Segmentation semi-automatique de signes à partir de corpus vidéo en langue des signes

Matilde Gonzalez, Christophe Collet

Université Paul Sabatier
118 Route de Narbonne
F-31062, Toulouse cedex 9
collet@irit.fr

RÉSUMÉ. *De nombreuses études sont en cours afin de développer des méthodes de traitement automatique des langues des signes. Plusieurs approches nécessitent de grandes quantités de données annotées pour l'apprentissage des systèmes de reconnaissance. Nos travaux concernent l'annotation semi-automatique de ces corpus de données vidéo. 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. Afin de montrer les avantages et limitations de nos contributions, nous avons évalué chacune des méthodes proposées à l'aide de corpus internationaux. Le système de segmentation des signes montre des résultats prometteurs.*

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 addresses 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 hands during occlusions and an approach to segment gestures using motion and hand shape features. In order to show the advantages and limitations of the proposed approaches, we have evaluated each one using international corpus. The full sign segmentation approach shows promising results.*

MOTS-CLÉS : *langue des signes, annotation, corpus.*

KEYWORDS: *sign language, annotation, corpora.*

DOI:10.3166/TS.29.333-358

Extended Abstract

Sign Languages (SL) are visual-gestural communication languages used by deaf communities. They are characterized by the motion of hands, arms, trunk, head among other parts of the body. SL sentences are composed of a sequence of signs. When one sign is performed after another the co-articulation phenomenon occurs between two signs. Co-articulation, in SL, corresponds to the meaningless gesture which is in the middle of two signs. In fact one sign is influenced by the previous sign and itself influences the following sign, then, a sign can be performed in a different way depending on its context.

In Sign Language (SL) video corpora, hands and head tracking is a challenging task because of the high dynamics of objects. Even though head movements remain small, hands move very fast. Sign segmentation is challenging, in addition to the lack of linguistic knowledge, because of the co-articulation problem. It is difficult to segment a video sequence into signs and transitions because one sign is influenced by the previous sign and itself influences the following sign. However, it is possible to approximate signs limits using manual and non-manual features. Although only manual features are considered in this work we are aware that lot of information can be extracted from non-manual features.

Nowadays many researches focus on the automatic analysis of sign language (Ong, Ranganath, 2005), especially, automatic sign language recognition (Imagawa *et al.*, 1998 ; Starner, Pentland, 1995 ; Zieren *et al.*, 2006). Many approaches consider that the training data set is composed of isolated signs performed by various signers several times (Grobel, Assan, 1997), however many signs are context-dependent and the co-articulation phenomenon is not considered. Co-articulation is an important problem since a large amount of data is required to train recognition systems and achieve high recognition rates. Generally these data are collected by annotating sign language video corpora. Annotation is, in general, manually performed by linguists and computer scientists. However this is time consuming, error prone and non reproducible. In addition the quality of the results depends on the annotator's knowledge. In this work these problems are addressed by focusing on the study and development of robust image processing techniques to assist the annotation task.

In this paper we present a novel approach to semi-automatically segment signs based on low level features ; motion and shape. The former allows to characterize gestures in terms of velocity to perform the segmentation and the latter is used to identify hand shape changes between segmented gestures. For the extraction of motion features we propose a body part tracking algorithm robust to hand over face occlusions. In Sign Language (SL) video corpora, hands and head tracking is a challenging task because of the high dynamics of objects. Even though head movements remain small, hands move very fast in a random way. In addition to the movement speed problem, several other problems are faced during body part tracking. Objects are highly deformable and their model is not easily determined. Although in some cases hand configuration could last during the whole sign, hand shape changes very fast even when only the

orientation of the palm has changed. Moreover objects appearance is very similar i.e. objects are similarly coloured. Also objects can partially or fully be either occluded or occlude other similarly coloured objects, which is often the case in SL performances.

Our tracking approach is based on particle filtering and a penalisation function that allows the interaction between filters. Body part tracking consists of three filters running simultaneously in the same frame ; one general filter for head and two annealed filters for hands. Since all of them are based on skin colour features, objects will influence weight computation of the three filter regardless of the filter associated to the object. In fact hands influence the head filter and the expectation is displaced to hands position. This problem is addressed using a penalisation function based on the exclusion principle (MacCormick, Blake, 2000). Filter observation of one target is penalised using the particles of other objects.

During SL performances hands could be placed in front a skin region. When the hand is not overlapping any other skin region, the segmentation can easily be performed using colour features. However the main problem raises when the hand overlaps the face. In this case any simple colour based technique fails in hand segmentation and other features have to be considered. Indeed when a hand overlaps the face the segmentation of the hand region becomes extremely challenging because of the colour similarity between objects. For shape features we propose a segmentation algorithm allowing to extract the hand region when this is placed in front of the face.

For this we propose a contour classification method in addition to the change of appearance. In addition to appearance, edges information is considered to define hand boundaries. Indeed we have noticed that in some places of the face where colour is smooth, e.g. forehead or cheek, it is possible to easily identify contours belonging to the hand. The opposite, in places where the colour is different from skin presenting many edges, e.g. lips or eyes, it is hard to distinguish contours belonging to hand or head, in this case we can use appearance which has significantly change to determine hand region.

Even though we use manual features to improve sign segmentation, i.e. features extracted from hand motion and shape, face expression or other articulators information could be useful to achieve this task.

The main contribution of this work is that low level features are used to detect events so that an annotator could correct segmentation and label sequences. In addition, since only low level features are used, this approach can be used for any sign language or any other gesture based application. Our methods have been evaluated using international corpora and have shown promising results. The proposed method allows to find sign boundaries for processing continuous SL. The results can be used for the annotation of glosses using a linguistic description of signs.

1. Introduction

La langue des signes (LS) est une langue gestuelle développée par les sourds. 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 parallèle 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 appelés *transitions*). La figure 1 montre un exemple de co-articulation qui correspond au geste entre les signes [ÉTATS-UNIS] et [TOUR]. 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'erreurs, non reproductible et extrêmement chronophage. De plus, la qualité des annotations dépend des connaissances en LS de l'annotateur. Ainsi, l'association de l'expertise de l'annotateur à des traitements automatiques facilite cette tâche et représente un gain de temps et de robustesse.

