



Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation

Matilde Gonzalez

► To cite this version:

Matilde Gonzalez. Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation. Computation and Language [cs.CL]. Université Paul Sabatier - Toulouse III, 2012. English. NNT: . tel-00768440

HAL Id: tel-00768440

<https://theses.hal.science/tel-00768440>

Submitted on 21 Dec 2012

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

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



THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)



Cotutelle internationale avec :

Présentée et soutenue par :

Matilde Gonzalez Preciado

Le 24/09/2012

Titre :

Computer Vision Methods for Unconstrained Gesture Recognition in the
Context of Sign Language Annotation

École doctorale et discipline ou spécialité :

ED MITT : Image, Information, Hypermedia



Unité de recherche :

Institut de Recherche en Informatique de Toulouse

Directeur(s) de Thèse :

Philippe Joly
Christophe Collet

Rapporteurs :

Annelies Braffort
Petros Maragos

Autre(s) membre(s) du jury :

Dominique Boutet
Frédéric Lerasle

*Men, because they lose their health to earn money, then lose money to restore health.
And by thinking anxiously about the future, forget the present, so that ultimately not live
in neither the present nor the future. And live as if they will never die, and die as if
they had never lived.*

Gandhi

Dedicated to my family and to my mother's memory

Acknowledgments

Tout d'abord je voudrais remercier les membre de mon jury, Annelies Braffort, Petros Maragos, Dominic Boutet, Frédéric Lerasle, Philippe Joly et Christophe Collet pour le temps consacré à la lecture de mon manuscrit et l'intérêt porté à mes travaux de recherche.

Something about three years ago I started a new challenge, my PhD. During this period I have learnt quite a lot about life. But I didn't do it alone, I met new people who helped me to grow up not only in a professional way but also in a personal manner. Indeed this work is the result of many contributions unlike what many people would think from my numerous human and professional skills ;-).

I will profit of this place where I can say whatever I want and thank who ever I wish. I write these words full of love and gratitude to those that have make each single moment in my life a memorable one.

First of all I would like to thank all the people from the IRIT. Particularly my PhD co-director, Christophe Collet, who has, during these years, been a support for me in many ways. He has lead me in a good direction listening to me and giving me suggestions from a technical point of view but also from a daily life one. We had nice time travelling in conferences, going to Mexico and eating spicy food... I have many wonderful souvenirs from the time we have spent together. Thank you for everything. More than a director you became a friend.

Thank to all the other PhD students and interns from the IRIT. Particularly to Arturo Curiel for standing me and being so available when I needed the most and also when I did not need it at all. You were there for anything and I will be here also for you. Pamela Carreño for being such a wonderful lady and such a distraction to me (I could have finish my PhD in two years instead of three). You and the wonderful Arthur have take care of me in an amazing way. Also thank to July with whom I shared good moments in home, thank you for teaching me Sign Language and taking care of Latosa while I was away.

Thank to other friends in Toulouse. Danielle Richaud for her support in difficult moments, you helped me and listening to me like a amazing friend. I really thank you for listening my "problemas existenciales". Jerome Lubin you have been an amazing guy and a wonderful friend, always there for anything I really thank you from the deepest place in heart. All my friend in Toulouse; Cris, Lea, Virginie. I love you guys!

I would also like to thank the wonderful people from LIMSI for their collaboration with my researches without you I could not have achieved this PhD thesis; Michael, Max, Annelies and Annik. Michael just can tell you thank you for those wonderful moments and our amazing talks. I found in you a very tender guy, a confident. Max thank you for the amazing moments we passed together in conferences. Annick and Annelies always there taking care of others. I loved how you are such a close team and thank you for letting me be a part of it. That is something I will miss in my new life.

Mexicans from #yosoy132 in Toulouse. After eight years in France this is the first time I feel part of a Mexican community. I felt fighting for my people in Mexico. Adri, Edna, Mario, Javi, Ali, Monica, Angel, etc. A ver, ahora si que vengan a decirme revoltosa ya ponte a estudiar!

My friends from l'Alliance française; Javi, Ana, Sophia, Mire, Dani y Auro (porque ademas de mi hermana eres mi amiga). Although it has been short (because you left such a bad friends) I have to tell you that I missed you during this last year. I had very good moments with you.

I could not have done it with the help of Sergio Rosino and his lovely family, they have make me to feel a member of the family. Muchas gracias por estos años que he podido pasar con ustedes, por su apoyo y su cariño. He podido sentirme parte de una familia a pesar de los miles de kilometros que me separan de la mia. Siempre seran parte de mi vida.

I would like to thank the lovely people that helped me at the beginning of my new life in Nice while finishing my PhD. Thank you Alberto Martinez, Wellington and Sergiu. You have helped at the end of this wonderful adventure by offering to me your support and by sharing this months with me while I was finding a flat.

I want to thank a my wonderful flatmate Grégory Poutrain, you have helped me in several ways when I was preparing my PhD defence... for example cooking :-). Always being there for listening, helping and talking. I love that you are always in good mood, smiling and joking. Always ok for a little beer and talk after work. But the most important is that you are someone I can rely on. You became a true friend and hope you will always be part of my life.

To my lovely family.. Papi, toño y mi Auro. Muchas gracias por todo el apoyo y el cariño que me han dado, sin ustedes no seria lo que soy y no estaria donde estoy. Esta tesis esta dedicada especialmente para cada uno de ustedes. Gracias de todo corazon!

The last but not least, thank Latosa, who moves her little tail every time I came back home. She gives me lot of smiles and happiness while coming back home from a hard day. Gracias bolita de pelos... miaaaauuu... (Many of you will think I am crazy.. and yeah maybe I am but that is the result of many months writing this thesis...)

I could write another 200 pages only acknowledging all those that have made it possible. I write these last words tearfully. I want to thank all of you for being born... Muchas gracias!!

Happiness is only real when shared!!!

Résumé

Cette thèse porte sur l'étude des méthodes de vision par ordinateur pour la reconnaissance de gestes naturels dans le contexte de l'annotation de la Langue des Signes. 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. 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 extrêmement chronophage. De plus, la qualité des annotations dépend des connaissances en LS de l'annotateur. L'association de l'expertise de l'annotateur aux traitements automatiques facilite cette tâche et représente un gain de temps et de robustesse. Le but de nos recherches est d'étudier des méthodes de traitement d'images afin d'assister l'annotation des corpus vidéo: suivi des composantes corporelles, segmentation des mains, segmentation temporelle, reconnaissance de gloses.

Au cours de cette thèse nous avons étudié un ensemble de méthodes permettant de réaliser l'annotation en glose. Dans un premier temps, nous cherchons à détecter les limites de début et fin de signe. Cette méthode d'annotation nécessite plusieurs traitements de bas niveau afin de segmenter les signes et 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 basée sur le filtrage particulaire. 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 temporelle des signes 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. Une fois les signes segmentés, on procède à l'extraction de caractéristiques visuelles pour leur reconnaissance en termes de gloses à l'aide de modèles phonologiques.

Nous avons évalué nos algorithmes à l'aide de corpus internationaux, afin de montrer leur avantages et limitations. L'évaluation montre la robustesse de nos méthodes par rapport à la dynamique et le grand nombre d'occultations entre les différents membres. L'annotation résultante est indépendante de l'annotateur et représente un gain de robustesse important.

Mots-clés : analyse de gestes, langue des signes, annotation automatique.

Abstract

This PhD thesis concerns the study of computer vision methods for the automatic recognition of unconstrained gestures in the context of sign language annotation. Sign Language (SL) is a visual-gestural language developed by deaf communities. Continuous SL consists on a sequence of signs performed one after another involving manual and non-manual features conveying simultaneous information. Even though standard signs are defined in dictionaries, we find a huge variability caused by the context-dependency of signs. In addition signs are often linked by movement epenthesis which consists on the meaningless gesture between signs. The huge variability and the co-articulation effect represent a challenging problem during automatic SL processing. It is necessary to have numerous annotated video corpus in order to train statistical machine translators and study this language. Generally the annotation of SL video corpus is manually performed by linguists or computer scientists experienced in SL. However manual annotation is error-prone, unreproducible and time consuming. In addition the quality of the results depends on the SL annotators knowledge. Associating annotator knowledge to image processing techniques facilitates the annotation task increasing robustness and speeding up the required time. The goal of this research concerns on the study and development of image processing technique in order to assist the annotation of SL video corpus: body tracking, hand segmentation, temporal segmentation, gloss recognition.

Along this PhD thesis we address the problem of gloss annotation of SL video corpus. First of all we intend to detect the limits corresponding to the beginning and end of a sign. This annotation method requires several low level approaches for performing temporal segmentation and for extracting motion and hand shape features. First we propose a particle filter based approach for robustly tracking hand and face robust to occlusions. Then a segmentation method for extracting hand when it is in front of the face has been developed. Motion is used for segmenting signs and later hand shape is used to improve the results. Indeed hand shape allows to delete limits detected in the middle of a sign. Once signs have been segmented we proceed to the gloss recognition using lexical description of signs.

We have evaluated our algorithms using international corpus, in order to show their advantages and limitations. The evaluation has shown the robustness of the proposed methods with respect to high dynamics and numerous occlusions between body parts. Resulting annotation is independent on the annotator and represents a gain on annotation consistency.

Keywords : gesture analysis, sign language, automatic annotation.

Contents

| | |
|---|------------|
| Acknowledgments | 5 |
| Résumé | i |
| Abstract | iii |
| 1 Introduction | 1 |
| Résumé: Introduction | 1 |
| 1.1 General introduction | 3 |
| 1.2 Context | 4 |
| 1.3 Goals | 5 |
| 1.4 Problem statement | 5 |
| 1.5 Specifications | 7 |
| 1.6 Thesis Organisation | 8 |
| 2 French Sign Language | 9 |
| Résumé: Langue des Signes Française | 9 |
| 2.1 Introduction | 11 |
| 2.2 History | 12 |
| 2.3 Sign language characteristics | 13 |
| 2.3.1 Manual and non manual features | 13 |
| 2.3.2 Temporal and Spatial Organisation | 14 |
| 2.3.3 Linguistic Elements | 15 |
| 2.3.4 Coarticulation | 17 |
| 2.4 Conclusion | 19 |
| 3 Sign Language Corpora and its notation: State of the art | 21 |
| Résumé: Corpus en langue des signes et leur annotation | 22 |

| | | |
|-------|--|----|
| 3.1 | Introduction | 23 |
| 3.2 | Data acquisition | 23 |
| 3.2.1 | Active Sensors | 24 |
| 3.2.2 | Passive Systems: Sign Language Video Corpus | 27 |
| 3.2.3 | Discussion | 30 |
| 3.3 | Corpus annotation | 31 |
| 3.3.1 | From a linguistic point of view | 31 |
| 3.3.2 | From a computer vision point of view | 32 |
| 3.3.3 | Annotation Tools | 33 |
| 3.3.4 | Architecture of a Annotation System | 34 |
| 3.3.5 | Discussion | 36 |
| 3.4 | Linguistic representation of signs | 38 |
| 3.4.1 | Parametric approaches | 39 |
| 3.4.2 | Temporal approaches | 42 |
| 3.4.3 | Discussion | 45 |
| 3.5 | Feature extraction | 47 |
| 3.5.1 | Hands and head location and motion | 47 |
| 3.5.2 | Hand shape | 55 |
| 3.5.3 | Sign boundaries | 60 |
| 3.5.4 | Discussion | 65 |
| 3.6 | Automatic Recognition of Signs | 66 |
| 3.6.1 | Speech recognition techniques for SL recognition | 67 |
| 3.6.2 | Classification methods | 68 |
| 3.6.3 | Discussion | 71 |
| 3.7 | Sign Language Generation | 73 |
| 3.7.1 | Manual generation | 73 |
| 3.7.2 | Automatic Generation | 75 |
| 3.7.3 | Discussion | 77 |
| 3.8 | Conclusion | 79 |

| | | |
|----------|--|-----------|
| 4 | Sign language automatic annotation by SL generation | 81 |
| | Résumé | 81 |
| 4.1 | Introduction | 83 |
| 4.2 | Skin Model | 84 |
| 4.2.1 | Colour space | 84 |
| 4.2.2 | Skin pixels learning data set | 85 |
| 4.2.3 | Specific model building | 86 |
| 4.2.4 | Skin segmentation algorithm | 88 |
| 4.2.5 | Conclusion | 89 |
| 4.3 | Body tracking from a mono camera | 90 |
| 4.3.1 | Particle filter | 92 |
| 4.3.2 | The model | 96 |
| 4.3.3 | Multiple object tracking | 100 |
| 4.3.4 | Tracking algorithm structure | 103 |
| 4.3.5 | Experimental Results | 104 |
| 4.3.6 | Conclusion | 108 |
| 4.4 | Hand Segmentation | 111 |
| 4.4.1 | Occlusion detection | 113 |
| 4.4.2 | Hand segmentation without occlusion | 116 |
| 4.4.3 | Hand segmentation during occlusion | 117 |
| 4.4.4 | Experimental results | 125 |
| 4.4.5 | Conclusion | 127 |
| 4.5 | Temporal segmentation | 129 |
| 4.5.1 | Motion Classification | 132 |
| 4.5.2 | Shape features | 134 |
| 4.5.3 | Experimental results | 136 |
| 4.5.4 | Conclusion | 141 |
| 4.6 | Semi-Automatic Annotation of Glosses | 142 |
| 4.6.1 | Gloss recognition from querying Zebedee | 146 |

| | | |
|----------|--|------------|
| 4.6.2 | Gloss recognition from ReZeBeDee | 150 |
| 4.6.3 | Experiments and results | 156 |
| 4.6.4 | Conclusion | 160 |
| 5 | Conclusions and perspectives | 163 |
| | Résumé | 163 |
| 5.1 | Conclusion | 165 |
| 5.1.1 | Summary: Our Contributions | 165 |
| 5.1.2 | Problem statement and specifications | 166 |
| 5.2 | Perspectives | 167 |
| A | Finger spelling | 169 |
| B | Zededee XML example | 171 |
| | Bibliography | 175 |

List of Figures

| | | |
|------|---|----|
| 2.1 | Example of sign [PARIS] in LSF. Source IVT [Girod 1997] centre and right. | 11 |
| 2.2 | Sign [SMALL] and [HUGE], left and right respectively, in LSF. | 14 |
| 2.3 | Signing space in front of the signer | 15 |
| 2.4 | Modelling of the signing space shared between two signer. Image extracted from [Lenseigne 2005] | 15 |
| 2.5 | Example sign language performance switching between the two ways of saying; iconicity and standard signs. Images are extracted from [LS-COLIN 2002] | 16 |
| 2.6 | Shape and size transfer (left), situational transfer (center) and personal transfer (right). Images extracted for [LS-COLIN 2002] | 17 |
| 2.7 | Example of <i>co-articulation</i> : gesture between the end of the sign [UNITED STATES] and the beginning of the sign [TOWER] in French Sign Language | 18 |
| 2.8 | Sign [DEAF] in French Sign Language in different context | 18 |
| 3.1 | Motion capture sensors [Moeslund 1999] | 24 |
| 3.2 | Hybrid systems: Hy-BIRD ¹ | 25 |
| 3.3 | Mechanic sensors: exoskeleton Gypsy 7 (left) and CiberGlove II (right) ² | 25 |
| 3.4 | Magnetic sensors: MotionStar ³ | 26 |
| 3.5 | Optical tracking: IMPULSE motion capture system (left), optical glove ⁴ (left) | 26 |
| 3.6 | Inertial sensors: Xsens's Moven ⁵ (right) and the AcceleGlove ⁶ (left) | 27 |
| 3.7 | Ls-Colin corpus configuration | 28 |
| 3.8 | Degels corpus (left), Ls-Colin corpus (centre) and Dicta-Sign (right) . . . | 28 |
| 3.9 | AnColin annotation software example | 34 |
| 3.10 | Annotation Tool Example: (a) Normal environment, (b) A ³ call and (c) A ³ result. | 35 |
| 3.11 | Distributed system architecture for assisted annotation of video corpus . | 35 |
| 3.12 | Sign [NAKED] and [CHOCOLATE] in French Sign Language. Notice that only the movement of right hand is different. Source IVT | 38 |

| | | |
|------|--|----|
| 3.13 | Finger configuration: (a) Four fingers bent, (b) all fingers extended, (c) two fingers curved and all other fingers closed and (d) index extended and all other fingers closed. | 39 |
| 3.14 | Hand configuration kind example. Image extracted from [Sandler 2012] . | 40 |
| 3.15 | Sign [DEAF] in French Sign Language | 40 |
| 3.16 | Sign [SEND] in French Sign Language involves a change on configuration while the movement is taking place. | 41 |
| 3.17 | Sign [SEA] in French Sign Language involves a complex movement to illustrate the movement of the waves. | 41 |
| 3.18 | Sign [RACIST](Source [Companys 2004]) and [SKIN] (Source IVT) in LSF respectively. Manual features are the same only the facial expression changes. | 41 |
| 3.19 | Sign [WHAT?] in LSF (Source IVT) and its notation in HamNoSys and SignWriting representation. | 42 |
| 3.20 | The two Zebedee description axes [Filhol 2008] | 43 |
| 3.21 | Key postures for the sign [BALL] in French Sign Language using Zebedee representation illustrating the elements used to describe the sign, e.g. [loc] corresponds to the imaginary centre, all the other parameters depending on it. Source [Filhol 2008] | 44 |
| 3.22 | Sign [BUILDING] in French Sign Language. Source [Braffort 2008] and different performance depending on what we image to add | 45 |
| 3.23 | HamNoSys "flat" and "bent" hand configurations | 45 |
| 3.24 | Customized colour glove example. Image extracted from [Wang 2009]. . . | 48 |
| 3.25 | Object representation. (a)centroid, (b) cloud of points, (c) rectangular shape, (d) elliptical shape, (e) articulated model, (f) skeleton model, (g) control points on object contour, (h) object contour and (i) object silhouette. Image extracted from [Yilmaz 2006]. | 49 |
| 3.26 | Articulated hand models for hand tracking and pose estimation. Image extracted from [Lu 2003]. | 51 |
| 3.27 | Hand tracking example using a 3D articulated model. Image extracted from [Wang 2009]. | 53 |
| 3.28 | Example of the 3D articulated model used with the Annealed Particle Filter. (a) shows the segments and joint angles and (b) the articulated geometrical model. Image extracted from [Duetscher 2000]. | 53 |
| 3.29 | Example of the results performed using image force field. Image extracted from [Smith 2007]. | 56 |

| | | |
|------|--|-----|
| 3.30 | Illustrates the hand extraction using a template for hand and head before occlusion. Image extracted from [Tanibata 2002]. | 56 |
| 3.31 | Geometric features. Image extracted from [Von Agris 2008]. | 59 |
| 3.32 | Segmentation comparison between three different teams segmenting the same sequence. Source [Braffort 2012]. | 61 |
| 3.33 | Boarder detection parameter. Source [Liang 1998]. | 63 |
| 3.34 | Word model recognition system components. Source [Von Agris 2008]. . . | 69 |
| 3.35 | Sub-unit model recognition system components. Source [Von Agris 2008]. | 70 |
| 3.36 | Rotoscoping for SL generation. Source [Braffort 2010] | 74 |
| 3.37 | Generation from motion capture [Elliott 2000] | 74 |
| 3.38 | Example of [WHEN?]. Source [Girod 1997] | 76 |
| 4.1 | Skin segmentation result using an explicitly defined skin model in RGB colour space | 86 |
| 4.2 | Bivariate normal distribution $C_b C_r$ | 87 |
| 4.3 | Skin segmentation algorithm | 88 |
| 4.4 | Skin probability map S_k obtained in RGB (left) and $YC_b C_r$ (right) . . . | 89 |
| 4.5 | Hands dynamics example | 90 |
| 4.6 | Hand over face occlusion and hand shape changing example | 91 |
| 4.7 | Head distance between two frames | 91 |
| 4.8 | Head and hands position without penalisation | 92 |
| 4.9 | Probability density with associated weights [Isard 1998] | 93 |
| 4.10 | Without an annealing effect, the particles get stuck in the local maximum (left). In order that the particles escape from the local maximum, the annealing effect is used (right). (To change) | 96 |
| 4.11 | (a) Illustrates the proposed face model. (b) shows the rectangular model lying over the skin probability map when ρ_{R_T} is maximal and (c) represents the best matching position for the template registration. | 97 |
| 4.12 | Skin probability map (left), head particles weight (middle) and expectation result (right) without particles penalization (up) and with penalization from hand samples (bottom). | 100 |
| 4.13 | Penalization coefficients computing | 102 |

| | | |
|------|---|-----|
| 4.14 | Head template updating principle | 102 |
| 4.15 | Normalize root mean square error for different threshold values. | 103 |
| 4.16 | Schematic representation of the proposed approach. | 103 |
| 4.17 | Skin probability map for each objet. Notice that other object are penalized. | 104 |
| 4.18 | Tracking results. For Lefebvre (top row), Gianni <i>et al.</i> (middle row) and the proposed approach (bottom row). | 105 |
| 4.19 | LS COLIN corpus example | 106 |
| 4.20 | Evaluation criteria for comparison with other tracking algorithms | 107 |
| 4.21 | Good tracking rate GTR achieved by Lefebvre, Gianni <i>et al.</i> and our approach. | 108 |
| 4.22 | Missed tracking rate GTR achieved by Lefebvre, Gianni <i>et al.</i> and our approach. | 109 |
| 4.23 | False tracking rate GTR achieved by Lefebvre, Gianni <i>et al.</i> and our approach. | 109 |
| 4.24 | Execution time for Lefebvre, Gianni <i>et al.</i> and our approach. | 110 |
| 4.25 | Signs where the hand is deliberately placed in front of the face. | 111 |
| 4.26 | Example of two regions of the face before and during occlusion; cheek and eyes. The former illustrates that the contours give more information about hand boundaries. The latter shows that contours classification is challenging but colour gives additional information to determine hand region. | 112 |
| 4.27 | Hand segmentation template and the template used for the segmentation. Template obtained from distance (left) and best template just before the occlusion (right) | 114 |
| 4.28 | (a) face model used for tracking, rectangles configuration without occlusion (b) and with occlusion (c) | 114 |
| 4.29 | Head signature: (left) head contours and (right) EOH in polar coordinates | 115 |
| 4.30 | Labelled skin map. Without (left) and with (right) occlusion. | 116 |
| 4.31 | Occlusion function results for a short sequence. | 118 |
| 4.32 | Head signature (right) and best position (left). Black rectangle corresponds to the initialisation and red rectangle to the optimal position. | 120 |
| 4.33 | Distance map between the researched area and the head signature | 120 |
| 4.34 | Head signature using gradient orientation | 121 |
| 4.35 | Edges and appearance information for hand segmentation. | 122 |

| | | |
|------|---|-----|
| 4.36 | Contours from the face template and the image during occlusion. | 123 |
| 4.37 | Edge matching and edge orientation difference | 123 |
| 4.38 | Edges orientation difference map | 124 |
| 4.39 | Luminance difference between the template and the image during occlusion. | 124 |
| 4.40 | Combination of edges orientation and luminance difference. | 125 |
| 4.41 | First row shows five consecutive frames in a sequence with hand over face occlusion. The second row shows the segmentation result. Pixels in dark grey or in black have been classified as belonging to the non hand class, otherwise they are shown in their natural colour. | 126 |
| 4.42 | This figure shows the segmentation results for 3 different sequences. Each row corresponds to a sequence. | 127 |
| 4.43 | Segmentation results: artefacts under the chin and over the collar | 128 |
| 4.44 | Annotation tool <i>Elan</i> . Manual Ann. is the segmentation results by an human annotator and Auto Ann. tier shows the expected segmentation results. | 129 |
| 4.45 | Event detection algorithm schema | 130 |
| 4.46 | Event detection algorithm schema | 131 |
| 4.47 | Illustrates velocity of right and left hand. <i>Left</i> : The sign 'shocked' in French Sign Language. <i>Right</i> : Velocity profile for right and left hand. . | 133 |
| 4.48 | Illustrates velocity profile of right and left hand superposed. | 134 |
| 4.49 | Sign [WHAT?] in LSF (top-left). Double arrow represent a repetitive gesture. Hands velocities and the relative velocity with the classification results are shown on the left and the results of the detected events are shown on the right. | 134 |
| 4.50 | Frames corresponding to the detected events. | 135 |
| 4.51 | Illustrates hand segmentation for each detected frame. | 136 |
| 4.52 | Eccentricity and equivalent diameter measurement. | 137 |
| 4.53 | Evaluation criteria. | 138 |
| 4.54 | Manual annotation illustration. | 139 |
| 4.55 | [PASSPORT] and [EXAM] in LSF are homosigns. Source IVT [Girod 1997] | 143 |
| 4.56 | [BOY] and [SURGERY] in LSF could be homosigns if is a head surgery. Source IVT [Girod 1997] | 143 |

| | | |
|------|--|-----|
| 4.57 | Classification of sign within the database in terms of the number of transitions. | 144 |
| 4.58 | Gloss classification tree | 146 |
| 4.59 | Sign [SHOCKED] and [BUILDING] in LSF. | 151 |
| 4.60 | Skeleton used for the automatic generation of signs. Source [Delorme 2011] | 152 |
| 4.61 | Sign [BUY] in LSF. Linguistically both hands are used for performing the sign but only one hand moves. Source IVT [Girod 1997] | 153 |
| 4.62 | Movement direction quadrants. | 154 |
| 4.63 | Signing space sectors. Hand is labelled according to the sector and its neighbours. | 155 |
| 4.64 | Sign [DEAF] in French Sign Language in different context | 157 |
| 4.65 | Sign <i>we/us</i> in FSL showing the potential glosses | 158 |
| 4.66 | Sign [FACE], [EUROPE] and [SPAIN] respectively. | 160 |
| 4.67 | Sign <i>we/us</i> in FSL showing the potential glosses | 161 |

List of Tables

| | | |
|------|--|-----|
| 3.1 | Corpus with available ground-truth | 30 |
| 3.2 | Filter command predicate | 46 |
| 4.1 | Object and observation model for head and hands. | 100 |
| 4.2 | Results of the evaluation rates trough several sequences | 127 |
| 4.3 | Motion classification | 133 |
| 4.4 | Similarity measurements used for hand shape characterisation. | 136 |
| 4.5 | Evaluation results for motion segmentation and motion with hand shape improvement for several tolerance values. | 139 |
| 4.6 | Segmentation results using the automatic tracking and our automatic hand segmentation approach for a tolerance $\delta = 2$ | 140 |
| 4.7 | Segmentation results for a full-automatic sign segmentation with automatic tracking and hand segmentation algorithm and for $\delta = 2$ | 140 |
| 4.8 | Movement structure statistics (%) | 145 |
| 4.9 | Movement structure statistics for 1T (%) | 145 |
| 4.10 | Feature classification results | 156 |
| 4.11 | Feature classification results | 157 |
| 4.12 | Number of potential glosses | 158 |
| 4.13 | Number of hands percentage | 159 |
| 4.14 | Movement direction results | 159 |
| 4.15 | Relative position results | 160 |

Introduction

Résumé: Introduction

La langue des signes (LS) est une langue visio-gestuelle utilisée 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. L'extrême variabilité et l'effet de co-articulation représentent un problème important dans les recherches en traitement automatique de la LS. Plusieurs contributions se trouvant dans l'état de l'art ciblent plusieurs problèmes dans le traitement de la LS ; acquisition des données, annotation de signes, extraction de caractéristiques, etc. Ils nécessitent, généralement, de nombreuses vidéos annotées qui sont réalisées manuellement par des linguistes ou experts en LS. Ceci est source d'erreur, non reproductible et extrêmement chronophage. De plus, la qualité des annotations dépend des connaissances 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. C'est pour ça que nous proposons une nouvelle méthode d'aide à l'annotation de la LS définie comme :

Annotation semi-automatique de la Langue des Signes, en terme des gloses, utilisant une description linguistique adaptée aux caractéristiques visuelles extraites à l'aide d'un système de génération automatique des signes.

En effet, nous proposons d'utiliser une description des signes développée, à la base, pour la génération automatique de la LS, et adapté ensuite pour la reconnaissance de la LS. Cette nouvelle représentation est créée automatiquement à partir de la génération automatique. En effet chaque signe est généré aléatoirement afin de prendre en compte la variabilité de signes. À l'issue de cette génération nous sommes en mesure d'extraire des caractéristiques visuelles compatibles à celle que l'on peut extraire à partir d'une vidéo. De cette façon nous proposons à l'annotateur des signes potentiels. Cette annotation peut être utilisée pour développer des systèmes de reconnaissance automatique de la LS.

Dès nos jours, l'accessibilité des sourds est un problème important. Les recherches de la LS ciblent le développement d'applications permettant aux personnes sourdes de communiquer au quotidien; outils d'apprentissage, systèmes de traduction, etc. Pour cela plusieurs projets existent. Le projet Européen Dicta-Sign cherche à rendre internet accessible en Langue des Signes. D'une part en cherchant à rendre les vidéos en LS

anonymes et d'autre part en développant un dictionnaire traducteur d'une langue des signes à une autre. Dans ce contexte des données annotées sont nécessaires. Cette tâche est attribuée à l'institut de Recherche en Informatique de Toulouse (IRIT), où cette thèse a été réalisée, plus précisément au sein l'équipe Traitement et compréhension d'images TCI.

Notre principal objectif est de proposer aux linguistes et informaticiens des données annotées de qualité afin de mieux analyser la Langue des Signes. Nous souhaitons étudier et développer des méthodes de traitement d'image permettant d'extraire des caractéristiques des vidéos en LS. Ceci nous permet de proposer des données d'annotation le but étant de diminuer le temps d'annotation, de rendre l'annotation reproductible et indépendante des connaissances et de l'expertise de l'annotateur. De plus nous voudrions rendre ces traitements accessible à la communauté scientifique.

Le développement des traitements automatiques de la LS sont complexes car cette langue a sa propre structure spatio-temporelle. En effet l'espace devant le signeur est utilisé pour positionner des entités et créer des relation entre elles. La difficulté dans les recherches de la LS ne concerne que la spécificité de la langue mais aussi les limitations d'un point de vue informatique. Afin d'extraire des caractéristiques dans des vidéos de LS nous avons besoin de suivre les composantes corporelles, de segmenter la main même en cas d'occultation. Ces informations permettent d'extraire des caractéristiques de mouvement et de forme afin de caractériser les signes.

Les problèmes rencontrés lors de notre étude concernent la dynamique du mouvement et la grande variabilité de forme de la main. De plus en LS il y a des nombreuses occultations entre les main et le visage. Ceci rend le suivi et la segmentation de la main challenging à cause de la similarité de couleurs de ces objets. Par ailleurs la représentation des signes et la reconnaissance en gloses est difficile dû à la grande variabilité des signes et leur dépendance au contexte.

Ces problèmes permettent de déterminer les spécifications et les contraintes de notre système. L'extraction des caractéristiques de forme et de mouvement doit être robuste à la dynamique du mouvement, à la variabilité de la forme de la main et à la similarité de couleur. La représentation de signes doit être générique, c.a.d. la représentation de chaque signe doit tenir en compte sa variabilité en fonction du contexte. Finalement pour la reconnaissance des signes, l'utilisation de caractéristiques de bas niveau sont souhaitables afin de rendre notre approche indépendante d'une LS. De plus il faut éviter l'utilisation d'apprentissage utilisant des données annotées pour ne pas biaiser nos résultats.

La suite de cette thèse est organisée de la façon suivante. Le Chapitre 2 présente les caractéristiques de la Langue des signes. Le Chapitre 3 décrit les approches dans la littérature concernant l'annotation de corpus vidéo et le Chapitre 4 détail nos contributions. Finalement le Chapitre 5 présente nos conclusions et nos perspectives.

1.1 General introduction

Sign Languages (SL) are visual-gestural languages used by deaf communities. They are characterised by the movement of hands, shoulders, head, etc. The performance of a sign is highly variable because of the strong context-dependency of signs, i.e. the same sign can be performed in a different way depending on the context. In addition, in continuous SL, i.e. performing one signs after another, one sign influences the following sign and itself is influenced by the previous sign. This phenomenon is called co-articulation effect or movement epenthesis. All these make the automatic processing of SL a challenging task.

Sign Language is a complex language, see Chapter 2, and its automatic processing is challenging. Numerous contributions in the state-of-the-art, see Chapter 3, intend to address sign language recognition issues; data acquisition, sign notation, features extraction, SL annotation, etc. Generally the automatic processing of SL needs high amounts of training data. Collecting these data is, in general, manually performed by linguists and computer scientists. This is time consuming, error prone and unreproducible. In addition the quality of the results depends on the annotators knowledge and experience. These problems are addressed, in this work, using a novel approach which is defined as:

Automatic Sign Language annotation, in terms of glosses, using a linguistic description of signs adapted to computer vision features through sign language generation

We propose to address SL annotation using linguistic description of signs, firstly developed for SL generation, lately extended for SL recognition. This new model is automatically created through the random generation, within the possible variations, of the same sign several times which allows to extract common characteristics of several productions of a sign. This leads to a model implicitly considering sign variability and context-dependency. Sign descriptions are stored in a database for further querying. Finally through Image processing techniques representative features are extracted, see Chapter 4.3, 4.4 and 4.5, e.g. number of moving hands, movement direction, etc., to query the database obtaining, then, a list of potential signs proposed to the annotator, see Chapter 4.6.

This annotation system has been evaluated at each stage in order to point out its limitations and performance. The correction (if required) by the annotator is then performed. This novel approach shows promising results using a different way of recognising signs.

In this chapter we, first, present the context in which this PhD thesis has been held, Section 1.2. Second the goals to achieve and the problem statement are detailed in Section 1.3 and Section 1.4 respectively. Afterwards we present the requirements that our system must meet, Section 1.5. Finally we describe the thesis organisation in Section 1.6.

1.2 Context

One of the main concerns in SL computer science research is the study and development of applications to improve the communication accessibility of deaf people in a daily life; teaching tools, recognition systems, translation machine, etc. For this, several projects have been founded by regional, national, European or international programs. This PhD thesis is part of the Dicta-Sign¹ project belonging to the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°231135. The main goal concerns the accessibility of deaf people to the internet in their native language through the study and development of computer vision techniques. The motivation of allowing people to interact using SL is that, mostly, deaf-born people neither know how to read nor how to write oral languages. In France about 80% of French deaf-born are illiterate [Gillot 1998].

On the internet users can leave their messages and comments in an anonymous way by writing them down. The problem arises when deaf people want to use this kind of tools to discuss about any subject. Unlike oral languages SL do not have a traditional or formal written form, many graphical or computational representations have been proposed, e.g. HamNoSys [Prillwitz 1989] or Sign Writing [Sutton 1995], see Section 3.4, but they are mainly used by linguists and computer scientists to study SL. Posting a comment, on the internet, in sign language means uploading a video of the signer. However this is not anonymous and the signer might feel uncomfortable and might not express himself as desired.

Contrary to what might be thought there is not one universal SL. Indeed SL is influenced by the culture and how deaf communities see the world. In oral languages many websites allow to translate words or sentences from one language to another. In sign language this is not possible, available dictionaries on the internet (Pisourd, Spread the sign, sematos, etc.) only translate from the written form of an oral language or the representation of a sign to SL which is not accessible to illiterate deaf people.

In the Dicta-Sign project the proposed solution to the problems mentioned before concerns image processing techniques, SL computational representations, signing avatar, etc. This is achieved by the collaboration of several research teams [Eleni Efthimiou 2010], among them the Institute of Computer Research of Toulouse (Institut de Recherche en Informatique de Toulouse, IRIT). The Image Processing and Understanding team (Traitement et compréhension d'images, TCI) at IRIT is in charge of the study and development of image processing techniques to assist the annotation of SL video sequence. The annotation of SL video sequences consists on furnishing any critical commentary, explanatory notes or extracted features. This annotation is then used by our collaborators to study the language, build models, etc. This PhD is carried out within this team and specifically focus on the annotation of SL video sequence.

1. <http://www.dictasign.eu/>

1.3 Goals

This PhD thesis aims to provide linguists and computer scientists with good quality annotated data for studying the language and performing learning tasks, avoiding as much as possible human intervention. Our main goal is to improve annotation results in terms of time-spent, reproducibility and robustness. However at some point it is necessary to think about what improving SL annotation means? We consider that it is not important if the total annotation time with the assisted tools remains the same as a fully manual annotation as long as it is robust and can be reproduced by several persons. For this we use image processing techniques to minimise the dependence of the results on the annotator's knowledge and experience. Since so far, manual annotation results are not reproducible even by the same annotator. Then, annotation performed by two different persons might have a large influence in the annotation results, thus, the representativeness and the quality of the data in the learning step of automatic recognition systems.

In short achieving our goals correspond to

- decrease the annotation time by using the proposed annotation algorithms.
- make the annotation reproducible and independent from annotators knowledge and experience.
- avoid the annotation results to be dependent to any training data.
- make available the annotation algorithms to a large scientific community.

At the end of this work we desire to provide, through the intermediate algorithms for gloss recognition, several annotation algorithms allowing to annotated different features such as the position of hands and head, hand or sign segmentation. Reaching our goals leads to a deep knowledge about the problem statement and the problems to face from a linguistic and a computer vision point of view.

1.4 Problem statement

The annotation of SL video sequences is needed to build recognition systems and linguistic models. From a linguistic point of view SL processing is challenging. Sign Language has its own temporal and spatial organisation. It uses the space in front of the signer, called signing space, to place entities and relationships between the different entities can be created, also the signing space can be shared with other signers. The complexity of the language make the development of computer vision approaches very difficult and might require linguistic models to have additional information. In addition to the challenges imposed by the language itself, many computational limitations have to be addressed. For example depending on the application the video recording conditions can significantly change. In the case of applications concerning SL research, video recording is, in general, carried out by linguists and computer scientists. For linguistic purposes, often, the recording conditions are very simple and consists of a single frontal view camera which fulfils their needs. Other applications such as 3D reconstruction or computer vision

applications can use a complex set-up with several cameras, a stereo camera or even more complex devices such as the Kinect (recently available for consumers). Since our goal is to assist linguists to perform the annotation task of SL video corpora, the study and development of our algorithms are restraint to a mono-camera view. The fact of using a mono-view makes the extraction of features from any sign or from the signing space much more challenging because the depth information is missing. This represents a very important problem because when an object is in front of another, e.g. the hand in front of the face, it is difficult, even impossible to know if objects are in contact or if one object is passing in front of the other.

Although any approach developed for SL recognition can be used for assisting SL annotation, the constraints are very different. In SL recognition computation time is important since, generally, we intend to achieve real time applications. In addition using learning data is common to train the system with possible performances of a sign according to the context. In the case of SL annotation, execution time is not really a constraint as long as it allows to save time during the annotation task which depends on the complexity of the SL performance and the information to furnish. For example for manually segmenting a two-minute video, which consists on selecting the beginning and the end frame of each sign, it is necessary 25 minutes which is more than 12 times the length of the video. Also, although learning data can be used for annotation purposes it is not suitable because the data at the first step might be manually annotated, then, it is preferable to avoid using any learning step.

Many problems to extract features from a video, either from the language or from the object features, are faced. For:

- **tracking:** The dynamics and high variability of hands as well as the total or partial occlusion between similarly coloured objects.
- **hand segmentation:** When hand is in front of the face it is very challenging to distinguish hand pixels from face pixels because of the appearance similarity between the two objects in addition to the complex shape of hand.
- **temporal segmentation:** It is difficult to determine where are the beginning and the end of a sign in continuous SL because of the coarticulation effect (Section 2.3.4).
- **sign representation:** High variability of signs and context-dependency. The same sign can be performed in a very different way depending on the context.
- **gloss recognition:** The same performance could correspond to various different signs. The only way to disambiguate this situation is using high level characteristics through grammatical models or other higher levels.

In this PhD is intended to study and develop image processing approaches to assist the annotation taking into account the complexity of the language and the computational limitations. For this some specifications have to be met by our system.

1.5 Specifications

The proposed system for the annotation of SL video sequences, in terms of glosses, requires the study of several algorithms addressing different problems. Three main tasks are pointed out; (i) developing robust feature extraction algorithms, for continuous SL, to obtain representative characteristics from a performance; (ii) proposing an extended sign representation for SL or gesture recognition; and (iii) filtering the glosses from the sign representation using the extracted features.

- **Feature extraction:** Features used here are motion and hand shape. For this a body limb tracking approach and a hand segmentation algorithm are required. These methods must be robust to high dynamics, hands shape variability and occlusions. Special attention has to be paid during occlusion either for tracking or for hand segmentation. Also to process continuous SL, a temporal segmentation algorithm is needed to determine limits; over segmentation is preferred than under-segmentation for a faster correction by annotators.
- **Sign representation:** A representation model describing signs, as generic as possible to consider context dependency and sign variability, is required. This description might allow to filter signs having similar visual characteristics as the ones extracted from computer vision techniques, e.g. the number of hand performing the signs, the movement direction, etc. A SL generation system is required in order to obtain from synthetic data common features for extending the formal model. The goal of this is to be able to generate several performances of signs with several parameters and only extracts what remain stable regardless the selected parameters.
- **Gloss recognition:** Low level feature are suitable to achieve gloss recognition and remain independent from any specific SL. Also using any high level grammatical model is out of the scope of this work. Then several glosses can be proposed to users for selecting the one corresponding to the video sequence.

The specifications described above are considered to justify any choice during the development of our work. For the remaining document we refers to Sign Language, notated SL, when our comments concern any SL regardless the country or region otherwise it will be specified whether it concerns only French Sign Language, notated LSF, or any other language. Also we refer to a sign in the text, i.e. the corresponding word in English, as written in upper-case and within square brackets [], e.g. [SIGN].

1.6 Thesis Organisation

This PhD thesis is accessible to linguists, computer scientist or any person interested in the automatic processing of SL video corpus and is organised in four parts:

- The first chapter, Chapter 2, concerns the description of French Sign Language in the society as well as the characteristics of SL performances. This part will allow readers to get familiarized with the linguistic vocabulary and to be aware of the current situation of the French Sign Language. Moreover this part will allow readers to understand why it is a challenging domain through the description of the specificity and complexity of SL.
- The second chapter, Chapter 3, presents the state-of-the-art, about the different works proposed in various domains; corpus acquisition and annotation, computational modelling, image processing techniques, automatic SL processing and SL generation. This part will allow reader to get known what already exists in terms of any problem dealt within this work to achieve our goals. This shows the limitations of current approaches and the importance of our research.
- The third chapter, Chapter 4, details the methods proposed in this PhD thesis as a result of our studies. We have dealt with several problems at different levels; body part tracking Section 4.3, hand segmentation Section 4.4, sign characterisation and segmentation , Section 4.5, and the annotation in terms of glosses, Section 4.6. It consists on feature extraction algorithms to extend the chosen computational representation for gloss recognition. We insist in the fact that our approach is not specific to the French Sign Language (LSF), although it has been applied to the LSF.
- The last chapter, Chapter 5, presents our main conclusions and the future work.

French Sign Language

Résumé: Langue des Signes Française

La Langue des Signes utilise le haut du corps pour transmettre de l'information. Contrairement aux langues orales qui utilisent le canal audio-vocal, la LS utilise la canal visuo gestuel. Ceci mène à deux différences très importantes; la quantité d'information qui peut être transmise simultanément et l'utilisation de l'iconicité. Ces différences rendent l'étude de la LS, indépendant des langues orales.

La LS a évolué aux fils des années et a passé par plusieurs étapes avant d'être reconnue une langue naturelle. Les sourds ont montré leur besoin de communiquer à travers les gestes depuis longtemps. Cependant, en France, c'est l'oralisme qui a été préféré à la place des langues gestuelles. En effet la LS a été interdite en milieu académique afin d'aider les sourds dans leur intégration mais la productions des sons sans pouvoir les entendre reste très compliqué. De ce fait ceci a seulement aidé à leur désintégration dans la société, où ils étaient considérés comme des handicapés mentaux. Ce n'est qu'en 2005 que la Langue des Signes Française (LSF) a été reconnue en tant que langue officielle en France et que les recherches de la LSF ont commencé. Ce qui explique le retard des recherches de la LSF par rapport à d'autre langues des signes, par exemple la langue des signes américaine (ASL).

La LS permet de transmettre l'information à l'aide des articulateurs manuels et non manuels qui sont réalisés en parallèle respectant de contraintes anatomiques. Les caractéristiques manuelles correspondent à l'orientation, la configuration, l'emplacement et le mouvement. Les caractéristiques non manuelles correspondent à des poses ou mouvement du corps et l'expression du visage. La plus part de l'information est transmise à l'aide des caractéristiques manuelles, néanmoins les caractéristiques non-manuelles transmettent des information lexicales, grammaticales, etc.

La structure des signes dans une phrase tient en compte le temps mais aussi de l'espace. En effet les signes ou entités, peuvent être placés dans l'espace devant le signeur pour ensuite créer des relations entre eux. De plus pendant une conversation entre plusieurs personnes l'espace et les entités peuvent être partagés. Ceci rends le traitement automatique de la LS très difficile.

De plus, dans un discours en LS les signes employés ne correspondent qu'à de signes standard comme ceux que l'on trouve dans les dictionnaires, mais aussi de signes iconiques.

En effet Cuxac [Cuxac 2000] présente dans son approche qu'un discours en LS correspond à une alternance entre signes standard et iconiques. L'iconicité représente le moyen de transmettre l'information grâce à la production de la perception. Plusieurs structures d'iconicité existent, étant les plus connues les structures de grande iconicité ; transfert de taille et forme (TTF), transfert situationnel (TS) et transfert personnel (TP).

Le TTF produit à l'aide des composantes manuelles la forme et/ou la taille d'une entité, éventuellement placée dans l'espace de signation. La configuration de la main montre la forme de l'objet alors que les caractéristiques non-manuelles donnent des informations concernant l'aspect de l'objet. Le TS montre l'emplacement et/ou le déplacement d'un objet. Ce type de transfert est utilisé pour illustrer la trajectoire d'un objet mobile dans l'espace de signation. Souvent une main joue le rôle d'une entité fixe, placée dans l'espace et l'autre de l'objet mobile. Le mouvement entre les mains représente l'interaction entre les objets. Le TP permet aux signeurs de jouer le rôle d'un agent dans le discours. Dans ce cas le signeur adopte le comportement de l'entité qui est représentée, e.g. une personne, un objet, un animal, etc. Ces structures montrent la richesse de la LS et illustrent la complexité du traitement automatique de la LS.

