au visage ou à la main, nous pouvons en trouver la limite. La détection des contours avant et après occultation est réalisée grâce au filtre de Canny. Pour chaque pixel appartenant à un contour dans l'image contenant l'occultation nous cherchons le contour le plus proche le long du vecteur normal dans l'image de la tête avant occultation appelé aussi le motif. Nous calculons la carte des différences d'orientations des contours à partir de l'image de la tête avant et pendant occultation grâce l'équation 8.

$$\Delta\theta = ||\theta_o(x + n_x, y + n_y) - \theta_p(x, y)||, \tag{8}$$

où θ_p correspond à l'orientation du contour dans le motif et θ_o à l'orientation du contour dans l'image avec l'occultation. Quand un contour n'est pas apparié il est fort possible qu'il appartienne à la main. La figure 3 montre la carte normalisée de la différence d'orientation des contours. Nous remarquons que les valeurs proches de 1 correspondent à la main et les faibles valeurs au visage. Cependant près de la bouche ou des yeux d'autres valeurs apparaissent en fonction de l'angle d'intersection et certains contours peuvent coïncider sans appartenir au même objet, e.g. contours des doigts alignés aux contours du nez. Nous calculons la différence de luminance entre les images avant et pendant



(a) (b) (c)

FIGURE 3 – (a) la carte de différence d'orientation des contours, (b) la combinaison de caractéristiques de couleur et de contours et (c) le résultat de segmentation.

TABLE 1 – Classification du mouvement

Statique	Une main	Deux mains
$v_r \approx 0$	$v_r > 0$	$v_r \approx 0 \quad v_r > 0$
$(v_1 \; et \; v_2 \approx 0)$	$(v_1 \approx 0 \oplus v_2 \approx 0)$	$(v_1 \; et \; v_2 \neq 0)$

l'occultation afin d'isoler les pixels susceptibles d'appartenir à la main. Nous remarquons de considérables changements de luminance dans les régions où la classification de contours est ambigüe. Par exemple bien que la couleur des yeux et de la bouche contraste énormément avec le reste du visage les contours peuvent avoir la même orientation en fonction de la configuration de la main. Cependant les contours de la main sont facilement identifiables dans des zones comme les joues et le front où les couleurs sont unies. C'est pour cela que la combinaison de ces caractéristiques (Fig. 3b) rend la segmentation plus robuste. En tenant compte de la connectivité des pixels nous sommes finalement en mesure de segmenter la main. (Fig. 3c)

5 Segmentation Automatique des Signes

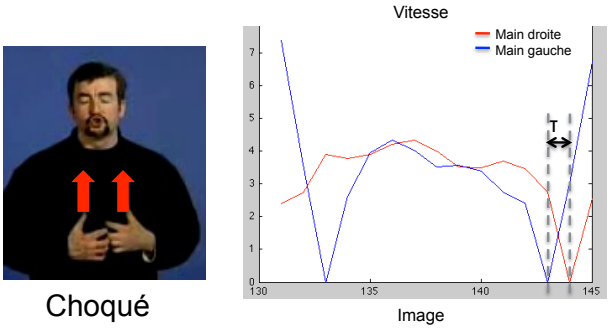
La segmentation des signes correspond à la détection du début et de la fin d'un signe. Pour cela nous exploitons les résultats obtenus dans les deux sections précédentes. D'abord nous utilisons les résultats de suivi de composantes corporelles afin de segmenter les signes grâce à des caractéristiques de mouvement. Ensuite la forme de la main est utilisée pour améliorer les résultats de segmentation.

5.1 Classification du mouvement

Les caractéristiques de mouvement sont extraites à partir des résultats du suivi des composantes corporelles. Les vitesses des mains droite et gauche, $v_1(t)$ et $v_2(t)$ sont calculées à l'aide d'une fenêtre glissante d'une taille comprise entre 3 et 5 images permettant de lisser le signal. La norme de la vitesse est utilisée pour le calcul de la vitesse relative $v_r(t)$ entre les mains comme suit :