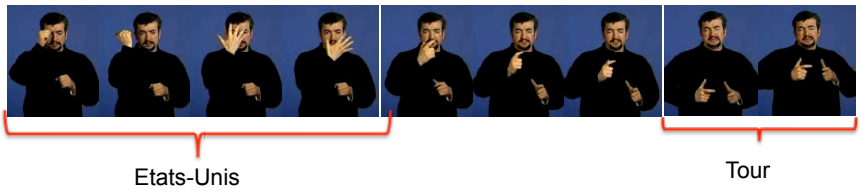

Figure 1. Exemple de co-articulation

Cet article présente une méthode permettant de segmenter semi-automatiquement des énoncés en LS, sans utiliser d'apprentissage automatique. Plus précisément, nous cherchons à détecter les limites de début et fin des signes. La figure 2 montre un exemple de résultats de segmentations manuelle, *Manual Ann*, et automatique, *Auto Ann*. Dans le cas de l'annotation manuelle l'annotateur sélectionne le segment correspondant au signe. Pour l'annotation automatique nous sommes uniquement en mesure de proposer des limites à l'annotateur pouvant correspondre au début ou à la fin d'un signe car aucune information linguistique n'est utilisée. Cette méthode de segmentation nécessite plusieurs traitements de bas niveau afin d'extraire les caractéristiques de mouvement et de forme. 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 quand elles sont devant le visage. Puis, les caractéristiques de mouvement sont utilisées pour réaliser une première segmentation qui est par la suite améliorée avec l'utilisation de caractéristiques de forme. En effet, elles permettent de supprimer les limites de segmentation détectées en milieu de signes (sur-segmentation). Cet article correspond à une version étendue de celui présenté à RFIA 2012 (Gonzalez, Collet, 2012). Il est structuré comme suit. La section 2 présente une synthèse des méthodes de suivi des composantes corporelles,

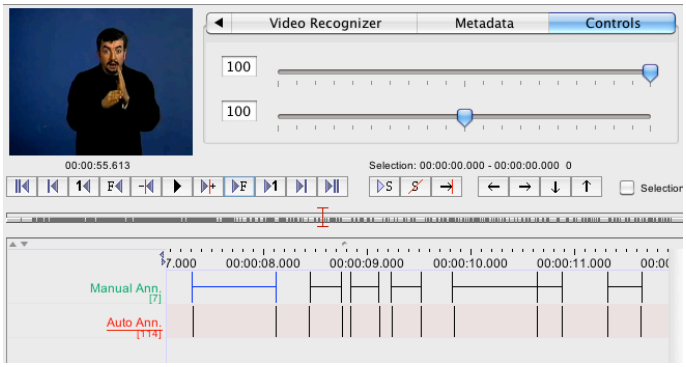

Figure 2. Visualisation de résultats de segmentation grâce à l'outil d'annotation Elan

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 présentés en section 6. Enfin, en section 7, une conclusion rappelle les principaux résultats obtenus.

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 différence entre l'image et un motif (Birchfield, 1998), soit sur des modèles dynamiques qui estiment la fonction de densité de probabilité *a posteriori* du système. Dans le cas de systèmes non linéaires ou non gaussiens, le filtrage particulaire est très populaire (Isard, Blake, 1998). Il nécessite un modèle d'observation qui en général tient compte de la couleur ou des contours (Gianni *et al.*, 2009 ; Micilotta, Bowden, 2004 ; Lefebvre-Albaret, 2010). 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 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 en utilisant des caractéristiques globales ou locales en plus de celles utilisées pour la détection (Gianni *et al.*, 2009 ; Tanibata *et al.*, 2002 ; Lefebvre-Albaret, 2010). Gianni *et al.* (2009) ont proposé une approche basée sur le filtrage particulaire et la couleur. Ils considèrent que chaque objet ciblé est un nuage de points et utilisent le principe d'exclusion (MacCormick, Blake, 2000) pour éviter que les filtres convergent vers le même objet, gérant implicitement les occultations. Cependant la position des objets pendant l'occultation n'est pas connue avec précision. Lefebvre (2010) a présenté une approche qui considère chaque cible comme une région rectangulaire et 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 (2002) 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 (Hamada *et al.*, 2002) ou encore la seule région de peau (Habili *et al.*, 2004 ; Ramamoorthy *et al.*, 2003). 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'autres méthodes basées sur des contours actifs (Ahmad *et al.*, 1997 ; Holden *et al.*, 2005 ; Diamanti, Maragos, 2008) ou sur le recalage de motifs (Tanibata *et al.*, 2002) ne donnent des résultats satisfaisants que si la forme change peu, or en LS ce n'est pas le cas. Dans (Smith *et al.*, 2007) est présentée 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 (Ong, Ranganath, 2005 ; Von Agris *et al.*, 2008 ; Cooper *et al.*, 2011), plus particulièrement de sa reconnaissance (Zieren *et al.*, 2006 ; Shiosaki *et al.*, 2008 ; Theodorakis *et al.*, 2009 ; Piater *et al.*, 2010). La reconnaissance de la langue des signes ne correspond pas uniquement à identifier les signes dans une séquence vidéo mais aussi à traduire une séquence de signes dans une phrase orale. Dans (Grobel, Assan, 1997) les données d'apprentissage sont des signes isolés réalisés plusieurs fois par un ou plusieurs signeurs. Nous avons expliqué en introduction que la réalisation des signes est dépendante du contexte et, donc dans le cas des signes isolés, la coarticulation 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 (Braffort *et al.*, 2004), Elan (Wittenburg *et al.*, 2006). Dreuw et Ney (2008) proposent une approche pour générer automatiquement l'annotation des corpus en gloses, c'est-à-dire la segmentation et l'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.* (2006) 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.* (2009) 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-Albaret, Dalle, 2009) 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 donc sur-segmentés.

Afin de résoudre certains problèmes émergents de l'état de l'art nous proposons un algorithme de suivi robuste aux occultations, une méthode d'extraction de la main pendant les occultations et une méthode de segmentation automatique des signes.