Précédemment nous avons expliqué qu'un discours en LS correspond à une séquence de signes standards et de signes iconiques. La transition entre ces signes donne lieu à l'effet de co-articulation. En effet il s'agit du geste qui va de la fin d'un signe au début de l'autre. De cette façon la réalisation d'un signe est influencée par le signe précédant, donc du contexte. De plus la réalisation d'un signe peut avoir une phase de préparation qui se fait en parallèle du signe.

La variabilité des signes et l'effet de co-articulation représentent un problème majeur dans le traitement automatique de la LS par des méthodes de vision par ordinateur. Il s'agit d'extraire des caractéristiques de bas niveau et de les interpréter, par exemple pour les systèmes de reconnaissance de signes. En effet la même caractéristique de bas niveau doit pouvoir être interprétée à plusieurs niveaux; sémantique, lexical, etc. Par exemple le mouvement dans les verbes directionnels des mains peut changer significativement le sens d'une phrase.

Dans ce chapitre nous présentons les caractéristiques de la LS afin de montrer les difficultés rencontrées lors du traitement automatique de la LS. La grande variabilité, l'effet de co-articulation et la dépendance au contexte nous demande des nombreuses données d'étude. Dans la littérature, des méthodes d'acquisition des données, de représentation de signes, d'annotation, etc. existent. Le chapitre suivant détaille les méthodes existantes avec leurs avantages et leur limitations.

2.1 Introduction

Sign Languages (SL) are visual-gestural languages used by deaf communities. They use the (whole) upper-body to produce gestures instead of the vocal apparatus to produce sound, like in oral languages. This difference in the channel carrying the meaning, i.e. visual-gestural and not audio-vocal, leads to two main differences.

- The amount of **information that is carried simultaneously**, though this is limited articulatorily and linguistically. Body gestures are slower than vocal sounds but more information can be carried at once, refer to Section 2.3.1.
- The visual-gestural channel allows sign languages to make a **strong use of iconicity** [Cuxac 2000], see Section 2.3.3. Parts of what is signed depends on its semantics. This makes impossible to describe lexical units with pre-set phonemic values. In addition SL are strongly influenced by the context and the same sign can be performed in different manners.

A common error is to think that SL are somehow dependent on oral languages, i.e. that they are oral languages spelled out in gesture. Sign Language has been developed by deaf communities to substitute the oral by the visual channel. Although part of oral languages can be used, e.g. finger-spelling, new signs are continuously being created. SL can be used to discuss any topic, from the simple and concrete to the complex and abstract. They are independent from oral languages following their own paths of development and evolving with the culture. Although the basis of several SL come from the same root (French Sign Language, Italian Sign Language, Quebec Sign Language, etc...) SL have changed through time. This leads to many differences between SL even from the same country but different regions. Also the way of using the space in front of the signer produces a different production of signs depending on many other factors. For example in Toulouse, south of France, Paris is performed locating the sign [PARIS] on the top to represent that is at the north from Toulouse, Figure 2.1 (left). In Paris the sign is located in the centre to highlight that they are already in Paris, Figure 2.1 (center). People from other countries sign [PARIS] differently depending on what is relevant to them, in many cases it is represented by the Eiffel tower, Figure 2.1 (right). This example shows how a simple gestures conveys more implicitly information than oral languages.

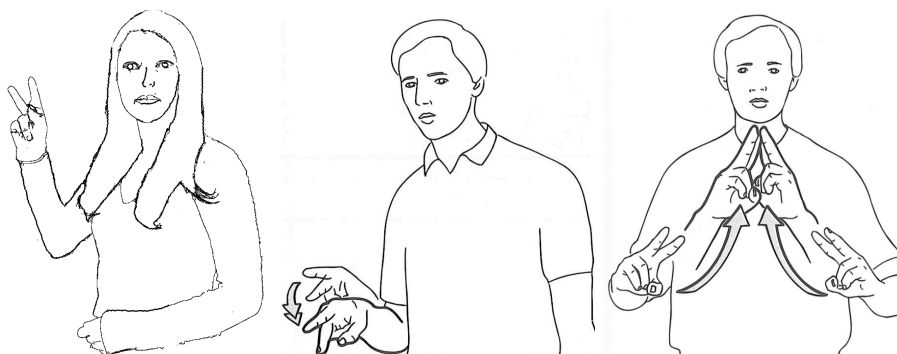


Figure 2.1: Example of sign [PARIS] in LSF. Source IVT [Girod 1997] centre and right.

This brief introduction shows the main differences between SL and oral languages and states why SL research is a complex domain. In the following section, Section 2.2, the history and the current situation of French Sign Language (LSF) is presented. In Section 2.3 are described the characteristics of SL and a discussion is finally held in Section 2.4.

2.2 History

Sign Language has a long history and has past for several phases before it has been recognised as a natural language. Since the beginning of times deaf people have shown their need to communicate with gestures. However it has been considered as a mental sickness and hearing people have chosen to push oralism instead of gestural communications. In the 18th century Charles-Michel de l'Epée, a.k.a. Abbé de l'Epée, has supported gestural communication and has affirmed that it was as useful as oral languages to communicate. In 1755 Abbé de l'Epée has founded the first school for deaf children in Paris.

Sign language became very common in schools until the late 19th century where the Milan International Congress of teachers for the Deaf in 1880 took place in which have participated 164 people with only two deaf people. They decided that oralist schools would be preferred rather than gestural schools after some demonstrations of the effectiveness of their methods. In addition they argued that sign language was a barrier for the integration of Deaf in the society unlike oralism which consists on producing sounds and reading lips. In France, LSF was banned at schools and remained that way during almost a hundred years.

Producing sounds without hearing them is difficult since the learning and the correction of sounds is achieved by hearing the produced sounds. Deaf communicating through oralism are hardly understandable. The main consequence of this is that Deaf were considered as mentally handicapped and have been somehow rejected from the society. Another important consequence concerned the education of deaf people, how to teach a science to a deaf person if the communication skills are scant. In addition to the fact that oralised deaf people can hardly hold a conversation, their culture knowledge is very low.

In [Simon 1908] has been presented the studies carried out by two psychologists who were curious on how a complex and delicate subject as speech can be taught to deaf people since the intonation is regulated thanks to the hearing. The conclusion of this work highlighted the failure of this method. They noticed that the socialisation expected by the fact of *talking* was poor. Also they argued that a deaf person cannot hold on a conversation with a person foreign to their environment or even to his relatives without using gestures. They also pointed out the difficulty of teaching oralism by highlighting the tiredness, hardness, sadness, etc.. of the method. Finally they proposed to combine oralism and manualism to improve communication skill.

Few years ago French Sign Language (LSF) was finally recognised as official language in France in the educational code for teaching, law *n°* 2005-102 from the 11th of February 2005. Since then LSF is used in teaching and public administration places. This law for *the equality of rights and opportunities* demands all public places to become accessible to handicap people, among them deaf and hearing impaired.

At the present time LSF remains a mysterious language since SL research is very recent. Numerous domains are under research for example ; sociology [Dalle-Nazébi 2006], history [Encrevé 2008], linguistics [Cuxac 2000, Boutora 2008], computer science and language processing [Ong 2005a, Cooper 2011], etc. In the last decade linguistics research considerably advanced leading to the development of Automatic Sign Language Processing (ASLP); computational SL modelling, SL recognition, SL generation, etc. French laboratories focusing on ASLP are: LIMSI¹, VALORIA² and IRIT³. These laboratories found an interesting and complex domain of research linked to the richness of Sign Languages, their characteristics are described in the following section.

2.3 Sign language characteristics

Using the visual-gestural channel to communicate permit to convey several information simultaneously. This characteristic leads to the use of numerous articulators in parallel; manual and non-manual features, organised not only temporally but also spatially in the virtual space in front of the signer. Linguistic elements are used to produce meaningful sentences consisting of several signs performed in a continuous way linked by a meaningless gestures, this is called co-articulation effect.

2.3.1 Manual and non manual features

SL organize elementary meaningless units, also called phonemes, cheremes, signemens or morphemes in the case of sign languages, into meaningful semantic units. These meaningless units, or articulators, are represented as combinations of features. Articulators used in SL are of two kinds; manual and non-manual. The former corresponds to features concerning hand. The latter involves all other articulators carrying meaning, e.g. face expression.

Characteristics belonging to the kind of articulators are:

- **Manual Features:** Hand orientation, configuration, location and motion.
- **Non manual features:** Postures or movements of the body, head, eyebrows, eyes, cheeks and mouth.

Linguists assume that manual features convey most of the information and base their models on these features, see Section 3.4. Information conveyed by manual features

1. www.limsi.com

2. www-valoria.univ-ubs.fr

3. www.irit.fr

combined with non-manual features shows lexical, grammatical, adjectival or adverbial information among others. For example the sign [SIZE] has different lexical meanings according to the cheeks configuration; huge or small, see Figure 2.2. Or another example when asking questions the eyebrows play an important role in the grammatical structure; rising eyebrows. Or even eyes gaze, generally, refers to any element previously positioned in the space during the discourse [MacLaughlin 1997, Thompson 2006].



Figure 2.2: Sign [SMALL] and [HUGE], left and right respectively, in LSF.

Manual and non-manual features cannot be combined randomly. For example the use of two manual articulators is subject to motor constraints, resulting in symmetric restrictions [Sandler 2012]. The temporal and spatial organisation of this articulators remain an interesting and challenging problem for computer vision researches.

2.3.2 Temporal and Spatial Organisation

The structure of signs in a sentence not only uses time but also space. Generally signs are signed in the half-sphere in front of the signer, Figure 2.3, called signing space. Entities can be located in the signing space or animated following a path illustrating their displacement. Later in the performance, these entities are referred by pointing or simply looking at them [Thompson 2006].

During a conversation between two or several signers the signing space can be shared. Then, entities placed by a signer can be referred and/or pointed by other signers without mentioning them again. The computational modelling of the signing space is addressed in several works [Lenseigne 2005, Braffort 2004] in order to study SL grammar for linguistic models building that combined with computer vision approaches handle complex SL productions. At the present time mostly works focus on one interlocutor discourse. In the case of a conversation between two signer (one in front of the other), Figure 2.4 illustrates an example of the modelling of the signing space with their inter-objects relationship.

Using the signing space makes signing productions much flexible with a high degree of inflection, and a topic-dependent syntax. This allows the use of *classifiers* which allow to spatially show size, shape, movement, or extent taking advantage of the spatial nature of the language.



Figure 2.3: Signing space in front of the signer

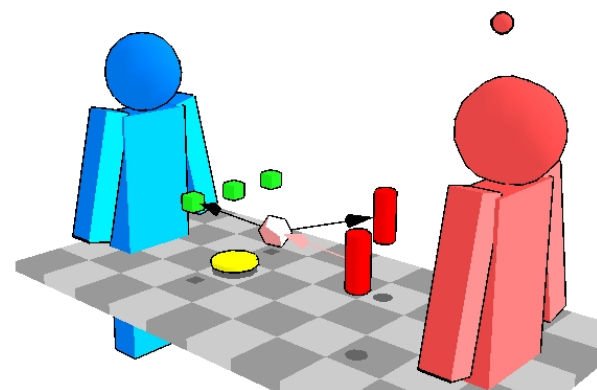


Figure 2.4: Modelling of the signing space shared between two signer. Image extracted from [Lenseigne 2005]

2.3.3 Linguistic Elements

Signs are conventional and make part of a vocabulary set which do not necessarily have a visual relationship to their referent. However the visual modality of SL leads to a preference for close connections between form and meaning. First research in Sign Language highlighted the iconic property of SL but believed that once a sign has been accepted in the vocabulary it does not play an actual role in perception and production of signs [Stokoe 1976]. Later in [Cuxac 2000] has been argued that some aspects in iconicity are semantically motivated. The author describes three iconic structures. Unlike other approaches, C. Cuxac led his research considering SL as an independent system without extrapolating any theory from oral languages. A brief description of the linguistic elements introduced in [Cuxac 2000] is presented in the following.

2.3.3.1 Two ways of saying

First of all, it has been introduced the "two ways of saying". Cuxac's theory basis correspond to the fact that it exists two ways for performing sentences in SL and that continuous SL is a combination of this two ways. One way is fully based on illustrating (iconicity) and the second way is the opposite by using standard lexicon which correspond to previously defined signs in the vocabulary. A signer switches from one way to the other to express himself. Figure 2.5 shows some iconic signs and some standard signs used by a signer during a discourse illustrating Cuxac's theory.

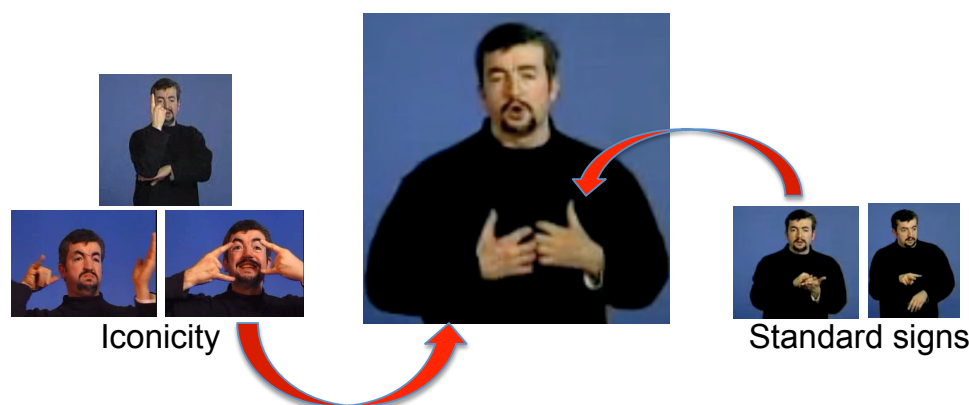


Figure 2.5: Example sign language performance switching between the two ways of saying; iconicity and standard signs. Images are extracted from [LS-COLIN 2002]

2.3.3.2 Iconicity

Iconicity, comparable to onomatopoeia for oral languages, is the likeness between a world object or behaviour and a sign [Cuxac 2000] and is largely confined to sign formation. It consists on telling something by illustrating it playing an important role in SL. It has been compared by C. Cuxac with an oral sentence "I have caught a fish big-like this-", showing his hands so that the gap between them correspond to the size of the fish. Sign Language exploits naturally this illustrating way of expressing because it uses the visual channel.

In LSF the basic structures of Iconicity are three, but other variants or other structures composed of the basic structures and standard lexicon [Sallandre 2003] exist. Here we will only present the basic structures.

The first corresponds to the **shape and size transfer (TTF)**. This kind of structure produces through a manual form the shape and/or the size of an entity, eventually, located in the signing space. Manual configurations allows to show the form of the object and non manual features give further information concerning the aspect of the object as shown in Figure 2.2. In this kind of transfer eyes gaze follows the placement of the hands in the signer space this is, generally, accompanied with a non manual gesture. An example of

this kind of transfer is illustrated in Figure 2.6 (left), showing the shape of the building; a pointed building.

The second structure described concerns the **situational transfer (TS)** which locates concepts or illustrates displacements of an element in the signing space. This kind of transfer is used to illustrate the trajectory of a mobile entity in the signing space. Often, one hand plays the role of the located entity and the other the role of mobile element, the movement between hands represents the interaction between the two elements. Each hand configuration intends to represent the object, called pro-forms. The movement of the strong hand is followed by the eye gaze. An example of this is illustrated in Figure 2.6(centre), one hand has the pro-form of [PLANE] and the other the [BUILDING], this shows that the plane is getting close to the building giving the relative location between two objects. Notice the face expressing that something sad is going to occur.

The last structure discussed in here is called **personal transfer (TP)**, this allow signers to plays the role of one agent in the sentence. In this case the signer adopts the behaviour of the entity (person, object, etc...) that is being represented. This structure is detected when the shoulders orientation and eyes gaze are orientated to another place than the receptor. The freedom obtained by this kind of structures is suitable to represent persons, animals, object or even concepts. Figure 2.6 (right) shows an example when the signer acts as the agent looking through the window.



Figure 2.6: Shape and size transfer (left), situational transfer (center) and personal transfer (right). Images extracted for [LS-COLIN 2002]

These basic structures are widely used in combination with a large set of standard signs in continuous SL. Producing signs one after another leads to the co-articulation effect or movement epenthesis.

2.3.4 Coarticulation

Co-articulation or movement epenthesis [Vogler 2001, Yang 2007], also referred in this work as '*Transition*' between two signs, corresponds to the meaningless gesture inter-signs. Indeed 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. Figure 2.7

shows an example of the co-articulation between two signs in French Sign Language (LSF); [UNITED STATES] and [TOWER].

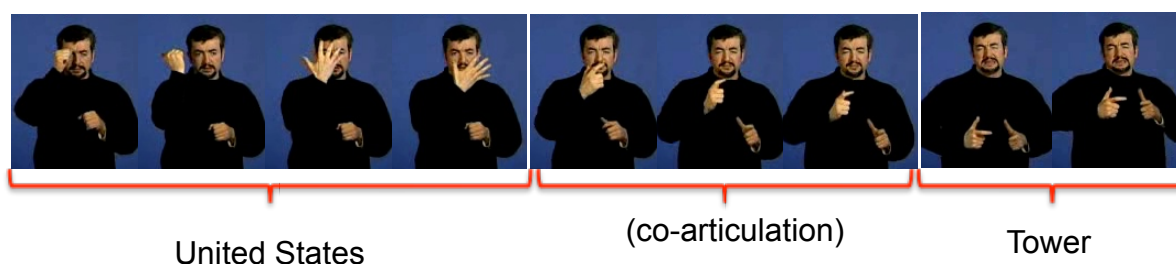


Figure 2.7: Example of *co-articulation*: gesture between the end of the sign [UNITED STATES] and the beginning of the sign [TOWER] in French Sign Language

Co-articulation represents a major problem in computer vision approaches since signs production, from low level features, is highly modified. In addition the preparation, locating hand and changing the configuration to the beginning of the following sign, influences de production. Thus movement epenthesis can take place during the performance of other signs. Figure 2.8 (left) shows the sign [DEAF] in French Sign Language (LSF). It corresponds to one-hand sign with an "arc" path. Figure 2.8 (right) shows the performance of the same sign in a different context, this time left hand moves straight. In this context signer prepares the following sign which corresponds to a sign performed with two hands. This example illustrates how the same sign give different motion features.



Figure 2.8: Sign [DEAF] in French Sign Language in different context

This characteristic of continuous SL is challenging to model [Segouat 2010] and to detect using computer vision, some approaches intending to segment signs (detecting the beginning and the end of a sign) are described later in Section 3.5.3.

2.4 Conclusion

In this chapter history and the characteristics of SL are described pointing out the richness and complexity of the language. All these particularities have numerous repercussions on the automatic processing of SL making it particularly challenging.

French Sign Language (LSF) research is very recent, only few years compared to oral language research, as consequence of its history, thus few linguistic models are yet available. In addition the vision modality of the language makes it a flexible language conveying lot of information simultaneously. Some features are more difficult to detect than others, particularly non-manual features because the motion is very subtle, e.g. cheeks configuration or trunk orientation. In addition the kind of data used make the detection, in our case mono-view, of out of plane rotation very difficult.

SL processing not only consists on the recognition of isolated features, e.g. motion, hand shape, eyes gaze, etc. but also about the integration of these heterogeneous features in the processing systems. Indeed processing systems might be able to interpret the same feature at different levels; semantic, lexical or syntactic level. For example the motion can furnish extra information and can radically change the meaning of a sign even in a standard sign, e.g. directional verbs. Also iconicity is very difficult to recognise from a computer vision system since it remains free in the way a shape is being described, i.e. the same situation can be differently signed using iconicity. In addition to all these reasons the co-articulation effect appearing during continuous SL influences the production of signs.

In the last decade numerous contributions in the literature appeared for data acquisition and annotation and for the automatic processing of S. this is described in the following chapters.

Sign Language Corpora and its notation: State of the art

Résumé : Etat de l'art

Actuellement plusieurs recherches s'intéressent au problème de l'analyse automatique de la LS [Ong 2005b, Von Agris 2008, Cooper 2011], plus particulièrement de sa reconnaissance. 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 des signes comme une phrase dans une langue oral. Ce type de systèmes nécessitent de grandes quantités de données représentatives du problème. La collection de données se fait généralement à l'aide d'un système de capture de mouvement qui consiste en un capteur de mouvement et un analyseur des données. La complexité de l'analyseur dépend de l'information obtenue par le capteur et est inversement proportionnel à sa complexité. Les capteurs peuvent être classés comme passifs ou actifs et peuvent être intrusifs s'il doivent être placés sur des membres du signeur. Ce type de capteurs ne sont pas souhaitables pour l'étude de la LS car ils influencent la réalisation des signes. Nous privilégions les capteurs qui ne font qu'observer et enregistrer les discours en LS.

Ce type de corpus est très utilisé par les linguistes qui annotent manuellement en général, des caractéristiques importantes de la LS. Plusieurs niveaux d'annotation peuvent être considérés à partir de deux points de vue ; linguistique ou informatique. D'un point de vue linguistique les caractéristiques peuvent être de type grammatical, iconique, sémantique, etc. Ces caractéristiques sont annotées par des experts en LS. D'un point de vue informatique il s'agit de caractéristiques de bas et de haut niveau. Le bas niveau correspond à des caractéristiques sans signification par elles même mais combinées avec d'autres caractéristiques donnent du sens à une phrase. Ces annotations sont généralement réalisées manuellement par des linguistes et informaticiens. Il existe plusieurs logiciels d'annotation pour assister cette tâche, e.g. AnColin [Braffort 2004], etc. Cependant ces logiciels correspondent à une interface ne permettant à l'annotateur que la manipulation de vidéos et des données d'annotation. L'annotation des signes d'un point de vue linguistique utilise systèmes de notation de signes qui correspondent à une représentation lexicale. Les plus connus sont de type paramétriques ou temporelles. Les systèmes paramétriques assument que les signes peuvent être décomposés en plusieurs articulateurs qui sont produits simultanément alors que les notations temporelles considèrent que les signes ont une structure séquentielle. Ces représentations des signes sont utilisées

pour décrire les caractéristiques manuelles et non manuelles des signes. Le système de représentation de signes ZeBeDee attire particulièrement notre attention car il s'agit d'un système générique et modulable. Il a été conçu pour la génération automatique des signes tenant en compte l'intention du signeur et non la production d'un signe. De ce fait la variabilité et la dépendance au contexte des signes sont considérées. Afin de s'aider de cette représentation nous étudions les méthodes d'extraction de caractéristiques, à partir d'une vidéo, utilisées pour la description des signes dans ZeBeDee: mouvement et configuration de la main.

Dans la littérature les caractéristiques de mouvement sont extraites à l'aide de méthodes de suivi des composantes corporelles. Ces méthodes sont principalement basées soit sur des mesures de différence entre l'image et un motif, soit sur des modèles dynamiques qui estiment la fonction de densité de probabilité *a posteriori* du système. Ils utilisent des caractéristiques représentatives des objets à suivre comme c'est la couleur de la peau pour les mains et le visage. L'inconvénient des approches ne considérant que la couleur, est le fait de représenter plusieurs objets avec le même modèle. 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 des caractéristiques de mouvement, la configuration de la main est aussi décrite dans la représentation ZeBeDee. Afin d'extraire des caractéristiques de forme de la main, nous devons segmenter les mains même quand elles se trouvent devant le visage. Ceci 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 ou encore la seule région de peau. 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 ou sur le recalage de motifs ne donnent des résultats satisfaisants que si la forme change peu, or en LS ce n'est pas le cas. Nous avons donc besoin d'une méthode de segmentation robuste aux changements rapides de configuration de la main. En plus des caractéristiques de mouvement et de forme de la main, la segmentation temporelle des signes est nécessaire pour déterminer leur structure temporelle ainsi que les limites des signes dans une phrase en LS. Les caractéristiques extraites des vidéos sont par la suite utilisées pour reconnaître les signes. Les systèmes de reconnaissance des signes dans la littérature utilisent ces données pour apprendre les caractéristiques des signes. En effet afin de reconnaître les signes, des informations de haut niveau sont nécessaires. L'utilisation de données d'apprentissage sont dans tous les cas nécessaires. Pour ça nous envisageons de créer des données synthétiques plutôt que des données obtenues dans des vidéos qui vont biaiser nos résultats. Dans ce contexte nous étudions les méthodes de génération automatique des signes afin d'identifier les méthodes dont on peut se servir pour générer nos données d'apprentissage.

3.1 Introduction

In Sign Language researches a large and structured set of data are used to perform statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules [Ong 2005b]. The data must be representative of the problem, this is generally collected through the study of native signers performing isolated signs or a discourse in continuous SL. The contents of the corpus are defined through the elicitation which defines the tasks to be performed by the signer or various signers to consider the variability of the performances depending on the signer morphology, educational background, region, etc. [Hanke 2010, Matthes 2010]. The information in the corpus could concern isolated signs, i.e. repeating the same signs several times by several signers or continuous sign language, i.e. from a given subject to the signer this one express himself freely in a continuous way. The instructions are generally given using visual support instead of written text to avoid the influence of oral languages during the performance.

Methods used for data acquisition can be classified in two kinds; active methods, e.g. sensor based, and passive methods, e.g. marker-less approaches or video based methods. Our work focuses on the use of passive methods based on a customized set of cameras for SL. A brief overview of motion capture systems is given herein Section 3.2.

SL corpus are annotated to furnish several information to the sequence, e.g. lexical or grammatical information, motion characteristics, etc. to perform learning or for evaluation purposes. Annotation is, in general, manually performed which is time consuming, error prone and unreproducible. This is addressed by studying and developing computer based methods to assist the annotation, e.g. editor tools, see Section 3.3. A distributed system architecture [Collet 2010] is detailed to facilitate the use of automatic computer vision algorithms with an editor tool, refer to Section 3.3.4. The annotation of SL corpus can be performed using linguistic representation of signs established by linguists that helps image processing techniques.

The remaining of this chapter is organised as follows. Section 3.2 introduces the data acquisition methods used in the literature to collect data to study. In Section 3.3 presents the methods for annotating SL video corpus from a linguistic and a computer vision point of view. Later Section 3.4 details existing representations of signs that could be added to the annotation of corpus at the linguistic level. Finally our main conclusions are presented in Section 3.8.

3.2 Data acquisition

In SL research motion capture systems are used to collect information concerning the posture and the hand configuration of signers in order to study SL characteristics. A motion capture system is composed of two parts: a motion capture sensor and an analyser. The complexity of the analyser depends on the information obtained by the sensor. The simplest sensor need a complex analyser and vice-versa. In Fig. 3.1 is plotted the sensor

complexity (intrusiveness and prize) with respect to the analyser complexity (higher level of data, lower complexity) [Moeslund 1999]. A brief overview of capture motion sensors is given below, detailed information can be found in [Meyer 1992, Frey 1996, Welch 2002, Vlasic 2007].

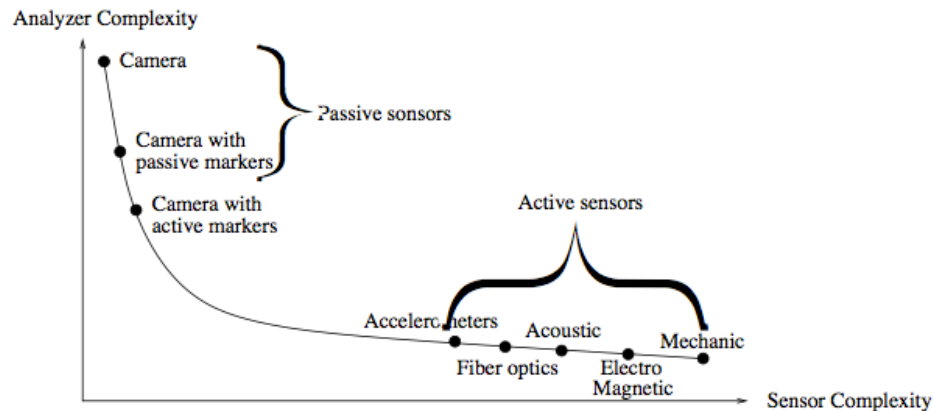


Figure 3.1: Motion capture sensors [Moeslund 1999]

Motion capture methods can be classified as; active sensors based on mechanics, magnetics, acoustics, fibre optics or inertial principles; and passive systems using camera or camera with passive markers; or even a combination of both e.g. camera with active markers. Several works using different motion capture systems to collect SL data exist; mechanic systems [Cox 2002, Vogler 2004]; hybrid systems (Fig. 3.2) using Cibergloves, magnetic sensors, opto-electronic devices and image processing techniques [Brashear 2005, Adamo-Villani 2008, Lu 2010b, Lee 2009]; a customized configuration of motion capture devices [Lu 2010a]; optical systems [Havasi 2005]; fibre-optic based data gloves [Kim 1996, Kim 2005]; inertial methods as accelerometers as the one in [Hernandez-Rebollar 2002].

3.2.1 Active Sensors

Using active sensors corresponds to place sensors transmitting or receiving signals. The sensors are placed in strategic places on the signers, e.g. fingers, shoulders, face, etc. see Fig. 3.6 (left) or Fig. 3.5 (right).

Mechanic sensors are use in SL applications to detect the hand configuration [Lu 2009] or the posture of the signer. They are the simplest approach in terms of conception, they are based on the attachment of a movable part to the body, when moved this outputs a signal reflecting the configuration of the movable parts. It consist on exoskeletons which are articulated series of interconnected rigid mechanical pieces that have to be worn by signers. An example of a mechanic system for the torso and for the hand configuration are

1. www.ascension-tech.com

Figure 3.2: Hybrid systems: Hy-BIRD¹

shown in Figure 3.3. The principle consists on the measurement of joint angles between the different parts of the device. The posture of the signer is straight forward obtained without further processing, e.g. inverse kinematics as is the case of estimating points on the body. The inconvenient is that exoskeletons constrain the motion of the signer and are uncomfortable to wear.

Figure 3.3: Mechanic sensors: exoskeleton Gypsy 7 (left) and CiberGlove II (right)²

Magnetic sensors use the Earth's magnetic field or a magnetic field generated by a transmitter to measure, through magnetometers or current induced in a electromagnetic coil, the local magnetic field vector at the sensor. A sensor is composed of three orthogonally oriented magnetic sensors, in this way the position and the orientation can be detected. Several sensors are strategically positioned on the part of the body to track, this is illustrated in Figure 3.4 where each sensor's size is 2.54cm x 2.54cm x 2.03cm cube. The position and orientation of each placed sensors are used to build the posture of signers through inverse kinematics. The cumbersome of the sensors makes it difficult to place them in small areas, e.g. fingers for hand configuration. This kind of devices are accurate but might be influenced by any ferromagnetic and conductive material close to the sensor. In addition they are expensive and have high power consumption.

2. www.metamotion.com

3. www.ascension-tech.com

Figure 3.4: Magnetic sensors: MotionStar³

Optical systems principle is to track retro-reflective marker or light emitting diodes placed on the part of the body to track. This method uses image processing techniques to obtain the 3D location of each marker/diode from the recorder video from surrounding cameras. An optical system is composed of light sources and optical sensors. These methods have the advantage to be very accurate and fast. Nevertheless the inconvenient is the price, the lack of portability and the need of having a clear line-of-sight from the optical sensor. Figure 3.5 shows a motion capture system (left) and a glove using some light emitting diodes (right) necessary to track hand configuration using an optical system.

Figure 3.5: Optical tracking: IMPULSE motion capture system (left), optical glove⁵(left)

Acoustic systems use audio signals to obtain the location of sensors. Acoustic sensors are attached to the signer, the emitted sound wave is received by a set of microphones. Using phase wave techniques or triangulation the position of the sensor is obtained. The inconvenient of this technique is that they are not portable and only few sensor can be handled.

Optic fibre systems are used along the limb of the body to track, a signal sent through the optic fibre is used to compute the bent of the fibre which is in this case the bent of the limb where it lays down.

Inertial sensing tracks body limbs using gyroscopes and accelerometers. The cumbersome of each sensor is important even if recently inertial sensing glove have been developed, see Fig. 3.6. The advantage of this systems is like the mechanical systems the

5. www.phasespace.com

no need to a clear line-of-sight. However sensors are sensitive to the Earth's gravitational field.



Figure 3.6: Inertial sensors: Xsens's Movien⁶(right) and the AcceleGlove⁷(left)

The price of the different active sensor based systems are very expensive. For example for a full body systems it starts at about \$8000 for electromechanical to less than \$60,000; for active-optical to about \$72,000; for inertial to \$100,000; and for multi-actor active-optical systems up to \$100,000⁸.

3.2.2 Passive Systems: Sign Language Video Corpus

Passive techniques principle is based in the non interference of any active-marker. This only observes and capture using a camera for further processing through image processing techniques where motion parameters are obtained. This involves the use of a camera [Zieren 2004], or a set-up with several cameras where the data consist on a sequence of 2D images without depth information. These approaches are less accurate but cheaper and portable. A challenging problem in these approaches is handling occlusions, i.e. any information is in the image any more. More complex cameras allow to have information about the depth; stereo cameras [Hasanuzzaman 2004]; the Kinect (recently affordable) [Zafrulla 2011, Keskin 2011] or the bumblebee [Elmezain 2009]. In this work we consider SL video corpus using a customize configuration of various cameras with one monocular frontal view where a person performs SL, see Figure 3.7. Numerous video corpus have been built for SL research to study several aspects of SL; grammatical, lexical, etc. Since SL are different in any country and also in different regions, representative corpus of the SL to study have to be created. American Sign Language (ASL), e.g. Boston corpora⁹, German Sign Language (DGS), e.g. Phoenix weather forecast corpora [Stein 2006] or Dicta-Sign corpus [Hanke 2010], French Sign Language (LSF), e.g. LS-Colin [Braffort 2001], Degels [Boutora 2011], Dicta-Sign¹⁰, Irish Sign

5. www.xsens.com

7. www.acceleglove.com

8. www.metamotion.com

9. www.bu.edu/asllrp/ncslgr.html

10. www.dicta-sign.eu

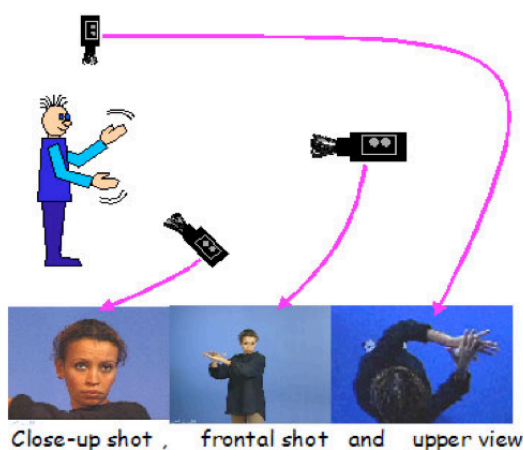


Figure 3.7: Ls-Colin corpus configuration

Language [Bungerot 2008] and many other SL corpora exist. These kind of corpora are used for the evaluation of automatic SL recognition approaches [Dreuw 2008a]; evaluation of isolated signs [Zahedi 2006] or continuous sign language processing [Dreuw 2007], head gesture [Erdem 2002], facial expression [Vogler 2008], body limbs tracking, generally hands and face, and hand shapes [Vogler 2004, Yuan 2005, Lefebvre-Albaret 2009, Gianni 2009], etc. A brief overview of the corpus used in this work for evaluation are described; LS-Colin, Degels and Dicta-Sign corpus.



Figure 3.8: Degels corpus (left), Ls-Colin corpus (centre) and Dicta-Sign (right)

These three corpus have been built not only for linguistic purposes but are also suitable for computer vision approaches. They consist on a recording of a frontal view video and some other views, with an homogeneous background, of a free performance of sign language by native signers, no constraints concerning speed, vocabulary, etc. have been given, Fig. 3.8. Also since signers are not wearing any cumbersome device, though signers wear a long-sleeve sweater, this corpus is representative of SL performances.

SL corpus have some hypothesis :

- **Background** is static and homogeneous with a different colour than the signer, generally, green or blue.
- **Signer clothes** are uni and dark, different than the background and the signer skin colour.
- **Illumination** is constant and does not change during the whole video.
- **Framing** is frontal upper-body and the hands are fully or partially visible during the whole video.

LS-Colin [Braffort 2001] is a corpus of French sign language (LSF). It consists of several sequences with a total length of 2 hours recorded with three cameras; frontal, top and bottom view, although only the frontal view is used in our work. It has recorded 13 native signers from different ages, regions and professions. The kind of discourse on the elicitation is narrative, e.g. telling a story, explicative and meta-linguistic. The corpus has been transcribed, 4929 glosses. In addition a video sequence [LS-COLIN 2002] of about 3000 frames has the ground truth for the tracking of hands and head and the segmentation of signs.

Degels corpus [Boutora 2011] is composed of two video sequences; one in French with co-verbal gestures and a second one in French sign language. It has been used to perform the annotation defining annotation criteria [Garcia 2011, Mlouka 2011, Devos 2011, Gonzalez 2012a] to discuss about the differences on the annotation of SL and gestures in oral languages. The sequence concerning the LSF, 1360 frames, is a conversation between two deaf people about some places to visit in Marseille, France. The cameras configuration is one frontal camera for each signer and a third camera for the side. This corpus has also the translation from LSF to French which is not aligned to the video. The video sequence in SL has been manually segmented by a native experienced annotator.

Dicta-Sign Corpus [Efthimiou 2009, Hanke 2010] is an international corpus available in four European sign languages: British Sign Language (BSL), German Sign Language (DGS), Greek Sign Language (GSL) and French Sign Language (LSF). This corpus consists of a conversation handled by two native signers, between 14 and 16 informants, and several sequences with a total length of at least 8 hours for each SL. Informants recorded from different regions, ages and educational background. Several cameras have been used among them two stereo cameras [Hanke 2010]. Several tasks have been given to signers concerning travelling subjects. A multilingual dictionary available on-line¹¹ of more than 1000 concepts has been created to compare sign in the sign languages previously mentioned. These concepts have been annotated using the sign descriptor HamNoSys [Prillwitz 1989], see Section 3.4.1. A transcription for all four languages has been carried out. This corpus has been used to evaluate several computer vision techniques [Gonzalez 2010, Elliott 2010]

11. http://www.sign-lang.uni-hamburg.de/dicta-sign/consign/demo/cs_list_eng.html

3.2.3 Discussion

Different capture motion methods have been described here, the following observations are pointed out:

- The high cost of motion capture sensors make these systems inaccessible to consumers, e.g. a full body system costs about \$8000. Although it is possible to use this kind of devices for SL research purposes, i.e. linguistic models building, signing avatar, etc.
- Invasive and cumbersome devices, e.g. cyber gloves, influence the motion and the performance of signs. SL production might not be realistic and natural using this kind of devices.

In short it is argued the inconveniences of active sensors in SL researches because of the non affordability, their influence on the motion and SL performance of a signer due to intrusiveness and the lack of interest for computer vision applications. For these reasons our research focuses on the motion capture using passive sensors in monocular configuration. These kind of corpus are accessible to consumer applications, e.g. using a web-cam, and maintain the SL production as natural as possible. The existing SL video corpus in frontal mono-view are briefly described in the following section.

Some corpus with different recording set-up and elicitation tasks have been mentioned. The three described corpus record native French signers performing French sign language (LSF). The choice of the corpus is not about the LSF or any other SL but about the available annotation and other works using the same corpora for evaluation purposes which allows a straight forward comparison between several works. Although some works are related to the SL in which the corpus has been built because it has been used for training systems, in our work these corpus are only used for evaluation of low level features and can be used for any SL unless otherwise specified.

Table 3.1 shows the number of annotated frames for body tracking, i.e. head and hands position, and for sign segmentation, i.e. beginning and end frames of a sign, as well as the number of glosses for the transcription sign by sign, standard signs (see Section 2.3.3). This annotation has been manually performed using some annotation tools, see 3.3. Notice that the available annotation is complementary and the evaluation of some parts of our work cannot be performed using all the three corpus.

Table 3.1: Corpus with available ground-truth

| Corpus | Tracking (frames) | Segmentation (frames) | Transcription (Nb. Glosses) |
|------------|----------------------|--------------------------|--------------------------------|
| LS-Colin | 2970 | 2970 | 4929 |
| Degels | X | 1360 | X |
| Dicta-Sign | X | X | 7000 |

3.3 Corpus annotation

Corpus annotation consists on furnishing extra information to the corpus for the study of SL. For this several features are annotated and aligned to the video sequence. Features which are meaningless on their own, e.g. eye gaze, hand configuration and motion, etc., become representative of the language at different levels, e.g. grammatical, lexical, etc. Indeed it is possible to perform some statistics on the data to build linguistic models through manual and non-manual features. The annotation can be performed from two points of view; linguistic and computer vision.

3.3.1 From a linguistic point of view

Annotation from a linguistic point of view can be performed at various levels through the annotation of manual and non-manual features for the analysis of SL [Koizumi 2002, Efthimiou 2007, Bungerot 2008, Chételat-Pelé 2008a, Chételat-Pelé 2008b]; grammatical, semantic, lexical, iconic and phonemic levels of annotation.

- **Grammatical annotation level:** Grammar in SL includes questions, negation, sentence boundaries and argument structure. Grammatical characteristics are conveyed by manual and non-manual features. For example questions are expressed through raised eyebrows.
- **Iconic annotation level :** In this level iconic structures are annotated since they make part of the discourse. Manual and non-manual features have to be annotated because they convey different information, see Section 2.3.3.2. For example manual features can represent shape and displacement of an agent in the discourse and non-manual features convey more information concerning the aspect of the agent or can even determine when a transfer structure is taking place, e.g. during personal transfer the informant turns shoulders and eyes gaze in a different direction than the receptor to show that the informant is acting as an agent in the sentence.
- **Semantic annotation level :** In this level, features giving the meaning of the sentence are annotated. For example agreement inflections for directional verbs, in fact hand motion as part of a directional verb can significantly change the meaning, e.g. the signs [GIVE] in LSF, let's say "*A gives to B*" the movement will be from A to B though the opposite "*B gives to A*" only the direction of the movement change from B to A. Also other features as adjectives and adverbs playing a role of qualifying nouns or verbs, giving more information about the object are annotated mainly through non-manual features.
- **The phonemic annotation level :** In linguistics this is the lowest level in which each feature alone is meaningless. Manual and non-manual features are combined to give meaningful semantic units. The phonemic aspects to annotate are described by linguists through various models, see Section 3.4. For instance

Stokoe [Stokoe 1980] defined three aspects composing a sign based exclusively on manual features; *tab*(location), *dez*(what acts) and *sig*(the action) which push us to the annotation of hand shape, position and motion. A different model, the sign structure introduced in [Liddell 1984] argues that signs are constructed, at the phonemic level, of temporal segments in this case sign segmentation has to be preformed. Other models include non-manual features at the phonemic level since some signs have the same manual components but the only difference concerns non-manual features, thus the same manual sign can be disambiguated at the lexical level.

- **The lexical annotation level :** The annotation at this level consists on annotate signs in terms of glosses. That means transcribing SL word-for-word by means of an oral language gloss written in all capitals in brackets. For annotating a video corpus in terms of glosses it is also needed to annotate the first and the last frame of the sequence where the sign takes place segmenting, then, signs.

Manual and non-manual features annotation play an important role at each single level of the linguistic analysis of SL. Improving annotation using computer vision methods leads to the study and development of image processing techniques for the detection of manual and non-manual features at a low level. In other words it consists on the detection of meaningless features which are later interpreted at a higher level for annotating more complex linguistic characteristics. For example the automatic annotation at the lexical level consisting on the recognition of glosses (high level) from the features extracted at the phonemic level (low level); hand shape, motion and location.

3.3.2 From a computer vision point of view

From a computer vision point of view the annotation is classified in the two levels previously mentioned: low and high level features. The former corresponds to the annotation of visual features, manual and non-manual, that characterise the behaviour or appearance of an object but that are, on their own, meaningless units without any linguistic information; hand location and motion, etc. The latter interprets a set of heterogeneous low level features with additional information, generally coming from linguistic models.

- **The low level features annotation :** At this level any computer vision algorithm extracts visual features that will be interpreted in a higher level. Often the recognition of some features is extremely challenging using image processing techniques and is constrained by the processing data. For example trunk orientation is harder to obtain from a mono-view than from a stereo-camera which have depth information. Another problem concerns the occlusion and self-occlusion of objects. For instance the same hand configuration leads to a high hand shape variability from a mono-view according to the palm orientation. For this reason hand configuration classification requires high amount of training data.

The extraction of these features is, generally, based on representative characteristics

of objects , e.g. colour, contours, etc. For example tracking hands and face relates often on the tracking of skin colour objects or even tracking objects with a respective shape. Numerous methods in the literature have been proposed to extract manual and non-manual features and are described in Section 3.5.

- **The high level features annotation :** This level concerns the interpretation of low level features to annotate meaningful units at different levels, e.g. lexicon recognition based on phonemic models. This requires deeper knowledge of the linguistic aspects of SL which is so far very difficult because of the recent study of SL. The lack of linguistic models make this level of annotation very challenging. Existing works intend to recognise signs [Cooper 2007, Vogler 2001] (see Section 3.6), facial expressions [Dat Nguyen 2011], or sign segmentation [Lefebvre-Albaret 2008, Sagawa 2000a]. Most of these algorithms have been designed for SL recognition and not for annotation purposes relying, then, on training data.

Automatic methods designed for SL annotation at linguistic levels are scant since linguists require lot of annotated data to build linguistic models and at the same time computer vision approaches need linguistic models to avoid using training data, thus mostly approaches are based on recognition using training data which is often manually annotated. At the present time some annotation tools have been proposed to assist the annotation improving this tedious task.