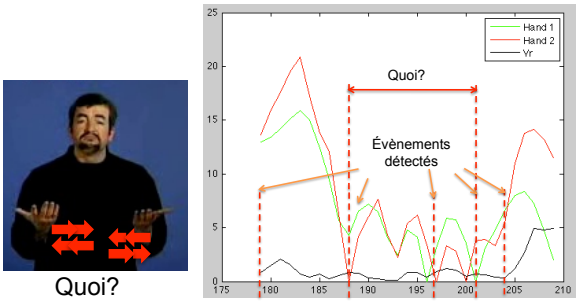
$$v_r = \|v_1(t) - v_2(t+\tau)\|\,, \tag{9}$$

où τ représente le décalage entre la main droite et gauche lors des mouvements symétriques. En fait quand les mains bougent ensemble nous remarquons un léger décalage entre les profils de vitesses des deux mains bien que leur



(a) (b)

FIGURE 4 – (a) illustre le signes "choqué" en LSF et (b) la vitesse des deux mains.



(a) (b)

FIGURE 5 – (a) *'Quoi ?'* en LSF et (b) les vitesses pour les deux mains, la vitesse relative et les événements détectés.

allure reste très proche comme on peut le voir avec le signe "Choqué" (Fig. 4). Grâce à la vitesse relative nous déterminons les séquences réalisées avec une ou deux mains. La classification en fonction du mouvement est détaillée dans le tableau 1. Il s'agit de trois classes : Statique , Une main et Deux mains. A partir de cette classification nous pouvons identifier les événements définis comme les début et fin potentiels de signes et détectés comme un changement de classe. Toutefois cette approche détecte des événements en milieu de signe. On dit alors que les séquences ont été sur-segmentées. Par exemple la figure 5a illustre la réalisation du signe *"Quoi ?"* en LSF. Il s'agit d'un signe symétrique répété où les deux mains bougent simultanément en direction opposée. La figure 5b montre les évènements détectés en fonction des classes définies précédemment. La segmentation peut être améliorée en tenant compte de la forme des mains car, pour ce signe comme pour beaucoup d'autres, la configuration des mains reste inchangée.

5.2 Caractérisation de la forme des mains

Dans cette étape nous introduisons des informations sur la forme de la main afin de corriger la sur-segmentation. La reconnaissance de la configuration de la main est un problème complexe du fait de la grande variabilité de la forme 2D obtenue à l'aide d'une seule caméra.
Afin d'extraire les caractéristiques de forme, nous devons d'abord segmenter les mains pour chaque événement. La

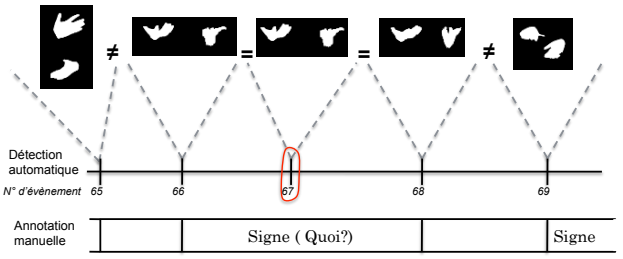


FIGURE 6 – Illustre les mains segmentées pour chaque événement détecté ainsi que la vérité-terrain.

forme de la main est systématiquement comparée avec celle des événements adjacents. Nous utilisons deux mesures de similarité : le diamètre équivalent ϵ_d et l'excentricité ϵ. La première mesure spécifie le diamètre d'un cercle ayant la même aire que la région de la main. La deuxième représente l'excentricité d'une ellipse avec le même moment quadratique que la région. L'avantage d'utiliser ces types de mesures est l'invariance en translation et en rotation. Cependant l'inconvénient est la sensibilité au changement d'échelle et au bruit. La figure 6 montre les résultats de segmentation du signe *"Quoi ?"* en LSF. L'étape précédente a segmenté le signe en tenant compte des caractéristiques de mouvement ce qui a entrainé la sur-segmentation du signe. Nous remarquons que la forme des mains reste similaire entre certains événements détectés. On supprime donc celui du milieu pour corriger la segmentation.

6 Évaluation et Résultats

Nous avons d'abord évalué le suivi des composantes corporelles, puis la segmentation de la main quand elle se trouve devant le visage et enfin la segmentation des gestes utilisant des caractéristiques de mouvement et de forme.