3. Suivi des composantes corporelles

Le suivi des mains et de la tête dans des corpus vidéo de LS est problématique à cause de la dynamique des mains et des nombreuses occultations entre les mains et la tête. Même si le mouvement de la tête reste faible, les mains bougent très rapidement et de façon aléatoire. De plus la variabilité de configuration de la main et sa similarité de couleur avec la tête rendent sa modélisation difficile. Afin de résoudre ces problèmes nous proposons un algorithme de suivi basé sur le filtrage particulaire, implémentation d'un filtre récursif bayésien.

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 obtenue 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 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'est-à-dire le modèle dynamique du système. D'autre part l'étape de mise à jour,

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

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 (Isard, Blake, 1998) 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* $p(\mathbf{x}_t \mid \mathbf{z}_t)$ 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'est-à-dire 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 :

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

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 rectangulaire où l'état $x_t^{head} = \{x, y\}$ ne représente que les coordonnées du centre du rectangle car sa taille est supposée fixe. Une observation correspond pour un rectangle, à la somme des probabilités de ses pixels d'appartenir à la classe peau. Pour les mains 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, $x_t^{hand} = \{x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}\}$. Une observation correspond à la probabilité du pixel d'appartenir à la classe peau.

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 3 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 (MacCormick, Blake, 2000) 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

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

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 $\mathbf{S}_k$,

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

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

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

où M représente le nombre total de cibles.

La figure 3 montre la carte de probabilité de la peau (d), associée à la tête, $\mathbf{S}_k^{head}$. 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 (figure 3e) dont l'estimation de la position (figure 3f) est moins influencée par la proximité d'autres objets.

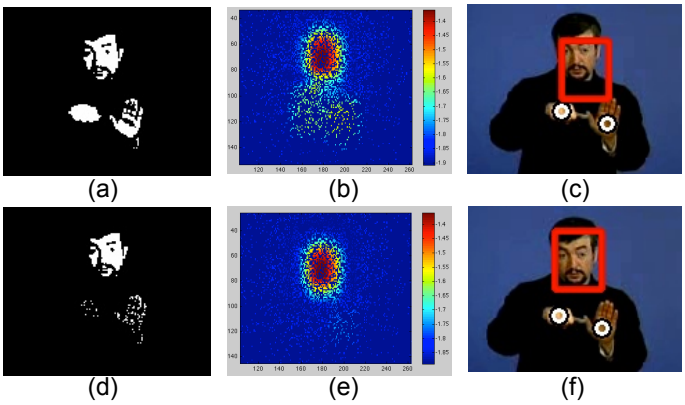

Figure 3. (a) et (d) carte de probabilité de la peau, (b) et (e) poids des particules de la tête et (c) et (f) le résultat sans et avec pénalisation des particules respectivement

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, telle que :

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

Ceci est pertinent car chaque particule a la même probabilité d'appartenir à la cible j.

Figure 4. (a) Résultat du suivi, (b) motif et (c) coefficients de pénalisation de la tête

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 4(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.

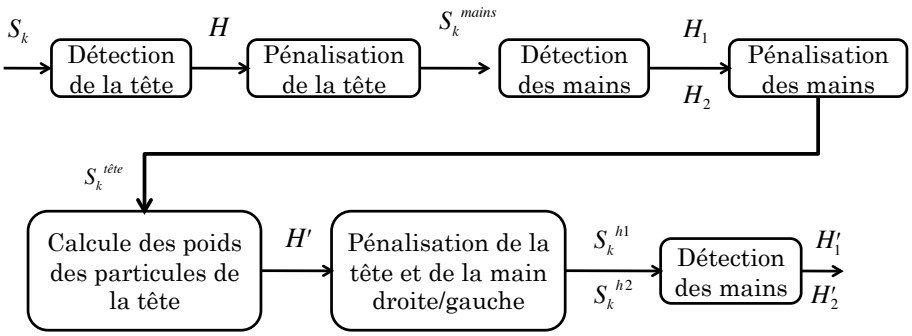

Figure 5. Diagramme de l'algorithme de suivi proposé

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

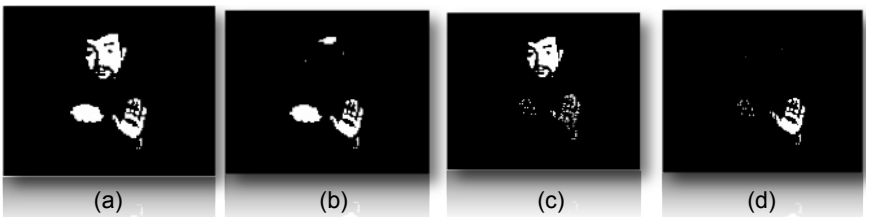

Figure 6. Carte de probabilité de la peau (a) avant pénalisation, (b) pénalisation de la tête, (c) pénalisation des mains et (d) pénalisation de la tête et d'une main

L'algorithme de suivi proposé est composé d'une séquence de détections et pénalisations (figure 5). En utilisant la carte de probabilité de la peau, figure 6(a), la tête est d'abord détectée et pénalisée à l'aide de son image avant occultation, figure 6(b), puis les mains sont détectées et pénalisées pour corriger leur influence sur la tête, figure 6(c) et enfin la tête et une main sont pénalisées afin d'éviter leur influence sur la détection de l'autre main, figure 6(d). Cet algorithme de suivi de composantes corporelles est utilisé pour la caractérisation de gestes afin de réaliser une première étape de segmentation temporelle. Ensuite il est nécessaire d'être en mesure de segmenter les mains même dans des configurations complexes, e.g. quand la main est devant le visage, afin de pouvoir utiliser des caractéristiques de forme par la suite.

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 effet nous remarquons de considérables changements de luminance dans les régions où les contours sont ambigus et *vice versa*. Par exemple, bien que la couleur des yeux ou de la bouche contraste énormément avec le reste du visage, leurs contours peuvent correspondre à ceux de la main en fonction de la configuration de la main. Cependant les contours de la main sont facilement identifiables dans des zones comme les joues ou le front. La figure 7 montre deux régions du visage avant et pendant occultation ; l'œil et la joue. Nous remarquons que dans la région de l'œil il est difficile de distinguer les contours qui sont apparus lors de l'occultation par rapport aux contours déjà présents dans cette région, alors que dans la joue c'est très facile grâce à l'absence de contours avant occultation. Cependant lors de l'occultation l'apparence dans la région de l'œil contraste énormément avant et pendant occultation ce qui permet d'enlever l'ambiguïté concernant les contours.