3.3.3 Annotation Tools

Computer scientists focus on the development of tools to assist the annotation task going from annotation software editors to automatic annotation methods using image processing techniques. Several annotation software exist, e.g. AnColin [Braffort 2004], Elan [Wittenburg 2006], Ilex [Hanke 2008], Anvil [Kipp 2001], etc. The goal of these annotation tools (AT) is to provide an interface to visualize videos and manipulate data structured in several tracks (or levels or tiers...) with a list of possible values associated to each frame sequence and aligned to a time-line. Figure 3.9 shows an example of the annotation software AnColin. Each track corresponds to the representative feature that we would like to annotate. However these tools are only an interface to manually perform the annotation, i.e. for temporal segmentation, annotator might select the beginning and the end frame for each sign.

Other approaches intend to incorporate image processing techniques into the AT to assist the annotation of SL video corpora either through an interface between AT and existing image processing algorithms [Dreuw 2008b, Collet 2010] or through the definition of systems with the possibility of carrying out image processing methods [Neidle 2001, Hruz 2008, Yang 2006b, Braffort 2004]. These systems argue that automatic video processing together with the annotator's knowledge facilitate the annotation task improving results and reducing the annotation time. Approaches focusing on interfacing AT to image processing algorithms, called Automatic Annotation Assistant (A^3), have the advantage

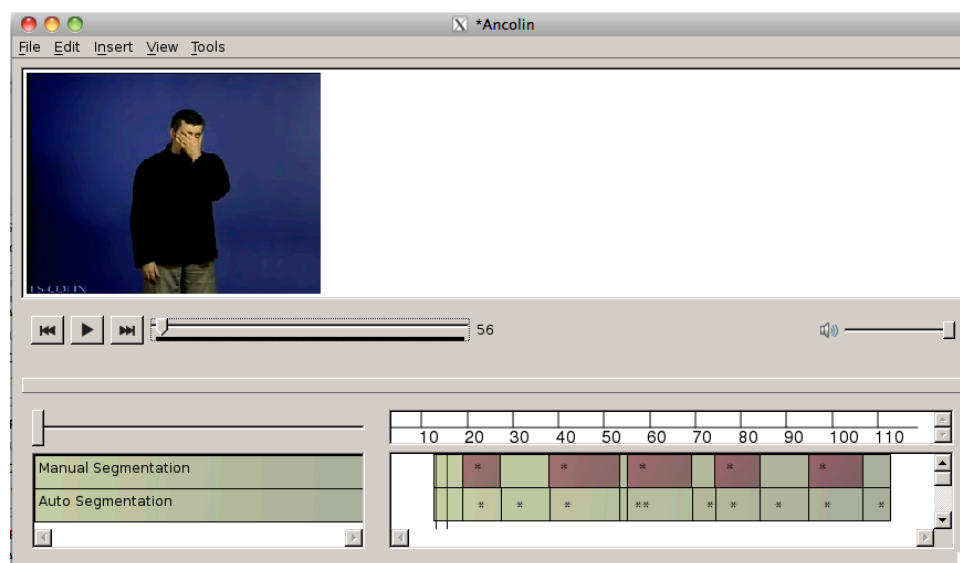


Figure 3.9: AnColin annotation software example

that any algorithm developed for SL processing can be used for annotation [Collet 2010], unlike systems integrating image processing approaches directly within the software. In [Dreuw 2008b] is presented an approach giving the possibility of importing results from image processing algorithms to the ELAN AT through an interface. These algorithms run independently from the AT. A different system intending to incorporate automatic processing is presented by Collet et al. [Collet 2010], it specifies a distributed system architecture to allow the use of any image processing algorithm as an external module of the AT, as long as the input and output format data is developed.

3.3.4 Architecture of a Annotation System

From the annotator's point of view, adding automatic processing must be easy to use, without adding complex extra work. The annotator should be able to extract a part of a video and to use a previously defined annotation as input parameter of the Automatic Annotation Assistant (A^3). For example, the annotator is working in the Annotation tool window (Fig. 3.10a), any modification done is saved on the two tiers AG1 and AG2. When the annotator executes an A^3 needing input parameters, e.g. for two input parameters, then, two additional tiers appear in the window (Fig. 3.10b). Filling in the two tiers could be done manually or using AG1 and/or AG2. Once the process is done the result is displayed as a new tier that the annotator can easily save or modify (Fig. 3.10c). This example shows how using automatic processing in this way can be easily performed from the AT instead of using intermediate files [Dreuw 2008b].

The complexity of integrating image processing techniques into the AT is not just about programming an efficient user friendly interface but also about making A^3 s and

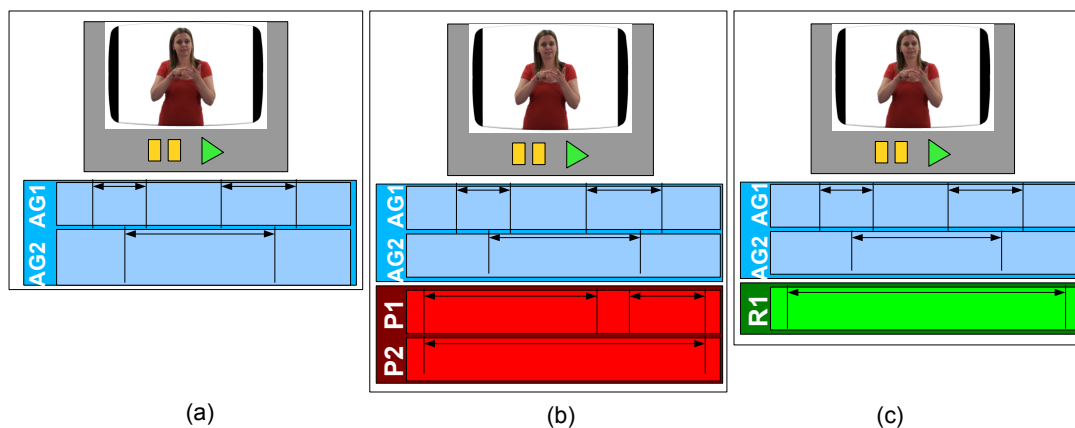


Figure 3.10: Annotation Tool Example: (a) Normal environment, (b) A^3 call and (c) A^3 result.

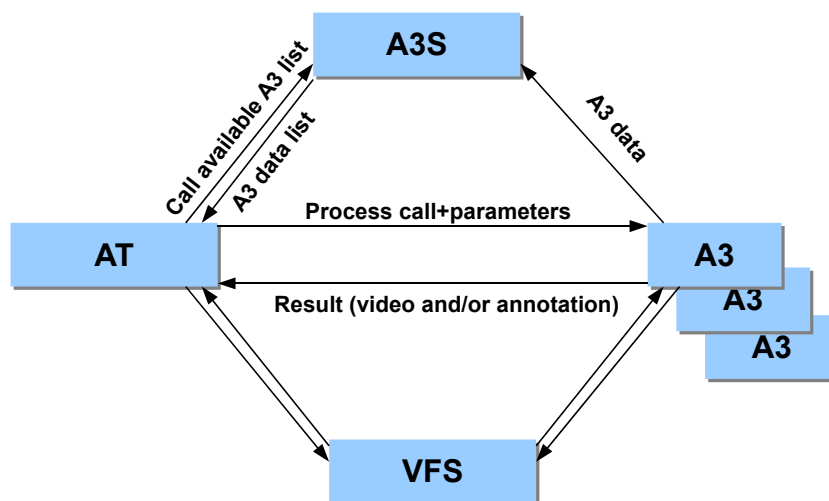


Figure 3.11: Distributed system architecture for assisted annotation of video corpus

ATs to communicate with each other knowing that the programming environment used to develop them is not generally compatible. The incompatibility of programming language, operative system and platform of development is the main problem about the integration of automatic processing into existing ATs. Nevertheless it is not possible to restrict unique development conditions to computer vision algorithms to assist the annotation. Moreover complex algorithms are preferable to be developed in a specific programming language or, even to be executed in adapted computers. That is why the Distributed System Architecture (DSA) considers that A^3 s are hosted in different computers where the communication and the data exchange are, then, done through the network using a protocol and an exchange data format understandable by all the parts of the system.

The structure of the DSA is illustrated in Figure 3.11. It considers the AT as a client and the annotation algorithms as remote servers to allow queries exchange. Since the

number of available algorithms and ATs can vary on time depending on new developments, another server called Automatic Annotation Assistant Supervisor (A^3S) is added to manage the information of the process at our disposal and to maintain an updated list of them. Thus at each time an A^3 is added it registers itself to the supervisor. Then when the AT requires an updated list of available process it requests the supervisor server since the AT can directly communicate with the A^3 . In addition the need of exchanging video files between ATs and A^3 s leads to introduce a Video File Server (VFS) to share videos in a simple and fast way.

This system allows to provide several automatic process to the scientific community without having to give the code as open source. All the algorithms obtained in this PhD thesis are adapted to integrate this system to make it accessible in an easy way.

3.3.5 Discussion

In this section we have described the different annotation levels from a linguistic and a computer vision point of view and the existing annotation tools for assisting the annotation of video corpus at these levels. The automatic annotation of SL video corpus is scant since many annotation levels require linguistic models which at the same time are built thanks to a high amount of annotated data. That is why current approaches use manually annotated data to train systems leading to recognition results dependent on training data which is to be avoided in this work. Annotation tools in the literature are; editors for assisting manual annotation, some of them including automatic annotation algorithms, or systems interfacing image processing methods to annotation editors. Editors only facilitate manipulating videos for manual annotation which remains time-consuming, unreproducible and error prone. Although some editors integrate image processing algorithms to assist the annotation this remains constrained to the annotation editor, development language and operative system. Systems interfacing processing algorithms to annotation editors are then more attractive in this works since any algorithm developed here can be easily integrated into annotation editors. Then in order to make our algorithm available to the scientific community, the advantages offered by the distributed system architecture push us to integrate our algorithm into this architecture by adapting our input and output parameters in the format needed by the AT [Dubot 2012].

At the present time annotating features at each level is challenging. This PhD thesis focus on the phonemic and lexical annotation levels based on manual features, though non-manual features give lexical marking to disambiguate sign lexical meaning they are not considered in this PhD thesis. At the phonemic level the computer vision annotation concerns low level features defined by linguistic representation of signs, see Section 3.4. The lexical level consists on determining through sign representations or learning based machine translation system the lexicon associated to a gesture, see Section 3.6, which corresponds to the high level from computer vision. Using sign recognition systems 3.6 to produce annotations does not address the problem of collecting data since the statistical machine translation or any trained system might use manually annotated learning data

in a basic step. Also these approaches are generally able to recognise only few signs from a controlled vocabulary simplifying the language. In fact being able to recognise a large amount of signs requires, with their approach, large amounts of training data. In this case weakly-supervised or unsupervised gloss recognition systems are required to address the annotation problem. That is why in this approach, designed for SL annotation, it is intended to use a computational model based on phonemic features. For this a brief presentation on linguistic representations of signs is given in the following section.

3.4 Linguistic representation of signs

Sign Language is a natural language without a writing form accepted by deaf communities. As described in Section 2.3 sign language is a complex visual language with lot of variability and a specific spatial-temporal structure. This makes very difficult the definition of a unique writing formalism which is able to register the richness of this language.

Linguists and computer scientists have proposed several models of signs in a phonemic level [Sandler 2012]. Generally, it consists on distinguishing basic components of sign gestures, also called phoneme subunits. Figure 3.12 shows how the meaning of a sign can completely change even when only one parameter has changed, for the same hand configuration, orientation and location but different movement signs are completely different; [NAKED] and [CHOCOLATE], Figure 3.12 left and right respectively.

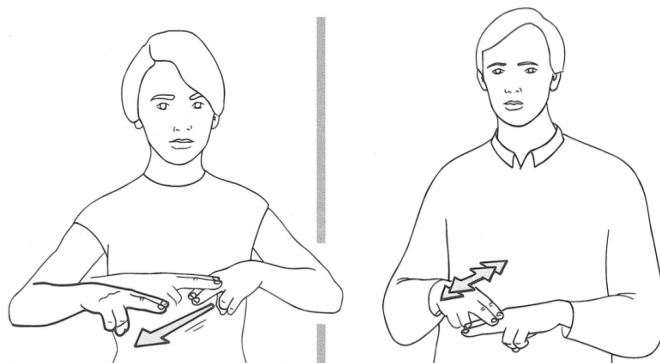


Figure 3.12: Sign [NAKED] and [CHOCOLATE] in French Sign Language. Notice that only the movement of right hand is different. Source IVT

Some approaches highlight the simultaneity of these subunits through the definition of parametric approaches while others the sequential organisation of sub-units. Sign representation can be classified as follows:

- **parametric approaches** describe signs as an ensemble of parameters taking place simultaneously defining a unique sign gesture through a sequence of symbols where each symbol represents a parameter.
- **temporal approaches** describe signs as a sequence of temporal units, e.g. hold gestures and transition, where each of them contains a phoneme subunit description.

Herein a brief description of the parametric and temporal approaches of sign representation is presented. Later in Section 3.4.3 is discussed the advantages and limitations of these approaches regarding the needs of our work.

3.4.1 Parametric approaches

Parametric approaches assume that all signs can be created from the combination of a set of parameters, performed simultaneously, where each parameter alone is meaningless [Stokoe 1960]. The combination of these parameters can not be performed randomly but some constraints have to be considered which are not violable [Sandler 2012]. For example the internal movement constraints states that if a finger position changes all selected fingers does [Mandel 1981]. Or for two-handed signs a symmetry constraint has been defined in [Battison 1978], it states that when both hands move they must have symmetric hand-shape, movement, and location. Many other constraints exist, for further information refer to [Battison 1978, Corina 1993, Brentari 1998].

Stokoe's [Stokoe 1960] parametric representation of signs defines three aspects of the structure of a sign; location, hand configuration and hand motion. These parameter have been called by the author designator (dez), tabula (tab) and signation (sig).

- **Designation (dez)** corresponds to the hand configuration which consists on positioning selected fingers in a particular position; extended, closed, curved or bent. Hand-shape changing in a sign means that selected fingers change in the same way [Mandel 1981]. However this is not respected [Fischer 2011] in some far eastern SL. Hand configuration is defined as "marked" depending on the difficulty of the production, an example is illustrated in Figure 3.14, refer to [Brentari 2011].

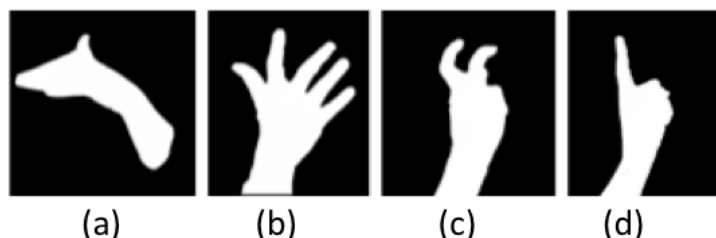


Figure 3.13: Finger configuration: (a) Four fingers bent, (b) all fingers extended, (c) two fingers curved and all other fingers closed and (d) index extended and all other fingers closed.

- **Tabular (tab)** is the aspect that specifies the proximity of the hand to a part of the signer's body, by position in space or by configuration of the non-moving hand. Signs can be described in terms of two locations, beginning and end of the sign, e.g. in the sign [DEAF] in LSF the index finger is in contact with the ear and then to the mouth, Fig. 3.15. Signs in contact with a body part; head, non-dominant hand or non-dominant arm, are somehow easier specified than signs where the location concerns a place in the space. Defining the location of signs without contact in the signing space is challenging and remains a discussion subject between researchers since signs can be defined in different ways. For example the same sign can be defined as: two locations determined from a major location and a distance parameter, e.g. *near* or only one location and a movement feature,

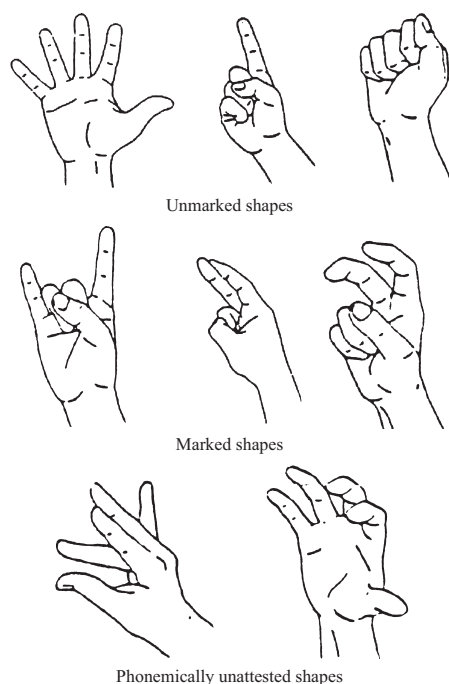


Figure 3.14: Hand configuration kind example. Image extracted from [Sandler 2012]

e.g. forward. This is a challenging problem that have to be considered for sign recognition systems, see Section 4.6.1.



Figure 3.15: Sign [DEAF] in French Sign Language

- **Signation (sig)** is the movement or change in configuration of the dez (hand configuration changing) in the same or in an other tab (location). This also corresponds to a changing only in hand orientation. These movements without a trajectory are called local movements and can occur simultaneously, e.g. the sign [SEND] involves a trajectory and a hand configuration changing, Figure 3.16. Trajectories from one location to another can have a defined primitive, e.g. *arc* or *circle*, though straight movements are most common when going from one location to another. Other path can be more complex such as directionally repetitions, trills, etc, re-



Figure 3.16: Sign [SEND] in French Sign Language involves a change on configuration while the movement is taking place.



Figure 3.17: Sign [SEA] in French Sign Language involves a complex movement to illustrate the movement of the waves.

fer to [Mak 2011]. For example the signs [SEA] in LSF intends to illustrate the movement of the waves Figure 3.17.

Stokoe [Stokoe 1960] proposed the use of some symbols to specify parameters belonging to the [Tab][Dez][Sig]. This sign representation has been first used in an English-ASL dictionary [Stokoe 1976]. This approach has been extended adding other parameters to the representation: hand orientation and contact area [Battison 1978, Klima 1980].

Later another parameter, non-manual, has been added to the description: facial expression, see Section 2.3.1. Indeed some signs contain facial expression to express the meaning [Bébian 1825], e.g. [RACIST] and [SKIN] where manual features are the same but the face expression is different, Figure 3.18.

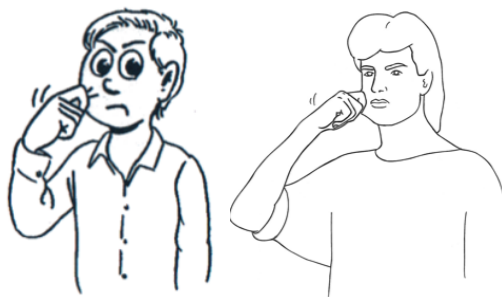


Figure 3.18: Sign [RACIST](Source [Companys 2004]) and [SKIN] (Source IVT) in LSF respectively. Manual features are the same only the facial expression changes.

From a computer science point of view some notations are based on parametric approaches. For example HamNoSys [Prillwitz 1989] and its compatible XML version SignML [Elliott 2004]. Figure 3.19 (middle) illustrates the HamNoSys notation of the sign [WHAT?] in LSF. Sign-Writing¹² is a notation system highly pictographic. Although it is considered as non-linear and non-phonemic system, this is a parametric description based on visual symbols and developed in 1974 by Valerie Sutton, a dancer who developed two years before a notation to write Dancing. Thus linguistic bases are not clearly defined mainly because it has been developed in a first place to write gestures for dancing purposes. In [Filhol 2008] it is considered as an hybrid approach because of the visual iconicity and graphical characteristics of symbols. Figure 3.19(right) shows the description of the signs [WHAT?] in LSF. The goal of this approach is to be able to write and read SL keeping the richness of the language since non-manual expression can be represented. Nowadays it has been used in educational environment to teach Sign Language in some schools [Flood 2002, Brugeille 2006]. It consists on a finite number of symbols describing hand configuration positioned relatively to a symbol representing the face. An XML version has been proposed SWML [Costa 2003] in order to use this description in dictionary access [Aerts 2004]. The XML version intend to capture the relative position between hands and face, as graphically represented, by adding the coordinates in the XML and hand configuration is encoded by a label, e.g. the symbol \Downarrow called Thumb-Index-Middle.

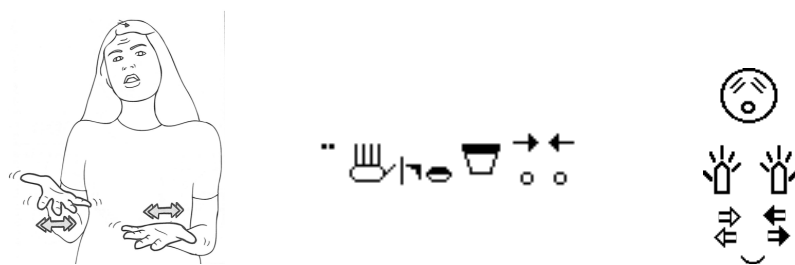


Figure 3.19: Sign [WHAT?] in LSF (Source IVT) and its notation in HamNoSys and SignWriting representation.

3.4.2 Temporal approaches

Unlike previously described approaches where the simultaneity on the organisation of signs is assumed, temporal approaches argue the sequential structure of signs [Liddell 1984, Sandler 1989, Perlmutter 1992, Newkirk 1998, Brentari 1998]. The hold-movement-hold structure model [Liddell 1989, Johnson 2010, Johnson 2011] considers two locations and one movement connecting them. The parameter described previously (location, configuration and orientation) are considered at a lower level of description. The higher level corresponds to a temporal axe representing timing units. Each timing unit give a loca-

12. www.signwriting.com

tion of a hold or a movement between two hold timing units. The sequence of hold and movement timing units defining a sign corresponds to the temporal structure of a sign.

- **Hold (H):** corresponds to a temporal segment in which the parameters describing the sign remain stable. Parameters involved in this segments represent location, configuration, and orientation. A duration can be associated to a temporal segment.
- **Movement (M):** represents a temporal segment where parameters can change. Herein is specified the trajectory of hand(s) movement (straight, arc, circle), movement dynamics (acceleration, slow motion, etc.).

Computer science models for SL generation are based on a linguistic representation of signs with a temporal structure [Filhol 2009a, Losson 2000, Gibet 2008] where the sign description is compatible to sign synthesis and signing avatar.

In [Filhol 2008, Filhol 2009a] is presented a formal model called Zebedee. This is a computational sign representation based on the temporal approach described previously where the structure is defined as a sequence of key posture (K) and transitions (T) timing units, equivalent to hold and movement in [Liddell 1989] respectively. During each timing unit the features describing the geometrical constraints, linguistically motivated, defining the behaviour of the skeleton are specified (illustrated in Fig. 3.20). In particular, key postures use primitive constraints to geometrically place and orient articulators of the body simultaneously, and transitions use various options to specify the shift from a key posture to the next.

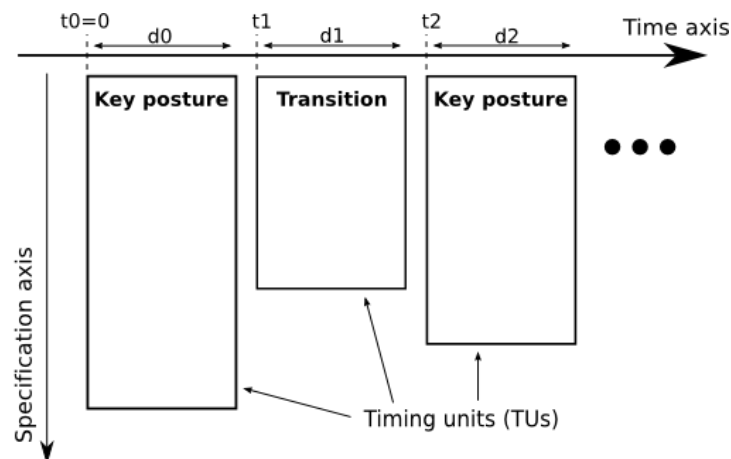


Figure 3.20: The two Zebedee description axes [Filhol 2008]

The novelty of this approach is that it is based on a spatial grammar and a geometrical representation of the lexicon. Moreover constraints account for a lexically relevant intention, not for an observation of a signed result even if it is invisible. For example for the sign [BALL] (Fig. 3.21), the intention of the signer is to move hands around an invisible point where the hand orientation is constrained by the path i.e. the palm orientation is aligned to the normal direction of the path. Thus geometrical constraints can be defined from "imaginary" locations in the signing space. Moreover, every object or value may depend on other objects or values.

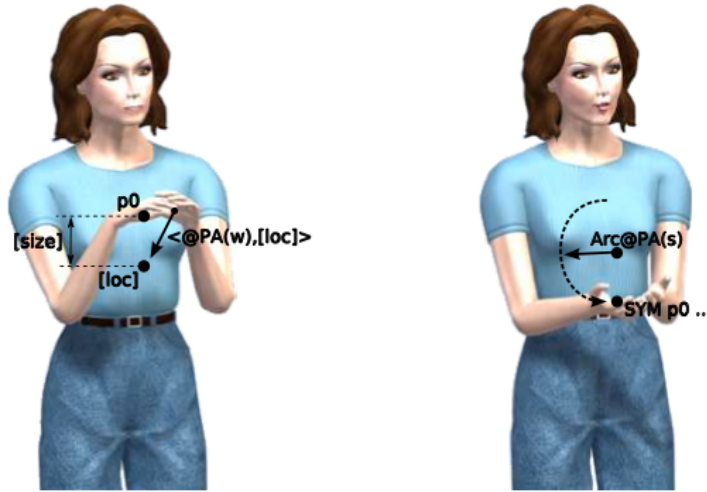


Figure 3.21: Key postures for the sign [BALL] in French Sign Language using Zebedee representation illustrating the elements used to describe the sign, e.g. [loc] corresponds to the imaginary centre, all the other parameters depending on it. Source [Filhol 2008]

Another significant improvement with respect to other descriptors concerns the adaptability to the context-dependency of signs. Indeed it is possible to refer to some contextual elements in the description. In particular, contextual dependencies allow descriptions to adapt to grammatical iconic transformations in context. For instance Figure 3.22 illustrates the sign [BUILDING] in LSF which involves three external parameters Loc, Size and Height (full description in Appendix B). Hand orientation constraint is defined using the orientation of the strong hand and the "imaginary" line L. It is therefore variable in a lot of ways, but all instances will fit the same description.

Some signs according to which is the more comfortable could use different hand configuration. For example \square and $\bar{\square}$ HamNoSys hand configurations can be used (see Fig. 3.23).

This sign representation, first of all designed for SL generation, allows a huge modularity in the description of signs since the same generic description can match all the performance of a sign. This approach is used for automatic sign generation, see Section 3.7, in a platform generation GeneALS [Delorme 2011]. A database of about 1600 signs in LSF have been described using this formal model and is stored in a PostgreSQL database. This uses a dedicated command-based interface to obtain informations from the database. Two main commands are pointed out "INFO" and "FILTER". The former queries a specific information from a sign. The latter, "FILTER" command, allows to narrow down the list of descriptions, given a predicate that accepts or rejects description entries. This allows to obtain a list of signs validating the predicate. Table 3.2 shows a list of features that can be queried from the database. For example to obtain all the sign whose structure is *KTK* means filtering all signs for which the number of transitions is equal to 1, the command line is, then, *FILTER transcount ~ "1"*.

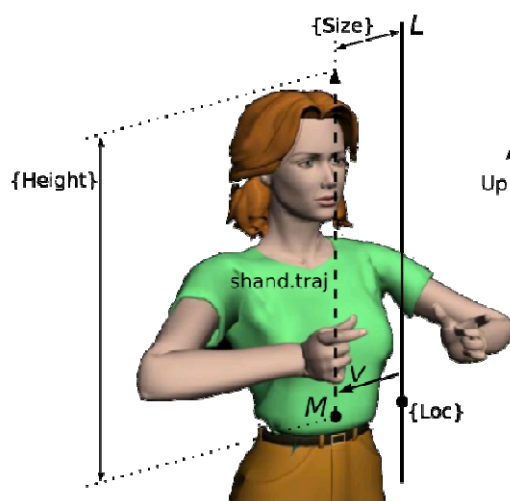


Figure 3.22: Sign [BUILDING] in French Sign Language. Source [Braffort 2008] and different performance depending on what we image to add



Figure 3.23: HamNoSys "flat" and "bent" hand configurations

In short this temporal approach considers the context-dependency of SL unlike parametric approaches. In addition a generic description of signs is able to describe all performances of signs considering dependencies between objects. Finally the already available database and the query interface make this approach very attractive for recognition purposes.

3.4.3 Discussion

Here we have presented some linguistic representations of signs. Parametric approaches describe signs as a combination of independent parameters simultaneously performed. In this kind of approach signs are over-specified because all parameters have a defined value independent from other objects or contexts. Some dependences are considered in HamNoSys, like the palm orientation relative to the path, however other signs might depend on other objects. Also the description of the same sign differently performed, e.g. placing an object in a different place in the signing space, leads to a different description of the sign. Temporal approaches make some assumptions about the structure of signs. The formal model in [Filhol 2009a] has the advantage of considering the high variability of signs, not only by allowing external parameters to modify the production of the sign but also by describing only what really matters in the sign, i.e. what remains stable regardless the context, in a sign, e.g. finger constraints instead of configuration.

Table 3.2: Filter command predicate

| Type of information | Description |
|---------------------|---|
| Name | Obtain descriptions whose name matches the predicate |
| Deps | Obtain the dependences expressed in the description |
| DepCount | Obtain the number of dependences |
| TransCount | Get all the signs having n number of transition "T" in the movement structure |
| TimeStruct | Obtain the time structure, i.e. the sequence of key postures "K" and transition "T" |
| MvtStruct | Obtain the movement defined for each transition "T". |

Synthesising the needs of our work from a sign recognition point of view, see Section 1.5. The need of having a computational model that allows, from the extraction of visual characteristics, to obtain the signs corresponding to visual features leads to a representation as generic as possible taking into account the variability of signs, for example allowing signs to be performed in different ways according to the context. This is the main advantage of the temporal representation Zebedee which leaves what can vary in a sign as a set of external parameters in the description. Our most important need concerns the correspondence of described features in the linguistic representation and the visual features extracted from the performance. None of the existing representations fulfils the needs of SL recognition, though Zebedee [Filhol 2009a] deals with body articulator simultaneity and integration of iconic dependencies at the lowest level of description and allows grouping all possible performances of one sign under a single parametrised description dealing with the variability of signs which is one major problem in SL recognition approaches, see Section 3.6. However the major problem is that the same sign can be described in several ways making very difficult the filtering of signs corresponding to some features. For example when the hand is in front of the face this can be describe as being located with respect to the nose or to the forehead. Even some signs are defined in terms of the palm orientation like in the example of [BUILDING] which is very difficult even impossible to obtain using image processing from a 2D video sequence. Other simple features as the number of hands performing the movement cannot be straight forward extracted from the description but could significantly reduce the number of potential signs fitting the performance.

Although the Zebedee representation model is not adapted for SL recognition, it can be extended to make visual features extracted from videos compatible to the features described in the representation.

3.5 Feature extraction

In order to analyse SL conversations using computer vision techniques, the extraction of representative characteristics has to be performed. Motion and shape features require specific approaches designed for SL purposes since any method for gestures recognition does not fulfil SL recognition needs. Gestures correspond to isolated hand, body and facial movements that are rarely broken down into primitives with few constraints unlike signs which are hand, body and facial movement as part of a sentence which are often broken down into primitives called phonemes and have a numerous phonetic and syntactic constraints.

As it will be described later in the following section, statistical classification methods need some features to be trained. Herein some methods for extracting features from video corpus are described. The features presented in this section are somehow related to the linguistic description of signs, e.g. location of hands, hand-shape, motion, etc., see Section 3.4. These approaches intend to extract information such as the position of head and hands for each video frame to compute motion features; to extract the hand silhouette to study hand shape; and to perform temporal segmentation allowing to identify the limits between signs and transitions for the processing of continuous SL.

3.5.1 Hands and head location and motion

Extracting motion features from head and hands involve the detection and tracking of body parts. Body tracking is challenging because of the presence of noise, occlusions, fast dynamic changes and background complexity. Particularly SL conversations involve high body limbs dynamics and high variability of shape that could take place simultaneously. In addition the interaction of hands and head produces occlusions which because of the appearance similarity of objects are difficult to handle. Many tracking algorithms in the literature have been proposed to deal with these problems, they use several representative features in combination with shape models.

3.5.1.1 Feature selection

Tracking quality results depend on the selected features to track according to the application [Shi 1994]. The selection of representative features depends on the characteristics of the object to track; appearance, shape, dynamics, rigidity, etc. Generally feature selection concerns colour, motion and edges information [Ong 2005a, Yilmaz 2006].

- **Colour** information is used in early works to simplify hand features using customized coloured gloves [Starner 1995a, Sutherland 1996, Bauer 2002, Kadir 2004, Holden 2001, Hienz 1999, Wang 2009]. This assumption not only simplified tracking but also hand segmentation and pose. Figure 3.24 shows an example of a recent customized colour glove which aims to detect hand pose. Although coloured

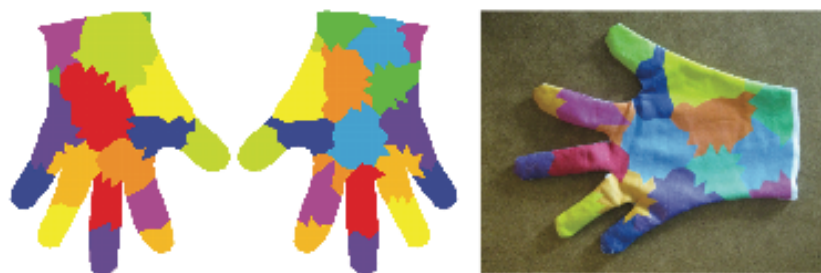


Figure 3.24: Customized colour glove example. Image extracted from [Wang 2009].

gloves are less cumbersome than active motion capture sensors it influences the performance of sign language mainly during the production of some marked hand configurations. A solution is to remove any wearable device but instead using skin-colour models [Fritsch 2002, Vezhnevets 2003, Phung 2005, Duan-Sheng 2006, Kakumanu 2007] to represent hands and head. The similarity of colours between objects make tracking very challenging during occlusions and other simplifying assumptions appear. The restriction of other skin-coloured objects either in the background or belonging to other parts of the body, e.g. signer is required to wear long-sleeved clothing in dark colour. Colour based techniques have the inconvenient that same object model represents various objects, e.g. skin regions represent head and hands, and additional processing is required to identify each object, for example using anatomical [Micilotta 2004, Lefebvre-Albaret 2010] or other data association techniques [Gianni 2009, Deutscher 2000].

- **Motion** cue based approaches [Cui 2000, Huang 2001, Lu 2003, Chen 2003] use techniques to define the translation of each pixel in a region, e.g. optical flow. In [Huang 2001] a motion detector is used to capture all the possible moving objects by examining the local grey level changes. Colour and motion cue [Habibi 2004] are combined and by using a change detector and a skin segmentation technique hands and face are tracked. These techniques consider that hands are constantly moving and are the only moving object in the scene, restricting to a static background and to other body parts motionless, e.g. head, trunk and shoulders. The assumption that other parts of the body are motionless is not adapted for sign language applications since non-manual motion conveys linguistic information, see Section 2.3.1.
- **Edges** information capture important properties of objects. Indeed they correspond to change in depth, surface orientation or scene illumination and are represented by strong image intensity changes. Edge detectors [Bowyer 1999, Martin 2004] gives a set of curves corresponding to objects boundaries. Generally edges are combined with colour information or other features to improve robustness [Birchfield 1998, Wu 2004, Yang 2005]. This feature is often used as edge orientation histograms for simplicity, efficiency and generalization [Freeman 1995, Dalal 2005, Zhou 2004, Lee 1999, Shakhnarovich 2003, Maung 2009]. Unlike colour cue edges are less sen-

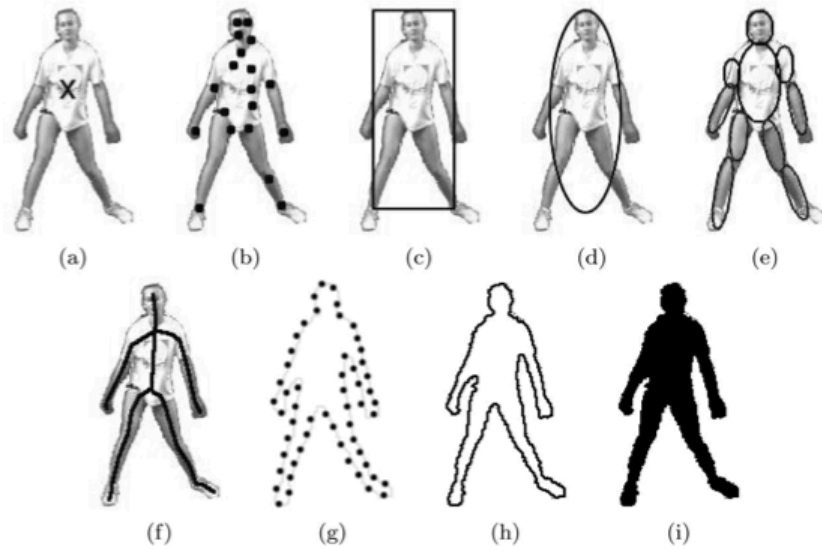


Figure 3.25: Object representation. (a) centroid, (b) cloud of points, (c) rectangular shape, (d) elliptical shape, (e) articulated model, (f) skeleton model, (g) control points on object contour, (h) object contour and (i) object silhouette. Image extracted from [Yilmaz 2006].

sitive to illumination changes but are more sensitive to clutter.

Generally body tracking approaches use a combination of features [Kulkarni 2010] overcoming some drawbacks concerning a feature by the advantages proposed by other features. For example colour features are sensitive to illumination changes unlike edges information or the sensitiveness to clutter for edges features which is not the case for colour features. In addition these features are generally used with a 2D or 3D models.

3.5.1.2 2D model based techniques

Shape based [Ong 2004, Tanibata 2002] techniques take into account spatial layout information. The complexity of the computation depends on the level of representation of the model. Using contours is computationally expensive and might not be adapted for fast deformable objects as is the case of hands. Other simplified models as a rectangular patch decreases the computation time but represents object shape with less detail than contours. Objects can be represented using numerous models depending on the needs of the application and the characteristics of objects, e.g. whether they are highly deformable or rigid. A classification of 2D object models [Yilmaz 2006] is illustrated in Figure 3.25. Some popular 2D models are listed and described below, however this list is not exhaustive and other models could be used.

- **Points model** represent either the centroid of the object or a sparse region of objects, e.g. a cloud of points [Gianni 2009], (Fig.3.25(a),(b)) respectively. The

former is suitable for objects that are small and only slightly deformable. The latter is more adapter for high deformable objects depending on the quantity of the points chosen to represent the object.

- **Geometric primitives models** are used as a simplification on the shape of the object. Assumptions are made to simplify object model, mainly rectangular [Ju 1996, Lefebvre-Albaret 2009] and elliptical [Holden 2005], see Figure 3.25(c) and (d) respectively. Rectangular patch is less accurate than using edges or the elliptical representation, but it allows to use fast computing techniques [Viola 2002]. Although elliptical approaches can better represent objects their parameters remains difficult to compute and, in the case of hands, it depends on the hand configuration and orientation. In works where the shape of the object [Micilotta 2004, Lefebvre-Albaret 2010] is a geometrical primitive with polygonal shape, often combine colour and geometrical cues directly in the observation model. These approaches use anatomical constraints to handle occlusions. Nevertheless objects depend on each other, e.g. elbow position depends on hand position and vice-versa, which is error prone. In addition when hand overlaps the head one skin region might be lost.
- **Articulated models** [Wu 2001, Lu 2003] intend to track an object as an ensemble of components tracked separately and relied by some constraints, Figure 3.25(e). In this case each component is modelled using other single 2D models for example elliptical or rectangular patches for each part of the body, then constraints are added by anatomical models. Figure 3.26 shows an example of hand modelled as a set of articulated rectangular patches estimating straight forward the pose of the fingers. The main limitation of this approach is in case of occlusions or self occlusions, often palm has to be oriented to the camera.
- **Skeletal models** represent a simplification on object modelling since it describes objects as a reduced set of segments and joint angles. These models can be obtained using the object silhouette and are adapted for articulated and rigid object tracking [Cheung 2005, Gall 2009], see Figure 3.25(f). In the case of objects like hands some works use the tracking of some coloured passive markers in order to build the skeleton and estimate hand pose [Holden 2001].
- **Contour models** define limits of the objects either by a continuous curve or by a set of points delineating object contours, see Figure 3.25(g),(h) respectively. It can be used to represent non-rigid objects however algorithms based on contours models are complex and time consuming. In addition the changing on the shape has to be performed slowly e.g. snake statistical based approaches [Heap 1995]. Also occlusions are difficultly handled since contours are sensitive to clutter.
- **Silhouette models** are represented by the connected pixels region inside the contours of objects, also called blobs [Imagawa 1998, Roberts 2004, Habili 2004, Soontranon 2005], see Figure 3.25(i). Methods based on this model are well adapted to the fast shape changing, however it requires the segmentation and labelling of

blobs. An important problem arises in the case of occlusions between similarly coloured objects. Hand and head models are, generally, equivalent in tracking systems, however unlike hands head shape variation is still and texture is slightly modified and can be detected easily using simple filtering techniques [Viola 2002] however the challenge appears when the head is occluded by the hands or in case of head rotation.



Figure 3.26: Articulated hand models for hand tracking and pose estimation. Image extracted from [Lu 2003].

This list of 2D models is not exhaustive. Indeed objects can be represented in numerous ways and tracking approaches generally use these models for the extraction of other features e.g. colour inside a rectangular patch. Hands and head tracking can be performed either separately [Piater 2010] or using simultaneous object labelling for dealing with the interaction between objects [Micilotta 2004, Gianni 2009]. The latter is more robust since it deals with significant overlapping and complex interactions between hands and head. These works use other constraints like anatomic models [Lefebvre-Albaret 2010] or probabilistic data association [Deutscher 2000, Gianni 2009].

Tracking approaches can be either deterministic or stochastic. The former are based on a similarity cost function between a template and the current image incorporating, then, *a priori* information is considered [Birchfield 1998, Tanibata 2002, Bradski 2002, Hager 2002]. The latter methods are based on a dynamic model of the system. In the case of linear-Gaussian model, a Kalman filter estimates the posterior probability density function [Jang 2002, Kiruluta 2002, Stenger 2001]. For non-Linear or non-Gaussian multi-modal distributions, the particle filter algorithm [Isard 1998] has become very popular.

Particle filter based tracking algorithms usually use contours, colour features and appearance models [Gianni 2009, YoungJoon 2010]. An important problem in body parts tracking remains the occlusion between objects. In fact the similarity of appearance between hands and head make the tracking challenging. In previous works, head and hands occlusions are usually handled using data association, body features or local features [Gianni 2009, Tanibata 2002, Lefebvre-Albaret 2010]. Gianni *et al.* [Gianni 2009] proposed a particle filter based tracker using colour cue. Their technique considers each target as a cloud of points and use a probabilistic exclusion principle to associate and interpret data in terms of various targets. This avoids filters to converge to the same object and occlusions are directly handled. However during occlusion, filters share the

same skin region and the position of each object cannot be accurately determined. Lefebvre [Lefebvre-Albaret 2010] uses colour cue and anatomical models. Body features extraction, torso and elbow recognition, aims to expect the position of other objects partitioning the searching space. This technique is fast in terms of computation time but makes targets prone to error since objects are dependants on each other. Tanibata and Shimada [Tanibata 2002] use a template matching technique to robustly handle occlusions. Texture templates of head and hands prior to the occlusion are used for template matching, thus occluded objects can be separated considering local features. However while head deformation can be considered small, 2D hand shape has a high variability and hand shape can change during occlusion.

These works use a single cue, pixel or polygonal region colour probabilities, in the observation model which has proved to offer good results. However in dynamic environments where changes on illumination, shape and occlusions occur the integration of several cues represents a solution. Some works consider the fusion of several cues as a linear combination of particle weights for each feature independently computed [Raducanu 2006, Zhao 2007], e.g. colour and geometrical features. However it is dependent on the coefficients used in the linear combination.

Herein were presented the most representative methods for tracking using 2D models, handling occlusions in this approaches is challenging because 3D information is missing. Other approaches consider 3D object representation using a single camera or multiple cameras systems.

3.5.1.3 3D model representation

Three dimensional model based approaches mainly focus on reconstruction and pose estimation [O'Rourke 1980, Delamarre 1999, Duetscher 2000, Stenger 2001, Horain 2002, Delamarre 2001, Ding 2009]. In Section 3.2 it has been presented several methods for human reconstruction using motion capture methods which give high quality results but are cumbersome and not affordable. Herein some methods using computer vision techniques either from single or multiple camera systems, are described.

Simple assumptions can be used to determine the 3D position of objects. For example the size of the objects, when it is close to the camera the greater is the size. However this is only valid for rigid object where the shape remains the same, in the case of hands this is not possible because of the high hand shape variability. The distance between the camera and the object can also been determined locating the camera in a raised position [Brooks 1997]. However this method constrain hands movement plane.

In the same way 2D model concern geometrical primitives, skeletal or articulated models this is also the case for 3D approaches. Using 3D models allows to obtain additional information concerning the three dimension of objects. In [Downton 1992] is used a 3D cylindrical model with a matching process allowing to estimate kinematic parameter for the model. Other approaches represent hands as 3D objects [Delamarre 1999],

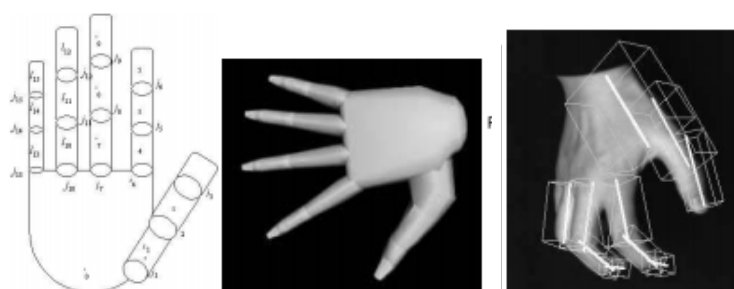


Figure 3.27: Hand tracking example using a 3D articulated model. Image extracted from [Wang 2009].