6.1 Suivi des composantes corporelles

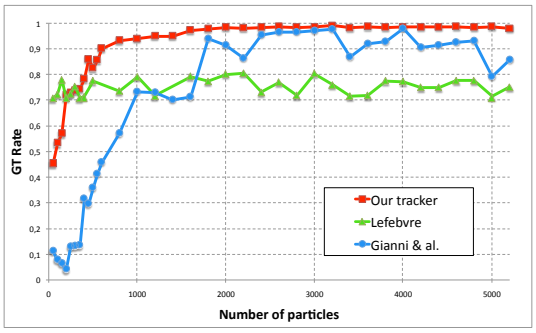
L'évaluation de l'algorithme de suivi a été réalisée sur le corpus vidéo LS-Colin (disponible sur demande). Une séquence de 3000 images a été sélectionnée et manuellement annotée. Dans cette séquence un sourd-né raconte une histoire en LSF. Nous insistons sur le fait qu'aucune contrainte qu'elle soit de type linguistique (lexique restreint) ou dynamique (vitesse de réalisation), n'a été imposée au signeur.
Notre algorithme de suivi a été comparé à ceux proposés par Gianni *et al.* [3] et Lefebvre [5]. La figure 7 montre les résultats de suivi pour une séquence pendant laquelle la main est devant le visage. Dans le cas des régions rectangulaires [5], quand la main se trouve devant le visage les pixels de la main et ceux du visage sont confondus et une région rectangulaire est manquante (Fig. 7, en haut). Quand les mains et le visage sont modélisés comme des nuages des points [3], au moment des occultations les pixels de la main et du visage sont partagés par le filtre et il est impossible de déterminer la position de la main avec précision (Fig. 7, au centre). Contrairement aux cas précédents, notre

FIGURE 7 – Résultats de suivi pour Lefebvre [5] (en haut), Gianni *et al.* [3] (au milieu) et notre méthode (en bas).

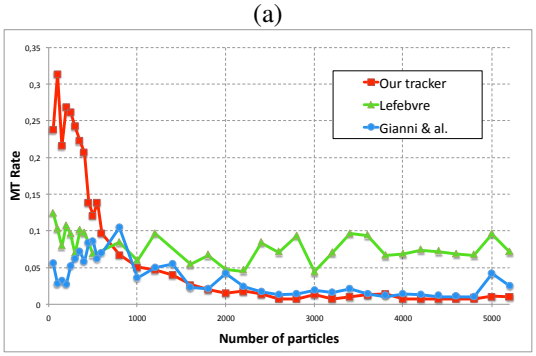
algorithme de suivi est capable de trouver la position de la main même en cas d'occultation, (Fig. 7, en bas).
Dans le but de montrer les performances de notre approche, nous avons quantitativement évalué les résultats de suivi. Nous utilisons le taux de suivi correct, (GTR :good tracking rate), le taux de suivi raté (MTR :miss tracking rate) et le taux de faux suivi (FTR : False tracking rate). Le GTR évalue le suivi correct des objets par le filtre assigné, le MTR évalue le suivi d'un objet par un autre filtre du même type, c.a.d les filtres des mains qui s'échangent, le FTR évalue les résultats des suivi pour les filtres qui ont été échangé mais qui suivent différents type d'objet, e.g. le filtre le la main suit la tête. La figure 8 montre les taux d'évaluation pour les algorithmes proposées en [3] et [5] ainsi que pour notre algorithme de suivi. Nous remarquons que notre méthode améliore significativement la stabilité en fonction du nombre des particules en comparaison avec les autres deux approches.

6.2 Segmentation de la main devant le visage

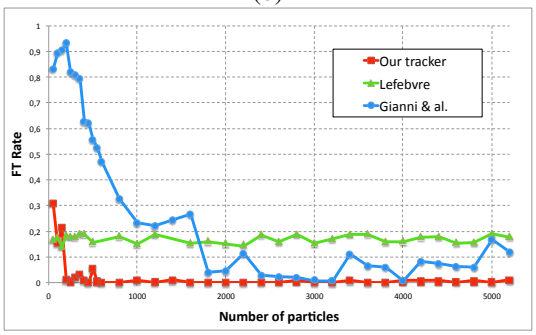
L'évaluation de la segmentation de la main devant le visage a été réalisée sur le corpus de LSF du projet Dicta-Sign annoté manuellement. Le corpus est composé de plusieurs conversations à deux signeurs en LSF avec un total de 8 sessions soit 16 personnes et cinq heures de vidéo. Nous avons sélectionné 5 séquences vidéo où la main occulte le visage ce qui correspond à près de 50 images. Dans ces séquences la configuration et l'expression du visage peuvent changer pendant l'occultation.
La figure 9 montre les résultats de segmentation pour différentes séquences. La segmentation de la main placée devant le visage est satisfaisante. Cependant les résultats montrent des artefacts et des trous. Les artefacts sont principalement dus à la rotation "hors plan" de la tête ou au changement d'expression du visage. D'un autre côté les trous sont dus au manque d'information. En effet le contour de la main peut coïncider avec celui de la tête où l'information sur le changement de couleur peut être manquante.