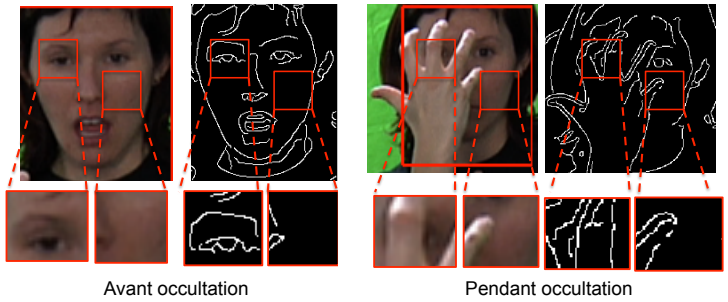

Figure 7. Caractéristiques d'apparence et de contours avant et pendant occultation

L'algorithme de segmentation de la main, figure 8, considère la classification de contours et le changement d'illumination. Ces caractéristiques sont combinées afin d'extraire la main devant le visage. Cette approche nécessite un motif de la tête avant

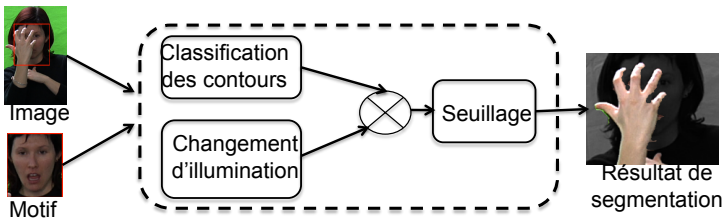

Figure 8. Segmentation de la main devant le visage

occultation afin de comparer les caractéristiques d'apparence et les contours avant et pendant occultation. Nous avons besoin d'un motif de texture de la tête juste avant occultation pour limiter les changements d'expression. Nous sélectionnons grâce à la distance entre la main et la tête, une image pour laquelle on est sûr qu'il n'y a pas d'occultation. En utilisant une fonction qui permet de détecter le moment où la main est devant le visage, figure 9, nous pouvons en utilisant le principe de dichotomie trouver le motif optimal.

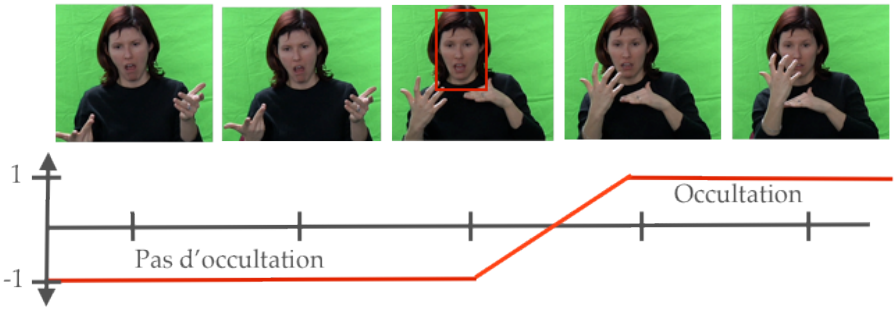

Figure 9. Fonction de détection d'occultation

La fonction de détection d'occultation se sert de la connectivité des pixels de la tête et de chacune des mains. Ceci est réalisé à l'aide des résultats de suivi de la tête, de la main droite et de la main gauche respectivement $\{x_h, y_h, x_{h_1}, y_{h_1}, x_{h_2}, y_{h_2}\}$. La figure 10 montre l'étiquetage de régions de peau. On note que quand la main est devant le visage l'étiquette pour la main et le visage est la même ; il y a donc occultation.

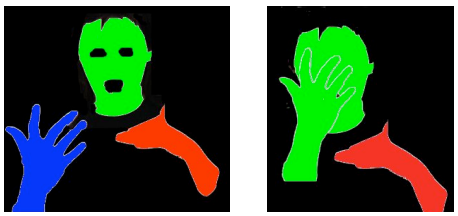

Figure 10. Étiquetage des régions de peau sans (a) et avec occultation (b)

La détection des contours avant et pendant 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 motif), figure 11. 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 probable qu'il appartienne à la main.

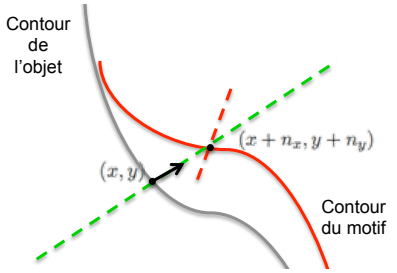

Figure 11. Appariement des contours

La figure 12 montre (en haut) 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. Par ailleurs nous calculons la différence de luminance entre les images avant et pendant l'occultation afin d'isoler les pixels susceptibles d'appartenir à la main (en bas). En tenant compte de la connectivité des pixels, par un seuillage par hystérésis, nous sommes finalement en mesure de segmenter la main (à droite).

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

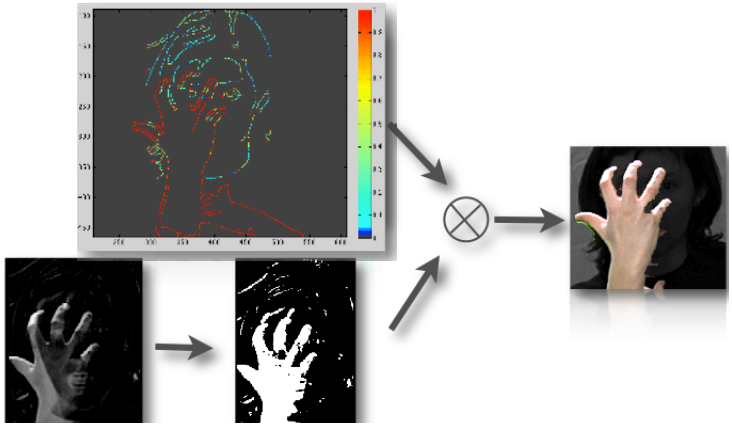

Figure 12. Carte des différences d'orientation des contours (en haut), différence d'illumination (en bas), résultat de segmentation (à droite)