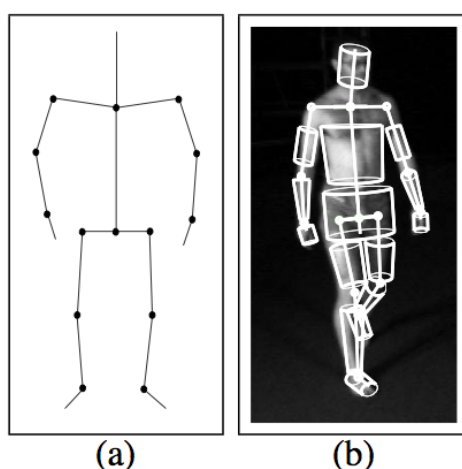


Figure 3.28: Example of the 3D articulated model used with the Annealed Particle Filter. (a) shows the segments and joint angles and (b) the articulated geometrical model. Image extracted from [Duetscher 2000].

e.g. ellipsoid [Drummond 2001] or cylinders [Deutscher 2000] but this is computationally expensive and does not take into account the high variability of hand shape.

Articulated models are popular for tracking human and body parts. This considers the high variability of shape, see Figure 3.27 but are computationally expensive. Stochastic algorithms as particle filter have been extended in [Duetscher 2000] to propose the Annealed Particle Filter (APF) with an articulated 3D model, see Figure 3.28. The limitation of this approach is that it uses multiple cameras to disambiguate object position. Works using multiple cameras to obtain 3D information [Vogler 1997, Bernier 2009] require complex recording set-up and time consuming calibration process.

Other methods employ prior information to avoid using multiple cameras [Mikic 2001]. This kind of methods use 3D motions as training data to model motion using methods like the Principle Component Analysis (PCA) [Sidenbladh 2000] or Gaussian Mixture Models built from several motions [Howe 1999]. The limitation of this kind of approaches is that

results depend on the training data and in the case of SL application motion is very complex with high dynamics and completely free.

Methods avoiding training data use additional information from anatomical models to find specific parts of the body [Lee 2002], e.g. torso, hands, head and elbow. Partitioning the searching space using known position. This is faster but make objects dependent on other objects.

Using 3D models is a good solution depending on the application, however the reconstruction of the body pose or the hand configuration using 3D models at each frame is time consuming and error prone.

3.5.1.4 Discussion

The main challenge in body parts tracking is the development of tracking algorithms robust to the presence of noise, occlusions and unconstrained and highly variable motion. Herein motion features extraction algorithms have been presented. The feature to track depends on the object characteristics. For example head and hands representative feature is the skin-colour. However colour is sensitive to occlusions between hands and head because of the similarity of appearance. The combination of colour features with other cues robust to occlusion improve tracking results. Generally tracking methods use local features in addition to a 2D or 3D model.

Body parts tracking methods are either determined by detecting each body part independently or using statistical models learned from annotated data. The latter needs a manual annotation steps which has been argued along this PhD thesis that is time consuming, error prone and unreproducible. In addition the statistical model depends on the representativeness of the training dataset. In SL performances motion is uncontrolled and this is impossible to build a representative learning dataset without getting the vocabulary and the context extremely constrained. For these reasons methods avoiding learning steps are privileged. Good results have been shown using stochastic methods which are based on a dynamic model of the system like Particle Filter approaches or the improved method Annealed Particle Filter.

The choice of the object model depends on the needs in terms of features to extract. Approaches based on linguistic models require motion and location features such as the relative location to the body and the trajectory within a sign. In other words the location of hands and head at each frame belonging to the sign sequence. Using 3D models for SL purposes concerns SL posture recognition and becomes more complex than extracting hand and head location using 2D models. Complex models like articulated or skeletal models are time consuming and are better adapted for pose recognition or human reconstruction problems. Simplest models as rectangular patches can improve execution time but are not adapted for highly shape variability objects as hands though it might be adapted for head since the shape changing is still. Contour based models consider the shape variability, however tracking contours for hands is challenging because of the

high shape variability. Silhouette models represent well the shape of the object but are difficult to segment. Using blob models for hands and head make it challenging without any further processing in the case of occlusions. A simplification of this consists on using a dense cloud of points which is adapted for shape variability and at the same time can improve execution time.

In the literature, methods tracking hands and head consider the same kind of model for head and hands. For example in [Gianni 2009] hands and head are considered as a cloud of points to take into account shape variability. In [Micilotta 2004, Lefebvre-Albaret 2010] head and hands are considered as rectangular patches. In fact hand and head motion features are quite different. Head move slower than hands and the possible paths are completely different in SL performances. In addition head shape is still compared to hand shape variability. For this reason using the same kind of model and the same tracking approach for completely different objects seem not adapted for SL motion analysis. That is why in this work it is proposed a different approach which uses adapted models and algorithms for hands and head, see Chapter 4.3.

All these approaches aim to track hands and head positions as a system and not independent objects for motion analysis of human gestures. Additional processing is necessary for hand configuration analysis, such as hand segmentation, in Sign Language (SL) or for a full SL recognition system.

3.5.2 Hand shape

Hand shape features are extracted to furnish information that combined with motion can better describe signs. Here first of all have to be considered the segmentation process which consists on isolating hand region for further processing. These regions are used for the extraction of geometric features and classification of hand shape.

3.5.2.1 Hand segmentation

Hand segmentation intends to isolate the hand from the background. This can be achieved once the hand position has been detected or as part of the detection process. Hand extraction can be straight forward performed where the hand background is different from skin-colour. The difficulty concerns hand segmentation when the background contains similarly coloured objects. For example when hand overlaps the face, hand segmentation becomes very challenging because other features than colour have to be considered. Early works use coloured markers to simplify hand segmentation [Davis 1994]. Other approaches assume that the hand is the only object in the image [Hamada 2002] or the only skin region [Cui 1995, Zhu 2000]. Skin models are popular for extracting skin regions in an image which can be labelled as hands or face [Cui 2000, Habili 2004, Ramamoorthy 2003, Awad 2006, Howe 2008]. However, these approaches do not handle skin objects occlusions. In [Awad 2006] an occlusion detection method is used but hands



Figure 3.29: Example of the results performed using image force field. Image extracted from [Smith 2007].

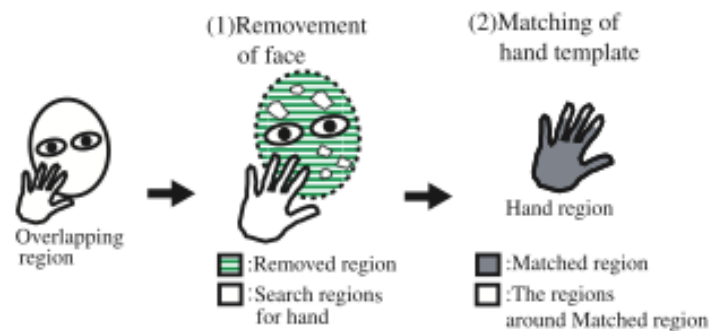


Figure 3.30: Illustrates the hand extraction using a template for hand and head before occlusion. Image extracted from [Tanibata 2002].

are not segmented when they are placed in front of the face. In [Diamanti 2008] hand segmentation during occlusion is addressed however they use a priori information of hand shape before occlusion. Thus when hand configuration change during occlusion cannot be handled.

Some methods address hand over face occlusions using active contours giving good results. But do not cover the fast change and variability of hand shape [Ahmad 1997, Holden 2005]. They assume that hand shape change is very small between successive frames which is, normally, not the case in sign language unless the video is acquired on specific recording conditions such as high-speed frame recording.

In [Smith 2007] an approach to solve hand over face occlusion is introduced using the concept of image force field. Results show that the hand is roughly segmented, an example of a sequence where the hand passes in front of the face is shown in Figure 3.29. In fact this only give a region where the hand region might be without really extracting hand region. This might not be enough for robust hand shape features extraction, classification or recognition.

In [Tanibata 2002, Von Agris 2008] is introduced a template based approach. They consider the face and hand template before occlusion. Even though face deformation remains small, 2D hand shape, during occlusion, can quickly change without any hand configuration changing. In theses approaches it is necessary to first find an approximated position of the face, using a tracking algorithm, and later to register the image template to perform the segmentation.

These approaches use mainly colour, edges and template based features. In the case of occlusions colour and edges are ambiguous, it is very difficult to distinguish which pixel region belong to the face and which to the hand. The combination of these features may greatly improve hand over face segmentation, see Chapter 4.4.

3.5.2.2 Geometric features extraction

Hand region is used to compute geometric features that characterise hand shape. These features can be used for training classification systems or for directly compare shapes. Describing objects using the shape require the extraction of dimensionless quantities independent of its size that are representative of the object shape also called shape factors. These are generally computed using measurements from the segmented object such as diameter, area, etc., and represent their similarity to ideal objects, e.g. ellipse, circle, etc. Often the shape factors are normalized so the similarity quantities varies from 0 to 1 when 1 corresponds to the maximum similarity. They are applicable to all geometric shapes.

In the literature many algorithms use shape factors to extract hand shape features in SL performances. They use factors as circularity and direction of the inertia principal axis [Shiosaki 2008] or the flatness of the hand region and its area [Tanibata 2002]. In [Von Agris 2008] a combination of several factors are used for orientation of main axis, ratio of inertia, compactness and eccentricity.

Herein we present some shape factors commonly used for hand shape representation, though this list is not exhaustive and other geometrical feature could be used.

- **Centre coordinates** x, y . This position can determine different spatial aspects of the object. For example the gravity centre of the region, the centre palm position, the centre of the bounding box containing the region, etc. This depends directly from the algorithm used for the detection and tracking of objects and the model chosen for their detection and tracking, see Section 3.5.1.
- **Area** a **and perimeter** p , the area quantifies the extent of a 2D surface. Ideal shaped object area is computer using predefined formulas. In image processing the area of an irregular region is computed using the pixels belonging to the region from a binary image. This is defined as follows :

$$a = \sum_{\{x,y\} \in I} p_{x,y} \quad (3.1)$$

where $\{x, y\}$ correspond to the position each pixel in the binary image I and $p_{x,y}$ the value of the pixel either 0 or 1. And the perimeter corresponds to the path length quantity surrounding an area. It is for a region the number of pixels constituting the contour of the object.

- **Aspect ratio** r describes the relationship between the largest diameter and the

smallest diameter orthogonal to it quantifying the proportionality between its width and its height, see Figure 3.31. Defined as

$$r = \frac{j_1}{j_2} \quad (3.2)$$

where j_1 and j_2 correspond to the diameters of the object.

- **Circularity** c of a region is the quantity that expresses the roundness of an object. This is somehow related to the eccentricity. This quantity varies from 0 to 1 where 1 corresponds to a circular shaped object. It is defined as

$$c = \frac{4\pi a}{p^2}, \quad (3.3)$$

where a and p correspond to the area and the perimeter of the region respectively.

- **Eccentricity** represents the quantity measuring how a region deviates from the circular ideal shape. This measurement gives similar information than circularity but here the eccentricity of a circle is zero. Region close to ellipse shaped is greater than zero but less than 1. The computation considers the central moments of the region and is defined as

$$\varepsilon = \frac{(\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}}{a} \quad (3.4)$$

where $\mu_{p,q}$ represent the central moments. The main advantage of this measurement is that it is invariant to the size.

- **Equivalent Diameter** corresponds to the diameter of a circle whose area is equivalent to the area of the region. The advantage of this quantity is that it is invariant to rotation and displacement and that it is robust to noise, however it is dependent on the object size.

Hand shape classification systems use representative features of objects like the ones described above. In [Von Agris 2008] presented a method for classifying shapes where an off-line database is built and later during the recognition procedure, this is queried using the features extracted from the region e.g. compactness, eccentricity, etc. Filtering gives potential shapes which are disambiguated using the continuity of the shape over time. Other hand detector and hand classifier approaches use a boosted cascade classifier [Ong 2004, Francke 2007] or hierarchical decision tree [Coogan 2006] where each leaf corresponds to a hand shape. Approaches using transition networks built by learning the different transitions from one shape to another [Fillbrandt 2003, Hamada 2004].

Herein it has been described the most common measurements used for expressing shape features. Although geometrical information can be used to describe an object for learning statistical models, the region or its contours can be directly used for building such a model for further hand shape classification.

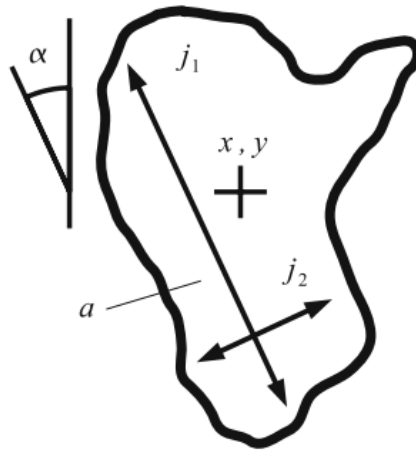


Figure 3.31: Geometric features. Image extracted from [Von Agris 2008].

3.5.2.3 Discussion

Manual features extraction concerning hand shape are challenging because of the high similarity between face and hands and the high amount of occlusion between objects. Even though hand characterization consists on a processing that has to be performed once the position and the hand region are known, hand segmentation is challenging. Approaches in the literature use colour, edges and template based features for performing the segmentation of the hand when it is in front of the face using different approaches. The limitation of these approaches concerns the high variability of hand shape. The quality of hand characterisation depends on the quality of the segmentation results, thus robust hand segmentation methods in complex configurations is required, e.g. when the hand is in front of the face. Existing approaches in the literature dealing with hand over face occlusion roughly segment hand [Smith 2007] which is not enough for further study of hand shape. A new approach has to be designed for dealing with this challenging problem. In Section 4.4 we present the proposed approach designed in this PhD thesis.

Hand shape could be, after segmentation, classified using statistical classification methods to recognise hand configuration. As we mentioned before hand configuration recognition is very challenging because the same configuration could leads to several shapes according to the point of view. Although several classification methods using the contours or the region exist they are constrained to the shapes on the training data, a frontal view of the hand and only few configurations [Ong 2005b]. In fact classification methods, e.g. Support Vector Machine (SVM), require an off-line training step making results dependent on the data in the learning set. Extracting geometrical features is, thus, more suitable for characterising hand shape without recognising configuration. In addition they allow to characterise shape through a finite number of features reducing space and allowing to easily compare shapes between them. Selecting representative and complementary features is essential for robustness without needing a huge list of features.

For example eccentricity is robust to scaling but sensitive noise, equivalent diameter is sensitive to scaling but robust to noise. This two complementary measurements are very attractive.

These features characterising hand shape in addition to motion features are then used to characterise signs for boarder detection or for its recognition.

3.5.3 Sign boundaries

Sign language recognition approaches focus on the recognition of isolated signs or continuous SL. Many approaches intend to recognise isolated signs since this is easier than continuous SL. Even though extending approaches designed for recognising isolated signs is not adapted for continuous SL recognition because of the context-variability of signs, some approaches intend to train recognition systems from isolated signs repeated several times either starting and ending in a neutral position or by exaggerating pauses between signs. In this case word boundaries are known unlike continuous SL. Continuous SL recognition problem considers word or sub-unit models, see Section 3.6 which have to be extracted from the continuous sequence. The explicit segmentation of signs gives an important information concerning word boundaries or in the case of sub-unit models, as the ones describe in Section 3.4, allows to identify the structure of signs. For example linguistic descriptions based on temporal approaches, see Section 3.4.2, involve the segmentation of signs to define the sequence of Holds and Movements for further processing.

The main advantage about explicit sign segmentation is that this allows to extract only important information at key frames, e.g. the beginning or the end of the sign, instead of obtaining all the information at the same time for each frame which is time consuming and unnecessary.

Identifying borders from a linguistic or from a computer science point of view is still challenging and lead to discussion between linguists and computer scientists. In fact it is difficult to define word borders [Brentari 2006]. Sign segmentation remains a subjective procedure for linguists and depends on the knowledge and appreciation of the language. An interesting experience¹³ has shown some differences on the manual segmentation of sign language [Braffort 2012] where several teams manually and semi-automatically segmented the same sequence and compared their results [Lefebvre-Albaret 2012, Millet 2012, Gonzalez 2012b]. In Figure 3.32 is presented an example of the results from this experience. Notice that the same sign is segmented differently for all the teams. This shows the main challenge during word segmentation which is the lack of high level information or any standardised boundaries definition. Also this makes unsuitable the use of any training data which is biased by the annotators experience and interpretation. The example shown corresponds to an isolated sign during the discourse where the borders should be defined with less ambiguity, however what we notice is that border selection remains

13. <http://degels.limsi.fr/>

quite different between all the teams.

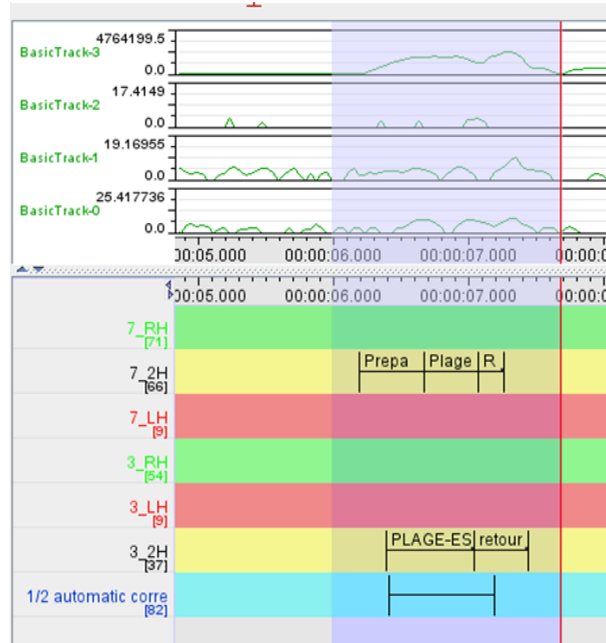


Figure 3.32: Segmentation comparison between three different teams segmenting the same sequence. Source [Braffort 2012].

Some approaches do not explicitly need the temporal segmentation of signs since this is implicitly and automatically considered during training, particularly using Hidden Markov Models (HMM) [Vogler 1999b]. This requires the use of training data which is to be avoided in this PhD thesis because of the annotation constraints. Then using motion and shape features that can be straight forward extracted are preferred to characterise word borders.

3.5.3.1 Segmentation features

Herein we present some features used for characterising word borders. Although non manual features could give important information concerning word boundaries this has not been deeply studied from computer vision approaches due to the complexity of automatically extracting non-manual features. Several features, particularly manual can be used for detecting boundaries such as velocity, change on the trajectory, curvature, directional angle, etc. In [Khan 2011] they seek for pause length detection, change on motion or shape repetition.

- **Pause length** is often associated to word boundaries. It consists on holding hands at the same position and with the same configuration for a few time. This could be detected either by verifying that hands positions remains stable during sometime or considering velocity zero. During continuous SL pause is difficult to distinguish since

sign are performed one after another. Approaches improve segmentation results by imposing an artificial pause between signs.

- **Motion features** are the most popular features used for characterising word boundaries. In fact motion feature extraction is easier to extract than hand shape or non-manual features. They are often used for detecting a change on the direction of articulators, to compute the velocity profile, etc. Velocity and acceleration are computed from the position of hands for each frame using a moving window. The magnitude and the direction can be used to compute rates and classify sequences. For example relative velocity between hands which corresponds to the difference on the velocity magnitude between two hands allows to classify sequences as two-hands, one-hand or static signs.
- **Shape features** can be used to determine word boundaries in combination with motion features. In fact because of the high shape variability the use of shape features alone to determine boundaries is not possible. Even if we intend to use the full configuration, which is difficult to determine from image processing techniques, often the configuration change during the performance of the sign.
- **Repetition** consists on performing the same gesture several times either for emphasising a sign in the discourse [Khan 2011] or because the signs consists on some motion path performed several times [Lefebvre-Albaret 2008]. This is generally detected through the articulator trajectory and direction by searching similarly repetitive patterns.

The features above described are used to identify word borders in continuous SL, though this list is not exhaustive. Most approaches use a combination of features, however they do not consider it as a weighted combination thus all features are considered to represent sign boundaries in the same way, though often some features are more significant than others.

Approaches in the literature herein described are classified according to the data acquisition methods used: device based or vision based approaches. The former corresponds to the use of motion capture systems, see Section 3.2.1, allowing to straight forward acquire motion and shape features. The latter instead uses image processing algorithms to extract features. Even though it has been argued before the inconvenient of using device based approaches it seems important to show the features used by this kind of approaches which could be eventually extended to be used within vision based methods.

3.5.3.2 Device-based approaches

Temporal segmentation methods, using motion capture devices for collecting data, are described. Although our work focuses on vision based techniques, it is interesting the way in which sign segmentation has been performed using a direct and accurate acquisition method. The features used in this kind of methods could be eventually used

in vision based approaches unless the features cannot be extracted using image processing techniques.

Boundaries detection has been addressed in several ways using motion and shape features [Liang 1998, Sagawa 2000b, Bauer 2002, Gibet 2007, Kong 2008, Han 2009]. In some approaches like in [Sagawa 2000b] boundaries are detected using motion features as minimum in velocity and large trajectory changing. In addition segments are labelled as signs or transitions using the acceleration and velocity ratio.

In [Liang 1998] is also taken into account hand shape features in addition to its motion. It considers time-varying parameters (TVP) which are: hand posture, position, orientation and motion. In this approach signs are considered as a sequence of hand shape linked by their movement, however this does not consider sign involving hand shape changing while the hand is moving which is often the case in SL.

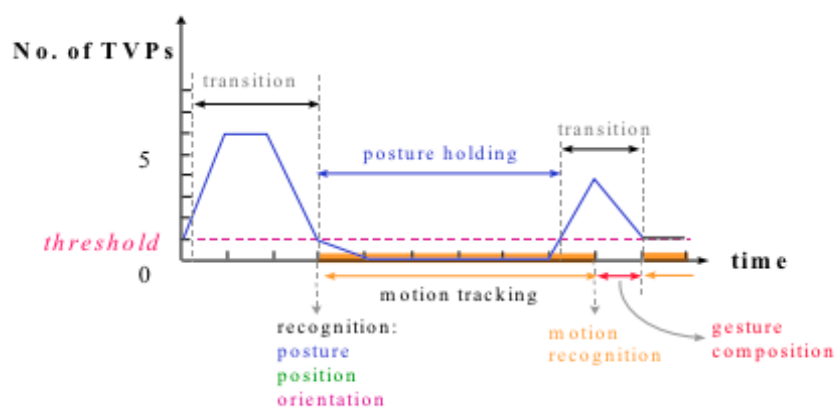


Figure 3.33: Boarder detection parameter. Source [Liang 1998].

Other measurements can be used for characterising gestures boundaries like curvature or directional angle in addition to velocity [Gibet 2007, Kong 2008, Han 2009] or involve a training step to learn representative features and perform temporal segmentation [Fang 2002, Yang 2006a]. In [Fang 2002] a self-organising map is used to automatically extract features for word boundaries segmentation. This has the advantage of learning appropriated features.

Approaches described here use DataGlove sensors, motion capture systems or colour gloves for simplifying the data collection. Some features are used in vision based approaches, particularly motion. However hand configuration and orientation is very complex and is generally avoided. Features used in vision-based approaches are detailed below.

3.5.3.3 Vision-based Approaches

Vision based approaches use tracking and segmentation algorithms to extract features from a sequence of images. The extracted features are less robust than the one obtained using motion capture devices. As we mentioned before segmentation is often avoided using learning recognition systems based on HMM. Other approaches intend to explicitly segment signs or part of signs that are representatives, sub-units.

Some methods instead of finding representative features for boundaries segmentation they model movement epenthesis to generate a stochastic model [Gao 2004, Yang 2010, Kelly 2009]. This is interesting for few signs and depends on the characteristics used to build the model. Indeed for continuous SL the context might strongly influence the performance of signs. Characterising movement epenthesis in terms of velocity and acceleration profile [Pitsikalis 2010] allows to decompose signs phases; preparatory, holding, transition, etc. Using features from movement epenthesis is representative since the motion corresponds to a ballistic movement which take the fastest trajectory for going from the end of a signs to the beginning of the following sign; a straight line.

Other approaches using image processing techniques have been proposed in the literature to explicitly segment signs. Nayak et al. [Nayak 2009] proposed an unsupervised approach to automatically segment signs by extracting parts of the signs that are present in most occurrences. They consider relative position between hands by using multidimensional time representation. However some characteristics of signs are influenced by the way in which signer place objects in the signing space and relative position between hands and head can be very different for the same sign depending on the context.

A motion-based approach has been introduced in [Lefebvre-Albaret 2008] to semi-automatically segment sign in the conyexy of SL corpora annotation. In this work only motion is considered to identify various kinds of symmetry. The initialisation step consists on asking the annotator to select one and only one frame for each sign, called "*seed*" frame. In this way segments that contains a seed frame are considered as a sign. However many signs are composed of several segments and kinds of symmetry, these signs will be over segmented.

Concerning vision based segmentation approaches we notice that they only use motion and relative position unlike device based approaches. This is because extracting shape features from a mono-camera video is very difficult though shape features conveys lot of information.

3.5.3.4 Discussion

Sign language recognition systems require to explicitly segment signs if a training step has to be avoided, this is the case of annotation in terms of glosses. Word segmentation is a challenging task because of the lack of linguistics information. So far this tedious task is performed manually and not standardisation of the criteria has been defined. In fact defining word borders remains a difficult problem.

Computer scientists have proposed several semi-automatic segmentation methods using several features extracted from the corpus in order to assist the annotation but also to propose to linguists objective features. We have noticed that they use manual features either by collecting the data using motion capture devices or by using image processing techniques. Approaches collecting data from special devices are able to use information that cannot be obtained for image processing techniques.

A fully automatic approach for word boundaries is challenging because specifying rules that are systematically respected with few features is very difficult and the effectiveness depend on the selected features. Indeed the same sign is performed differently depending on the context and even sometime during the performance of a one-hand sign the other hand moves to prepare the following sign.

3.5.4 Discussion

In this section we have detailed the state of the art for feature extraction methods concerning manual features. As we mentioned before these features correspond to the phonemic annotation level defined by linguists which will be used later for annotating at a lexical level.

We have detailed what exists in the literature in terms of hand and head location and motion, hand segmentation and characterisation and word boundaries detection. We have also discussed the advantages and limitations of existing approaches according to our specifications.

- **Location and motion** of hands and head is a complex task because of the high dynamics and the presence of occlusions between objects. It is important to select wisely significant feature to be used; colour, edges, motion, etc. Each features has its own advantages and limitations and a combination of complementary features is suitable for improving robustness. Head and hands representative features is the skin colour, however this is sensitive to illumination unlike shape which is robust to illumination changes but sensitive to cluster. In the literature numerous shaped models have been proposed, the choice of using one model depends directly on the object shape and on the level of detail desired. For example choosing a rectangular model for face is a trade-off between representation detail and speed. In addition adapted tracking algorithms have to be used according to object dynamics which are very different for hands and head. These has to be taken into account for designing a robust tracking algorithm, see Section 4.3.
- **Hand shape** features give additional information. For extracting hand shape features it is needed to obtain hand region from images even in complex configurations, e.g. hand in front of the face. This is specially challenging because of the appearance similarity between objects. Instead of detecting hand for segmentation, it is possible to use results from the tracking methods at the initialisation step. However results will depend on the quality of the tracking results during occlusions, thus a

tracking robust to occlusions is required. For segmenting hand in front of the face, appearance alone cannot be used. Other features, in combination to appearance, for finding hand borders accurately are required. In fact this is the case of edges, the main challenge is the manner in which edges can be classified as belonging to the hand among all the edges in the image. This is discussed in Section 4.4. Once the segmentation has been achieved the characterisation of shape is needed. For this geometrical measurement are proposed in the literature. According to the level of characterisation a set of features has to be selected. In our case we do not intend to describe shape accurately, thus only few representative features are needed.

- **Sign boundaries** detection correspond to the detecting the beginning and the end frames of a sign in a sequence. Here we face a very important problem the definition of word limits. We have shown that temporal segmentation from a linguistic point of view is not well defined and remain purely subjective. In order to make it objective computer scientists work on the automatic selection of features characterising the performance. Motion features are the most well-known for detecting words boundaries. Although hand shape has been used for word segmentation, this has only been performed using device based approaches since recognising hand configuration or classifying shape is challenging. So far image based approaches only use motion features, the introduction of hand shape can significantly improve word detection but has to be used wisely.

These features correspond to the phonemic level of annotation. Going to a higher level, e.g. a lexical level, needs further processing and the interpretation of the features extracted at this level; location, motion, shape and word boundaries. The lexical level consists on recognising the lexical meaning of a sequence, this is also called glossing. For this the existing methods in the literature for gloss recognition are presented below even though they might not be adapted for SL annotation but for recognition purposes.

3.6 Automatic Recognition of Signs

In the previous section it has been discussed automatic feature extraction methods which are a basis for SL recognition systems. In this section we present existing works in the literature for recognising signs from a video sequence by describing approaches concerning the combination of features describing signs. Further information is found in recent SL recognition reviews [Ong 2005a, Von Agris 2008, Cooper 2011].

Early works on SL recognition considers the extension of spoken languages recognition approaches. However this is not suitable because the data to process is completely different and processing videos is more challenging than acoustic signals. Recent works generally use statistical classification methods SL recognition which needs a model, previously built, of signs in the vocabulary to be recognised. This is addressed using either word or sub-unit models. The former considers signs as a whole set and the latter as a

set of subunits, called cheremes or phonemes. The choice on the sign model depends on the number of signs composing the vocabulary and the availability of data for building the model.

Often the training data is composed of isolated signs performed by various signers several times and are used to recognise isolated signs [Grobel 1997, Vogler 2001, Bowden 2004], however many signs are context-dependent and the co-articulation phenomenon is not considered. Co-articulation and context-dependency is an important problem since large amount of data is required to train recognition systems and achieve high recognition rates. This leads to a main problem while using word models. Since each model represents a word, then the same word in a different context, thus differently performed, requires the building of its own word model unlike sub-unit models. Herein we describe spoken techniques for SL recognition as well as approaches for the recognition of signs using word or sub-unit models. This will give an overview of what exists in the literature and how well adapted this is for annotation purposes.

3.6.1 Speech recognition techniques for SL recognition

Automatic speech recognition approaches have seen many advancement in this domain over the last 30 years. Speech signal are produced from the vibration of the vocal cords modulated by an articulatory system whilst sign languages use the visual-gestural channel adding another dimension with the use of the signing space. Processing of two dimensional signals, e.g. video frame, are significantly more complex than one dimensional acoustic signals.

The acoustic signal is, generally, composed in terms of fundamental frequency, energy and spectral frequency. Feature extraction for acoustic signals consists on finding some transformation to dissociate frequency and energy parameters. These parameters are lately used in classification methods such as Dynamic Time Warping [Myers 1980], Hidden Markov Models or neural networks [Lippmann 1987] for speech recognition. Early works proposed to adapt speech language recognition [Rabiner 2010] techniques for SL recognition [Dreuw 2007]. They are hardly transposable to the study of SL [Bruno 2002, Dreuw 2008c] mainly because sign languages and spoken languages are very different concerning the kind of signal to process.

One main problem concerns the high difference between the acoustic signal and the video data. But other important problems involve the characteristics of the language. Main differences between speech and sign language recognition, other than the data to process, consist on the

- **simultaneousness** of articulators while speech is sequential. Sign language uses several manual and non-manual articulators in parallel, see Section 2.3.1 giving the whole meaning to the sentence, e.g. adjectives are indicated using the facial expression.
- **variability** on the performance of signs according to the context. Same signs can

be placed anywhere in the signing space according to the context.

- **signing space** in which persons or objects are placed at some point of the discourse to then be referred creating relationships [Braffort 2005], refer to Section 2.3.2
- **iconicity** which makes the production of iconic signs completely free. Thus a standardised vocabulary database cannot be easily collected. Generally speech recognition algorithms use a large set of vocabulary.

For these reasons spoken language techniques are not straight forward adapted for SL recognition. Other methods using models of signs are preferred. Below we describe approaches considering word models.

3.6.2 Classification methods

Statistical methods are used for training and for classification of representative features extracted from a set of learning data [Von Agris 2008, Cooper 2011]. For this significant features are required. Common classification methods are, as in the case of speech recognition, Neural Nets (NN) [Vamplew 1998, Huang 1998, Munib 2007], Dynamic Time Warping (DTW) [Heloir 2006a, Lichtenauer 2008b, Kim 2009] and Hidden Markov Model (HMM) [Starner 1995b, Assan 1998, Vogler 2003, Al-Rousan 2009, Theodorakis 2009].

- **Neural Networks** are mainly used in hand configuration recognition than in sign recognition. For sign recognition approaches existing methods process images before using them in NN. For example [Munib 2007] uses a Hough transform and the results is then used in a NN achieving 96% on the recognition rate over 14 signs. NNs based approaches use a very small vocabulary to achieve around 90% of recognition rate.
- **Dynamic Time Warping** has the advantage of requiring few training data. In fact one sign can be used as reference model. For example in the case of hand configuration classification in [Darrell 1993] one image is compared to several reference images using the correlation measurement. This kind of methods have also been used for aligning signs [Heloir 2006a] or for finding a sign in a SL sentence [Alon 2006]. The main inconvenient concerns the complexity of the model for each sign leading to high execution time.
- **HMM** aim to the reduction of the size of the model in a minimal number of states. The hypothesis of HMM considers that the sign is a sequence of gestures. States in HMM use the features extracted from videos corpus. Recognition rates depends on the quality and representativeness of the data and the vocabulary size. In addition they are generally signer dependent. For example in [Assan 1998] a recognition rate of 94% over 262 signs dropping to 73% for an unseen signer.

Here a brief description of some classification methods is presented though this is not exhaustive. In fact several variants of these approaches have also been used for SL

recognition, e.g. Parallel Hidden Markov Models (PaHMM) [Vogler 1999a] or Time Delay Neural Network [Yang 2002]. These statistical approaches can be used at a word level or a sub-unit level. The former considers for each whole sign a model and the latter decompose sign in sub-units at a phonemic level so that the sign is a combination of sub-units.

3.6.2.1 Recognition using word models

Recognition systems use the whole word models for training. Features are extracted from video sequences at a word model level in order to build a model database for classification purposes. Figure 3.34 shows the components of the classification process [Von Agris 2008]. Systems are trained using a set of know signs which are used to build statistical models through the extraction of representative features of the performance. Each sign leads to a word model in a database vocabulary. Later during the classification models in the database are compared to the unknown sign through the selected features to identify signs.

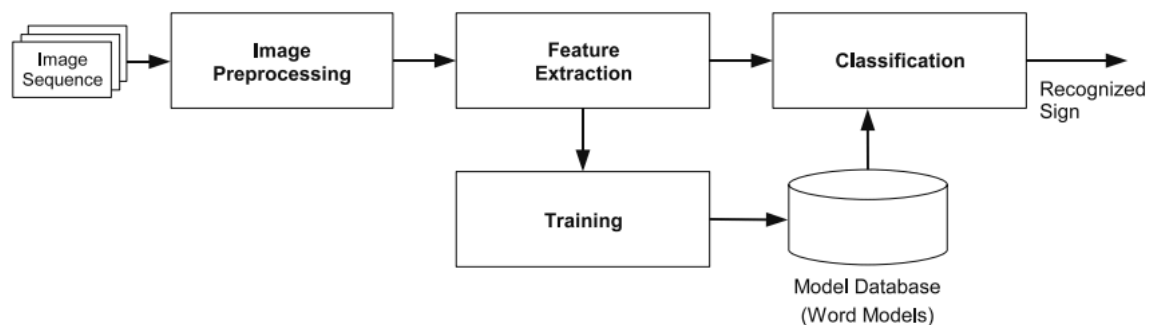


Figure 3.34: Word model recognition system components. Source [Von Agris 2008].

This kind of approaches have the disadvantage that the same sign in different context leads to different models in the vocabulary leading to high amounts of training data as well as lot of word models in the database though several models represent the same word. Complexity of the training step increases with the vocabulary size. Also adding unknown signs is very difficult. These problems are addressed using sub-unit models instead word models.

3.6.2.2 Recognition using subunit models

A different approach than the one previously described consists on breaking down words in sub-units which are considered to be the smallest unit in language. Unlike word models, signs are represented as the concatenation of several subunits. Using sign decomposition in sub-units leads to decrease the amount of training data since the number

of sub-units is smaller than the number of signs. Also new signs are easier to define since they can be represented as the combination of several subunits which are already in the system. Finally this allows to go from visual features to a meaningful semantic higher level. Some sub-unit approaches [Cooper 2007, Paulraj 2010] are based on linguistic representations of signs defined by linguists, see Section 3.4.

Figure 3.35 shows the components of sub-unit recognition system. The main different with respect to the word level concerns the description of signs and what is used in the training and classification stage. For example let's one sign *Sign 1* be described by the sub-unit sequence $\{SU_4 SU_7 SU_3\}$ and another sign *Sign M* by the sub-unit sequence $\{SU_2 SU_7 SU_5\}$. Notice that both signs have the same sub-unit SU_7 . This could for example correspond to two sign with the same kind of trajectory.

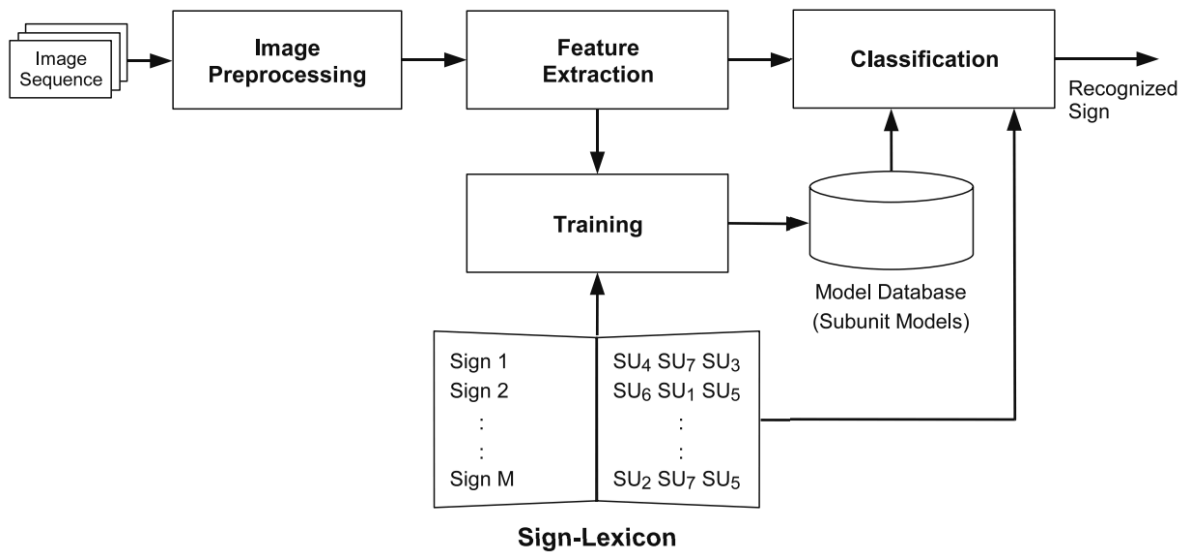


Figure 3.35: Sub-unit model recognition system components. Source [Von Agris 2008].

Many approaches in the literature concerning SL recognition using sub-unit models exist [Vogler 1997, Vogler 1999b, Vogler 1999a, Yeasin 2000, Bowden 2004, Kadir 2004, Fang 2004, Cooper 2007, Lichtenauer 2008a, Han 2009, Pitsikalis 2011]. These kind of approaches use first the detection of sub-unit features that could be automatically detected by other methods or defined by linguistic models for the recognition of signs.

Sub-units could represent any features extracted from video corpus and do not have to be necessarily motivated at a linguistic level. Some approaches intend to automatically segment motion using trajectory changing and acceleration to build a dataset of possible trajectories into sub-units [Kong 2008, Han 2009].

Using linguistic models give the advantage of covering large vocabularies and taking into account variation performances. The linguistic representation of signs mostly used is the sequential model introduced by Liddell and Johnson [Vogler 1997] based on hold and movement sequences, see Section 3.4.2 or the simultaneity model introduced by

Stokoe's [Vogler 1999a] argued the simultaneousness of the articulators, see Section 3.4.1.

Unlike word level approaches using sub-unit representations of signs make the introduction of unknown signs to the recognition systems much easier. In fact if the sub-units constituting the sign are already in the system only the sequence of subunits will be added to the database leading to high recognition rates using only few occurrences of a sign [Bowden 2004, Kadir 2004, Cooper 2007, Lichtenauer 2008a]. For example in [Cooper 2007] they achieve 74% over a random vocabulary of 164 sign using only 5 training examples.

Using sub-unit models based on linguistic representations of signs is very interesting because this allows to link features extracted from video corpus and corresponding to the parameters in the linguistic description. However to build the data base a training step is used. The ideal is to use an model database that would not involve the use of video training data since collecting SL video corpus requires a complex recording set-up, good illumination and various native signer to perform signs.

3.6.3 Discussion

SL recognition is achieved by extracting representative features from signs for a set of examples. These features are then used to train classification methods which can be done at word or sub-unit level. The former has the disadvantage of requiring one model for each sign differently performed regardless if different performances correspond to the same sign. Thus this is not suitable for an unconstrained vocabulary. The latter use the decomposition of sign in terms of sub-unit. Thus what is learned are the sub-units which can be common to several signs reducing then the vocabulary size. Although sub-units could be obtained using any sign decomposition, this is wiser to choose a linguistically motivated decomposition which is called phonemic representation.

Recognising signs requires information from a higher level which is, in the literature, introduced using learning step from a set of annotated examples. The quality of the recognition depends on the selected features and the representativeness of the training data. Each sign in the vocabulary has to be trained from several examples. This requires lot of data which is not straight forward collected, e.g. need of a recording setup and several native signers, see Section 3.2.2.

In short we will use the sub-unit decomposition but this also needs training data. Generally high amounts of annotated data are not available, particularly at the very first step for annotation process. Nevertheless we are aware that SL recognition cannot be achieved without any additional information than the one extracted from video corpus. Our very first idea was to directly use the linguistic representation of sign proposed by [Filhol 2009b] in addition to their database of ≈ 1600 sign already annotated. This was firstly motivated by avoiding training data and the advantages offered by Zebedee (see Section 3.4.2) however we have argued in Section 3.4.3 why this is not possible to straight forward use this sign description and the filters already implemented. In short

the same feature can be annotated in various ways.

Avoiding the use of several example sequences of the same sign performed several times in different contexts by various signers pushed us to the idea of using synthetic data to perform the training. This can be done using the generation of signs, methods in the literature are described below.

3.7 Sign Language Generation

Several approaches have been developed for SL generation [Phan 2009, Cox 2002, Wells 1999, Pezeshkpour 1999] through virtual signers also called signing avatar. These generation techniques can be either manually or automatically performed. Manual approaches need the intervention of human designers during the generation process which is time consuming, unreproducible and error prone. In addition, as any manual technique, the quality of the results depends on the designer experience. Manual approaches are not only used for SL generation but also in many other applications such as animated films and cartoons. Automatic methods use gesture models to pilot a virtual signer, this is faster than manual methods but the results remain less natural. The models used in this kind of approaches consider anatomical and linguistic aspects.

Manual and automatic SL generation is briefly described herein. A discussion between the performances and limitations of these approaches justifying our choices according to our needs is carried out at the end of this section.

3.7.1 Manual generation

The most common manual approaches for SL generation are rotoscoping [Fleischer 1917, Filhol 2007, Chen 2002] and motion capture [Heloir 2006b, Adamo-Villani 2008]. The former is a fully manual method patented in 1917 by the Max Fleischer and the later is an assisted technique using motion capture devices, see 3.2.

Rotoscoping is a technique mainly used for cartoons since it gives human and animal realistic dynamics to a character. The device propose by Fleischer is called "Rotoscope" which helps to produce realistic animation. First an actor performing the movements to be obtained for the character is recorded. Later, on the underside of a glass the film is played back while on the topside the animator copies the silhouette of the actor for each frame. The resulting drawings are added to the character clothing to obtain a very realistic animation. Although nowadays the rotoscope has been replaced by computers, the technique of manually creating a live-action element to merge with another background is called "Rotoscoping".

Focusing on SL generation this technique is used with advanced computer software such as Maya¹⁴ [Adamo-Villani 2004], 3Ds Max [Filhol 2007] or Motion Builder¹⁵ to give natural and realistic movements to a virtual signer. A corpus of the signer from several views is required to perform the rotoscoping for 3D animation, generally two cameras (front and lateral) are enough for the generation [Braffort 2010]. Then the animator, and specialists on computer graphics, synchronizes the corpus and the virtual signer. For this the different views of the corpus are used as background and the skeleton of the virtual signer is positioned so that the body gesture corresponds to the signer's gesture 3.36. This

14. <http://usa.autodesk.com/maya/>

15. <http://area.autodesk.com/motionbuilder2012>



Figure 3.36: Rotoscoping for SL generation. Source [Braffort 2010]

is not performed for each frame but only for key frames which correspond to significant changes in the gestures, e.g. movement direction changing. Transitions between key frames are directly interpolated by the software after some dynamic settings to remains as natural as possible.

Generating SL sentences might consider the movement epenthesis also coarticulation effect, see Section 2.3.4, [Chen 2002, Segouat 2010, Braffort 2011] to give a natural and realistic SL performance.

Although this technique gives very good quality results it is time consuming, expensive, e.g. 3Ds Max is about 4500 euros¹⁶, and unreproducible, it requires to adapt the gesture from the virtual signer to the source signer considering the morphological difference between them. In addition the quality of the generation depends on the animator's experience on computer graphics and Sign Language because of the need of selecting significant frames from a linguistic and an artistic point of view. Results are not adjustable to context, i.e. producing the same sign differently located leads either to the whole generation process or to decompose signs [Chen 2002] to compose signs previously unknown by the system.