(a)



(b)



(c)

FIGURE 8 – (a) montre le GT, (b) le FT et (c) le MT obtenus par Lefebvre [5], par Gianni *et al.* [3] et par notre méthode.

Notre approche de segmentation a été évaluée quantitativement à l'aide du taux de vrai-positifs (TP : True positive) et du taux de de faux-positifs (FP : False positif). Le tableau 2 présente les résultats pour chaque séquence. Le taux de TP moyen pour toutes les séquences est de 96%. Le taux de FP est de 8% et correspond à des pixels qui peuvent être facilement supprimés par des traitements a posteriori, e.g. lignes sous le menton ou sur le cou.

6.3 Segmentation Automatique des Signes

Nous avons réalisé l'évaluation à l'aide de deux séquences vidéo sans aucune contrainte sur la langue : LS Colin et DEGELS. L'algorithme de segmentation a été appliqué sur 2500 images. Les vérités-terrain pour les deux séquences ont été manuellement réalisées par un signeur sourd-né. L'évaluation consiste à à compter les événements correctement segmentés en tenant compte d'une tolérance (TP : true positifs) et les événements détectés mais qui ne correspondent pas à une limite annotée (FP : False positif). La tolérance δ pour le calcul du taux de TP a été déterminée expérimentalement. Un signeur expérimenté a annoté une séquence vidéo plusieurs fois afin de déterminer sa variabilité qui s'élève dans notre cas à 1,7 images en moyenne. La segmentation est considérée comme correcte si le nombre d'images entre l'annotation et l'événement détecté est proche à la variabilité du signeur. Le tableau 3 montre les résultats pour les deux séquences vidéo avec une tolérance de deux images. On remarque qu'à l'introduction des caractéristiques de forme de la main le taux de TP reste le même alors que le taux de FP diminue de 3% pour LS-Colin et de 10% pour le corpus Degels.

FIGURE 9 – Résultats de segmentation de la main.

TABLE 2 – Résultats de segmentation des mains

	No. de séquence				
	2	4	7	9	11
TPR (%)	96.71	96.51	96.59	95.07	98.15
FPR(%)	3.62	6.48	13.91	6.44	8.27

7 Conclusion

Dans cet article nous présentons une méthode de segmentation temporelle de séquences vidéo en LS. La segmentation a été réalisée en ne considérant que des caractéristiques de bas niveau, ce qui rend notre méthode généralisable pour toutes les LS. Nous utilisons d'abord les caractéristiques de mouvement extraites à l'aide de notre algorithme de suivi qui est robuste aux occultations. Ensuite grâce aux caractéristiques de forme de la main nous sommes capable de corriger la segmentation. Cette méthode a montré des résultats prometteurs qui peuvent être utilisés pour la reconnaissance de signes et pour l'annotation en gloses des séquences à l'aide d'un modèle linguistique de la LS.

Remerciements

Ces recherches sont financées par le 7ème programme cadre Communauté Européenne (FP7/2007-2013) accord no 231135.

TABLE 3 – Résultats de segmentation de gestes

	Motion		Motion + Hand Shape	
	TP(%)	FP(%)	TP(%)	FP(%)
LS- Colin	81.6	46.2	81.6	44.9
DEGELS	74.5	54.2	74.5	44.7

Références

[1] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. Int. Conference CVPR*, 1998, pp. 232–237.

[2] M. Isard and A. Blake, "Condensation-conditional density propagation for visual tracking," *Int. Journal of Computer Visional*, vol. 29, no. 1, pp. 5–28, 1998.

[3] F. Gianni, C. Collet, and P. Dalle, "Robust tracking for processing of videos of communicationÂŠs gestures," *Gesture-Based Human-Computer Interaction and Simulation*, vol. 5085/2009, pp. 93–101, 2009.