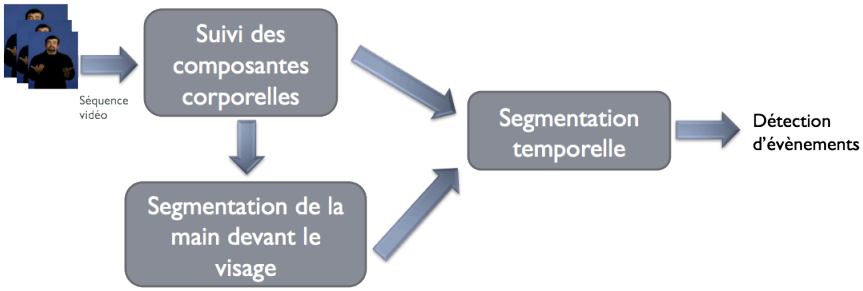

Figure 13. Schéma de segmentation temporelle

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 effet, quand les mains bougent ensemble nous remarquons un léger décalage entre les profils de vitesses des deux mains, bien que leur allure reste très proche comme on peut le voir avec le signe Choqué (figure 14). 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 aussi des événements en milieu de signe. On dit alors que les séquences ont été sur-segmentées.

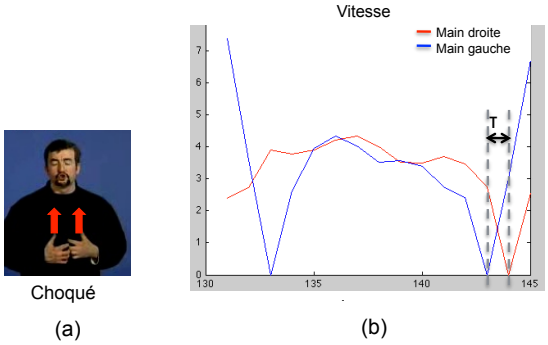

Figure 14. (a) illustre le signes choqué en LSF et (b) la vitesse des deux mains

Par exemple la figure 15a 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 15b montre les événements 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 la plupart des signes avec mouvements répétés, la configuration des mains reste inchangée.

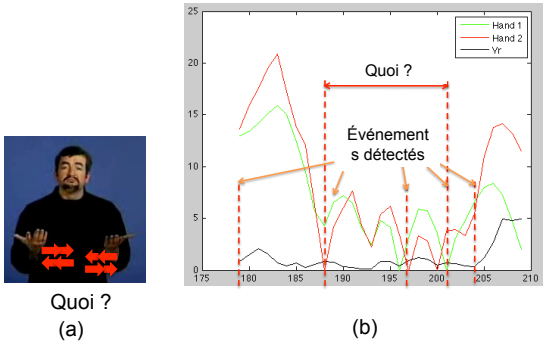

Figure 15. (a) Quoi ? en LSF et (b) les vitesses pour les deux mains, la vitesse relative et les événements détectés

Tableau 1. Classification du mouvement

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

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 pour une même configuration 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 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 16 montre les résultats de segmentation du signe *Quoi ?* en LSF. L'étape précédente a segmenté le signe en tenant compte des caractéristiques de mouvement menant à 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.

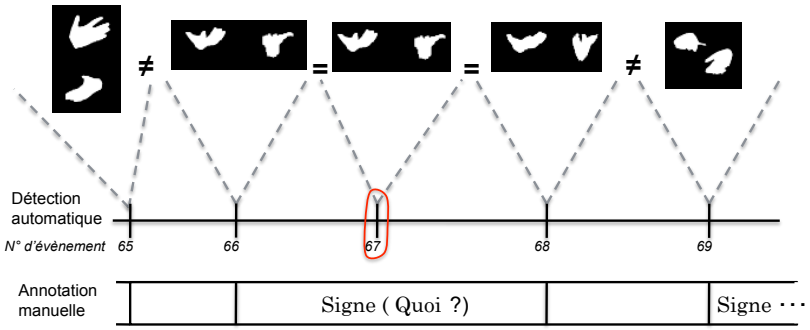

Figure 16. Illustration des mains segmentées pour chaque événement détecté ainsi que la vérité-terrain

6. Évaluations 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 signes utilisant des caractéristiques de mouvement et de forme afin de montrer les avantages et limitations de notre approche.

6.1. Suivi des composantes corporelles

L'évaluation de l'algorithme de suivi a été réalisée sur le corpus vidéo (LS-COLIN, 2002). Une séquence de 3 000 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.* (2009) et Lefebvre (2010). La figure 17 montre les résultats de suivi pour une séquence pendant laquelle la main est devant le visage. Dans le cas des régions rectangulaires (Lefebvre-Albaret, 2010), 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 (figure 17, en haut). Quand les mains et le visage sont modélisés comme des nuages des points (Gianni *et al.*, 2009), 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 (figure 17, au centre). Contrairement aux cas précédents, notre algorithme de suivi est capable de trouver la position de la main même en cas d'occultation, (figure 17, 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*), voir figure 18. Le GTR évalue le suivi correct des objets ; le MTR évalue le suivi d'un objet par un autre filtre du même type, c'est-à-dire les filtres des mains qui s'échangent ; le FTR évalue les résultats des suivis pour les filtres qui ont été échangés mais qui suivent différents types d'objets, e.g. le filtre de la main suit la tête. La figure 19 montre les taux d'évaluation pour les algorithmes proposés dans (Gianni *et al.*, 2009) et (Lefebvre-Albaret, 2010), ainsi que pour notre algorithme de suivi.

Nous remarquons que notre méthode améliore significativement la stabilité en fonction du nombre de particules en comparaison avec les deux autres approches. De plus lorsque plus de 300 particules sont utilisées, notre méthode montre un taux de suivi correct (GTR) > 0.8 et un taux de suivi raté (FTR) ~ 0 donnant des meilleures performances par rapport aux deux autres approches. Cependant on observe aussi une tendance plus grande à échanger les filtres des mains tant que le nombre des particules n'a pas atteint un seuil de l'ordre de $1\ 000$, contrairement aux deux autres méthodes. Avant ce nombre de particules il est très difficile d'identifier la main droite de la main gauche et d'autres traitements sont nécessaires.