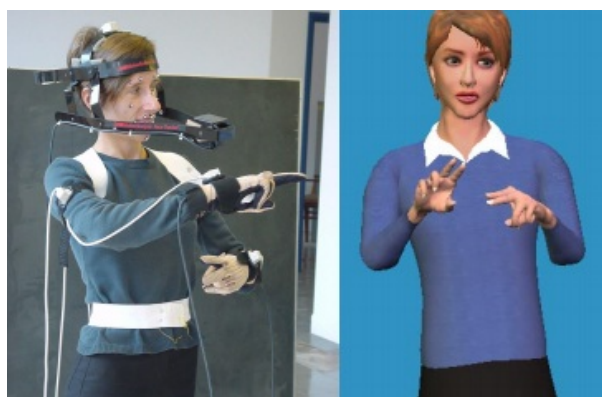


Figure 3.37: Generation from motion capture [Elliott 2000]

A different approach, a semi-automatic method, in which animator intervention is less needed corresponds to use motion capture devices to collect the generation information.

16. <http://www.cadline.fr>

This allows to collect high amounts of data in a very short time, see Section 3.2 for further information on Motion Capture Methods. Motion capture is used quite common in films and video games since it allows to build a 3D skeleton using sensors strategically placed on the body to automatically animate the virtual character through some animation software [Elliott 2000, Adamo-Villani 2008]. This technique is widely used for SL generation [Elliott 2000, Duarte 2010, Lombardo 2010, Lu 2009]. For this, a huge amount of generated sign is stocked and the synthesis of SL is then performed using the gloss. However this do not consider the context dependency of signs and the high usage of the signing space, see Section 2.3.2. Then being able to produce any sign in any context means either stocking all the variations of a sign -and even doing this the problem about accessing the sign using only the gloss raises- or using complex approaches to modify the position of hands in the frame [Gleicher 1998]. Moreover including new signs in the database represents lot of work in terms motion of capture set up and calibration (skeleton, marker placement, etc.) and the intervention of a native signer performing the signs.

In Section 3.2 are described several motion capture methods for SL corpus collection, the inconvenient of using such a methods have been discussed. In short it has been argued the high cost of motion capture devices, the cumbersome influencing the performance of signs and the set up and calibration complexity. In addition to this, in SL generation the error added by the motion capture devices leads to a generation result with poor quality according with the accuracy of the device. This noise has to be post-processed, manually or automatically, to improve the generation results which is time consuming. Although the data collection is carried out quite fast once all the capture environment has been adjusted, the set up complexity can last very long and must be performed by an expert.

Manual approaches require human intervention in different levels. The generation results is high quality, however they are time consuming, error prone and unreproducible. Good quality results require experienced people from a computer science and a SL knowledge. The animation result is highly biased to the signer educational and professional background, gender and morphology.

3.7.2 Automatic Generation

Automatic generation is, unlike rotoscoping or motion capture techniques, mostly used for SL generation [Karpouzis 2007, Fotinea 2008, Suszczańska 2002, Kipp 2011]. It considers linguistic and anatomical models corresponding to the information to generate and the input parameter which coupled pilot a virtual signer. The complementarity of both models make the generation of signs much more human feasible.

Linguistic models are generally formal models of signs which decompose signs such as the ones described in Section 3.4. Mostly works use parametrized descriptions of signs which allow to represent features as input parameter, e.g. hand location, configuration and motion, of the signing avatar to achieve SL synthesis such as HamNoSys [Hanke 2002, Kennaway 2003, Kennaway 2007, Marshall 2003] or Zebedee [Filhol 2009b, Delorme 2009].



Figure 3.38: Example of [WHEN?]. Source [Girod 1997]

These models describe the required information for generating a sign but without specifying the configuration of all the articulators of the body. Firstly because adding such a information does not add any extra information to better understand signs. Indeed for a sign to be understandable, few constraints defined by the formal model are needed regardless the position of other articulators, e.g. the sign [WHEN?] in LSF, Figure 3.38 the index finger must be in contact with the palm but the position of the elbows does not change the meaning of the sign. Secondly because the same sign can be performed slightly different depending on the signer background and morphology regardless the context which already adds more variability to a sign. Finally because over-representing signs leads to reduce the generation possibilities, thus the adaptability of the generation to different contexts.

Anatomical model adds the missing information, i.e. other articulators location, for the generation. This allows to take into account constraints or movements in gestures that are impossible to be preformed by a human although linguistically it remains correct. In addition to anatomical constraints to obtain a more human realistic generation, so far unnatural and robotic, some approaches intent to add comfort measurements to obtain more natural performances [Delorme 2011]. For example generating a sign where what matters is the contact between both hands, as the [WHEN?] example, comfort measurements give the best location of elbow among all the possibilities of performances respecting linguistic and anatomic constraints.

Automatic generation needs a formal model to describe signs. The quality of the generation results depends on the goodness of the linguistic and anatomic models and the generation algorithm, e.g. inverse kinematics [Wang 1991]. Unlike manual methods, automatic generation is fast and the adaptability to context and variability depends on the formal model chosen, for example Zebedee is highly adaptable contrary to HamNoSys (Sec. 3.4). In addition it does not require any expensive devices and adding an unknown sign to the database means adding the description in the formal model chosen. However the generation is quite unnatural and not realistic mainly during transitions. Indeed while in the sign representation a transition is described as straight or circle, the generation might produce a perfect straight line or a perfect circle which is impossible to perform

by humans in a natural way because of several anatomic constraints.

3.7.3 Discussion

Manual and automatic approaches have some advantages and inconvenient. Manual approaches have the advantage of a high quality of the results in terms of realism and naturalism of the production. However they need the intervention of human animators which make the generation process time consuming and unreproducible. Moreover a huge database is required to carry out the generation of SL sentences which needs interpolation methods and remains low adjustable to the variations of signs according to the context, e.g. placing the sign in a different position in the signing space. Thus this kind of systems are suitable for applications with a constrained vocabulary. For example the system ATLAS [Lombardo 2011] or Octopus [Braffort 2011] developed for forecasting announcement and train information (SNCF) respectively. Automatic generation address the problem of sign variability using the adapted sign representation, but gives a more unnatural result which is not important in this work because this is considered as an intermediate step in SL gloss annotation.

The optimal generation approach depends on the needs and the application. The application in this work is about annotating the gloss from a performance in a video. For this it is needed to query a database from the visual features extracted from the video. It has been already argued in Section 3.4.3 why this is not possible to use a formal model to straight forward query it. Summarizing, so far parametric approaches have been developed for SL synthesis, then the sign representation is described in terms of constraints from parts of the body that cannot be detected from a mono-camera, e.g. index direction, or high variability on the description of a sign. Then an extended version, developed explicitly for SL recognition, of a formal model is needed, see Section 4.6.1.

In order to automatically extend the formal model chosen in this work, Zebedee, see Section 3.4, and the automatic generation system using this formal model of signs is required, GeneALS [Delorme 2011]. The specifications required in our system, see Section 1.5, are to extract visual features from the production; features that are compatible to what can be extracted from videos, e.g. the number of moving hands, movement direction, etc, see Section 4.6.1, using image processing techniques. For this a general description of a sign is used to generate several times the same sign with different parameter, e.g. for the sign [BUILDING] in LSF, Figure 3.22, several generations for random parameters [loc][size][height] are carried out, to extract visual information that is constant in all the generations but that is not explicitly described in the linguistic model.

3.8 Conclusion

In this chapter we have presented the state-of-the-art for SL corpora and its annotation. Data acquisition consists on the collection of high amounts of data for training or evaluation. Motion capture based devices (Sec. 3.2) have been discussed. We have argued the inconvenient of using motion capture systems concerning their high cost, their invasiveness and their cumbersomeness. In order to collect representative data which remains natural we prefer using passive sensors: video camera. In addition video corpus are used by linguists for studying the language using, generally, only one camera. For these reasons we have decided to focus on studying methods from a mono-camera. The main difficulty about using this kind of data concerns its annotation. Corpus annotation (Sec. 3.3) can be addressed from two points of view: linguistic and computer science. The former concerns linguistic information such as lexical, semantic, phonemic, etc. The latter concerns low and high level feature such as motion, location, velocity, etc.

In order to assist the annotation, computer scientists have proposed annotation software for manually manipulating and annotating data. Other approaches propose automatic processing for extracting features, but only concerning the phonemic level. In this PhD thesis we intent to go further focusing on a lexical level in addition to the phonemic one. Phonemic features are described by linguists through phonemic representations of signs (Sec. 3.4). Using sign representations we are able to go from a phonemic level (meaningless units) to a lexical level (meaningful units). The ZeBeDee lexical description of signs is very attractive because of its advantages on the variability and context-dependency of signs. However it has been designed for SL synthesis and it is not straight forward usable for SL recognition. These linguistic descriptions point out features corresponding to that level of annotation and that have to be extracted from video. We investigate existing methods for the extraction of features (Sec. 3.5) using image processing techniques for the automatic annotation at the phonemic level. Hand and head location and motion, hands shape and temporal segmentation methods are detailed.

Since annotation has not been specifically addressed at a lexical level, we investigate existing SL recognition methods. Approaches focusing on SL recognition (Sec. 3.6) use learning data to train classification methods. However their results depend on the representativeness of the data. Training data is unsuitable for SL annotation since the annotation of the training data is also required. We prefer to collect training data differently by generating it. Thus we might learn sign performances from synthetic data. SL generation can be performed manually or automatically (Sec. 3.7). Automatic generation uses linguistic models for piloting the generation. Thus the generated data can be used.

In short this study about the state-of-the-art pointed out the advantage and the inconvenient of exiting methods and their adaptability to the context of of automatic SL processing. This push us to propose a novel approach using synthetic data generated from linguistics models. This allows us to have as much data as we wish in different performance thanks to the modularity of the linguistic representation used in the generation. This data is used for extracting motion features for gloss recognition. The proposed approach with our contributions for SL annotation are detailed in the following chapter.

Sign language automatic annotation by SL generation

Résumé: Annotation assistée de la LS à l'aide de la génération des signes

Dans ce chapitre nous présentons les contribution issues de nos travaux de recherche. Nous avons adressé plusieurs problématiques concernant le traitement automatique de la LS, particulièrement en ce qui concerne l'annotation de la LS. Dans ce cas nous avons besoin de méthodes qui ne contraignent pas le contexte ni la taille du vocabulaire. Ici nous proposons un système d'annotation proposant une liste de signes potentiels à l'annotateur qui n'utilise pas de données d'annotation. Notre système consiste en l'extraction et l'analyse des caractéristiques de bas niveau dans une séquence vidéo.

Afin d'extraire des caractéristiques représentatives de la réalisation d'un signe à partir d'une séquence vidéo. Nous avons besoin de suivre des caractéristiques des mains et de la tête dans la vidéo. La couleur de la peau est très populaire pour faire le suivi des composantes corporelles ou pour segmenter la main dans une séquence vidéo. La dynamique des mains et des nombreuses occultations entre les mains et la tête rendent particulièrement difficile le suivi des composantes corporelles. 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 et nous introduisons une fonction de pénalisation permettant de gérer les occultations. Les résultats ont montré une robustesse aux occultations et à la dynamique du mouvement supérieure à celle d'autres méthodes de suivi dans la littérature. Les résultats de suivi nous permettent d'extraire des caractéristiques de mouvement comme la vitesse et l'accélération qui peuvent être par la suite exploitées pour la segmentation temporelle.

En plus des caractéristiques de mouvement, la forme de la main nous permet d'avoir des informations complémentaires. Pour ceci la segmentation de la main même pendant occultation est nécessaire ce qui est une tâche difficile. 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, leur 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.

Le suivi et la segmentation de la main permettent d'extraire des caractéristiques représentatives des signes qui sont utilisées pour la segmentation de signes dans un discours en LS. La segmentation temporelle correspond à la détection du début et de la fin d'un signe. 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. 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. Pour cette raison nous préférons utiliser la forme 2D afin d'extraire de caractéristique géométriques. La forme de la main est systématiquement comparée avec celle des événements adjacents. Le but de la segmentation semi-automatique est de proposer des limites à l'annotateur à l'aide de mesures objectives.

Afin d'éviter l'utilisation de données d'apprentissage nous proposons l'utilisation d'un représentation informatiques des signes. Nous avons choisi l'utilisation de Zebedee qui permet de représenter les signes et qui pilote un système de génération. L'utilisation de ce système n'est pas directement adapté à la reconnaissance des signes. En effet l'annotation d'un signe peut être réalisée de plusieurs façon étant le résultat de la génération similaire. Par exemple l'annotation de la position de la main devant le visage peut être annoté comme en utilisant le front ou le nez.

Nous proposons d'ajouter des caractéristiques visuelles qui peuvent être extraites directement de la vidéo et qui correspondent à des caractéristiques annotées dans la représentation informatique. Nous proposons d'utiliser des données synthétiques afin d'extraire des caractéristiques visuelles. Pour ça nous utilisons ZeBeDee pour générer plusieurs réalisations d'un signes.

4.1 Introduction

In this chapter we present the approaches developed in this PhD thesis. We address several problems encountered in the literature for the automatic processing of SL. These problems are mainly due to the application: SL annotation. Indeed we need methods that are general and well adapted for SL processing, i.e. not constrained to a context or a small vocabulary. The automatic gloss recognition system consists on a

- **skin model** for classifying pixels as skin class and non skin class (Sec. 4.2). Indeed much of our work is based on the skin colour of objects since this is a representative characteristic of hands and face. Other methods are then required for identifying face and hands for performing tracking and segmentation.
- **body tracking** method specifically designed for SL corpus (Sec. 4.3). Tracking results correspond to the position of face and hands for each frame. From this we are able to extract motion features for characterising signs. The main challenge of body parts tracking consists on robustly handle occlusions because of the colour similarity between hands and face.
- **hand segmentation** algorithm for extracting hand region (Sec. 4.4). Hand segmentation when the hand is in front of the face requires an adapted method using a template before occlusion. It considers in addition to skin colour, edges for finding hand limits. An occlusion detection method is described for selecting the adapted segmentation algorithm and the optimal template before occlusion.
- **temporal segmentation** approach for detecting sign borders in continuous SL as well as the temporal structure of signs (Sec. 4.5). Indeed we have argued that using the phonemic sign description, ZeBeDee, for the phonemic annotation needs the temporal structure of signs, i.e. the sequence of key postures and transitions. This approach uses the results from our tracking algorithm and our hand segmentation methods for extracting features and characterising limits.
- **gloss recognition** method using a SL generation approach (Sec. 4.6). This novel method uses synthetic data from SL generation in order to extract visual characteristics that can be easily extracted from video. The generation approach is piloted by a linguistic model. Using the linguistic representation of signs is the way for linking low level features at a lexical level. As a result we propose a list of glosses to the annotator. In this way we assist the annotation at a lexical level without using any training data.

The proposed methods in this work composing a gloss recognition system has been designed for SL annotation and has been evaluated for pointing out the performances of our approach. These methods are detailed below.

4.2 Skin Model

Skin models are used to perform skin segmentation of body members where the most representative feature is the colour, such as hands and head. Generic models need a training step and make the model dependent to the skin-colour samples, illumination and environment conditions in the training set. Specific models have a better accuracy but need a robust initialisation.

We propose to use a specific model which is built using the skin pixels from the subject in the video. Even though in our work, annotation purposes, human intervention is allowed, we will avoid any intervention in this case to make our algorithm more easy-to-use. First of all it is needed to chose the colour space to become illumination and environment conditions independent. The initialisation of the training data set corresponds to some skin pixels sampled from the video to process. This is achieved using other explicitly defined skin models to select a few pixels and use them to obtain the specific model. Finally skin regions segmentation is performed using a simple decision rule.

Since the model is specific to each signer and recording condition from the video to process, the sampling and building step have to be preformed at the beginning of each video. These steps are very fast $\approx Xms$ but have to be taken into account in the processing time. The segmentation is more robust since the skin model is specific to the subject and to the recording environment.

4.2.1 Colour space

The main difficulty achieving high segmentation rates in colour based segmentation algorithms consist on the selection of the colour space. In RGB (Red, Green, Blue) colour space all the three channels are highly correlated mixing chrominance and luminance data. As a result segmentation is strongly dependent on the illumination conditions.

Other colour spaces have the advantage of decoupling luminance from chrominance. Here we have decided to use the YC_bC_r colour space. The transformation function from RGB to YC_bC_r is the weighted sum of the *Red*, *Green* and *Blue* channels, equation 4.1.

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ C_r &= R - Y = 0.701R - 0.587G - 0.114B \\ C_b &= B - Y = 0.886B - 0.299R - 0.587G \end{aligned} \tag{4.1}$$

The simplicity of the transformation and the decoupling of luminance channel make this colour space very attractive for our skin colour modelling. Also the perception of different skin colours by the human eye is strongly dependent on the luminance component which can be rejected using this colour space.

4.2.2 Skin pixels learning data set

The collection of skin pixels for building our specific skin model can be performed either manually or automatically using a generic skin model. Manually the annotator selects a skin region from the first frame in the video, however it is preferable to save annotators time to be spent in more complex tasks e.g. correcting tracking results. Here we propose to automatically select skin pixels for the learning data set and build the specific model avoiding annotators intervention.

Several approaches can be used to determine the first skin region for the learning step. However it is wiser to avoid any skin model requiring any training since that will represent additional work without having better results for our model. An explicitly defined skin model based on simple detection rules would fulfil our needs. In fact few skin pixels are required to build our model as long as the detection is robust and most of the detected pixels correspond to the skin class.

Even though it has been argued before that a $YCbCr$ colour space is preferable than the RGB colour space because of the decoupling of luminance and chrominance components, we can use an explicitly define skin model in the RGB colour space as long as the luminance and chrominance decoupling is used latter in our model.

The explicitly defined skin model used in our work has been introduced by Kovac *et al* [Kovac 2003]. The boundaries skin cluster in RGB colour space have been defined through a number of simple rules,

$$\begin{aligned}
 & (R, G, B) \text{ is classified as skin if:} \\
 & R < 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and} \\
 & \max\{RGB\} - \min\{RGB\} > 15 \\
 & |R - G| > 15 \text{ and } R > G \text{ and } R > B,
 \end{aligned} \tag{4.2}$$

where (R, G, B) correspond to the components value for each pixel. Pixels value that verify all these conditions are classified in the skin class.

Since the luminance and the chrominance components are not decoupled, classification is dependent on the illumination conditions and some shadowed regions might not be well classified. For example using Eq. 4.2 for each pixel of Fig. 4.1(left), the result, Fig. 4.1(center), shows that in regions where a shadow appeared the detection became inaccurate, e.g. neck and fingers. However the found skin region is enough for our building step. Also the advantage of this approach is that there is no need of any learning stage, it is easily implemented and it gives a rough skin sample region. The sample pixels for the model are those belonging to the greatest area connected component, head or hand(s) depending on the frame, as shown in Fig. 4.1(right).



Figure 4.1: Skin segmentation result using an explicitly defined skin model in RGB colour space

4.2.3 Specific model building

The specific model is built with the sample skin pixels obtained from the explicitly defined model in the RGB colour space. The sample pixels are transformed into the YC_bC_r colour space using Eq. 4.1. Since the Y component reflects the luminance, it is rejected to address shadow problems.

Consider C_b and C_r two random variables define by:

$$\begin{aligned} C_b &= aB + bY \\ C_r &= cR + dY \end{aligned} \quad (4.3)$$

where a, b, c, d are some constants. Since each of the chrominance components is a normal distribution, the joint probability density function PDF , defined in Eq. 4.4, takes the form of a bivariate normal distribution [Bertsekas 2002].

$$PDF(C_b, C_r) = \frac{1}{2\pi\sigma_{C_b}\sigma_{C_r}} \exp \left[-\frac{z}{2(1-\rho^2)} \right] \quad (4.4)$$

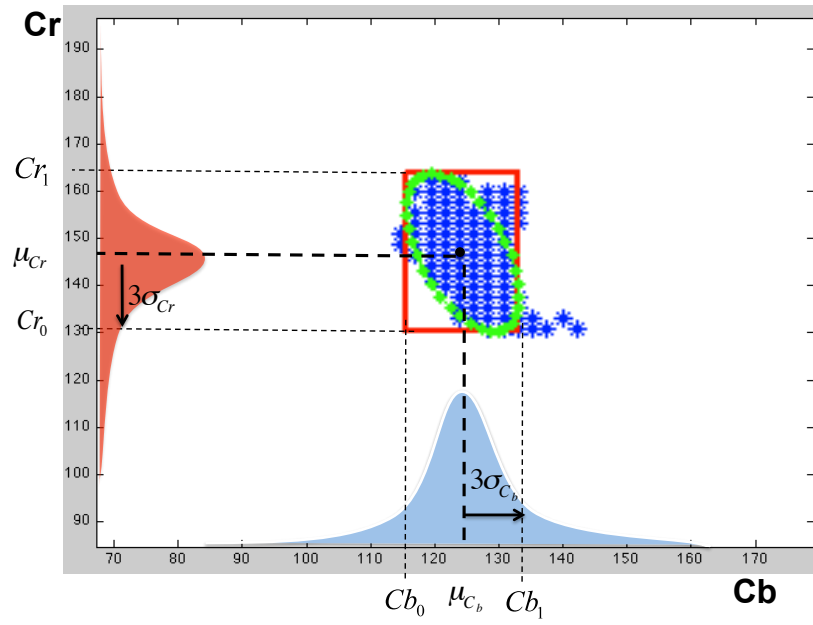
where

$$z = \frac{(C_b - \mu_{C_b})^2}{\sigma_{C_b}^2} - \frac{2\rho(C_b - \mu_{C_b})(C_r - \mu_{C_r})}{\sigma_{C_b}\sigma_{C_r}} + \frac{(C_r - \mu_{C_r})^2}{\sigma_{C_r}^2} \quad (4.5)$$

and

$$\rho = \text{cor}(C_b, C_r) = \frac{E[C_b C_r]}{\sigma_{C_b}\sigma_{C_r}} \quad (4.6)$$

The mean vector μ_S and the covariance matrix Σ_S , Eq. 4.7, of the distribution are estimated from the skin training pixels. Figure 4.2 shows the bivariate normal distribution

Figure 4.2: Bivariate normal distribution $C_b C_r$

in the $C_b C_r$ plane. This distribution is, generally, used to determine the distance of the (C_b, C_r) values of a testing pixels to the mean, Mahalanobis distance. Afterwards a threshold is used to keep pixels close to the mean [Habibi 2004]. Mahalanobis distance computation is time consuming and the threshold value is difficult to determine. Instead we propose to automatically define adapted threshold values to the $C_b C_r$ components. Thus the skin classification is a binary decision and the cut-off values are automatically computed for each signer.

$$\mu = \begin{pmatrix} \mu_{\sigma_{C_b}} \\ \mu_{\sigma_{C_r}} \end{pmatrix}, \quad \Sigma_S = \begin{pmatrix} \sigma_{C_b}^2 & \rho \sigma_{C_b} \sigma_{C_r} \\ \rho \sigma_{C_b} \sigma_{C_r} & \sigma_{C_r}^2 \end{pmatrix} \quad (4.7)$$

Threshold computation is performed using both normal distributions. Since $C_b C_r$ are correlated when the bivariate distribution is plotted in the (C_b, C_r) -plane the distribution appears to be squeezed. Considering both distributions as uncorrelated, the ellipse axes are aligned to $(C_b C_r)$ -axis, the main axes of the bivariate distribution will correspond to the borders of the red rectangle in Fig. 4.2. This allows to simplify the model building. Computing the thresholds using the normal distribution is faster and the error from the rectangle corners is neglected in our approach. For a single normal distribution the 99% of the distribution is between $[-3\sigma, 3\sigma]$. The thresholds considered in our model are expressed in Eq. 4.8

$$\begin{aligned} C_{b_0} &= \mu_{C_b} - 3\sigma_{C_b}, & C_{b_1} &= \mu_{C_b} + 3\sigma_{C_b} \\ C_{r_0} &= \mu_{C_r} - 3\sigma_{C_r}, & C_{r_1} &= \mu_{C_r} + 3\sigma_{C_r} \end{aligned} \quad (4.8)$$

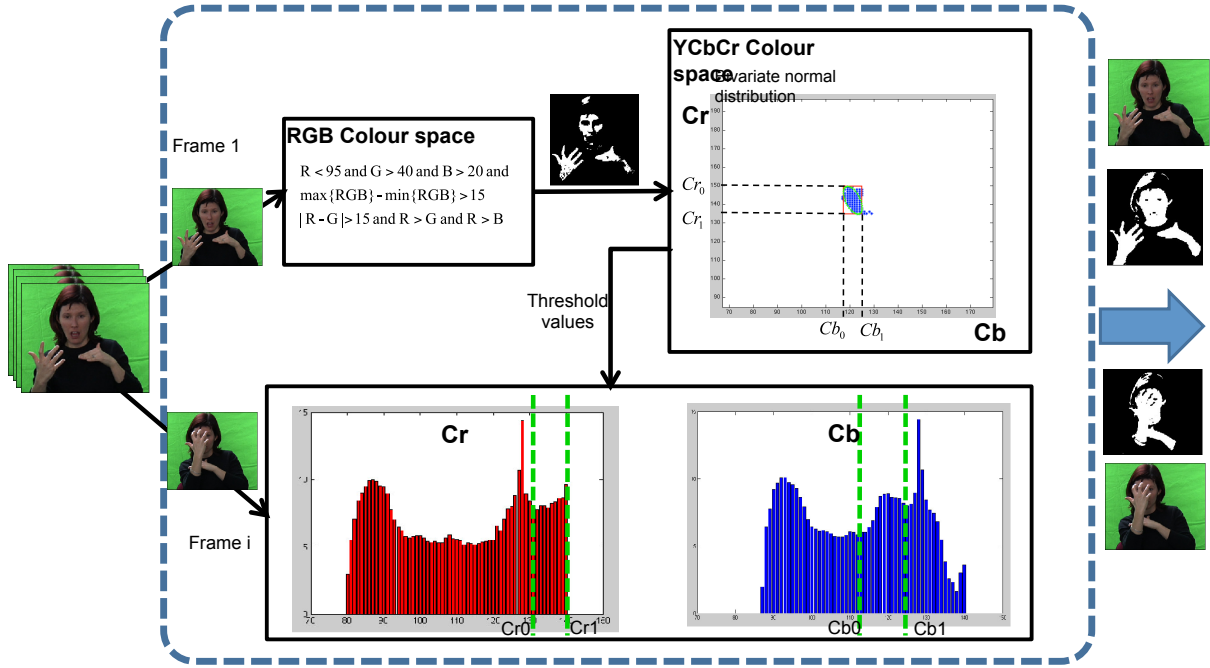


Figure 4.3: Skin segmentation algorithm

4.2.4 Skin segmentation algorithm

The segmentation algorithm, Algo. 4.1, is illustrated in Fig. 4.3. From the video to process, the first frame is used to collect the skin pixels in the RGB colour space. The learning skin pixel data set is used to build the bivariate normal distribution and to compute threshold values. For any other frame in the video, a transformation to the YC_bC_r from the RGB colour space is performed. The luminance component is rejected and only C_b and C_r are used. The pixel is classified as skin as expressed in Eq. 4.9.

$$(Y, C_b, C_r) \text{ is classified as skin if:} \\ C_{b0} < C_b < C_{b1} \text{ and } C_{r0} < C_r < C_{r1} \quad (4.9)$$

Figure 4.4(left) shows the result from the explicitly defined model in the RGB colour space. Notice that because the luminance component, neck and fingers are not well detected. Figure 4.4(right) shows the obtained result after thresholding the chrominance components. This is the skin probability map S_k used for tracking purposes. We notice that the skin pixels are better detected in shadowed regions. However, because of the simplicity of the decision rule, some pixels belonging to the hair were wrongly assigned to the skin class. We do not expect this to pose a significant problem since most of the skin pixels have been detected.

Algorithm 4.1 Skin segmentation algorithm**Initialisation**

1. Get skin region from the first frame in RGB using the equation 4.2
2. Build the bivariate distribution rejecting the luminance component Y .
3. Determine the threshold values using equation 4.8

Segmentation

1. Transform frame to segment from RGB to YC_bC_r using Eq. 4.1
2. Classify pixels using Eq. 4.9



Figure 4.4: Skin probability map S_k obtained in RGB (left) and YC_bC_r (right)

4.2.5 Conclusion

Skin colour is the most representative feature of hands and face. This feature is widely used in this work for hands and face tracking and hand segmentation. Then a skin segmentation approach, independent from the signer and robust to illumination changes, is required. For these reasons we have presented a specific model built from the skin pixels belonging to the signer in the video to process. In addition for robustness to illumination changes, it is important to choose a colour space where luminance component can be rejected. Although we have chosen YC_bC_r colour space any other space where the luminance is decoupled can be used.

In this section we have presented the skin model used for skin segmentation. We have proposed a specific model built from the skin pixels of the subject in the video. For this we have detected the face in the first frame, and skin pixels in the face are segmented using an explicit skin model in RGB colour space. The detected pixels are used for training the skin model in YC_bC_r colour space. It consists on a bivariate Gaussian from which we determine the threshold values for segmenting following frames. In this way the specific model is adapted to each signer. In addition this skin segmentation approach is fast in terms of computation time since the segmentation rules are very simple.

The skin map obtained from this segmentation is used in the following section for tracking hands and head. The problem faced concerns the dissociation of pixels belonging to hands and to the face so that the tracking is robust to occlusions, i.e. when hand is in front of the face.

4.3 Body tracking from a mono camera

In this research gestures are characterised using motion features in order to extract signs. For this it is necessary to know the position of hands and head at each frame of the processing video.

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. For example in Fig. 4.5 some frames for a short sequence are shown. This sequence is extracted from a natural performance of SL by a native signer. Notice that from one frame to the following one, right hand has significantly moved. The time between frames corresponds to approximatively $40ms$, according to the recording speed rate, generally 25 frames per second (fps). In our example right hand distance illustrated between the first and the last frame is about $50cm$. Thus right hand velocity is close to $18km/h$ which is more that three times the human average walking speed.

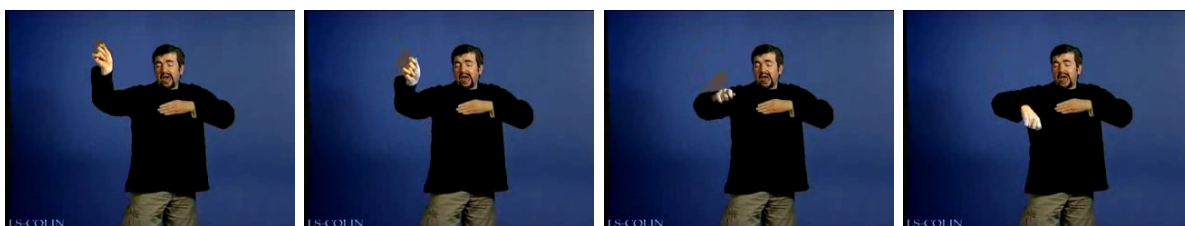


Figure 4.5: Hands dynamics example

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. Figure 4.6 shows an example of a sequence where the signer spells a word while the hand is in front of the face. In this case the hand is not only occluding the face but it also changes the configuration very fast. We notice that the similarity of colour makes hand shape hard to distinguish from the face background.

Since the most object representative feature is the *skin colour*, several approaches in the literature consider only colour to track head and hands. The main problem remains in how distinguish head from hands and vice-versa when the only feature considered is colour. Errors are easily introduced and tracking becomes very inaccurate. For example if head and hands are modelled in the same way, as a cloud of points, when the hand is in front of the face the same skin region will be used to determine hand and head position. Without further information the algorithm will not be able to accurately determine hand

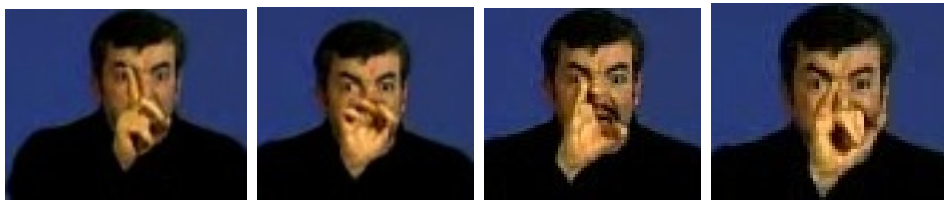


Figure 4.6: Hand over face occlusion and hand shape changing example

and head position. Figure 4.7 shows the plot of the distance between the head position in two consecutive frames for the approach in [Gianni 2009] which only uses skin colour features and models all objects, head and hands, in the same way.

Two main problems are pointed out :

1. The instability of the head position when the face does not move.
2. The head position displacement when an occlusion occurs without any head movement. Neither head or hand position are accurately determined.

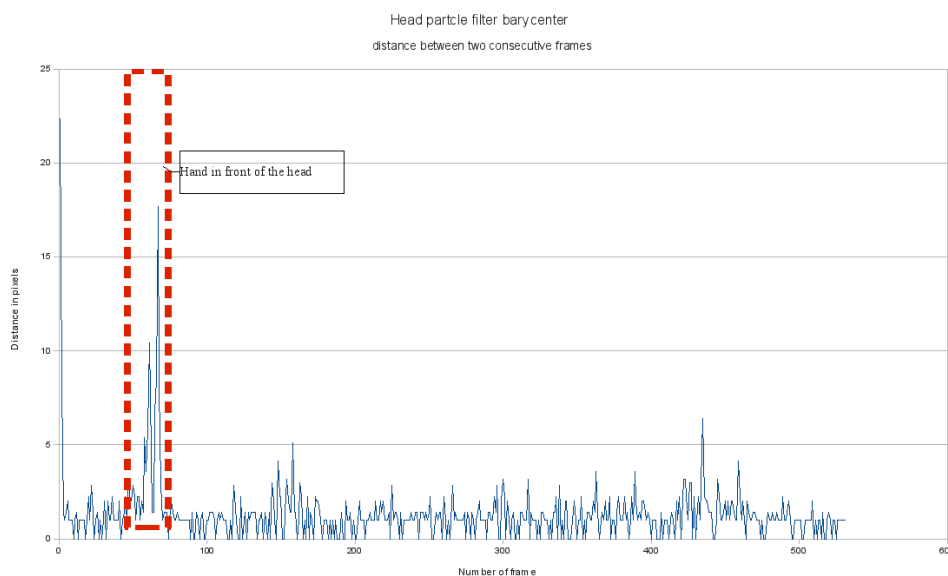


Figure 4.7: Head distance between two frames

In order to address these problems, we have decided to use a particle filter based approach. Unlike other methods in the literature (Sec. 3.5.1) we take into account that hands and head dynamics are very different as well as their shape variability. We propose to use a particle filter approach adapted to each object. A general particle filter based approach is used for head which normally moves very few and an annealed particle filter based algorithm is used for hands since it is adapted to hands movement. Also a different object and observation model are needed for hands and head. For hands it is wiser to chose a model where the shape remains free because of its huge shape variability e.g. cloud of points. However for head, we can assume that it is a rectangular grouping of



Figure 4.8: Head and hands position without penalisation

pixels since it is only slightly modified when the face expression changes. Although we consider head shape unchanged head texture does change according to the face expression. The observation model considered is mainly based on the skin colour of object but also considers shape thanks to the chosen model.

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 the filter associated to the object. Figure 4.8 shows the results obtained without any further processing. In fact hands influence head filter and the expectation is displaced to hands position. This problem is addressed using a penalisation function based on the exclusion principle [MacCormick 2000b] detailed in section 4.3.3. Filter observation of one target is penalised using the particles of other objects. For head coefficients are computed using the luminance difference before and after occlusion.

4.3.1 Particle filter

Visual tracking intends to estimates the state of the system that changes over time by using a sequence of noisy measurements. Bayes filter computes the posterior probability density function $p(\mathbf{x}_t | \mathbf{z}_{1:t})$ of the current state \mathbf{x}_t conditioned on all observations $\mathbf{z}_{1:t} = \mathbf{z}_1 \dots \mathbf{z}_t$ with \mathbf{z}_t the observation vector obtained at time t . For a first-order Markov process, i.e. the state \mathbf{x}_t depends only on \mathbf{x}_{t-1} , the probability density function $p(\mathbf{x}_t | \mathbf{z}_{1:t})$ can be obtained in two stages: prediction and update. It is derived as

$$p(\mathbf{x}_t | \mathbf{z}_{1:t}) = k \cdot p(\mathbf{z}_t | \mathbf{x}_t) \cdot p(\mathbf{x}_t | \mathbf{z}_{1:t-1}) \quad (4.10)$$

$$p(\mathbf{x}_t | \mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t | \mathbf{x}_{t-1}) \cdot p(\mathbf{x}_{t-1} | \mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1}, \quad (4.11)$$

where k corresponds to a normalization term independent of \mathbf{x}_t . Eq. 4.10 represents the update stage where the posterior probability density is computed using the observation likelihood $p(\mathbf{z}_t | \mathbf{x}_t)$ and the temporal prio distribution, $p(\mathbf{x}_t | \mathbf{z}_{1:t-1})$, over \mathbf{x}_t given past observations. Eq. 4.11 corresponds to the prediction stage where the prior distribution

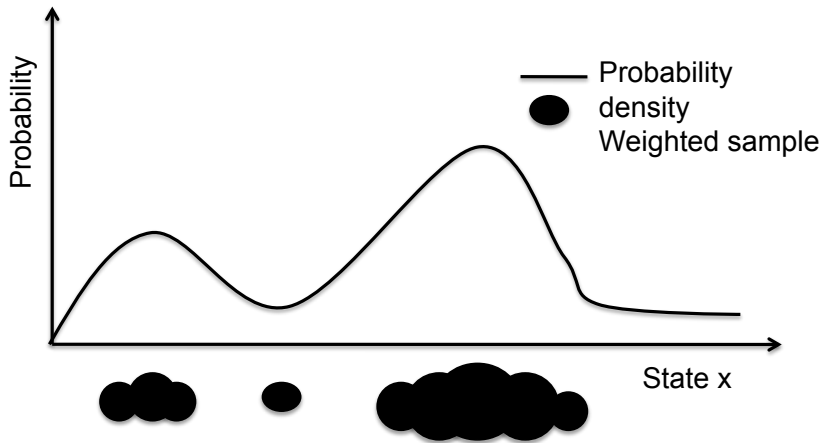


Figure 4.9: Probability density with associated weights [Isard 1998]

for $t + 1$ is estimated by the convolution of the posterior distribution $p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1})$ and the transition probability distribution $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, i.e. the dynamic model of the system.

Particle filter (PF) [Isard 1998] is a method based on a dynamic model of the system that estimates the posterior probability density function for non-Linear or non-Gaussian problems. It presents a good solution framework for tracking stochastic movements since it sequentially estimates, using random sampling to approximate the optimal solution, the states \mathbf{x}_t of the system by implementing a recursive Bayesian filter by Monte Carlo simulations.

The posterior probability density $p(\mathbf{x}_t/\mathbf{z}_t)$ of the current state \mathbf{x}_t is approximated by a weighted sample set, $\{\mathbf{s}_t^n, \pi_t^n\}_{n=1}^N$. PF maintain multiple hypothesis, i.e. each particle is a hypothetical state of the object, weighted by a discrete sampling probability $\pi_t^n \propto p(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{s}_t^n)$. Particle weights correspond to the observation generated by the hypothetical state and reflects the image feature relevance associated to each particle. The state \mathbf{x}_t is finally estimated using the particle set and the associated weights. Figure 4.9 illustrates the probability density function $p(\mathbf{x}_t/\mathbf{z}_t)$ with the associated weights π_t^n . Each particle corresponds to a state, the coordinates of the centre blobs in the example. The associated weight corresponds to the probability density and is illustrated by the size of the blobs.

The hypothetical states correspond to the object characteristics that can vary from time t to $t + 1$. For example the position of the object $\{x, y\}$, the orientation θ , the size s , the shape or several parameter defining the object changes. The state components is defined by the chosen object model, e.g. a point (pixel) or any other shaped model.

Particles weight is defined by the observation model of the system. Particle filter tracking algorithms usually use colour features and contours [Gianni 2009, Micilotta 2004, Lefebvre-Albaret 2010], but any feature characterising the object can be used to determine particles weight.

Algorithm 4.2 Resampling algorithm

-
1. **Compute** cumulated sum of weights $S_1 = \pi_t^1$ and $\{S_i = S_{i-1} + \pi_t^i\}_{i=2}^N$;
 2. **Generate** an uniform and random number $u^j \in U[0, t]$ with $t = \frac{S_N}{N}$
 3. **Find** i so that $i < N$ and $u^j > S_i$ for $j = 1 \dots N$
 4. **Define** the new particle state $\{s_t'^j\} = \{s_t^i\}$
-

4.3.1.1 Simple particle filter

The basic particle filter algorithm consist on three steps: resampling, propagation and weighting. Re-sampling is needed to avoid degeneration of the algorithm, particles are selected according to the associated weights. The performance of the tracking algorithm is attached to the resampling method [Kitagawa 1996]. It is important that the new states are representatives of the probability density function. We use the stratified resampling algorithm proposed by [Kitagawa 1996] and described in algorithm 4.2.

Particles are propagated with the dynamic model of the system i.e. first order autoregressive process model, $\mathbf{x}_t = \mathbf{x}_{t-1} + \eta$, where η is a zero-mean Gaussian random variable and \mathbf{x}_t is the state of the model time t . After propagation particles are weighted according to features in the observation model, see section 4.3.2. Particle filter algorithm is an iterative process of the steps described above, see algorithm 4.3.

The expectation is computes using the weighted particle set, as expressed in Eq. 4.12.

$$E[\mathbf{x}_t] = \sum_{n=1}^N \pi_t^n \mathbf{s}_t^n. \quad (4.12)$$

The basic particle filter approach is used to track head. Hands tracking particle filter consist of an annealed process to improve results.

Algorithm 4.3 Simple particle filter algorithm

-
1. **Resample** N particles from the set $\{\mathbf{s}_{t-1}^n, \pi_{t-1}^n\}_{n=1}^N$ to $\{s_t'^n, \frac{1}{N}\}_{n=1}^N$. As described in algorithm 4.2
 2. **Propagate** each particle using the dynamic model $s_t^n \sim p(\mathbf{x}_t | \mathbf{x}_{t-1} = \mathbf{s}_{t-1}^n)$ to obtain $\{\mathbf{s}_t^n, \frac{1}{N}\}_{n=1}^N$.
 3. **Weight** particles with the image feature \mathbf{z}_t as $\pi_t^n \propto p(\mathbf{z}_t | \mathbf{x}_t = \mathbf{s}_t^n)$ and normalize so that $\sum_{n=1}^N \pi_t^n = 1$.
 4. **Estimate** the tracking result of the object at time t by $E[\mathbf{x}_t] = \sum_{n=1}^N \pi_t^n \mathbf{s}_t^n$.
-

4.3.1.2 Annealed particle filter

Iterating the particle filter algorithm leads to a better representation of the posterior probability, however particle could get stuck in a local maxima. Depending on the weighting function at each iteration, the local maxima may be over-represented. In [Gall 2007] has been proposed a generic formulation of the annealed effect introduced in [Deutscher 2000] which applies a weighting function to the smoothly sample set and allow particle to converge to the global maxima.

Figure 4.10 illustrates the comparison from an iterating process without and with annealing effect. Notice that particles with annealing effect are able to scape from the local maxima Fig 4.10.

The **weight** and **propagate** steps in algorithm 4.3 are modified by the following iterated steps:

1. **Weight** the new particles sample set with the image feature \mathbf{z}_t as $\pi_{t,m}^n \propto p(\mathbf{z}_t | \mathbf{x}_{t,m} = \mathbf{s}_{t,m}^n)^{\beta_m}$ for $\beta_0 > \beta_1 > \dots > \beta_M$. Normalize so that $\sum_{n=1}^N \pi_{t,m}^n = 1$.
2. **Resample** N particles from the set using the dynamic model $\mathbf{x}_{t,m}^n \sim \mathbf{x}'_{t,m} = \mathbf{s}_{t,m}^n$ to give $\{\mathbf{s}_{t,m}^n, \pi_{t,m}^n\}$ and $\mathbf{x}_{t,m-1}^n \sim \mathbf{s}_{t,m-1}^n \sim \mathbf{s}_{t,m}^n + \mathbf{B}_m$.

Repeat steps (1) and (2) M iteration times from $m = M$ to 1 where M corresponds to the number of annealed layers. \mathbf{B}_m is a multi-variate Gaussian random variable with mean 0 and a vector variance $\mathbf{P}_m = \{\sigma_M \sigma_{M-1} \dots \sigma_m\}$.

For the last iteration, $\pi_{t,0}^n \propto p(\mathbf{z}_t, \mathbf{x}_t = \mathbf{s}_t^n)$ and normalize so that $\sum_{n=1}^N \pi_{t,0}^n = 1$. The expectation result is computed using Eq. (4.12) and weights $\pi_{t,0}^n$. After the number of annealing iterations is achieved M , the new states at time $t + 1$ at layer M are produced from the states at layer 0 at time t .

The rate of annealing at each iteration is determined by the value of β_m . Large β_m represents a high rate of annealing leading to a peaked weighting function. The problem about having a very large value of annealing rate is that local maxima will distort the estimation. The opposite will not allow to find the global maxima with enough resolution.