[4] A. Micilotta and R. Bowden, "View-based location and tracking of body parts for visual interaction," in *Proc. of BMVC*, 2004, pp. 849–858.

[5] F Lefebvre-Albaret, *Traitement automatique de vidéos en LSF, modélisation et exploitation des contraintes phonologiques du mouvement,*, Phd thesis, University of Toulouse, October 2010.

[6] N. Tanibata, N. Shimada, and Y. Shirai, "Extraction of hand features for recognition of sign language words," in *International Conference on Vision Interface*, 2002, pp. 391–398.

[7] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *Int. Journal of Computer Vision*, vol. 39, pp. 57–71, 2000.

[8] Y. Hamada, N. Shimada, and Y. Shirai, "Hand shape estimation using sequence of multi-ocular images based on transition network," in *Proceedings of the International Conference on Vision Interface*, 2002.

[9] N. Habili, C.C. Lim, and A. Moini, "Segmentation of the face and hands in sign language video sequences using color and motion cues," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 8, pp. 1086–1097, 2004.

[10] A. Ramamoorthy, N. Vaswani, S. Chaudhury, and S. Banerjee, "Recognition of dynamic hand gestures," *Pattern Recognition*, vol. 36, pp. 2069–2081, 2003.

[11] T. Ahmad, CJ Taylor, A. Lanitis, and TF Cootes, "Tracking and recognising hand gestures, using statistical shape models," *Image and Vision Computing*, vol. 15, no. 5, pp. 345–352, 1997.

[12] E.J. Holden, G. Lee, and R. Owens, "Australian sign language recognition," *Machine Vision and Applications*, vol. 16, no. 5, pp. 312–320, 2005.

[13] P. Smith, N. da Vitoria Lobo, and M. Shah, "Resolving hand over face occlusion," *Image and Vision Computing*, vol. 25, pp. 1432–1448, 2007.

[14] S.C.W. Ong and S. Ranganath, "Automatic sign language analysis : A survey and the future beyond lexical meaning," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, pp. 873–891, 2005.

[15] K. Imagawa, Shan Lu, and S. Igi, "Color-based hands tracking system for sign language recognition," in *Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 462–467.

[16] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," in *Proc. International Symposium on Computer Vision*, 1995, pp. 265–270.

[17] J. Zieren, U. Canzler, B. Bauer, and K.F. Kraiss, "Sign language recognition," *Advanced Man-Machine Interaction*, pp. 95–139, 2006.

[18] K. Grobel and M. Assan, "Isolated sign language recognition using hidden markov models," in *IEEE Int. Conference on Systems, Man, and Cybernetics*. IEEE, 1997, vol. 1, pp. 162–167.

[19] A. Braffort, A. Choisier, C. Collet, P. Dalle, F. Gianni, B. Lenseigne, and J. Segouat, "Toward an annotation software for video of sign language, including image processing tools and signing space modelling," in *Proc. of 4^{th} LREC Conference*, 2004, pp. 201–203.

[20] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan : a professional framework for multimodality research," in *Proc. of the 5^{th} LREC Int. Conference*, 2006, pp. 1556–1559.

[21] P. Dreuw and H. Ney, "Towards automatic sign language annotation for the elan tool," in *LREC Workshop on the Representation and Processing of SL : Construction and Exploitation of SL Corpora*, 2008.

[22] R. Yang, S. Sarkar, B. Loeding, and A. Karshmer, "Efficient generation of large amounts of training data for sign language recognition : A semi-automatic tool," *Computers Helping People with Special Needs*, pp. 635–642, 2006.

[23] S. Nayak, S. Sarkar, and B. Loeding, "Automated extraction of signs from continuous sign language sentences using iterated conditional modes," *CVPR*, pp. 2583–2590, 2009.

[24] F. Lefebvre-Albaret and P. Dalle, "Body posture estimation in sign language videos," *Gesture in Embodied Communication and HCI*, pp. 289–300, 2010.

[25] M. Gonzalez and C. Collet, "Robust body parts tracking using particle filter and dynamic template.," in *18th IEEE ICIP*, 2011, pp. 537–540.

[26] M. Gonzalez and C. Collet, "Head tracking and hand segmentation during hand over face occlusion in sign language," in *Int. Workshop on Sign, Gesture, and Activity (ECCV)*, 2010.