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 (Hanke *et al.*, 2010) annoté manuellement. Le corpus est composé de plusieurs conversations à deux signeurs en LSF, avec un total

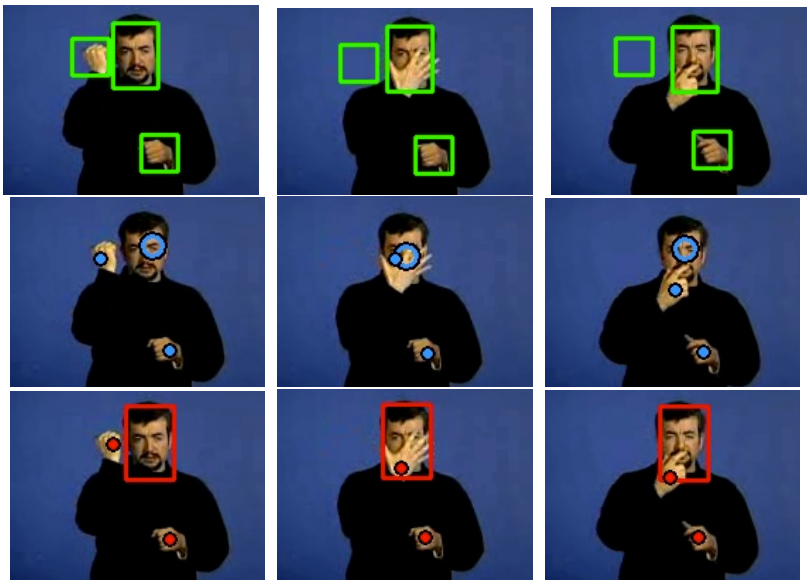

Figure 17. Résultats de suivi pour Lefebvre (en haut), Gianni et al. (au milieu) et notre méthode (en bas)

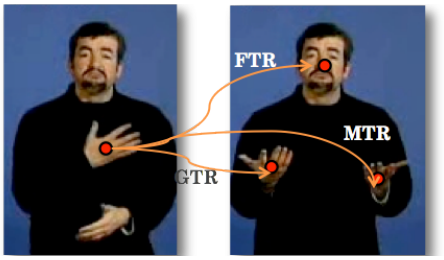

Figure 18. Critère d'évaluation

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 20 montre les résultats de segmentation pour différentes séquences. Notre approche de segmentation a été évaluée quantitativement à l'aide du taux de vrais-positifs (TP), correspondant aux pixels détectés appartenant à la main, et du taux de faux-positifs (FP), correspondant aux pixels détectés n'appartenant pas à la main.

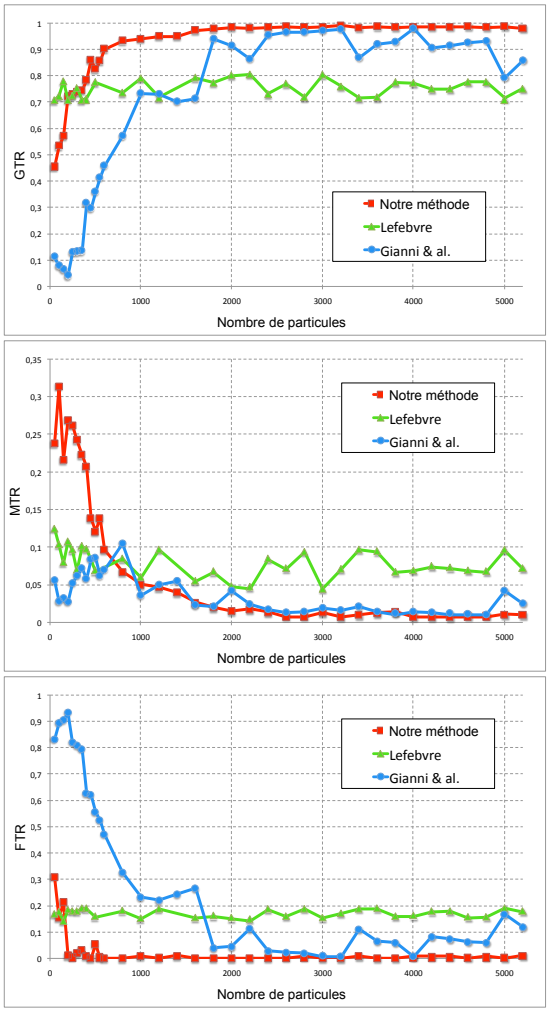

Figure 19. (a) montre le GTR, (b) le FTR et (c) le MTR obtenus par Lefebvre, par Gianni et al. et par notre méthode

$$TPR = \frac{\textit{Nombre des pixels correctement détectés}}{\textit{Nombre des pixels de la main}} \tag{10}$$

$$FPR = \frac{\textit{Nombre des pixels mal détectés}}{\textit{Nombre total des pixels - Nombre des pixels de la main}} \tag{11}$$

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 0,32 % et correspond à

Figure 20. Résultats de segmentation de la main

des pixels qui peuvent être facilement supprimés par des traitements *a posteriori*, e.g. lignes sous le menton ou sur le cou, figure 21.

Tableau 2. Résultats de segmentation des mains

	No. de sequence				
	2	4	7	9	11
$TPR(\%)$	96.71	96.51	96.59	95.07	98.15
$FPR(\%)$	0.14	0.58	0.3	0.32	0.26

Figure 21. Artefacts sous le menton et sous le cou

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. Les trous sont quant à eux 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. La rotation hors plan de la tête pendant occultation est très rare dans le contexte de la langue des signes. Cependant le cas opposé, changement de configuration pendant occultation, reste très fréquent.