The propagation of the effective number of particles from one layer to the following is chosen using the survival diagnostic defined in [MacCormick 2000c], Eq. 4.13.

$$\wp = \left(\sum_{n=1}^N (\pi^n)^2 \right)^{-1} \quad (4.13)$$

The particle survival rate α [MacCormick 2000a] represents a good measure for the annealing rate β_m and is derived for the survival diagnostic \wp

$$\alpha = \frac{\wp}{N} \quad (4.14)$$

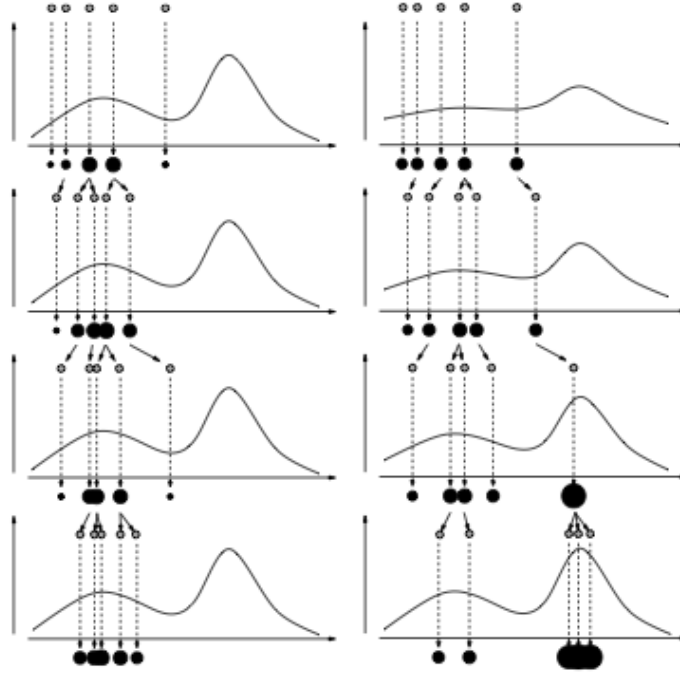


Figure 4.10: Without an annealing effect, the particles get stuck in the local maximum (left). In order that the particles escape from the local maximum, the annealing effect is used (right). (To change)

From α_i it is possible to derive $\beta_{t,m}$ at time t for all the annealing layers M . At the layer m , $\beta_{t-1,m}$ is used to determine the first set of particles. The annealing rate can be derived from Eq. 4.13 and Eq. 4.14.

$$\sigma_i = \frac{1}{N \cdot \sum_{n=1}^N (\pi^i)^{2\beta_m}} \quad (4.15)$$

with

$$\beta_m = \frac{1}{2}\beta_{m-1} \quad (4.16)$$

This time the $\mathbf{s}_{t-1,m}^i = \mathbf{s}_{t,m} + \mathbf{B}_m$ with \mathbf{B}_m a random Gaussian variable with variance \mathbf{P}_m and zero mean [Deutscher 2000]. This particle filter is used to track hands in a robust way.

4.3.2 The model

Particle filter tracking algorithms usually use colour features and contours [Gianni 2009, Micilotta 2004, Lefebvre-Albaret 2010]. In the state-of-the-art, see Section 3.5.1.2, it has been mentioned that colour based algorithms have the inconvenient that same model



Figure 4.11: (a) Illustrates the proposed face model. (b) shows the rectangular model lying over the skin probability map when ρ_{R_T} is maximal and (c) represents the best matching position for the template registration.

could be used to represent different objects, e.g. skin blobs represent head and hands, and further work is needed to label each object target. In addition occlusions between similarly coloured targets are hardly handled since it ignores spatial information. Contour based techniques take into account spatial layout information but it is not suitable for high deformable objects and is computationally expensive. Occlusions are difficultly handled since contours are sensitive to clutter.

For these reasons we have decided to use a different model, adapted to each object, for hands and head. For head we have chosen a shaped model, rectangle, since head shape changes very few. For hands we use a cloud of points to leave the shape free because of the huge hand shape variability.

Particle filter observation model can be composed of the entire set of visible features. In the case of head and hands the feature that characterises targets is the skin colour. However thanks to the object model shape is also considered.

4.3.2.1 Head models

We propose to use an image template of the subject in addition to a shaped model, i.e. a rectangle R_T divided in two regions of equal area, Fig. 4.11(a). R_{int} and R_{ext} define the sign of the weighted pixel colour probability. Thus the weighted sum of skin probabilities inside R_T ,

$$\rho_{R_T} = \sum_{\forall(x,y) \in R} R_T(x,y) \cdot p(c(x,y) | skin), \quad (4.17)$$

is minimal when most of the pixels with high probability are inside R_{int} , Fig. 4.11(b). Since the use of a complex shaped model, e.g. ellipse, has been avoided, this representation of the face increases the processing speed. In addition, considering an image template of the face, updated up to time, will help us to handle occlusion between similarly coloured objects, see Section 4.3.3.

A face detection technique using Haar-like features [Viola 2002], Ref:Appendix X, is used to initialize the model size and the face template. This technique has shown

robustness against illumination changes, scale and variation on facial expression for frontal faces. However as soon as the face is fully or partially occluded, detection tends to fail.

The heads state represents the rectangle centre coordinates, $\mathbf{x}_t^{head} = \{x, y\}$ and uses the simple particle filter algorithm 4.3. Thus each hypothetical state, particle, represents the position of the rectangle and the observation will be sampled at each position.

The observation measurement takes into account a rectangular shaped skin blob. Thus ρ_t^n for a particle sample $\mathbf{s}_t^n = \{x, y\}$ is expressed as

$$\rho_t^n = \sum_{(x', y') \in R} f_s(x + x', y + y') p(c_{(x+x', y+y')} | skin) \quad (4.18)$$

where,

$$f_s(x, y) = \begin{cases} -1 & \text{if } (x, y) \in R_{int} \\ 1 & \text{if } (x, y) \in R_{ext} \end{cases}. \quad (4.19)$$

In order to speed up the algorithm and achieve real time, ρ_t^n is computed using integral images [Viola 2002] of the skin probability map obtained through our skin model, see section 4.2. An integral image is an intermediate image representation allowing fast rectangular feature computation. Let $c(x, y)$ be the pixel intensity in the skin probability map \mathbf{S}_k at the coordinates (x, y) . The value of the integral image II at (x, y) corresponds to the sum of $c(x, y)$ and all pixels above and to the left. It is expressed as

$$II_{(x, y)} = \sum_{i=0}^x \sum_{j=0}^y c(i, j). \quad (4.20)$$

Using this representation any rectangular region can be easily computed performing basic mathematical operations. Let R_i be a rectangle defined by (x_1, y_1) and (x_2, y_2) , the sum of the pixels inside the rectangle is computed using Eq. 4.21 and ρ_t^n is easily computed for each particle, Eq 4.22.

$$R_i = II_{(x_2, y_2)} + II_{(x_1, y_1)} - II_{(x_2, y_1)} - II_{(x_1, y_2)} \quad (4.21)$$

$$\rho_t^n = R_{ext} - R_{int} \quad (4.22)$$

This measurement implicitly considers geometric information and colour feature since the best hypothetical state (particle) correspond to a maximum of skin pixels inside R_{int} . Finally the particles weight is computed using equation 4.23

$$\pi_{t,j}^n = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-\mathbf{z}_{t,j}^n}{2\sigma}} \quad (4.23)$$

where

$$\mathbf{z}_{t,j}^n = \rho_t^n \quad (4.24)$$

Each rectangular particle is propagated using an auto-regressive first order process, $\mathbf{x}_t = \mathbf{x}_{t-1} + \eta$ where η is a zero-mean Gaussian variable.

4.3.2.2 Hands Model

Hands model has to consider the huge hand shape variability. Thus it is wiser to chose a model that do not constraint the shape of the hand. For this reason a cloud of points has been chosen. Each particle is in this case a point in the cloud, i.e. a pixel in the image.

Unlike head, hands can move really fast. This is considered in the model as allowing each pixel to move at its own speed and velocity. The hypothetical states represents the position, velocity and acceleration for each particle [Gianni 2009], $\mathbf{x}_t^{hand} = \{x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}\}$. In this case the first order auto-regressive process is defined,

$$\mathbf{x}_t = \mathbf{T}\mathbf{x}_{t-1} + \eta \quad (4.25)$$

where the transition matrix is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{3}{4} & 0 & 1 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \quad (4.26)$$

Thus the new state at time t is done by

$$\begin{aligned} x_t &= x_{t-1} + \dot{x}_t + \eta & y_t &= y_{t-1} + \dot{y}_t + \eta \\ \dot{x}_t &= \frac{3}{4}\dot{x}_{t-1} + \ddot{x}_{t-1} + \eta & \dot{y}_t &= \frac{3}{4}\dot{y}_{t-1} + \ddot{y}_{t-1} + \eta \\ \ddot{x}_t &= \frac{1}{2}\ddot{x}_{t-1} + \eta & \ddot{y}_t &= \frac{1}{2}\ddot{y}_{t-1} + \eta \end{aligned} \quad (4.27)$$

The observation model considered is the probability of the pixel to belong to the skin. Let's (x, y) be to the position coordinates of the hypothetical particle state \mathbf{s}_t^n and \mathbf{S}_k the skin probability map. The observation is defined as

$$\mathbf{z}_t^n = \mathbf{S}_k(x, y) \quad (4.28)$$

and the particle weight associated to \mathbf{s}_t^n is defined in Eq. 4.29

$$\pi_{t,j}^n = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-\mathbf{z}_{t,j}^n}{2\sigma}} \quad (4.29)$$

Hands tracking uses the particle filter with annealed updated explained in section 4.3.1 to be more robust in case of high displacements. Table 4.1 synthesise the particle model chosen for head and hands, the particle states as well as the observation model.



| | Observation Model \mathbf{z}_t^n | Particle State \mathbf{s}_t^n |
|--|--|--|
| Head  | $\sum \text{sgn}(p) \cdot \mathbf{S}_k(p)$ | $\{x, y\}$ |
| Hand  | $\mathbf{S}_k(\mathbf{s}_t^n)$ | $\{x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}\}$ |

Table 4.1: Object and observation model for head and hands.

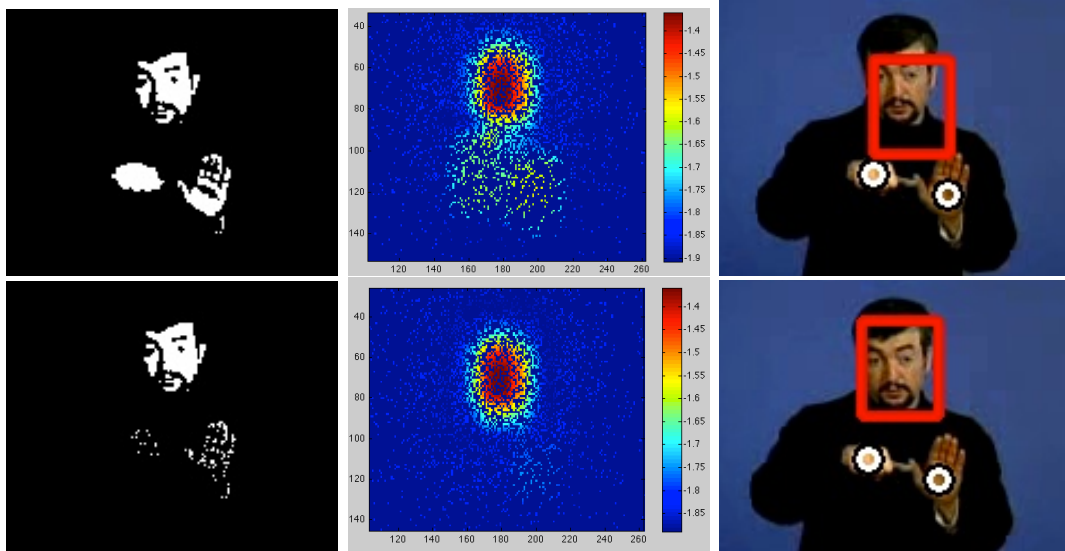


Figure 4.12: Skin probability map (left), head particles weight (middle) and expectation result (right) without particles penalization (up) and with penalization from hand samples (bottom).

4.3.3 Multiple object tracking

Multiple object tracking is challenging because of the presence of occlusions between similarly coloured targets. Since the observation model of our particle filter is based on skin colour, when similarly coloured targets are close or overlap other targets, filter observations (weights) are influenced by the presence of skin pixels not belonging to the target. For example, Figure 4.12 (top row) shows the skin probability map \mathbf{S}_k for a frame where hands get close to the head (left), the particle weight map obtained (middle) shows that hands pixels are considered on weight computation and the result (right) is then slightly displaced.

In order to avoid the influence of other object targets in weight computation, we use the exclusion principle introduced by MacCormick and Blake [MacCormick 2000b]. The

exclusion principle states that the observation for a same sample can belong at most to one filter. However for overlapped similarly coloured targets, observations may partially belong to various filters. Let's call f_j the particle filter associated to target j defined as

$$f_j(\mathbf{S}_k) = \sum_{n=1}^N \pi_{t,j}^n \mathbf{s}_{t,j}^n, \quad (4.30)$$

where \mathbf{S}_k is the skin probability map used to compute particles weights associated to target j for frame k . Using the adapted \mathbf{S}_k for each target increases the robustness of the system, e.g. a skin probability map \mathbf{S}_k^j where largest values represent the probability of skin pixels to belong to target j . For this, \mathbf{S}_k is penalized using the samples of other targets to obtain \mathbf{S}_k^j . Let's $g(\mathbf{S}_k, j)$ be the penalization function of the skin probability map \mathbf{S}_k using the samples $\mathbf{s}_{t,j}^n$ of target j ,

$$g(\mathbf{S}_k, j) = \mathbf{W}(\mathbf{s}_{t,j}^n) \cdot \mathbf{S}_k(\mathbf{s}_{t,j}^n), \quad (4.31)$$

where \mathbf{W} is a positive matrix of values between 0 and 1.

For particle n and target j , the weight $\pi_{t,j}^n$ is recomputed using the penalized skin probability map \mathbf{S}_k^j , defined as

$$\mathbf{S}_k^j = \prod_{i=0}^M g(\mathbf{S}_k, i) \quad i \neq j, \quad (4.32)$$

where M corresponds to the total number of targets.

Figure 4.12 (bottom row) shows the skin probability map for the head \mathbf{S}_k^{head} after penalization using hand particle samples (left). When head particle weights are recomputed, hands pixels are not considered and largest weights are concentrated in the head (middle) and the expectation result (right) is less influenced by other targets.

In the case of a point particle model (pixel) the penalization matrix \mathbf{W} can be filled up with a constant value since each sample pixel for target j has the same probability of belonging to the target. However when particles have a rectangular shape this is not possible to handle occlusions since pixels could belong to various skin coloured object. We propose to use a dynamic template to locally penalize \mathbf{S}_k .

The luminance difference between the found head and the template are used to determine \mathbf{W} values. Figure 4.13 shows the found image, the template saved before occlusion and \mathbf{W} used to penalize hand observation. Notice that \mathbf{W} represents well different values for head and hand.

The template associated to the particle model is updated considering the distance between head and hands, while the distance is greater than a specified threshold the template is updated otherwise an occlusion may be taking place and the head image before the occlusion is saved. This is illustrated in figure 4.14. An occlusion flag is used to determine if the template has to be updated, Eq. 4.33

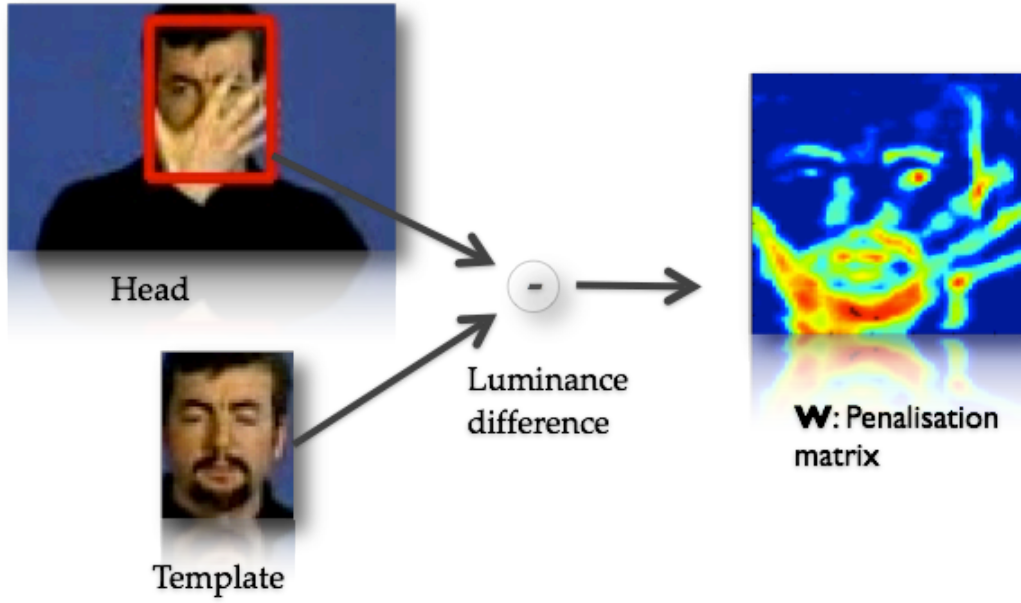


Figure 4.13: Penalization coefficients computing

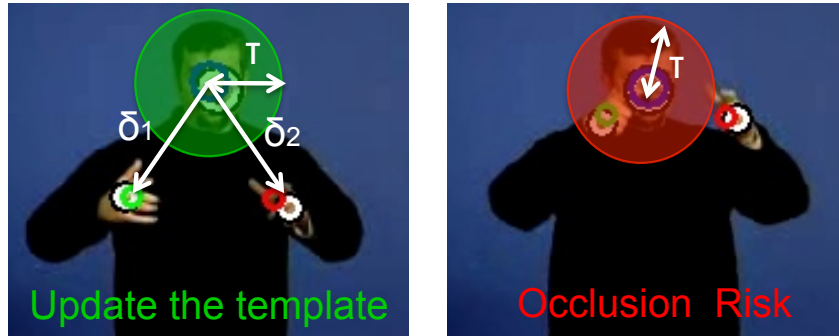


Figure 4.14: Head template updating principle

$$O_r = \begin{cases} 1 & \delta_1^t < \tau \parallel \delta_2^t < \tau \\ 0 & \text{otherwise} \end{cases} \quad (4.33)$$

with

$$\delta_j = \text{dist}(E_h^i, E_{h_j}^i) \quad \text{with } j = 1, 2 \quad (4.34)$$

where $E_h^i, E_{h_1}^i$ and $E_{h_2}^i$ are the expectation results from the particle filter at frame i , for head, right and left hand, and τ the specified threshold.

The main difficulty consist on determining the value of the threshold τ . For this we have computed for different values of τ , the normalised root-mean-square error (NRMS),



Figure 4.17: Skin probability map for each objet. Notice that other object are penalized.

Algorithm 4.4 Tracking algorithm

1. Compute skin probability map \mathbf{S}_k
 2. Head position expectation $H = f_{head}(\mathbf{S}_k)$
 3. Head penalization $\mathbf{S}_k^{hands} = g(\mathbf{S}_k, H)$
 4. Hands position expectation $H_1 = f_{hand_1}(\mathbf{S}_k^{hands})$ and $H_2 = f_{hand_2}(\mathbf{S}_k^{hands})$
 5. Hands penalization from particle samples in previous step, $\mathbf{S}_k^{head} = \prod_{i=1}^2 g(\mathbf{S}_k, H_i)$
 6. Recompute head expectation $H' = f_{head}(\mathbf{S}_k^{head})$
 7. Penalization using samples from all particle filters
 - (a) For *hands*, $\mathbf{S}_k^{hands} = g(\mathbf{S}_k, H)$
 - (b) For *hand₁*, $\mathbf{S}_k^{h1} = g(\mathbf{S}_k^{hands}, H_2)$
 - (c) For *hand₂*, $\mathbf{S}_k^{h2} = g(\mathbf{S}_k^{hands}, H_1)$
 8. Find $H'_1 = f_{hand_1}(\mathbf{S}_k^{h1})$ and $H'_2 = f_{hand_2}(\mathbf{S}_k^{h2})$
-

skin probability map for frame k , Fig. 4.17(a). Head position H is determined using the particle filter associated to this target. Since other targets may have influenced the result, head template cannot be used at this step to determine \mathbf{W} . Instead \mathbf{W} is completely filled up with a small value (~ 0) to avoid head influence in hands filtering, Fig. 4.17(b). Samples from hands filtering are used to penalize hands and to correct their influence in head expectation, Fig. 4.17(c). Optimal head position H' is determined recomputing weights after hands penalization. Head penalization coefficients \mathbf{W} are obtained from the luminance difference between the found head and the saved template. Finally hands penalize each other to avoid influence between them, Fig. 4.17(d). This procedure is detailed in Algorithm 4.4.

4.3.5 Experimental Results

The evaluation of our tracking algorithm has been performed to point out the performances and the limitations of our approach. It has been carried out in the LS-Colin corpus. Hands and head position have been manually annotated to obtain the ground truth.

The quality of the tracking results obtained using our approach has been evaluated through various rates, good tracking rate (GTR), false tracking rate (FTR) and missed tracking rate (FTR), used also in other works [Gianni 2009, Lefebvre-Albaret 2010]. Finally we have computed the execution time to evaluate the computational time require by our approach related to the quality results.

A comparison between our method, the algorithm proposed by Gianni *et al.* [Gianni 2009] (REF State of the art) and the approach introduced by Lefebvre [Lefebvre-Albaret 2010] (ref state of the art), has been carried out. These other two approaches have been also developed for sign language applications and allow us to directly compare to our results.

Figure 4.18 shows the tracking results for a sequence where hand overlaps the head. In the case of rectangular skin regions with an anatomical model [Lefebvre-Albaret 2010], when the hand fully occludes the face one skin blob is missed, Figure 4.18 (top row). For a cloud of points model with a penalisation process [Gianni 2009], the head and hand filter share the same skin region, thus filters do not accurately determine head and hand position, Figure 4.18 (middle row). On the other hand the proposed approach accurately find the position of the head and the hand when hand overlaps the face thanks to our improved penalisation process, Figure 4.18 (bottom row).



Figure 4.18: Tracking results. For Lefebvre (top row), Gianni *et al.* (middle row) and the proposed approach (bottom row).



Figure 4.19: LS COLIN corpus example

A sequence of about 3000 frames containing fast dynamics and several hand over face occlusions, has been used to perform the evaluation. For this head and hands positions have been manually annotated. In this corpus signer wears a long-sleeve sweater, Fig. 4.19, to avoid any skin region in the frame other than face and hands.

4.3.5.1 Good, false and missed tracking rate

In order to show the quality of the tracking results obtained by the proposed approach, we have quantitatively evaluated results through three rates: Good tracking rate (GTR), Miss tracking rate (MTR) and False tracking rate (FTR), Fig. 4.20 .

- **GTR** evaluates that the filter tracks the assigned skin object. For example that head filter tracks the head and hands filters track the hands.
- **MTR** quantifies the times that a filter tracks another skin object represented by the same model e.g right hand filter tracks left hand.
- **FTR** determine the times that a filter tracks none skin object or another object represented by a different model, e.g. hand and head filter exchange.

Figure 4.21 shows the GT rate achieved by the proposed tracking approach in comparison to the algorithms proposed in [Gianni 2009] and [Lefebvre-Albaret 2010]. We notice that the proposed method significantly improves stability with respect to the other two methods.

The GTR obtained for the method in [Lefebvre-Albaret 2010] are always between 0.7 and 0.8. This is because the algorithm needs very few particles since object model consist of rectangular shapes. On the other hand approach in [Gianni 2009] needs a high amount of particles to achieve the best rate since the particle model for each object is a cloud

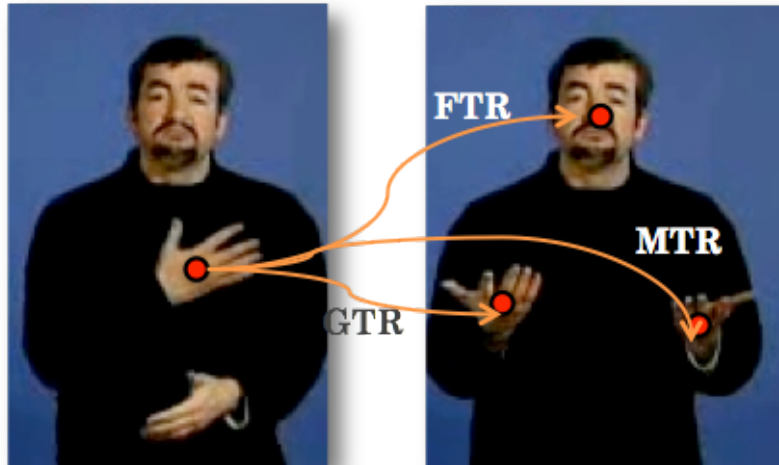


Figure 4.20: Evaluation criteria for comparison with other tracking algorithms

of points. In our case we need more particles than the first approach but less than the second to achieve a higher GT rate. In addition our method achieves the best GT rate beyond approximately 300 particles.

Unlike the approach proposed in [Lefebvre-Albaret 2010], our approach and the one presented in [Gianni 2009] shows similar results for the MTR, Figure 4.22. In fact this validates the model chosen for hands, cloud of points, which is better adapted to the high hand shape variability.

The FT rate results is shown in Fig. 4.23. We notice that our approach has the lowest FTR beyond approximately 50 particles. This is because the model for hands and head are very different and adapted to each object.

4.3.5.2 Execution Time

The execution time needed to process the LS-Colin sequence [LS-COLIN 2002], about 3000 frames, has been used to compute the number of frames per second processed by the three tracking algorithms; our approach, Lefebvre [Lefebvre-Albaret 2010] and Gianni *et al.* [Gianni 2009]. The computing rate has been plotted for various particle number, Fig. 4.24.

We notice that our approach is slower than the one proposed in [Lefebvre-Albaret 2010] which is real time until the number of particles is beyond 2000 particles. For this approach increasing the number of particles is meaningless because the highest GTR (0.8) is achieved with very few particles. However our tracker is a bit faster than the one presented in [Gianni 2009] and with better quality results.

The choice of a tracker will depend on the needs. For example if what matters is the

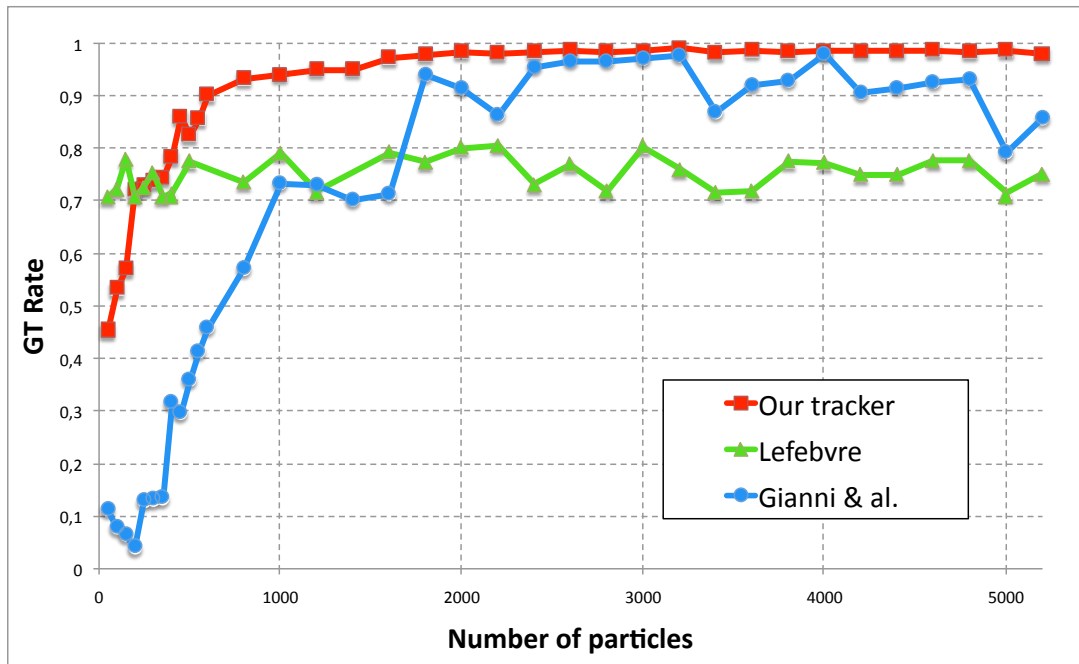


Figure 4.21: Good tracking rate GTR achieved by Lefebvre, Gianni *et al.* and our approach.

speed to be real time or to save time for post-processing performances regardless the quality of the tracking we could use the tracking in [Lefebvre-Albaret 2010]. However if we prefer a robust tracking as our approach, we can find other solutions to optimize the execution time.

4.3.6 Conclusion

We have presented in this chapter different problems faced during body part tracking. From the initialisation step until the final head and hands position.

The tracking algorithm is based in the skin colour since it is the most characterising features of hands and head. Thus the tracking results depend directly on the quality of the skin model. We have presented a method to automatically build the specific skin model from the first frame. The main advantages are that the initialisation and the building is completely automatic. Also it is independent of the illumination conditions in the video. In addition it is adapted to the signer for each video to process. Finally the simplicity of the decision rule makes the skin segmentation a very fast process. The main drawback, also because of the simplicity of the decision rule, consist on classifying some pixels as skin but not belonging to the skin class. Moreover the final skin model depends on the initialisation results.

Particle filter based tracking has the advantages of giving good results for high dy-

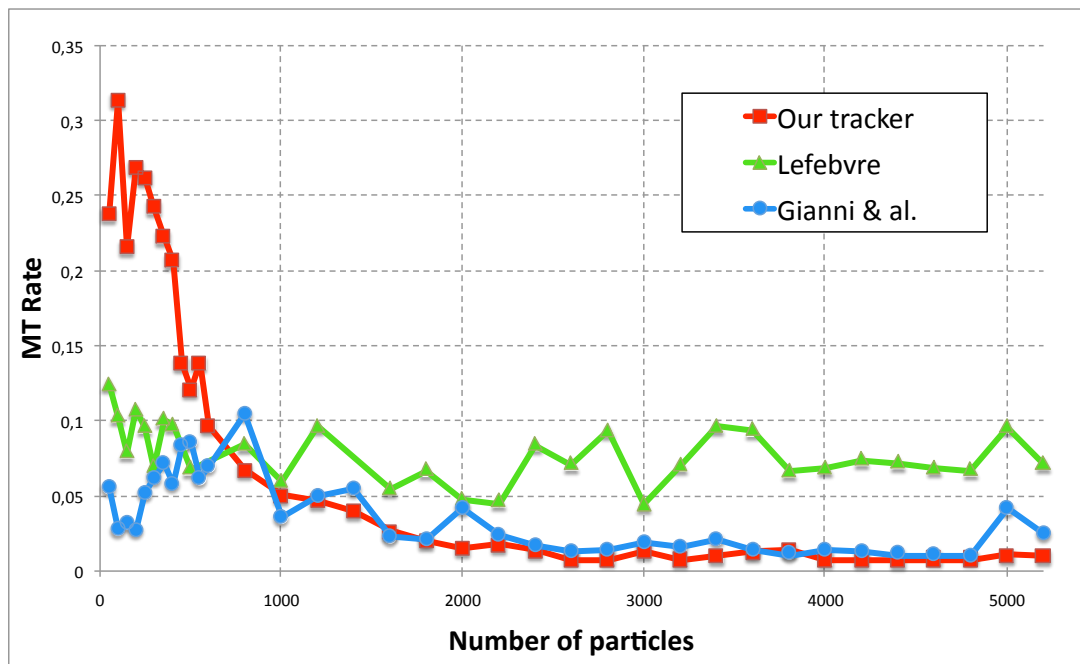


Figure 4.22: Missed tracking rate GTR achieved by Lefebvre, Gianni *et al.* and our approach.

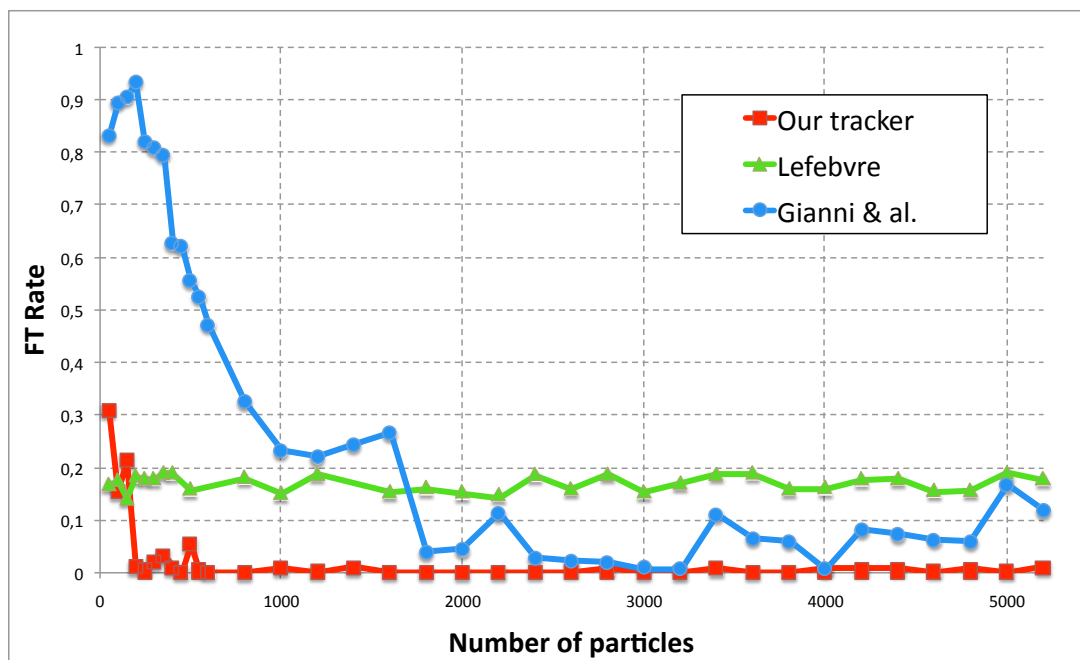


Figure 4.23: False tracking rate GTR achieved by Lefebvre, Gianni *et al.* and our approach.

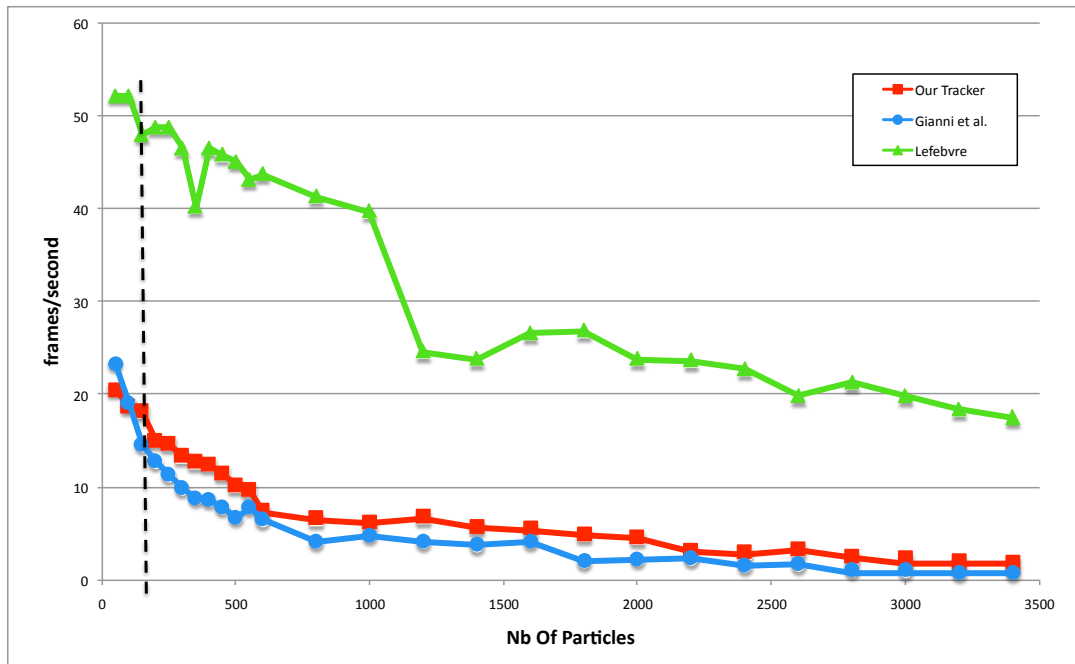


Figure 4.24: Execution time for Lefebvre, Gianni *et al.* and our approach.

dynamic objects. Results depend mainly on the selection of the object and the observation models. In our case we have decided to use a model for hands and a different model for head since shape variability and dynamics are very different. In addition a different particle filter algorithm is also chosen to avoid doing extra computation for objects that do not need it. For example using the same annealed particle filter for head, which does not move very fast, would only provide higher execution time without improving results.

Occlusions between similarly coloured objects are handled using a penalisation process. The main difficulty is to compute penalisation coefficients. We have proposed to use a template before occlusion to compute luminance difference between objects. This gives very good results since we are able to determine accurately the position of the hand event when it is in front of the face. The main limitation of our approach is when there is out-of-plane rotations during occlusions since the luminance between the face and the template changes significantly. The penalisation coefficients might be erroneous and introduce an error for tracking results. In sign language only few cases where the head rotation changes during occlusion occurs but the opposite, hand shape changing during occlusion occurs often.

We have compared our tracker to other tracking methods, also developed for SL purposes, using various rates. The chosen rates allow us to quantify the quality of the results. Although our method is computational consuming the tracking results are more robust, the missed tracking rate, which quantifies the times that the hands filter exchanges, is very few for a high number of particles our algorithm cannot label right from left hand. Other post-processing methods might be required for a fully automatic gesture processing.

4.4 Hand Segmentation

Manual features, in sign language, are characterized by the motion and the configuration of hands. Although all features in SL are important, manual features conveys most of the information in a sentence.

In computer vision approaches hand configuration recognition is a challenging task. From a mono camera view, the same hand configuration leads to different hand shapes depending on the palm orientation. Often it is more suitable to study hand shape instead of hand configuration. For this it is necessary to extract the hand region from an image by using image segmentation techniques, refer to Section 3.5.2.1, which generally use skin colour feature since it is the most representative characteristic of hands [Habibi 2004, Ramamoorthy 2003]. This has shown good results as long as the background colour is different from the skin colour. Nevertheless in sign language performance hands often cross the face area or even hands can be, deliberately, placed in front of the face. Figure 4.25 shows some signs where the hand is explicitly located with respect to the face; "sight", "dragon" and "to whistle" respectively.



Figure 4.25: Signs where the hand is deliberately placed in front of the face.

Hand segmentation algorithms have to be adapted depending on the background. During SL hands could be placed in front of a skin region or not. When the hand is not overlapping any other skin region, the segmentation can easily be performed using colour features, see 4.4.2. However the main problem arises when the hand overlaps the face, refer to 4.4.3. In this case any simple colour based technique fails in hand segmentation and other features have to be considered.

In this Phd thesis a novel approach to handle hand over face segmentation is proposed. 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

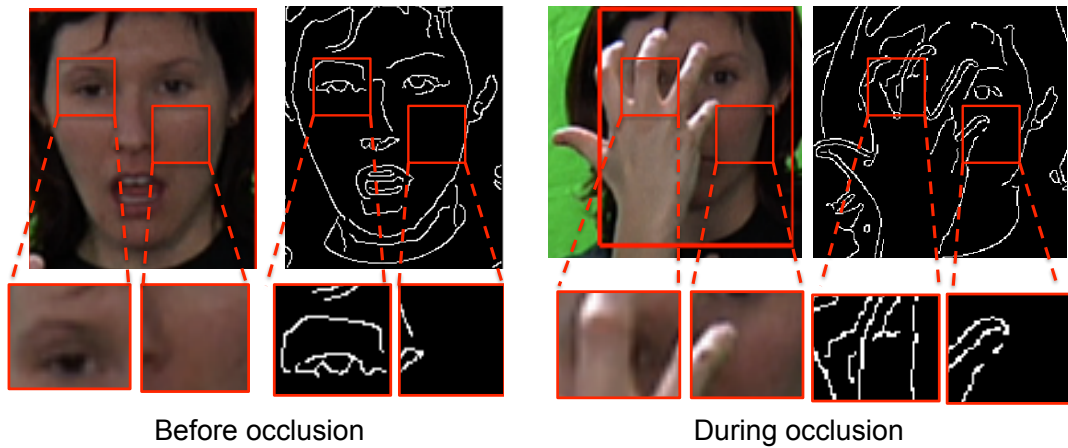


Figure 4.26: Example of two regions of the face before and during occlusion; cheek and eyes. The former illustrates that the contours give more information about hand boundaries. The latter shows that contours classification is challenging but colour gives additional information to determine hand region.

appearance which has significantly change to determine hand region. Figure. 4.26 shows an example of appearance and contours, before and during occlusion, for two different regions from the face. Notice how in the cheek region it is easier to determine edges belonging to the hand, however in the eye region this is more complicated because of the eyes contours. In this case colour has significantly changed and can be used to identify the hand region.

Edges and appearance features can be merged to perform hand segmentation. From edges information it is possible to identify boundaries belonging to the hand among all the contours in the image, and from appearance features we can extract regions from the face where appearance between hands and face is very different.

We propose a method that seeks for the differences between a template before occlusion and the image during occlusion. First of all the best template has to be found using an occlusion function based on skin pixels. The zero-crossing of this function corresponds to the limit between an image without occlusion and the following with. To speed up the algorithm using a dichotomy process is proposed. Afterwards the template is registered to the image during occlusion using either Edges Orientation Histogram (EOH) or local Gradient orientation. Finally we can compare the image before and during occlusion in terms of edges and appearance changes.

This section is organized as follows. First in 4.4.1 is detailed the occlusion detection function for determining if an adapted segmentation algorithm has to be used. Then in 4.4.2 and 4.4.3 are described, respectively, the hand segmentation approach without and with occlusion. Afterwards we present our evaluation framework, see 4.4.4, pointing out the performances and limitations of our approach. Finally our main conclusions are discussed in 4.4.5.

4.4.1 Occlusion detection

Occlusion detection consists on finding out whether the hand is occluding the face in a frame or not. It does not mean that hand is in contact with the face but only that from the camera view point hand occludes the face. Indeed from a mono-camera view, the depth information is missing thus, any distinction between overlapping and contact cannot be made.

This can be used either for annotating directly frames where the hand occludes the face or as a pre-processing step for hand over face segmentation. The latter will allow the extraction of the hand for further characterisation of shape. The occlusion detector is mainly needed in two cases; in the selection of the hand segmentation algorithm and in the head template finding procedure. The former is to know which algorithm to use for hand segmentation since an adapted one is needed during occlusion. The latter is about finding the best head template, extracted from the image just before the occlusion, to compare to the occluded image.

A straight forward solution will be to consider the distance between the face and the hand and choose a threshold value large enough to be sure there is no occlusion, as we did in the tracking algorithm, see section 4.3.3. That solution is not suitable in this case because for tracking we were looking for a global changing before and during occlusion, to roughly detect the area where the hand has produced a large modification in the illumination of face region.

Now we want to determine accurately the boundaries of the hand using local features and the result depends on the template used. An example of the template influence on the segmentation result is shown in Figure 4.27. On the left it is shown the segmentation result and the template used. This template has been determined by the distance between the hand and the face. Notice that many pixels belonging to the face have been classified in the hand class. On the right is shown the segmentation result with the respective template used for the segmentation. This one corresponds to the optimal template which belongs to the image just before the occlusion. Notice that face expression and orientation is different though in this example the hand moved very fast and there is only one frame gap between both templates. This example shows why it is important to select the best template to perform the segmentation.

Occlusion detection can be performed easily using several approaches. Here we address two methods. The first method uses measurements from the face model used in the tracking algorithm which consists on two rectangles enveloping the face. This is very fast but not very accurate. The second method considers pixel connectivity and the tracking results. This requires more computation, though it does not involve computationally expensive techniques. The basis consists on determining if the pixel corresponding to the hand position and the one corresponding to the face position resulting from the tracking algorithm are connected.



Figure 4.27: Hand segmentation template and the template used for the segmentation. Template obtained from distance (left) and best template just before the occlusion (right)

4.4.1.1 Pixel amount based approach

This method is based on measurements already performed in the tracking algorithm. It makes profit of the shaped model used for head which consists on two rectangles, see Section 4.3.2.1, enveloping the face (Figure 4.28). For particle weight it computes the number of skin pixels inside each one of the rectangles, R_{int} and R_{ext} , using integral images which allows the fast computation of the number of pixels.

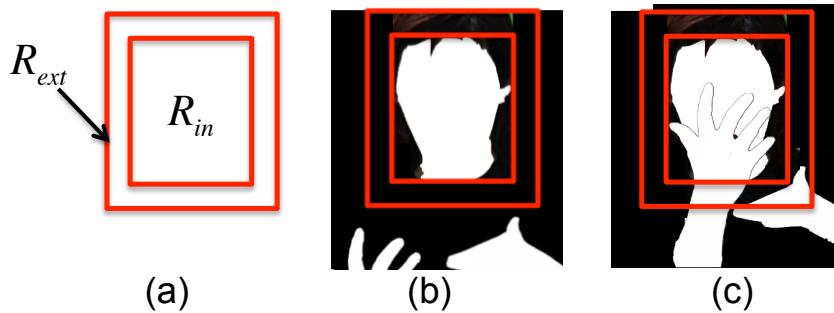


Figure 4.28: (a) face model used for tracking, rectangles configuration without occlusion (b) and with occlusion (c)

In the case without occlusion the amount of pixels from the skin probability map S_k inside R_{int} and R_{ext} is the same. When the hand overlaps the face the among of skin pixels in R_{ext} is greater that the one in R_{int} . Figure 4.29 shows the among of pixels in R_{ext} and R_{int} for the sequence [LS-COLIN 2002]. Notice that both graphs have the same form, differences between graphs correspond to the frames when hand overlaps the face. Determining whether there is an occlusion considers the difference between the amount of pixels of both rectangles and a selected threshold.

The difference between the number of pixels between both rectangles is systematically

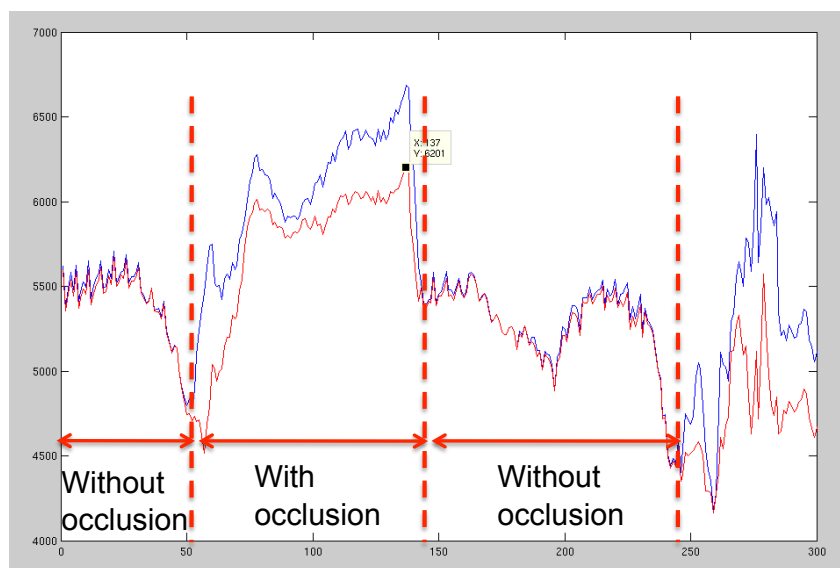


Figure 4.29: Head signature: (left) head contours and (right) EOH in polar coordinates

performed during tracking. This does not involve any additional computing since the integral images are already used during tracking. This is a very fast way of determining if there is an occlusion by referring to the obtained plot. The main inconvenient is the need of a threshold.