Figure 22. (a) Corpus Ls-Colin et (b) corpus DEGELS

6.3. Segmentation automatique des signes

Nous avons réalisé l'évaluation de la segmentation automatique des signes à l'aide de deux séquences vidéo sans aucune contrainte sur la langue : LS Colin (LS-COLIN, 2002) et DEGELS (Boutora, Braffort, 2011), figure 22. L'algorithme de segmentation a été appliqué sur 2 500 images. Les vérités-terrain pour les deux séquences ont été manuellement réalisées par un signeur sourd-né. L'évaluation consiste à compter les événements correctement segmentés en tenant compte d'une tolérance δ (TPR : *True positive rate*), équation (12) et les événements détectés mais qui ne correspondent pas à une limite annotée par rapport au nombre d'événements détectés (FDR : *False discovery rate*), équation (13). Ceci est illustré dans la figure 23.

$$TPR = \frac{1}{N}\sum_{n=1}^{N} c_n \quad avec \quad c_n = \begin{cases} 1 & if \quad l_n - \delta \leq e_m \leq l_n + \delta \\ 0 & otherwise \end{cases} \tag{12}$$

$$FDR = 1 - \frac{1}{M} \cdot \sum_{n=1}^{N} c_n \tag{13}$$

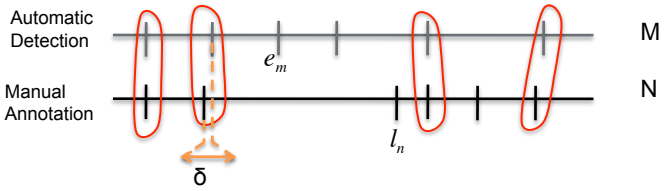

Figure 23. Critères d'évaluations

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 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 de la variabilité de l'expert annotateur.

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 FD diminue de 3 % pour LS-Colin et de 10 % pour le corpus Degels. Pour que l'on puisse utiliser ces résultats

Tableau 3. Résultats de segmentation de gestes

	$Mouvement$		$Mouvement + Forme\ de\ la\ main$	
	$TP(\%)$	$FD(\%)$	$TP(\%)$	$FD(\%)$
$LS-Colin$	81.6	46.2	81.6	44.9
$DEGELS$	74.5	54.2	74.5	44.7

pour l'annotation, notre algorithme génère des données au format ELAN. Il consiste en deux pistes (une pour chaque étape) et trois graphiques qui permettent l'alignement de limites en fonction de maxima ou minima de vitesse.

Afin de montrer que nos résultats de segmentation correspondent bien aux attentes des linguistes nous avons comparé notre segmentation à l'occasion d'un atelier (Braffort, Boutora, 2012), après correction de l'annotateur, aux résultats de segmentation manuelle de (Lefebvre-Albaret, Segouat, 2012) et (Millet, Estève, 2012), figure 24. Cette comparaison a été réalisée pour le corpus DEGELS correspondant à une séquence vidéo de près de 100 signes. La correction consiste à supprimer les événements qui ne correspondent pas aux limites d'un signe. Nous remarquons que nos résultats de segmentation automatique (piste en bas) correspondent ; pour le début à la segmentation d'une équipe (piste au milieu) et pour la fin à la segmentation manuelle de l'autre équipe (piste en haut). La segmentation du début et de la fin du signe ne correspond pas entre les deux annotations manuelles. En effet ces annotations sont subjectives et varient énormément d'un annotateur à l'autre. Or notre algorithme reste objectif et est basé sur des mesures qui permettent d'identifier des changements dans l'intention du signeur, donc l'étape de correction de nos résultats de segmentation reste indépendante de l'annotateur. Dans cette partie aucune évaluation qualitative ne peut être envisagé car il est très difficile de décider laquelle des segmentations correspond mieux aux limites des signes. Cependant ceci nous permet de montrer que nos résultats, issus de mesures objectives, correspondent aux limites choisies par plusieurs annotateurs.

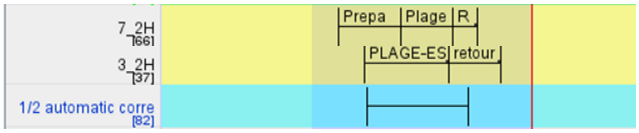

Figure 24. Comparaison d'annotations manuelles (piste en haut et au milieu) et semi-automatiques (piste en bas)

7. Conclusion et perspectives

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 capables d'améliorer la segmentation. Nous avons évalué nos algorithmes sur des vidéos sans contraintes linguistiques : vocabulaire libre, mouvements naturels, signes enchaînés et réalisés en contexte. Cette méthode a montré des résultats prometteurs qui seront utilisés par la suite pour l'annotation en gloses des séquences à l'aide d'un modèle linguistique de la LS. Notre méthode n'utilisant pas d'apprentissage ni d'information linguistique, l'utilisation de modèles de représentation de la langue des signes permettrait de mieux définir les frontières des signes et en même temps d'éliminer des faux positifs.

Remerciements

Ces recherches sont financées par le projet Dicta-Sign, 7^{e} programme cadre Communauté Européenne (FP7/2007-2013) accord no 231135.

Bibliographie

Ahmad T., Taylor C., Lanitis A., Cootes T. (1997). Tracking and recognising hand gestures, using statistical shape models. *Image and Vision Computing*, vol. 15, n° 5, p. 345–352.

Birchfield S. (1998). Elliptical head tracking using intensity gradients and color histograms. In *Proc. cvpr*, p. 232–237.

Boutora L., Braffort A. (2011). *DEfi Geste Langue des Signes. Corpus DEGELS1. Corpus ID oai:crdo.fr:crdo000767 : Video en LSF, informateur A*.

Braffort A., Boutora L. (2012, June). Défi d'annotation DEGELS2012 : la segmentation. In *Jep-taln-recital 2012*, p. 1-8.

Braffort A., Choisier A., Collet C., Dalle P., Gianni F., Lenseigne B. (2004, mai). Toward an annotation software for video of sign language, including image processing tools and signing space modelling. In *Proc. of 4th international conference on language resources and evaluation - lrec 2004*, vol. 1, p. 201–203. Lisbon, Portugal.

Cooper H., Holt B., Bowden R. (2011). Sign language recognition. *Visual Analysis of Humans*, p. 539–562.

Diamanti O., Maragos P. (2008). Geodesic active regions for segmentation and tracking of human gestures in sign language videos. In *Image processing, 2008. icip 2008. 15th ieee international conference on*, p. 1096–1099.

Dreuw P., Ney H. (2008). Towards automatic sign language annotation for the elan tool. In *Lrec workshop on the representation and processing of sign languages: Construction and exploitation of sign language corpora.*

Gianni F., Collet C., Dalle P. (2009). Robust tracking for processing of videos of communications gestures. In Springer-Verlag (Ed.), p. 93–101. Springer.

Gonzalez M., Collet C. (2012, janvier). Segmentation semi-automatique de corpus vidéo en Langue des Signes. In *Actes de la conférence RFIA 2012*, p. 978-2-9539515-2-3. Lyon, France. http://hal.archives-ouvertes.fr/hal-00656505 (Session "Articles")

Grobel K., Assan M. (1997). Isolated sign language recognition using hidden markov models. In *Ieee int. conference on systems, man, and cybernetics*, vol. 1, p. 162–167.

Habili N., Lim C., Moini A. (2004). 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, n° 8, p. 1086–1097.

Hamada Y., Shimada N., Shirai Y. (2002). Hand shape estimation using sequence of multi-ocular images based on transition network. In *Proceedings of the international conference on vision interface.*

Hanke T., König L., Wagner S., Matthes S. (2010). Dgs corpus & dicta-sign: The hamburg studio setup. In *4th workshop on the representation and processing of sign languages: Corpora and sign language technologies (cslt 2010), valletta, malta*, p. 106–110.

Holden E., Lee G., Owens R. (2005). Australian sign language recognition. *Machine Vision and Applications*, vol. 16, n° 5, p. 312–320.

Imagawa K., Lu S., Igi S. (1998). Color-based hands tracking system for sign language recognition. In *Proc. 3rd ieee international conference on automatic face and gesture recognition*, p. 462–467.

Isard M., Blake A. (1998). Condensation-conditional density propagation for visual tracking. *International journal of computer vision*, vol. 29, n° 1, p. 5–28.

Lefebvre-Albaret F. (2010). *Traitement automatique de vidéos en lsf, modélisation et exploitation des contraintes phonologiques du mouvement,*. Phd thesis, University of Toulouse.

Lefebvre-Albaret F., Dalle P. (2009, Feb). Body posture estimation in a sign language video. *In Proc of The 8th International Gesture Workshop*.

Lefebvre-Albaret F., Segouat J. (2012, June). Influence de la segmentation temporelle sur la caractérisation de signes. *DEfi Geste Langue des Signes, JEP-TALN-RECITAL*, p. 73-83.

LS-COLIN. (2002). *13:08 - 15:15 le 11 septembre 2001 par Nasredine Chab http://corpusdelaparole.in2p3.fr/spip.php?article30&ldf_id=oai:crdo.vjf.cnrs.fr:crdo-FSL-CUC020_SOUND.*

MacCormick J., Blake A. (2000). A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, vol. 39, n° 1, p. 57–71.

Micilotta A., Bowden R. (2004). View-based location and tracking of body parts for visual interaction. In *Proc. of british machine vision conference*, vol. 2, p. 849–858.

Millet A., Estève I. (2012, June). Segmenter et annoter le discours d'un locuteur de lsf : permanence formelle et variablité fonctionnelle des unités. *DEfi Geste Langue des Signes*, p. 57-72.

Nayak S., Sarkar S., Loeding B. (2009, June). Automated extraction of signs from continuous sign language sentences using iterated conditional modes. *IEEE Conference CVPR*, p. 2583-2590.

Ong S., Ranganath S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, n° 6, p. 873–891.

Piater J., Hoyoux T., Du W. (2010). Video analysis for continuous sign language recognition. In *Proceedings of 4th workshop on the representation and processing of sign languages: Corpora and sign language technologies*, p. 22–23.

Ramamoorthy A., Vaswani N., Chaudhury S., Banerjee S. (2003). Recognition of dynamic hand gestures. *Pattern Recognition*, vol. 36, p. 2069–2081.

Shiosaki T., Matsuo T., Shirai Y., Shimada N. (2008). Motion segmentation using hand movement and hand shape for sign language recognition.
The Fourth Joint Workshop on Machine Perception and Robotics(MPR2008). Beijing, China.

Smith P., Vitoria Lobo N. da, Shah M. (2007). Resolving hand over face occlusion. *Image and Vision Computing*, vol. 25, p. 1432–1448.

Starner T., Pentland A. (1995). Real-time american sign language recognition from video using hidden markov models. In *Proc. international symposium on computer vision*, p. 265–270.

Tanibata N., Shimada N., Shirai Y. (2002). Extraction of hand features for recognition of sign language words. In *The 15th international conference on vision interface*, p. 391–398.

Theodorakis S., Katsamanis A., Maragos P. (2009). Product-hmms for automatic sign language recognition. In *Proceedings of the 2009 ieee international conference on acoustics, speech and signal processing*, p. 1601–1604.

Von Agris U., Zieren J., Canzler U., Bauer B., Kraiss K. (2008). Recent developments in visual sign language recognition. *Universal Access in the Information Society*, vol. 6, n° 4, p. 323–362.

Wittenburg P., Brugman H., Russel A., Klassmann A., Sloetjes H. (2006). Elan: a professional framework for multimodality research. In *Proc. of the 5^{th} international conference on language resources and evaluation (lrec 2006)*, p. 1556–1559.

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

Zieren J., Canzler U., Bauer B., Kraiss K. (2006). Sign language recognition. *Advanced Man-Machine Interaction*, p. 95–139.

Matilde Gonzalez a obtenu une thèse de doctorat en informatique à l'Université de Toulouse, Université Paul Sabatier en 2012, après avoir obtenu son diplôme d'ingénieur en Génie électrique et de master recherche en traitement d'image, en 2009 à l'Institut National des Sciences Appliquées de Lyon (INSA-Lyon). Ses recherches portent sur les méthodes d'aide à l'annotation des corpus vidéo de la langue des signes par traitement d'image.

Christophe Collet est Maître de conférences à l'IRIT, Université de Toulouse - Paul Sabatier depuis 2004. Il était auparavant chercheur à l'ENS-Cachan et au LIMSI-CNRS (Orsay) où il a obtenu une thèse de doctorat en informatique en 1999. Il a ensuite eu un poste de MdC à l'IUT de Montreuil-Université Paris 8 de 2000 à 2004. Ses recherches concernent le traitement automatique des gestes et particulièrement de la langue des signes, et elles tendent à faire collaborer le traitement d'image, la reconnaissance des formes et le traitement automatique des langues.