This method is more robust than computing the distance between face and hand which only warns up when an occlusion could occur. A drawback of this approach is when the hand gets so close that is inside R_{ext} without being inside R_{int} , in this case we might consider it as an occlusion. Using the amount of pixels, Figure 4.29, it is straight forward to determine if a frame contains an occlusion as well as the optimal face template without further processing. A more robust approach but more time consuming concerns pixels connectivity.

4.4.1.2 Pixel connectivity based approach

The second occlusion detection function proposed here consists on finding out if the face pixels are connected to the hand skin pixels. In this way we are able to determine if hand overlaps the face at any frame. For this the skin map is labelled in terms of connectivity. Figure 4.30 illustrates two labelled images; one without occlusion (left) and one with occlusion (right). Each colour represents the label assigned to the region. In case of occlusion the hand and the face have the same label.

Tracking results are used to determine if the label of the hand region corresponds to the label of the face region. Let $\mathcal{L}(x, y)$ be the label corresponding to the pixel coordinates



Figure 4.30: Labelled skin map. Without (left) and with (right) occlusion.

Algorithm 4.5 Occlusion detection algorithm

Skin segmentation

1. Transform frame to segment from RGB to YC_bC_r using Eq. 4.1
 2. Classify pixels using Eq. 4.9
 3. Label regions in terms of connectivity
 4. Find if hand and face pixels are connected using Eq. 4.36
-

$\{x, y\}$. The occlusion function is expressed as

$$O(i) = \begin{cases} 1 & \text{if } (\mathcal{L}(E_{h_1}^i) \parallel \mathcal{L}(E_{h_2}^i)) = \mathcal{L}(E_h^i) \\ -1 & \text{otherwise} \end{cases}, \quad (4.36)$$

where E_h^i , $E_{h_1}^i$ and $E_{h_2}^i$ correspond to the expectation result from our tracking for head, right and left hand respectively, see Section 4.3. Algorithm 4.5 details the occlusion detection algorithm. This approach is more robust than the previous one but requires the labelling step in terms of connectivity.

The aim of the occlusion detection algorithm allows us to determine which segmentation algorithm has to be used. A simple one when there is not occlusion or a more complex for extracting the hand when it is placed in front of the face. This occlusion function is preferred since this is more robust though this is more time consuming.

4.4.2 Hand segmentation without occlusion

The segmentation of the hand when the background is different from the skin colour can be straight forward performed using the skin model described in section 4.2. From the skin map we can extract the skin region belonging to the hand. For this the expectation result from our tracking is used. The hand skin region corresponds to all the connected pixels to the pixel position corresponding to the hand expectation.

This simple procedure is used when the occlusion function detection is -1 . In this case we know that the hand is not occluding the face and there is no need of performing more complex procedures.

This is a very simple way of extracting the hand in a frame. The quality of the results depends directly from the quality of the skin segmentation and the accuracy of the tracking algorithm. This allows to segment all the skin pixels connected to the hand without making a distinction between hand and forearm. This is a limitation and leads to constraint signer to wear long sleeves.

4.4.3 Hand segmentation during occlusion

The algorithm to extract hands when the hand is in front of the face can not be preformed by only considering skin pixels since face and hands pixels are mixed up. Thus the use of a head template/signature will help us to identify skin pixels belonging to the face. Extracting hands when there is an occlusion consists on introducing prior information before occlusion for processing images during the occlusion. Thus during occlusion it is needed to:

- find the **optimal head template/signature** which consists on the face template extracted from the frame just before occlusion, see 4.4.3.1 . We have shown before the influence of the template on the segmentation results. Indeed modification in the face expression or out of plane rotation leads to worst results since in our approach we consider that everything that has changed during occlusion is consequence of the hand. However in SL language this is not often the case. For finding the optimal template/signature we use the occlusion detection functions described in 4.4.1.
- perform the **template/signature registration**, see 4.4.3.2. Although tracking results give the position of face in an accurate way, registration is requires to align the template/signature using local features other than skin. Also during tracking we have simplified head shape using a rectangular model to speed up the algorithm. Now a more accurate position of head is needed to well classify changes before and during occlusion. Registration could be performed using several features and different approaches. Here we propose two approaches; Edges Orientation Histogram (EOH) and the local Gradient orientation.
- classify features for **hand extraction**; appearance and edges features. For appearance changes we propose to use the luminance component which gives information about any illumination change before and during occlusion. In addition to appearance we use edges for delimiting hand boarders. We intend to classify edges into two classes: edges belonging to the hand and those belonging to the head. For edges identification we propose to compare their orientation between the template and the image with occlusion. Both features are then merged to extract all the information about hands. We first compute a pixel-to-pixel edge orientation difference map. Then we use the colour information from pixels that have considerably changed to determine hand pixels. Finally using pixels connectivity we are able to extract the hand.

This method allows to accurately segment hand unlike other approaches described in the literature (Sec 3.5.2.1). Here we have considered local features that allows the extraction of hand pixels delimited by the contours. Since we use a template before occlusion a limitation concerns the out-of-plane rotation and significant face expression changing. Although the opposite, hand configuration changing is very often during occlusions, face expression remains stable. Herein the first stage of our segmentation algorithm which finds the optimal template.

4.4.3.1 Optimal Template before occlusion

In order to segment the hand when it is in front of the face, we propose an approach in which local features are classified as belonging to the hand or to the face. This is achieved by a comparison between the image before and during occlusion. Obtaining the best classification results depends on the selected template. For this reason it is important to find the closest template just before occlusion to avoid any face expression modification to influence our segmentation result. For this we use the occlusion detection function, see section 4.4.1. Figure 4.31 illustrates the occlusion function value for a short sequence. The optimal template corresponds to the frame before the zero-crossing of the occlusion detection function.

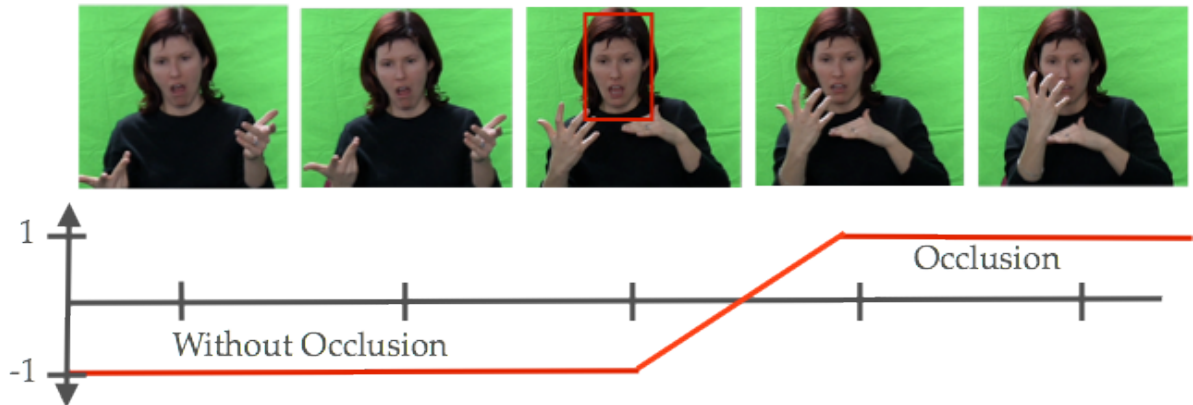


Figure 4.31: Occlusion function results for a short sequence.

The main idea is to go backward using the occlusion detection function until the first frame in which the hand is not occluding the face. Then from that frame we can extract the face template needed for the classification. The problem is that going backward and testing each frame can be time consuming. In order to optimise the searching procedure we propose to use the dichotomy principle which splits the searching timeline. For initialising the searching timeline we consider the distance between the face and the hand. If the distance is larger than a threshold, e.g. once the size of the head, then we are sure that there is not occlusion and that the occlusion detection function has a zero-crossing. The dichotomy procedure will be performed between this frame a and the frame that we are processing b . Then the occlusion function $O(i)$ is evaluated in the middle position of

Algorithm 4.6 Optimal template searching algorithm

-
1. Start from the values (a, b) . a corresponds to the frame number where the hand and the head distance is large enough to be sure that there is no occlusion and b is the frame during occlusion being processed.
 2. Evaluate the occlusion function $O((a + b)/2)$, Eq. 4.36.
 3. If $O((a + b)/2) < 0$ the $a = (a + b)/2$ otherwise $b = (a + b)/2$
 4. Repeat 2 and 3 while $abs(a - b)/2 < 1$ Otherwise a corresponds to the frame number containing the best template.
-

the segment (a, b) . Then the new searching timeline is the half where there is a change on sign. And this is split again until the searching segment length is of unitary value. This is described in Algorithm 4.6.

This approach allow us to find the optimal template which is used later for registering and classifying local features. Using dichotomy for finding the best template speeds up the algorithm and can be used with any occlusion function as long as the zero-crossing represents the limit between occlusion and non-occlusion. The results is registered to the image where the hand occludes the face. This is described below.

4.4.3.2 Face template registration

Once we have the optimal template we have to register it to the image with occlusion. Registration could be performed using several features, here we use the edges since this gives important information about the position of face. A global and a local approach are described: Edges orientation histogram (EOH) and local gradient orientation. Herein we discuss both methods and argue the advantages and limitations of both approaches.

Edge orientation histogram (EOH) is a feature descriptor used for the purpose of object detection. The technique counts occurrences of edge orientation in a selected area on the image. This method is scale-invariant and shape contexts. Head signature consists on the EOH counted into K number of bins and computed using a Sobel filter. The closest sample is searched around the head expectation result E_h^i by scanning the area in a neighbourhood of N pixels. Samples EOH are computed within a moving constant size window and compared to the *head signature* using the Euclidean distance. The best position of the head is then when the distance between both EOH is minimal: head signature and the sample.

Figure 4.32 shows an example of a frame with the head signature in polar coordinates (left) and the best found position (right). Notice how the EOH for both face region is very similar. Differences are introduced by the hand in front of the face and the change on face expression.

The distances map between the researched area and the head signature is shown in figure 4.33. As we can see it has a unique minimum, then to speed up the registration a

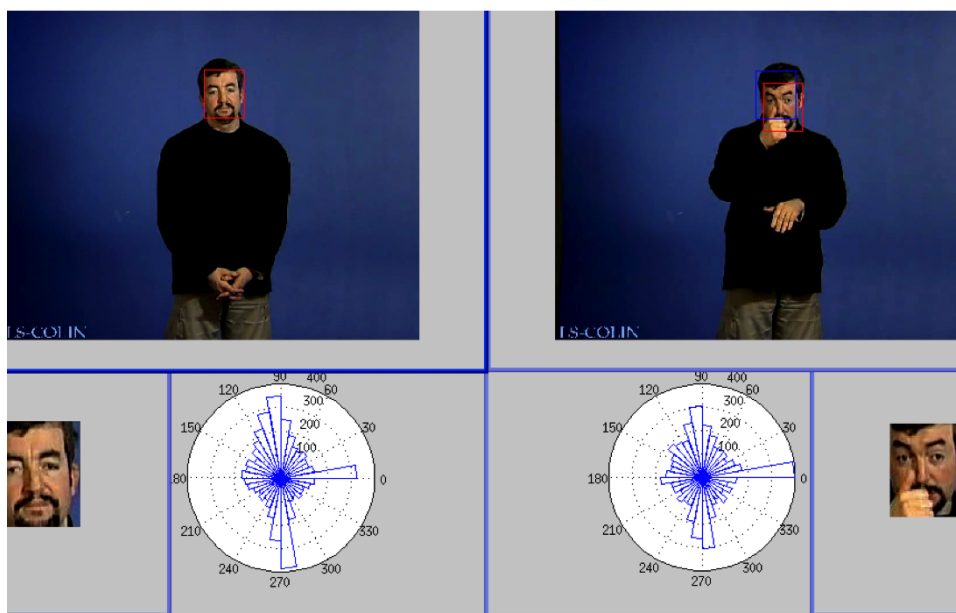


Figure 4.32: Head signature (right) and best position (left). Black rectangle corresponds to the initialisation and red rectangle to the optimal position.

gradient descent optimization could be performed. This is not implemented in this PhD thesis.

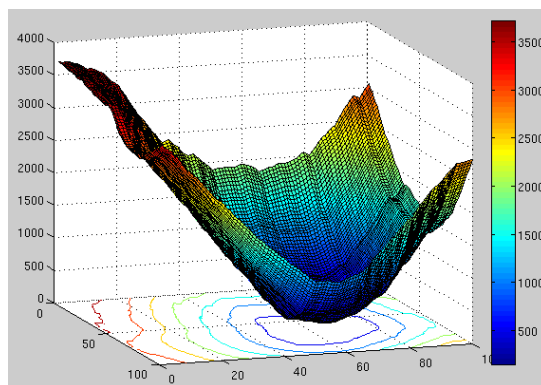


Figure 4.33: Distance map between the researched area and the head signature

Registration results depend on the configuration of the hand and the orientation of its contours. This method is called global since this uses the totality of the contours to build the histogram. A better solution is to consider only local features, this could be done by splitting the region into several sectors and build the histogram for each sector or by using the gradient orientation.

Gradient orientation is computed using the face image and is orthogonal to edges. Here the template is registered using the direction of the gradient. Figure 4.34 illustrates

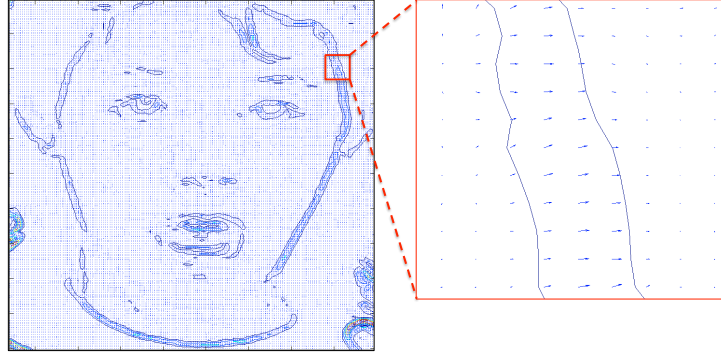


Figure 4.34: Head signature using gradient orientation

the face signature from gradient orientation. Contours have been added to the image for better appreciation of orientation.

Registration is performed in a searching window around the expectation results from the tracking algorithm, $\mathbf{E}_h = \{x_h, y_h\}$. The image transformation matrix considers displacements in \mathbb{R}^2 and rotation in the plane $X \perp Y$. Out-of-plane rotations are not handled in this work. The optimisation function is defined as

$$d(x, y, \theta) = \arg \min_{\theta \in [\theta_{min}, \theta_{max}]} \sum_{(x', y') \in I} \min(\theta_T'(x + x', y + y') - \theta_I(x, y)), \quad (4.37)$$

where $(x', y') \in N \times N$ with N the size of the searching window centred in \mathbf{E}_h , θ_{min} and θ_{max} are, respectively, the minimal and maximal rotation angles, θ_T' represents the edge orientation of the image template rotated by the angle θ .

The optimisation search is performed in a small neighbourhood inside a window since substantial head movements have already been considered by the tracking. Even though this searching algorithm is time consuming, local minima are avoided. The alignment of the template face handles local face deformation (e.g. lips, eyes, etc.) and partial occlusions leading to good results as long as the out-of-plane face rotation remains small.

Both methods use orientation features since our tracker have only used skin colour for tracking. The first method, uses the histogram of the edges orientation which corresponds to take into account the global information losing spatial information. The second uses the gradient orientation, using then local information and considering their spatial distribution. We prefer the second method which keeps the spatial distribution of pixels. Once the template has been registered we can proceed the feature classification step to distinguish pixels belonging to the hand from those belonging to the face.

4.4.3.3 Feature classification

Features used for the classification within the hand and the head classes concern edges and illumination. From previous preprocessing steps we have obtained the optimal template and the best position from registration. Now this information is used to compare edges and appearance before and during occlusion. Figure 4.35 shows the steps for the segmentation. It consists on an edges classification algorithm and the luminance difference between both image.

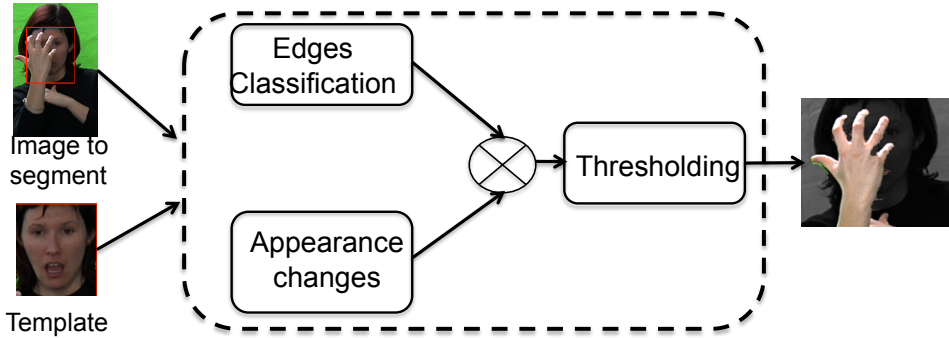


Figure 4.35: Edges and appearance information for hand segmentation.

- **Edge classification** is performed in distinguish edges as belonging to the face or to the hand. First we use the Canny edge detector in both the template image and the occluded image to detect edges. The result is illustrated in Fig. 4.36. For each pixel belonging to edges in the image during occlusion we seek for the closest edge, along a normal profile, in the template, Fig 4.37. When an edge is found in the template image the orientation difference is computed. It is defined as the angle between the edges normal direction from the image during occlusion and the template.

$$\Delta\theta = ||\theta_o(x + n_x, y + n_y) - \theta_p(x, y)||, \quad (4.38)$$

where θ_p corresponds to the edge orientation in the image template and θ_o to the edge orientation in the image with occlusion. When no edge has been found in the defined neighbourhood the difference is considered the highest orientation difference value ($\pi/2$). That means that this pixel in the occluded image has an large probability to belong to the hand. The orientation difference map is built in this way for each edge pixel in the image with occlusion.

The normalized edges orientation difference map θ_{map} is expressed as

$$\theta_{map}(x, y) = \begin{cases} \Delta\theta(x, y)/(\pi/2) & \text{if matched} \\ 1 & \text{otherwise} \end{cases} \quad (4.39)$$

Fig. 4.38 shows an example of the normalized orientation difference map θ_{map} . Notice that most of the edges belonging to the hand have high values, close to 1,



Figure 4.36: Contours from the face template and the image during occlusion.

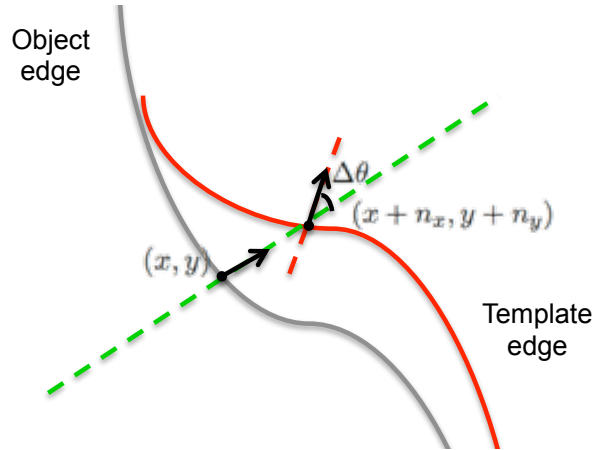


Figure 4.37: Edge matching and edge orientation difference

and many edge pixels from the face have low values, close to 0. In places where edges from the hand intersect edges from the face, e.g. over the mouth or eyes, other values appear depending on the intersection angle (low values if edges coincide). The separation of the edges is not straight forward from this difference map. For low values it is not easy to define if edges coincide or if they really belong to the face. Then if we try to classify edges with no further features we can miss important information. That is why the next step is to use appearance features to remove this ambiguity.

- **Appearance information** is used to complement edges information. We noticed that during occlusions, in places where there is an ambiguity concerning edge classification pixels colour have considerably changed. For example in the mouth or eyes area hand colour is very different, however edges could coincide depending on the shape and position of the hand over the face.

$$L_{map}(x, y) = \begin{cases} 1 & \text{if } \|I(x, y) - T(x, y)\| > th \\ 0 & \text{else} \end{cases} \quad (4.40)$$

This stage is very simple and consists on performing the pixel to pixel difference using a threshold th to obtain the skin region that has changed, Eq. 4.40. The luminance map allow us to determine changes between the template and the image

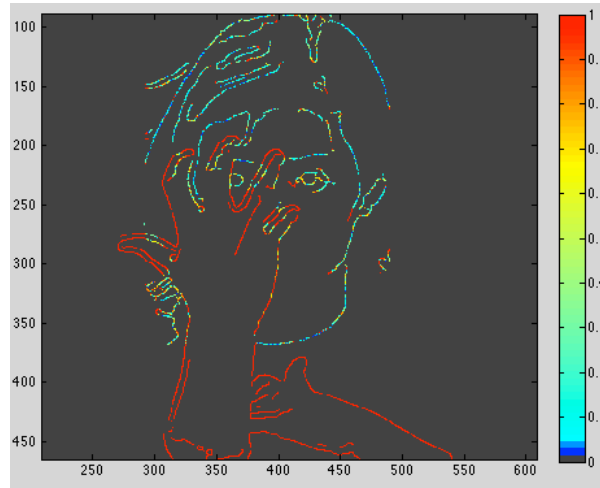


Figure 4.38: Edges orientation difference map

during occlusion in terms of illumination. Indeed luminance components changes significantly when the hand occludes the face. One limitation of this approach is that the threshold has to be adapted to the illumination conditions of the video. Also the out-of-plane rotation or a significant change in face expression introduces some artefacts in the binary image after thresholding. If those artefacts are connected to hand region they will be kept in the result.



Figure 4.39: Luminance difference between the template and the image during occlusion.

Information extracted from this two features are complementary. The combination of both features permit to classify pixels as belonging to the face or to the hand. Both maps: edges orientation and luminance are normalized and combined into a new map, Fig. 4.40. Hysteresis threshold is used to extract the skin region that is connected having a greatest probability of belonging to the hand, thus hand is segmented. We have filled the holes and extracted the largest connected area (hand).

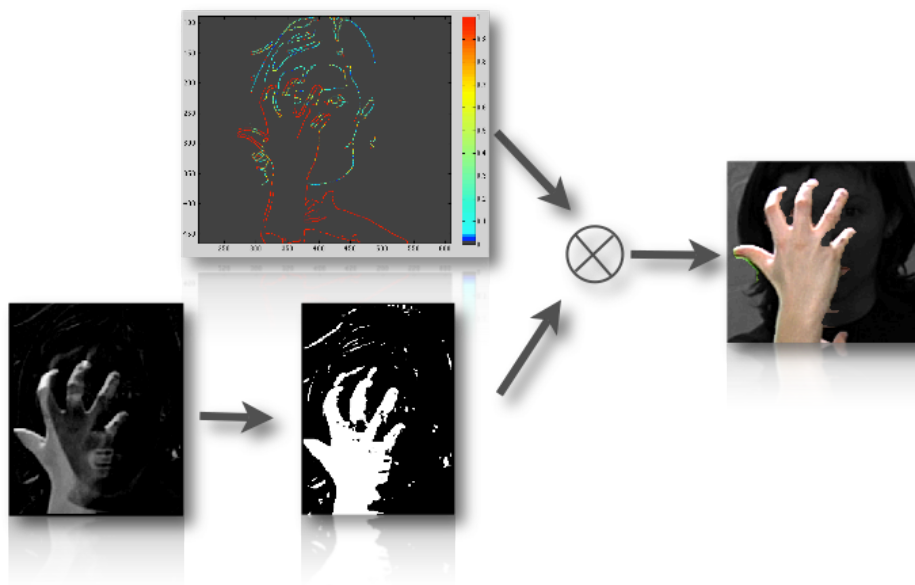


Figure 4.40: Combination of edges orientation and luminance difference.

4.4.4 Experimental results

In order to point out the performances and limitation of our approach an evaluation of the hand over face segmentation methodology has been presented. Because of the lack of publicly-available annotated corpus our approach cannot be compared with existing methods. We have used several frame sequences from the LSF Dicta-Sign 3.2.2 corpus where the pixels belonging to the hand have been manually annotated. We have selected some sequences that contain several hand over face occlusion. In these sequences, hand shape and face expression can change during the occlusion. Our approach has been tested on 5 sequences, around a total of 50 images with face occlusion.

The evaluation has been performed using the pixel connectivity based occlusion detection function and the gradient orientation registration. Although other methods have been described their evaluation is noted in our perspectives.

Fig 4.41 shows in the top row the images to segment and in the bottom row the segmentation results. Notice that in the segmentation results, pixels that have been classified in the hand class are shown in their natural colour, otherwise they are shown in dark grey or in black. Fig 4.42 shows the segmentation results for several frames of various sequences. These two figures indicate that the hand over face segmentation was performed reasonably well, however we can see some artefacts and some holes in the results. The artefacts are mainly due to large pixel changes, e.g. out-of-plane rotation or/and substantial face expression changing, or wrong skin segmentation. On the other hand holes are caused because the luminance change is very few. In fact sometimes an edge from the hand can coincide to an edge from the face, in terms of orientation and position. In that case the hand edge might be classified as belonging to the face. Then if



Figure 4.41: First row shows five consecutive frames in a sequence with hand over face occlusion. The second row shows the segmentation result. Pixels in dark grey or in black have been classified as belonging to the non hand class, otherwise they are shown in their natural colour.

there is no colour information because the pixel colour remains very similar, some pixels will be wrongly classified. In any case the overall hand shape is well defined.

The performance of the proposed segmentation approach has been qualitatively evaluated. Now to quantitatively evaluate this method, we have manually generated ground truth segmentation for all the frames in the sequences. Since the evaluation performed is pixel wise, the ground truth is a binary image of hand pixels. These ground truth images are used as reference to compare the automatically segmented images. The true positive (TP) and the false positive (FP) percentages are evaluated for each image of the sequences by

$$TP(\%) = \frac{\text{Number of correctly detected pixels}}{\text{Total number of hand pixels}} \times 100 \quad (4.41)$$

$$FP(\%) = \frac{\text{Number of wrongly detected pixels}}{\text{Total number of hand pixels}} \times 100 \quad (4.42)$$

where $TP(\%)$ corresponds to the rate of correctly detected pixels with respect to the total number of pixels to be detected and $FP(\%)$ to the wrongly detected pixels with respect to the total number of pixels that should not be detected. Since $FP(\%)$ depends on the number of non hand pixels, this rate becomes dependent of the background and size of the image. For this reason we decided to compute the $FP(\%)$ rate with respect to the total number of hand pixels. Thus this rate is representative to the number of hand pixels on the image. Table 4.2 presents the rates evaluated for each sequence, we notice that the $TP(\%)$ rate is in average about 96%, reaching until 99% for some frames. The $FP(\%)$ rate is about 8% and corresponds to pixels that can be easily detected and eliminated by post-treatments, e.g. thin lines under the chin and/or over the collar in Fig. 4.43. Moreover we plan to extract geometrical features (e.g. eccentricity, equivalent, etc.) to characterize the hand and these measurements should not be extremely corrupted by the remaining artefacts.



Figure 4.42: This figure shows the segmentation results for 3 different sequences. Each row corresponds to a sequence.

Table 4.2: Results of the evaluation rates trough several sequences

| Rates | Sequence | | | | | Average |
|-------|----------|-------|-------|-------|-------|---------|
| | 2 | 4 | 7 | 9 | 11 | |
| TP(%) | 96.71 | 96.51 | 96.59 | 95.07 | 98.15 | 96.61 |
| FP(%) | 3.62 | 6.48 | 13.91 | 6.44 | 8.27 | 7.74 |

4.4.5 Conclusion

Hand segmentation is performed in order to extract hand region for characterising or classifying shapes. Indeed hands shape conveys lot of information and is very important for SL processing. The main problem encountered in hand segmentation concerns the case when the hand overlaps the face. This configuration is very often present in SL performances, hand not only passes in front of the face but it is explicitly placed near the face. Then distinguishing pixels belonging to the hand or to the face is challenging.

In order to process this case with adapted algorithms an occlusion detection function is required. For this we have proposed two methods one considering the amount of pixels in the area neighbouring the face and a second one taking into account the connectivity of pixels. The former is faster since uses some results from our tracking method but the latter is more accurate. We have used the second one for performing our experiments since for annotation we prefer accuracy that speed.

From the occlusion detection function we can determine whether a simple procedure



Figure 4.43: Segmentation results: artefacts under the chin and over the collar

such as skin segmentation, can be used for extraction hand or an adapted algorithm is required, in the case that hand is placed in front of the face. For the segmentation of hand over the face we propose a novel approach taking into account two local features; colour and edges. It considers the pixels colour changing and the edges orientation. When the hand occludes the face, both features information complements each other. That means that in some places where the pixels colour remains quite similar, we still have the edges information and vice-versa. By merging these features and by comparing the image before and during occlusions, we are able to well segment the hand.

Our proposed approach uses a face template before occlusion in order to compare what have changed. For this we propose a method for finding the optimal template, i.e. the head image corresponding to the first image just before occlusion. Then the template is registered to the image during occlusion and the luminance change and edges orientation are classified as belonging to the hand or to the face. Finally hand is extracted using the connectivity of pixels.

Experimental results have been performed in an international corpus (Sec 3.2.2) and indicate that this method is able to extract the hand effectively. Indeed, qualitatively, our results are better than approaches in the state-of-the-art where hand region is roughly obtained, see Section 3.5.2.1. A straight forward comparison with other studies in the literature could not be performed because of the lack of available annotated data. Our results showed the limits of the approach regarding the quality of the segmentation: artefacts and holes, which are due to the out-of-plane rotation. This is not often the case in SL, however this has to be taken into account if this approach is to be used in other context.

Once hand region can be extracted from any frame before or during occlusion, we can do further processing for characterising hand shape. This is used for the temporal segmentation of signs in combination of motion features (see the following section).

4.5 Temporal segmentation

In this section we present a novel approach for segmenting signs. We have argued (Sec. 3.5.3) the need of detecting word boundaries for processing continuous SL. In addition to boundaries detection the temporal structure of signs is required since we have decided to use the Zebedee (Sec. 3.4.2) sign description for gloss recognition. Then a temporal segmentation allowing to describe the sub-units composing the signs is required.

Our goal is to automatically segment signs based on low level features avoiding any learning step. Since linguistic information is not used it is not possible to label segmented sequences as signs or transitions¹. Thus our segmentation approach intends to detect the limits corresponding to the beginning or to the end of a sign or to a key frame K which is equivalent to Hold H in the temporal representations (Sec. 3.4.2).

Figure 4.44 shows an example of the segmentation results differences between a manual segmentation and an automatic approach. This is illustrated in the annotation software *Elan* where two tiers have been created: *Manual Ann.* and *Auto Ann.* The former, Fig. 4.44 *Tier: Manual Ann.*, represents the results of manual sign segmentation where the annotator has selected the first and the last frame of a sequence corresponding to a sign. The latter, Fig. 4.44 *Tier: Auto Ann.*, illustrates the expected results from an automatic annotation approach. Since no linguistic information is used to perform the segmentation, classifying sequences into *Signs (S)* and *Transitions (T)* is not possible. Sign segmentation is, then, modelled as an *Even Detection* approach where the limits between S and T are detected.

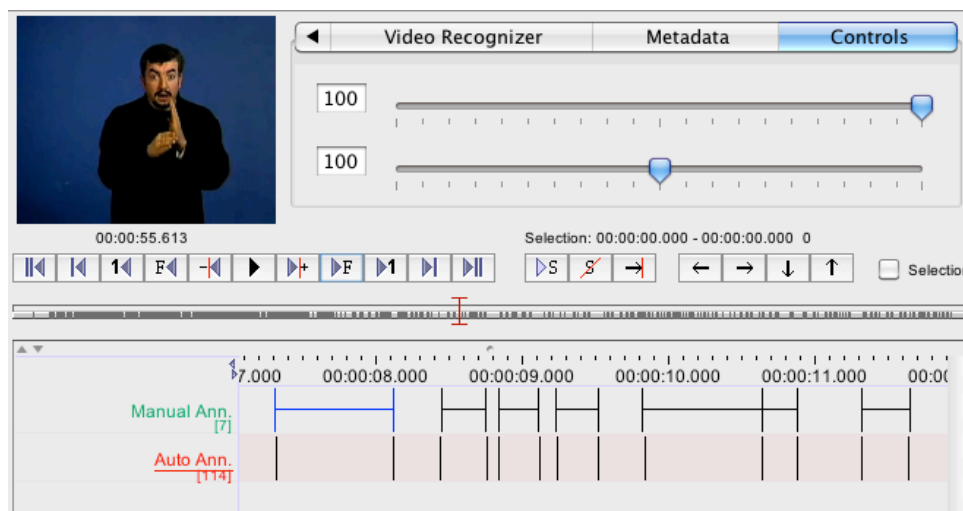


Figure 4.44: Annotation tool *Elan*. *Manual Ann.* is the segmentation results by an human annotator and *Auto Ann.* tier shows the expected segmentation results.

1. The word *Transition* can refer to the meaningless gesture between two signs as a results from the co-articulation effect or to the sequence between two key frames in the linguistic description of signs ZeBeDee. In this whole section it concerns the former definition

As we mentioned before (Sec. 3.5.3) we face a complex problem without any support from linguists. Indeed there is not agreement on any definition of word segmentation remaining a fully subjective procedure dependent on the annotators experience and on their knowledge and interpretation of the language. For this reason in this PhD we do not intend to propose a fully-automatic segmentation completely unsupervised. Instead we would like to propose limits that from objective measurements could correspond to what linguists manually select as word boundaries.

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.

Generally, in the literature (Sec. 3.5.3), motion features are used for segmenting signs. Although hand shape features can be used this only concerns approaches using motion capture devices since collecting hand shape information is challenging from computer vision approaches. In this PhD thesis we propose a novel approach considering, in addition to motion features, hand shape using geometrical shape measurements as the ones described in Section 3.5.2.2. Motion is used to characterise gestures in terms of velocity, to perform the segmentation. Later hand shape features are used to improve the previous segmentation step. Even though we use manual features to improve sign segmentation, face expression or other articulators information could be useful to achieve this task.

Our temporal segmentation approach uses the results obtained by our tracking algorithm for the computation of motion features and the segmentation approach for extracting the hand region even when it is placed in front of the face. This is illustrated in Figure 4.45.

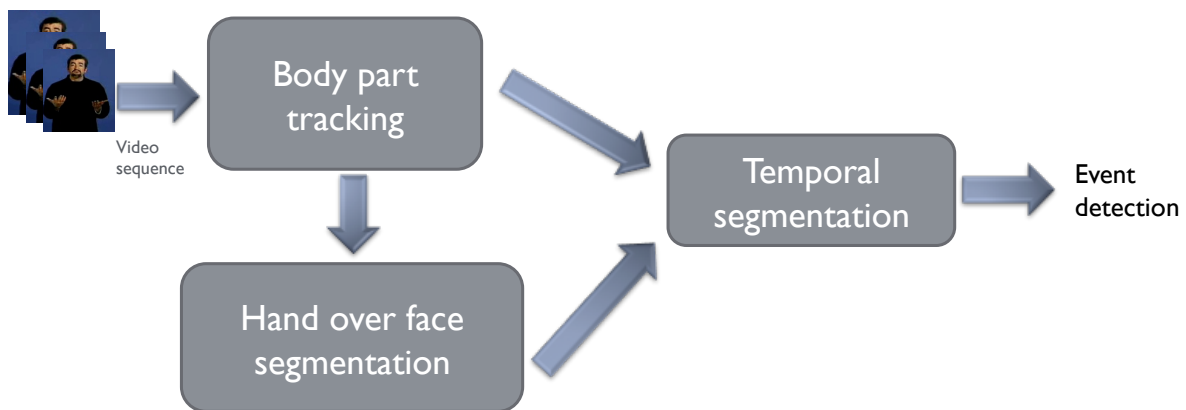


Figure 4.45: Event detection algorithm schema

These methods are used to extract motion and shape feature,

- **Motion features** correspond to the computation of hands velocity, though head velocity can be also computed this is not used in this work. Velocity gives information about the number of moving hands so that sequences can be classified. Also significant change on trajectory direction can be detected with the magnitude of the velocity.
- **Shape features** are extracted from hand region for each detected event obtained using motion features. Hand segmentation is performed using an adapted algorithm depending if hand is occluding the face or not. Hand shape is systematically compared to the adjacent detected events. This second step removes events for sequences that have been over segmented.

This gives potential limits that could correspond to the boundaries. The result consists on an Elan file which can be directly opened and modified. It consists of two tiers one with the segmentation result from motion features and a second one after the introduction of shape features. In addition three graphics corresponding to the velocity between hands (top) velocity of right hand (middle) and left hand (bottom), are also added to help annotators choosing limits in terms of motion measurements. The computation of these results is described below.

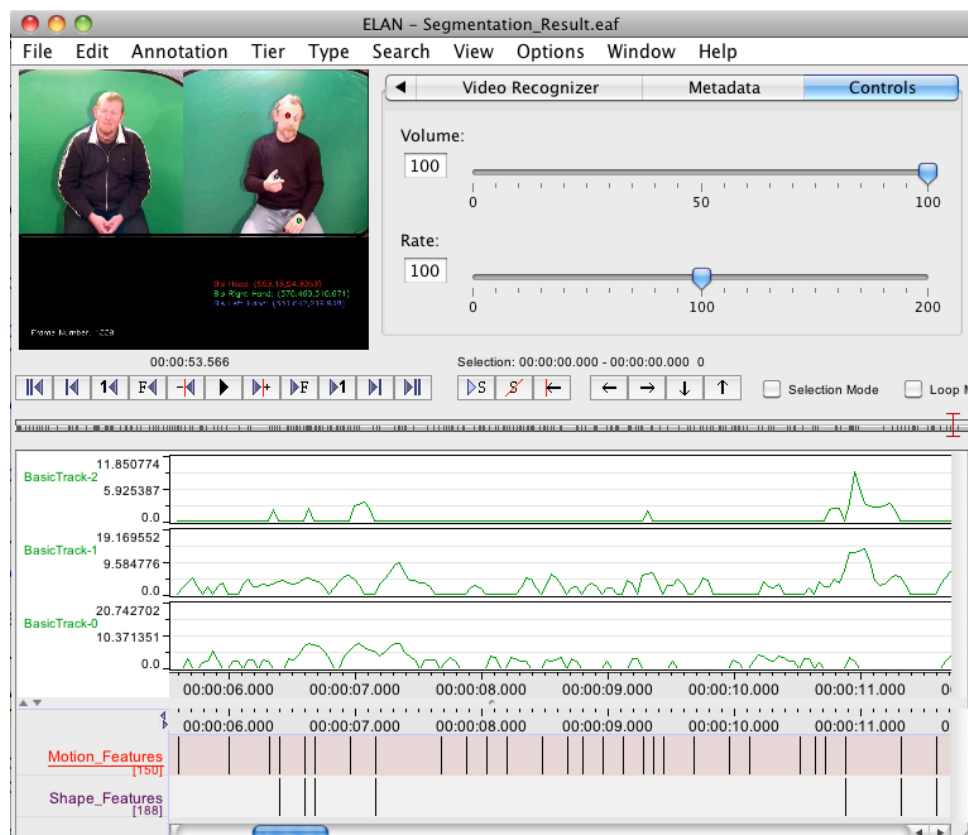


Figure 4.46: Event detection algorithm schema

4.5.1 Motion Classification

Velocity is used for classifying sequences in terms of number of moving hands and change of trajectory. Velocity for right and left hand, $v_1(t)$ and $v_2(t)$ respectively, are computed using a moving window of small size to avoid large signal smoothing (between 3 and 5 frames). Centring the window of size W on the number of frame k for which the velocity has to be computed, the velocity magnitude is defined as

$$v_j(k) = \frac{1}{2 * w} \cdot \{E_j(k + w) - E_j(k - w)\}, \quad (4.43)$$

where $w = \frac{W-1}{2}$ and W an odd integer corresponding to the size of the moving window, j the corresponding object h_1 or h_2 right or left hand respectively, and E the expectation result from our tracking algorithm (Sec. 4.3).

Velocity magnitude is used to determine relative velocity $v_r(t)$ between hands, i.e. velocity difference between right and left hand. It is derived as

$$v_r = \|v_1(t) - v_2(t + \tau)\|, \quad (4.44)$$

where τ represents the gap between hands velocity in symmetric movements. Indeed when hands move together there is a small gap between right and left hand velocities. Figure 4.47 shows (left) the sign 'shocked' in French Sign Language which corresponds to a two-hand movement. Hands move together in a symmetric way from bottom to top. The velocity profile for right and left hand (right) is plotted. Although superposed velocity profile (Fig. 4.48) for both hand looks very similar, one hand remains behind the other. In this example τ is about 1 frame.

Considering relative velocity $v_r(t)$ and hands velocities, $v_1(t)$ and $v_2(t)$, we propose to classify motion in three classes: static pose (S), one hand ($1H$) and two hands gestures ($2H$) (Table 4.3).

- **Static pose** (S) : consists on the sequences without any modification on hands location or configuration. This is detected when $v_r(t) \approx 0$ and right $v_1(t)$ and left $v_2(t)$ hand velocity are close to zero.
- **One Hand** ($1H$): gestures consist on sequences where only one hand moves and the other remains stable. This allows to identify signs performed only with one hand. However in continuous SL we have noticed that sometimes one-hand signs might be classified in the two hands class, this is corrected using further processing. Indeed during the performance of the sign the other hand moves to prepares the following sign when it concerns a two-handed sign. Sequences in this class are identify when $v_r(t) \approx 0$ as well as the velocity of one of the hands is close to zero.
- **Two hands**: gestures concern both hands moving either symmetrically or not. For two hands classification, $v_1(t)$ and $v_2(t)$ are different to zero. Within this class we can derive two subclasses. When $v_r(t)$ is close to 0, both hands move at the same speed 'symmetric gesture' otherwise both hands move but there is no symmetry.

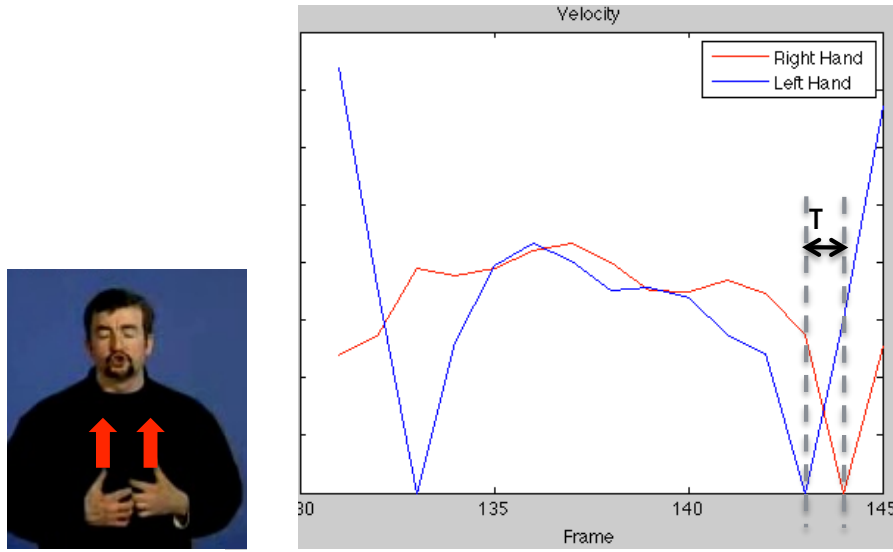


Figure 4.47: Illustrates velocity of right and left hand. *Left:* The sign 'shocked' in French Sign Language. *Right:* Velocity profile for right and left hand.

Table 4.3: Motion classification

| Static Pose (S) | One Hand(1H) | Two Hands (2H) | |
|--|--|----------------------------------|-----------|
| $v_r \approx 0$ | $v_r > 0$ | $v_r \approx 0$ | $v_r > 0$ |
| $(v_1 \approx 0 \ \& \ v_2 \approx 0)$ | $(v_1 \approx 0 \oplus v_2 \approx 0)$ | $(v_1 \neq 0 \ \& \ v_2 \neq 0)$ | |

From this classification events are detected as the changing from one class to another. This approach over-segments repetitive signs. This is not a limitation since this corresponds to the key postures K from the linguistic description of signs (Sec. 3.5.3).

Figure 4.49 shows an example of the results using our motion classification. On the top left the sign 'What?' is illustrated. It is a repeated symmetrical sign. Figure 4.49 shows the classification results and the detected events on the left and on the right respectively. Notice that the sign is over-segmented because magnitude velocity comes to zero in the middle of the sign when the trajectory changes the direction. In this case it is classified as a static pose.

In order to obtain the segmentation at a sign level hand shape features are introduced. For deleting events detected in the middle of a sign we consider that the hand shape has not changed. Thus over-segmentation can be corrected considering other features that remain constant during repetitive gestures as is the case of hand shape, see Figure 4.50.

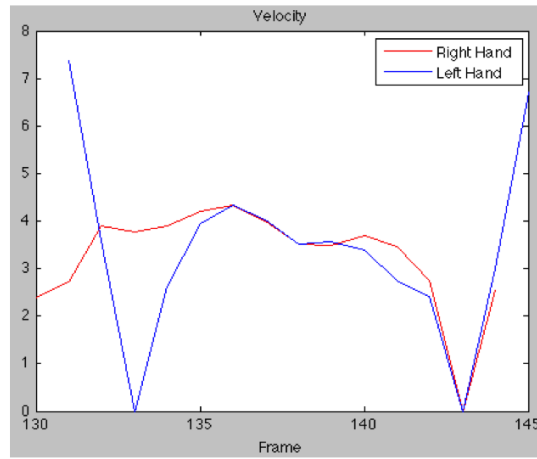


Figure 4.48: Illustrates velocity profile of right and left hand superposed.

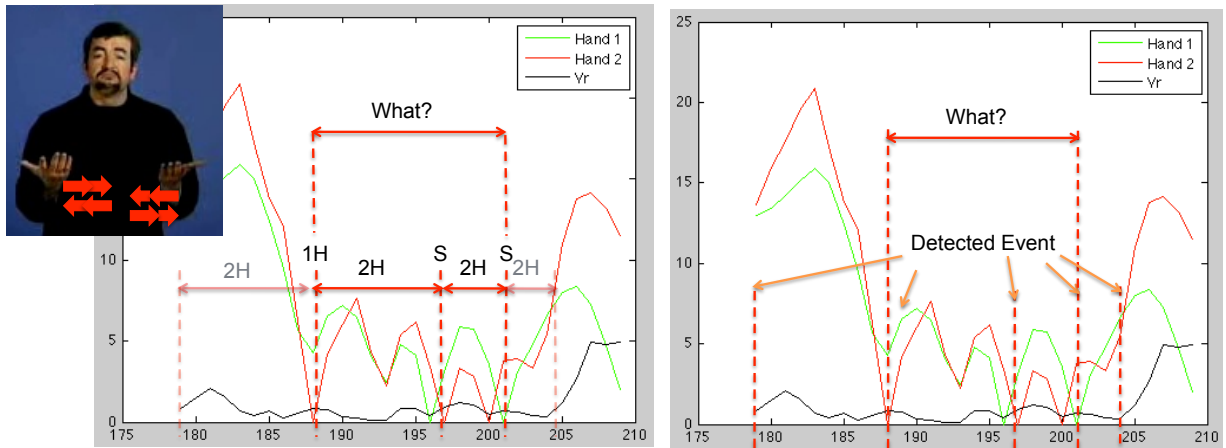


Figure 4.49: Sign [WHAT?] in LSF (top-left). Double arrow represent a repetitive gesture. Hands velocities and the relative velocity with the classification results are shown on the left and the results of the detected events are shown on the right.

4.5.2 Shape features

The next stage of our sign segmentation algorithm consists on the introduction of hand shape to improve over-segmentation. As we mentioned before hand configuration recognition is a challenging task because of the high 2D hand shape variability from a mono-camera view. Indeed same hand configuration may produce different hand shapes. Thus we prefer to characterise hand shape using geometric measurement such as eccentricity and equivalent diameter.

For characterising hand shape we need first of all to extract hand region from a frame using the algorithm described in Section 4.4. Segmentation is performed for each frame corresponding to each detected event from the motion classification stage (Figure 4.51). The adapted segmentation algorithm is selected regarding whether the hand is in front of the face or not.

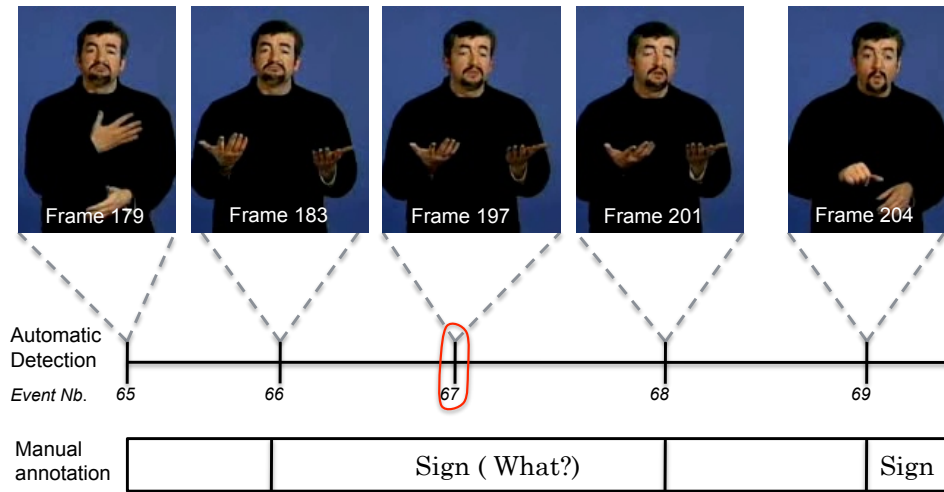


Figure 4.50: Frames corresponding to the detected events.

Hand shape is systematically compared with the hand shape from the previous and following event. If hand shape is similar to the shape of previous and following events, we assume that this event has been detected in the middle of a sign and it is possible to remove it. This assumption is validated during our experimental results.

Figure 4.51 show the *Automatic Detection* aligned to the *Manual Annotation* for the sign [WHAT?] in French Sign Language. Notice that the event (*Nb.67*) has been detected on the middle of the sign. We intend to remove this event considering that hand shape has not changed. As it has been explained before the segmentation step using motion features over segmented this sign. Now to correct the over-segmentation, hands are segmented for each detected event, Fig. 4.51 *on the top*. Notice that the hand shape is similar to both shapes neighbouring the event. It is, then, possible to remove this limit to correct the segmentation.

In order to compare hand shape between detected events it is necessary to extract shape features. Geometric measurements are used to determine hand shape similarity between shapes. Several geometric measurements have been described in Section 3.5.2.2. We propose to use two measurements: equivalent diameter, ε_d and eccentricity ε , Table 4.4. The former specifies the diameter of a circle with the same area as the region. The latter represents the eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The advantages of these measurements is that they are invariant to rotation and translation. However the inconvenient is the sensibility to scaling and noise.

Figure 4.52 shows the eccentricity and the equivalent diameter obtained for each detected event for the segmentation example of the sign [WHAT?]. Notice that the equivalent diameter remains constant for the events neighbouring the event *Nb.67*. Thus following our assumption which states that one event having the same shape that the

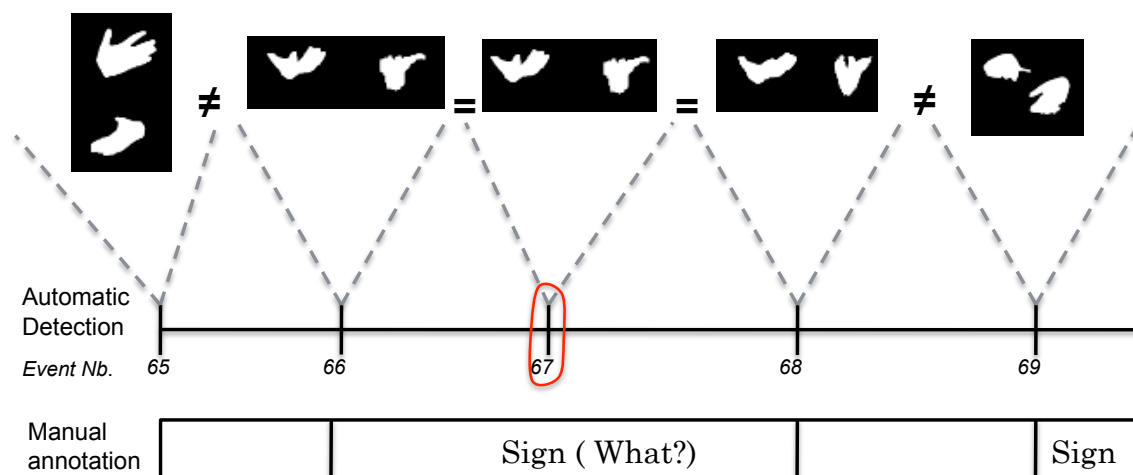


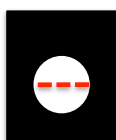


Figure 4.51: Illustrates hand segmentation for each detected frame.

Table 4.4: Similarity measurements used for hand shape characterisation.

| Similarity measurements | Eccentricity ϵ | Equivalent diameter Φ |
|---|---|---|
| Description | Eccentricity from ellipse with second moment equal to region second moment. | Diameter from a circle with same area than the region |
| e.g.  |  |  |

previous and following event corresponds to an event in the middle of a sign, this event can be removed.

4.5.3 Experimental results

Herein we intend to evaluate our approach to show its performance and limitations. We have argued in the introduction that a lack of sign border definition from a linguistic point of view makes manual annotation full subjective and dependent on the annotators knowledge and experience. Evaluation is challenging since a ground truth for comparison is required. Generally the ground truth is obtained by performing the manual segmentation by an expert on SL. However evaluation results will also depend on the annotator. Then our approach cannot be straight forward compared with other methods since so far,

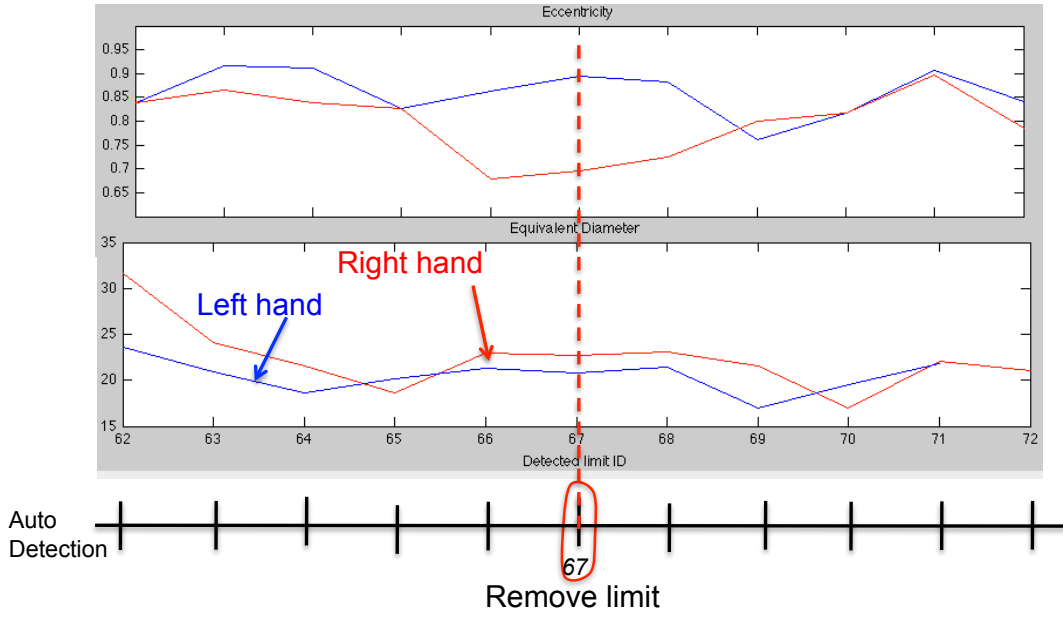


Figure 4.52: Eccentricity and equivalent diameter measurement.

there is a lack of consistency on the ground truth and on the boarder limits definition.

In this work we have asked a deaf native signer to manually segment sequences according to the criteria we have used for defining limits; stable motion and configuration. Thus sign preparation movement is not considered within the sign and the comparison between the automatic and the manual annotation is consistent.

The evaluation has been performed on two sequence without any language or performance speed constraints: LS Colin [LS-COLIN 2002] and DEGELS [Boutora 2011], see Section 3.2.2. Performances remains very natural and representative of the language. Our segmentation algorithm has been tested on 2500 frames, about 100 signs.

Our evaluation criteria consists on two measurements; the true positive rate (TPR) and the false positive rate (FPR). The TPR corresponds to the number of detected events e_m that do match a limit l_n manually annotated over the total number of limits N selected by the annotator. This is derived as

$$TPR = \frac{1}{N} \sum_{n=1}^N c_n \quad (4.45)$$

where c_n represents the matching between the limit l_n and the closest event e_m , defined as

$$c_n = \begin{cases} 1 & \text{if } l_n - \delta \leq e_m \leq l_n + \delta \\ 0 & \text{otherwise} \end{cases} \quad (4.46)$$

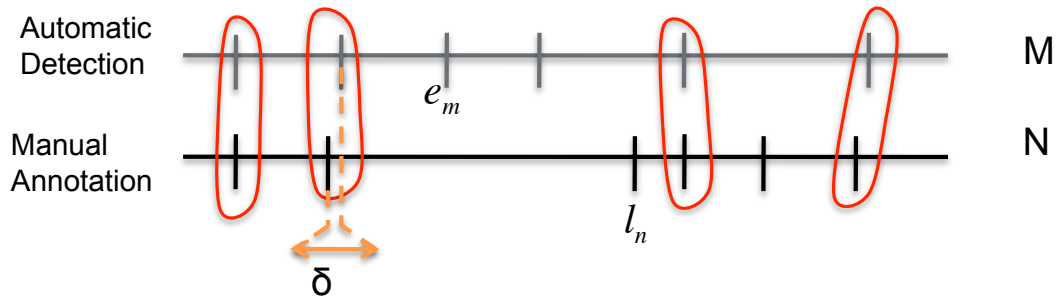


Figure 4.53: Evaluation criteria.

with δ a tolerance corresponding to the number of frames between the detected event and the manual selected limit so that the event is considered as correct.

The FPR corresponds to the number of detected events that do not match any limit on the ground truth segmentation over the total number of detected events M . This is derived as

$$FPR = 1 - \frac{1}{M} \cdot \sum_{n=1}^N c_n \quad (4.47)$$

Figure 4.53 illustrates the detected events e aligned to the manual limits l . The matching between the automatic detection and the manual segmentation according to a tolerance δ is encircled in red. The TPR is proportional to the number of matchings and the TPR to the not matched events.

In order to validate our segmentation approach we have computed the evaluation rate for a sequence of 1000 frames (100 limits) from the corpus LS-Colin [LS-COLIN 2002]. Here the tracking ground truth and the manual segmentation of hands have been used. Table 4.5 shows the results obtained for different δ values using only our motion classification method and also with the introduction of hand shape.

Notice that as expected the TPR increases with δ while the FPR decreases. What is important from these results concerns the assumption made for the introduction of hand shape. Indeed we assumed that if the hand shape remained the same during at least three events only the bordering events will be conserved since other events are considered to be the over-segmentation in the middle of a sign. In fact before and after the introduction of hand shape TPR remains the same while the FPR decreases after hand shape is considered. This validates our hypothesis since the TPR has been conserved.

The problem faced now is to define the value of δ for which the segmentation is as good as the segmentation performed by a manual annotator. The tolerance δ has been determined through an experience where we have asked a native signer to perform the manual segmentation of the same sequence several times. In fact this will allow us to define the variability of a same annotator, thus if our results are within this variability they

Table 4.5: Evaluation results for motion segmentation and motion with hand shape improvement for several tolerance values.

| δ | Motion | | Motion + Hand shape | |
|----------|--------|--------|---------------------|--------|
| | TPR(%) | FPR(%) | TPR(%) | FPR(%) |
| 0 | 33.3 | 79.3 | 33.3 | 78.0 |
| 1 | 73.7 | 54.3 | 73.7 | 51.4 |
| 2 | 86.8 | 46.2 | 86.8 | 42.8 |
| 3 | 91.2 | 43.5 | 91.2 | 39.9 |
| 4 | 93.9 | 41.8 | 93.9 | 38.2 |
| 5 | 96.5 | 40.2 | 96.5 | 36.4 |

are considered as good as the manual segmentation. We have noticed, as expected, that the same annotator have not selected exactly the same frame for each limit. Figure 4.54 illustrates the manual annotation obtained for three tiers. We have computed the average gap $\bar{\Delta}$, Eq. 4.48, from the gap at each limit Δ_i . In this example this corresponds to $\bar{\Delta} = 3.7$ frames, thus the tolerance $\delta \approx 2$.

$$\bar{\Delta} = \frac{1}{N} \sum_{i=1}^N \Delta_i \quad \delta = \frac{1}{2} \bar{\Delta} \quad (4.48)$$

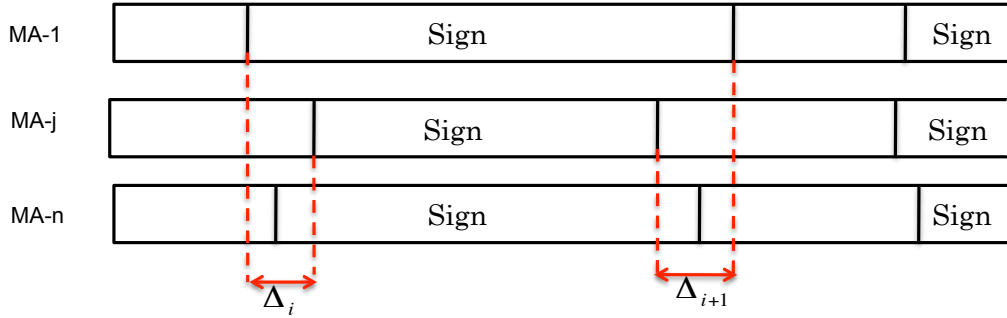


Figure 4.54: Manual annotation illustration.

Now that we have validated our assumption, we would like to point out the influence of our automatic tracking algorithm (Sec. 4.3) and automatic hand segmentation approach (Sec 4.4). The results from these methods are not corrected since we want to point out their robustness for the automatic segmentation of signs. Table 4.6 shows the segmentation results using motion from ground truth and automatic tracking. Notice that the influence of errors in the tracking decreases the TPR af about 5% and increases the FPR of about 1.5%. In other words something about 5 limits have not been detected and about 2 events have been detected but do not correspond to any manually segmented

limit. Introducing automatic hand segmentation does not degrade our TPR rate and reduces the FPR of about 3% being lower than the FPR using the motion classification with the ground truth. This show that the errors introduced by our tracking and hand segmentation algorithm is very low.

Table 4.6: Segmentation results using the automatic tracking and our automatic hand segmentation approach for a tolerance $\delta = 2$

| | Motion | | Motion + Hand shape |
|----------------|-----------------------|---------------------|--|
| | Ground Truth Tracking | Automatic Tracking* | Automatic Tracking and Hand Segmentation |
| TPR (%) | 86.8 | 81.6 | 81.6 |
| FPR (%) | 46.2 | 47.8 | 44.9 |

Table 4.7 shows the True positive rate *TPR* and the False positive rate *FPR* for both sequences for $\delta = 2$; LS Colin [LS-COLIN 2002] and DEGELS [Boutora 2011], see Section 3.2.2. Notice that the *TPR* for motion and motion plus hand shape remains stable, 81.6% for LS-Colin and 74.5% for DEGELS, while the *FPR* decreases of about 3% for LS Colin and 10% for the DEGELS corpus when hand shape features are introduced. In these results we notice that over 74% of the events are detected with an accuracy of ± 2 frames which corresponds to the variability of a manual annotation.

Table 4.7: Segmentation results for a full-automatic sign segmentation with automatic tracking and hand segmentation algorithm and for $\delta = 2$

| | Motion | | Motion + Hand Shape | |
|-----------|--------|--------|---------------------|--------|
| | TPR(%) | FPR(%) | TPR(%) | FPR(%) |
| LS- Colin | 81.6 | 46.2 | 81.6 | 44.9 |
| DEGELS | 74.5 | 54.2 | 74.5 | 44.7 |

We have evaluated the usefulness of our approach by determining the time needed by a native signer to manually segment a video sequence in comparison to the time needed for correcting the automatic segmentation. The time needed is quite similar (10 min for LS-COLIN and 4.5 min for Degels). The manual annotation and the correction have been performed using the annotation Tool ELAN [Wittenburg 2006] which has been developed to perform manual annotation and is not adapted for correcting semi-automatic

segmentation, since it is needed to select the beginning and the end of each sign. After correction of the automatic segmentation results we noticed that only 11.25% for LS-Colin and 7.1% for Degels from the FPR correspond to events detected in the middle of signs. This means the remaining over-segments transitions and not signs.

4.5.4 Conclusion

Temporal segmentation is a difficult task mainly because of the co-articulation effect. In addition the lack of linguistic information concerning sign borders definition make the segmentation challenging. Here we propose a method using objective measurements for proposing sign limits to annotators. This novel approach use only low level features for detecting events so that annotator could correct segmentation and label sequences as signs.

In this section we have presented a method for temporal segmentation addressing the problem of automatically segmenting large amount of continuous SL video corpora. Sign segmentation is addressed into two levels of detail, segmentation and improvement, based on low level features. The former concerns hand motion features automatically extracted using the proposed robust tracking algorithm. The latter is a correction step that uses hand shape to remove wrong limits and correct the segmentation from the first level. For this we assume that contiguous events having the same hand shape corresponds to an event in the middle of signs. This is verified in the evaluation. The introduction of hand shape for temporal segmentation from image processing techniques, has not been seen in the literature since only methods using motion capture techniques were able to extract hand shape features.

This approach has shown promising results. Although the time needed to correct the automatic segmentation results is equivalent to the time needed to fully manual annotate, the annotation becomes less dependent to the annotator's knowledge and is reproducible and these are some of the drawbacks of the manual annotation.

The proposed method allows to find sign boundaries for processing continuous SL. In addition the temporal structure of signs can be determined using only the first segmentation level. The results from this stage are used for the recognition of glosses using a linguistic description of signs. Indeed the linguistic representation chosen is based on a temporal description of signs and this feature can be used for filtering glosses that could correspond to the performance. The proposed approach can be used for any sign language or any other gesture based application since only low level features are used. The next step of our research concerns the labelling of the sign in order to perform a semi-automatic annotation in terms of glosses.

4.6 Semi-Automatic Annotation of Glosses

Semi-automatic annotation of glosses consists on transcribing one language into another² using linguistic knowledge. The problem faced concerns the way in which linguistic information could be used, so that in combination with features, annotated at a phonemic level, we are able to recognise glosses without any vocabulary restriction.

Gloss annotation could be addressed as a sign recognition problem (see Sec. 3.6), however we have mentioned that approaches leading with sign recognition require high amounts of training data making results dependent on the representativeness of this data. Training datasets are collected from the annotation of SL video corpora. This points out the need of avoiding any training data for gloss annotation.

In this PhD thesis we propose a novel approach using a linguistic representation of signs. The chosen approach, Zebedee (Sec. 3.4.2), is based on a temporal representation. Although this sign descriptor has been developed for SL synthesis we intend to use it for SL annotation. The main advantage of this descriptor is its modulation capability. Indeed the same description can fit a sign independently of its context and variability. This avoids the annotation of each different performance of a sign. The inconvenient for SL recognition is the lack of consistency on the description since the same characteristic can be annotated in several ways, making the matching between image features and described features in ZeBeDee a challenging task. This problem is addressed by extending ZeBeDee sign representation with features that are straight forward comparable to features extracted from videos using image processing techniques.

Expected results from the assisted annotation of glosses consist on a list of potential glosses corresponding to the performed sign. From this list, annotator can select the gloss not only for the preformed sign alone but considering its semantics on the discourse. Thus assisted annotation methods could only suggest some glosses that could potentially match the performance instead of seeking for proposing only one gloss. In fact this would require information at a higher level, e.g. semantic, grammatical, etc. which is out of the scope of this PhD thesis. Here only a lexical recognition of signs is performed. In this case homosigns³ cannot be distinguished unless other linguistic models are used.

Figure 4.55 shows an example of the sign [PASSPORT] and the sign [EXAM] in LSF which are homosigns. In addition to these kind of signs there are locative signs as [SURGERY] which can be located according to the place in which the surgery has been performed. Performance of this sign can be similar to [BOY] if the surgery concerns the head. At our level of recognition we cannot make any distinction, see Figure 4.56. For this reason we find much more appropriated to propose a list of glosses to annotators who can easily select the one corresponding not only to lexicon but also to the context.

2. In this PhD LSF is transcribed sign-by-sign to written French but our approaches are not constrained to LSF.

3. Signs that are performed equally but correspond to different meanings as homonyms in oral languages [Girod 1997].



Figure 4.55: [PASSPORT] and [EXAM] in LSF are homosigns. Source IVT [Girod 1997]



Figure 4.56: [BOY] and [SURGERY] in LSF could be homosigns if is a head surgery. Source IVT [Girod 1997]

The sign database annotated in ZeBeDee is composed of ≈ 1600 entries. This database has been annotated at LIMSI⁴ who has kindly put it at our disposal for performing this work. In addition to the database, some tools for filtering are at our disposal. These filters implemented at LIMSI allow to obtain a list of signs verifying a define predicate. This has been described in Section 3.4.2.

Features that can be used for filtering are:

- **Name** : Obtain descriptions whose name matches the predicate
- **Deps** : Obtain the dependences expressed in the description
- **DepCount** : Obtain the number of dependences
- **TransCount** : Get all the signs having n number of transition "T" in the movement structure
- **TimeStruct** : Obtain the time structure, i.e. the sequence of key postures "K" and transition "T"
- **MvtStruct** : Obtain the movement defined for each transition "T".

4. Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI) www.limsi.fr
Group: Information, Language, writtEn and Signed Group (ILES), Team Sign Language (LS)

Using these available tools we have first of all studied the annotated data in order to point out representative features that can be used for classifying glosses. Features that can be used are the *Transcount* and *MvtStruct*. Other tools described before only allow to get information that cannot be extracted from video. For instance the number of dependencies is not relevant since the performance in the video sequence corresponds to a well defined set of DEPS which is unknown by looking only at the performance. Other filtering tools are hardly developed because of the inconsistency on the annotation (same feature described in several ways), thus features that could be easily obtained from image processing techniques cannot be extracted from ZeBeDee, e.g. relative position of hands with respect to the body, motion direction, symmetry, etc.

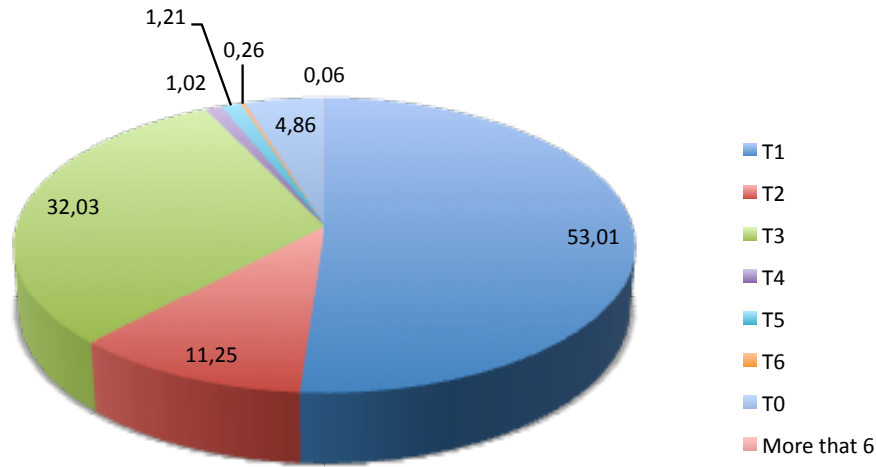


Figure 4.57: Classification of sign within the database in terms of the number of transitions.

Figure 4.57 shows the distribution of signs, in the database, according to the number of transitions obtained using the *TransCount* filter. Notice that more than the 50% of signs correspond to one transition $1T$ so a time structure KTK , followed by 30% and 10% for 3 and 2 transitions respectively. The command to obtain signs corresponding to n number of transitions is for example *FILTER transcount ~ "n"*. The temporal structure of signs *TimeStruct*, sequence of key postures and transitions, or which is equivalent the number of transitions *TransCount* gives important information that allows to reduce significantly the number of signs belonging to a class. Using this feature at some level of detail we are able to reduce potential glosses.

The second feature that can be used directly for Zebedee corresponds to the *MvtStruct*. For generation all signs are composed of one or various basic paths: Arc (A), Straight (S) or Circle C , which are defined at each transition for each hand. Table 4.8 shows the number of signs for each basic path for signs belonging to the one transition class. Notice that only about 15% of signs have a defined path in the transition. However most of the signs, about 83%, do not have any defined path. This is because what matters in this kind of signs is the initial and final position and it has been decided at LIMSI not to explicitly annotate it.

Table 4.8: Movement structure statistics (%)

| | S | C | A | - | 0 | N/A | Total |
|--------------|------|-----|------|-------|------|------|-------|
| Nb. of signs | 22 | 12 | 68 | 618 | 20 | 5 | 745 |
| (%) | 2.94 | 1.6 | 9.12 | 82.95 | 2.68 | 0.67 | 100 |

Table 4.9: Movement structure statistics for 1T (%)

| | | Strong Hand | | |
|-----------|---|-------------|------|------|
| | | S | A | C |
| Weak Hand | S | 35.7 | 0 | 0 |
| | A | 0 | 60.8 | 0 |
| | C | 0 | 0 | 3.46 |

For two hands we have looked to the movement structure combination for strong and weak hand. According to what has been stated in [Battison 1978], in ASL two-hand signs have the constraint of being symmetric on hand-shape, movement and location, though this is not respected in some SL. We verified that in our database it has been respected for LSF, Table 4.9. This is quite logical since performing one kind of movement with one hand and another with the other is very difficult from a human motion point of view. However this can be done subconsciously when one hand performs a sign and the other moves to prepare the following sign (Sec. 2.3.4).

Description of signs in ZeBeDee is performed offline and only once per sign regardless the context or any variation since the most important advantage of this descriptor is its generality. Thus adding a new sign to the vocabulary requires the introduction of its description in ZeBeDee.

Recognising signs performed in a video sequence using the sign database described in ZeBeDee requires the extraction of features from video that correspond to the features described in ZeBeDee. The features that can be extracted from :

- **video** correspond to the visual features obtained from our image processing algorithms previously described. It is possible to extract the location, the path, the temporal structure, direction of hands motion, symmetry, hand shape, etc.
- **sign description (ZeBeDee)** correspond to what is consistently annotated and that can be filter with the available tools such as the time and movement structure. Other features have to be introduced for matching visual features from videos.

Although we have argued (Sec. 3.4.3) that directly recognition from ZeBeDee is difficult because of the high variability concerning the annotation. Herein we present a classification methods using few features that can be directly extracted from the description in order to show that the vocabulary size is highly reduced. Other features cannot be directly extracted from ZeBeDee and will be introduced in an extended version ReZeBeDee for recognition purposes.

4.6.1 Gloss recognition from querying Zebedee

Visual features are extracted using image processing techniques to query the database of signs described in ZeBeDee. In this way glosses whose description match the performed sign can be proposed to the annotator.

Using ZeBeDee, straight forward for gloss recognition might not be the optimal solution since it has been designed for SL synthesis. Herein we describe an approach for querying directly ZeBeDee using the available filter at our disposal which are very few; *TransCount*, *TimeStruct* and *MvtStruct*, being *TransCount* and *TimeStruct* equivalent. We have shown in the introduction the distribution of signs concerning the number of transitions *TransCount* which reduces significantly the number of potential signs at least by half for *T1*. The number of transitions can also be obtained using image processing techniques, particularly with our temporal segmentation approach, see Section 4.5. The second feature that could be used from ZeBeDee corresponds to the movement structure *MvtStruct*. Three kinds of paths are described; Straight, Arc and Circle, which can be detected from video using the body tracking algorithm proposed in this work, see Section 4.3. From the movement structure we can determine the number of moving hands which can be also detected using velocity features. We propose a descending classification method composed of three levels where each level corresponds to a feature extracted from a video sequence and explicitly described in ZeBeDee. The proposed classification tree is composed of three levels and is illustrated in Figure 4.58.

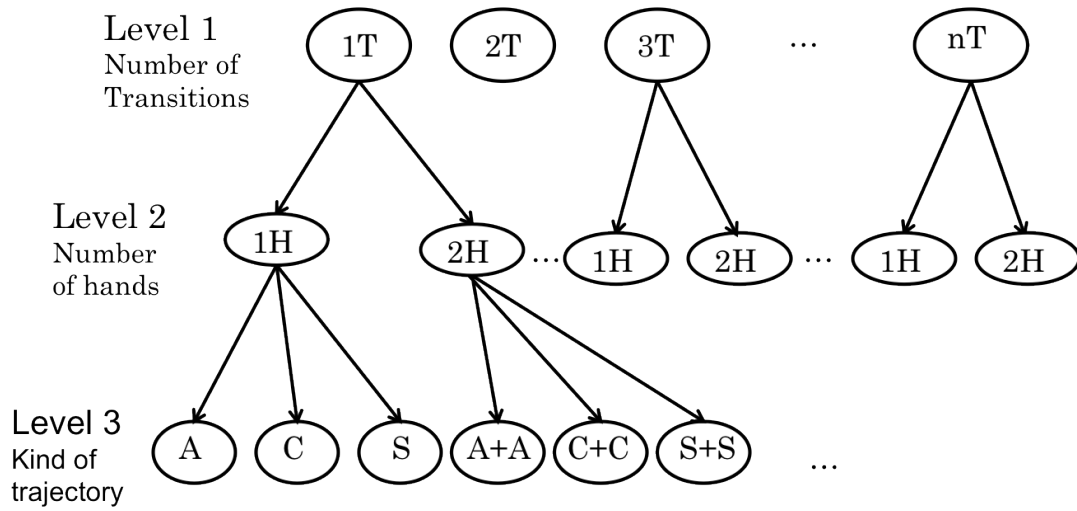


Figure 4.58: Gloss classification tree

These three classification levels are used differently from ZeBeDee and from image processing techniques. Herein we describe how signs can be filtered using the features at each classification level and how sign characteristics are extracted from video also at each classification level. Indeed we have to be able to filter signs within our database verifying a predicate which is defined from visual features.

4.6.1.1 Described signs from ZeBeDee

Sign descriptions are stored in a PostGres database. To filter the descriptions, we use a dedicated command-based interface to more complex SQL queries. Its *FILTER* command allows to narrow down the list of descriptions, given a predicate that accepts or rejects description entries. The command syntax used for obtaining signs is

$$FILTER\ f \sim "RE" \quad (4.49)$$

where f corresponds to the considered feature label, e.g. *timestruct*, and RE represents the regular expression to be matched corresponding to the predicate. The levels defined in our classification tree from ZeBeDee are defined and filtered as follows.

- **Level 1** corresponds to the number of transitions composing descriptions. Indeed we have mentioned that in ZeBeDee signs are described as an alternating sequence of key postures K and transitions T called *Time Structure*. Thus counting the number of transitions in the sequence allow us to reduce the list of potential signs. The defined label is *transcount* and the command for obtaining the list of signs composed by n number of transitions is

$$FILTER\ transcount \sim "n". \quad (4.50)$$

- **Level 2** corresponds to the number of moving hands. Although there is not a explicitly defined filter for obtaining signs performed with one or two hands this can be done using the movement structure in the description. Indeed signs performed with two hands are linked by the symbol $+$ in the middle of the path corresponding to the right and left hand. For instance for the sign [SHOCKED] in LSF, corresponding to a straight movement for both hand, the *MvtStruct* is $S + S$. This means that right and left hand have the same straight S path for a given transition T . Thus filtering descriptions performed by two hands is performed using the command line

$$FILTER\ mvtstruct \sim ". * +. *". \quad (4.51)$$

For one hand signs the regular expression is then $"not \sim ". * +. *"$ which means that the symbol $+$ is not in the movement structure.

- **Level 3** takes into account the kind of trajectory inside a transition T . Trajectories are of three kinds: Arc A , Straight S and Circle C . Some signs might not be associated to a trajectory if the intention is not to perform a straight movement. For example the intention of going from one point to another the trajectory is not described though it corresponds to a straight line. The regular expression for this feature correspond to the movement structure, e.g. one hand moving in a circle trajectory C . The command for this feature is

$$FILTER\ mvtstruct \sim "C" \quad (4.52)$$

These three filters are used for obtaining the list of descriptions matching the predicate. The regular expression to match is extracted from the performed sign in the video for which we want to propose potential glosses. Defining the predicate from video, corresponds to the extraction of visual features using image processing techniques.

4.6.1.2 Visual feature extraction

The number of transitions, the number of hands and the kind of trajectory can be detected by image processing techniques and used to filter signs in the description database. The image processing methods developed to define the filter predicate have been described in previous sections. Motion features such as the path, can be extracted using the results from our body part tracking algorithm. Tracking results are used to compute hands velocity and acceleration. Also our temporal segmentation approach is used to determine the number of transitions T composing a sign. Predicate at each classification level is defined as follows.

- **Level 1** corresponds to the number of transitions T in the sign. This is determined using our temporal segmentation approach (Sec. 4.5). For obtaining the number of transitions only the first segmentation step is used. In fact we have argued that the over-segmentation performed using only motion features correspond to the key postures. Since the time structure of a sign is a sequence of key postures K and transitions T where the number of transition T_{nb} is

$$T_{nb} = K_{nb} - 1 \quad (4.53)$$

with K_{nb} the number of key postures K .

- **Level 2** corresponds to the number of hand performing the sign. It is determined using the ratio between the average velocity of one hand and the greatest average velocity for right or left hand. The number of moving hands is determined using the ratio r between the difference of average velocities of right \bar{v}_1 and left \bar{v}_2 hand and the maximal average velocity, see Eq. 4.54.

$$r(v_1, v_2) = \frac{\|\bar{v}_1(t) - \bar{v}_2(t)\|}{\max\{\bar{v}_1(t), \bar{v}_2(t)\}} \quad (4.54)$$

If this rate is low that means that both hands move together, otherwise one hand moves much faster than the other. The main problem raises when we process continuous sign language. In this case signs are influenced by the previous sign and itself influences the following sign. For example when a two-hand sign follows a one-hand sign, signers tend to prepare the following sign by moving the weak hand to the beginning location of the two-hand sign. This is addressed using some statistics performed in our description database, Table 4.9. For instance for 17 no sign performed by two hands have different kind of trajectory for right and left hand, e.g. the movement structure A+S, where A corresponds to an arc for right hand and S to a straight movement for left hand, is not inside our database. Indeed it is hardly performed by a person. Using this we can deal with the preparation movement done during continuous SL.

- **Level 3** corresponds to the path followed by hands during the transition. Here the expectation results from our tracking algorithm E are used for determining the trajectory followed by hands. We intend to identify the basic paths described in ZeBeDee: Circle, Straight and Arc.

A circular trajectory is detected using the distance d_n between the first f_0 and the last point f_N of the trajectory normalized by the total length of the curve. This is derived as

$$d_n = \frac{E_{h_i}(f_1) - E_{h_i}(f_N)}{\sum_{j=1}^N E_i(f_j)}, \quad (4.55)$$

where $E_i(f_j)$ represents the tracking result for object h_i either right or left hand and f_j the frame j with $j \in [1, N]$ and N the number of frame in the transition T . For a circle C , d_n is a low value and for an arc A or a straight S movement is close to 1. This allows to distinguish the signs with a circular trajectory but not arc or straight trajectories can be classified from this measurement.

Straight S and Arc A trajectory have to be differentiated in another way. For this we perform a linear regression and compute the ratio r^2 which give some information about the quality of the fitting. Good quality leads to r^2 close to 1 and means that the fitting has been well performed otherwise the trajectory corresponds to an arc.

Using the features extracted from a video sequence we are able to determine predicates for classifying signs according to our classification tree. Then a list of potential glosses can be proposed to the annotator. Decreasing the number of proposed signs leads to improve the classification tree which depends on the descriptions of signs. For example image processing techniques are able to classify hand shape, however a hand shape Zebedee filter is difficult to implement because the same hand configuration can be described in several ways. The same problem is faced for signs described in terms of a relative position. For instance placing a finger close to the face could be described using the front or the nose position.

We have presented an approach to assist the annotation using a lexical description of signs. Here the description of signs is straight forward used for gloss recognition. This approach extracts image features from video corpora to query a sign description database and propose the potential glosses that could correspond to the performed sign. This approach can only use features that are consistently annotated and that can be directly extracted from computer vision approaches. Only few signs can be processed, e.g. for signs in the one transition class $1T$ only about 17% of signs have explicitly described the path. This could be improved if some annotation rules are defined. For example explicitly defining the number of hands performing the movement event though this is not required for SL generation. Instead of that we propose a novel approach which uses the generated data, i.e. synthetic data, for extending the description adapting it to SL recognition.

4.6.2 Gloss recognition from ReZeBeDee

The approach described above uses descriptions of signs to narrow down a list of potential glosses that verify a predicate. Although this represents a good and easy solution, the features that can be used are very few and the vocabulary size is reduced to signs where the kind of trajectory has been described. However we remain that ZeBeDee has been designed for SL synthesis, thus consistency on the description is not considered as a major problem as long as the resulting generation is equivalent.

Unlike the previous approach which uses very few features directly extracted from ZeBeDee, herein we intend to extend ZeBeDee, called ReZeBeDee adding visual characteristics that are easily detected from automatic methods but that are hardly manually annotated, e.g. relative distance between hands or the path trajectory direction in degrees. For this automatic image processing methods have to be developed for extracting visual features from representative training data. Extracted visual features from the data can easily be added to the description. A major problem concerns the collection of the training data. Indeed collecting representative training data for each sign in our database is very difficult because it requires a complex recording set-up, various native signers and the signs performed several times in several contexts. Thus having enough data for each sign for adding representative information to the ReZeBeDee description is challenging. Also this would represent that for each unknown sign we would require, in addition to the manual description in ZeBeDee, a high amount of data for extending the description.

In this PhD thesis we propose a novel approach for avoiding the need of video corpus for extending the description. Here we propose to use synthetic data for extracting visual features and extending the descriptions in ZeBeDee. The automatic extension is carried out using synthetic data collected through the automatic generation from ZeBeDee. We have argued the advantage of using a linguistic representation of signs, particularly the use of ZeBeDee because of its modularity. It considers the variability of signs through some dependencies parameters. The generation of the data is randomly performed several number of times for different values assigned to the dependencies. Unlike approaches using training data which are dependent on the representativeness of the data. Here we have the advantage of being able to generate any sign. Thus integrating an unknown sign to the database only needs the description in ZeBeDee since the extension is automatically performed to the ReZeBeDee database.

Automatic SL generation from ZeBeDee takes into account comfort measurement and gives us the position of hands and head at each key posture K composing the sign. This is what we obtain from our tracking algorithm. From generation we extract possible performances of a sign to extract common features adding it to the description. Later extracting the same features from video and querying the ReZeBeDee database we propose a potential list of glosses to annotators. The advantage of this is that the vocabulary database is neither restricted nor constrained to a context.

4.6.2.1 Visual features from synthetic data

Synthetic data is collected using the automatic Sign Language generation software GeneALS [Delorme 2011] which uses the linguistic representation ZeBeDee for piloting the software. Generation results consist on the coordinates of hands and head which are used for extracting visual features from synthetic data. The same features can be extracted from video using the results of our tracking approach. This allows to define the predicate to be verified by the list of potential signs.

For extending descriptions in ZeBeDee we propose to generate numerous times a sign using its description with random values of the dependencies, if dependencies have been described. From this we can add a set of possible features seen from generations which allows to consider sign variability and context-dependency. Although the Zebedee description of signs has to be manually performed by an expert on SL, the extension ReZeBeDee, for SL recognition is fully automatic. Features proposed from generation are however verified by an expert in SL and corrected if required.

We consider that this is a good solution for avoiding collecting high amounts of video corpus which might not be representative for the high variability of signs. Here we are able to generate representative data for extracting all the characteristics known for a sign which are added to the description. For example in Figure 4.59 are illustrated the signs [SHOCKED] and [BUILDING] in LSF. The motion and location of both signs are similar, what distinguish them is hand configuration being this one not considered in this work both signs could be proposed to annotators. In addition the sign [BUILDING] is a locative sign which can be placed anywhere in the signing space. The description of both signs in ZeBeDee is presented in Appendix B. Notice that for sign [BUILDING] three parameter dependencies are presented while the sign [SHOCKED] have no dependency. This means that this sign is always performed in the same way independently from the context. Thus a sign with similar characteristics as [BUILDING] and [SHOCKED] performed in another place could correspond to a [BUILDING] in context but cannot be the signs [SHOCKED]. In this way we can narrow down the list of potential signs considering the variability of signs and their context-dependency.

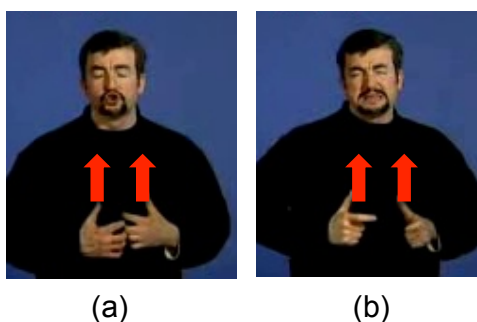


Figure 4.59: Sign [SHOCKED] and [BUILDING] in LSF.

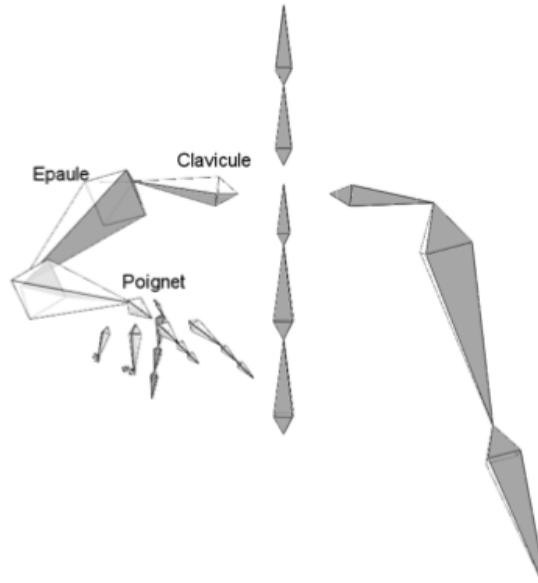


Figure 4.60: Skeleton used for the automatic generation of signs. Source [Delorme 2011]

The characteristics to be added to the description can correspond to any feature extracted from hands and head location. In order to point out the advantages of the proposed approach, here few but representative features are used. From the added features we can narrow down the list of potential signs. This approach could become a full recognition system by adding more characteristics. For example the introduction of hand shape can be used for distinguishing signs with the same motion characteristics but different hand configuration like the example above, Figure 4.59.

In this PhD thesis we only consider features that could be extracted from the position of hands and head, though the position of other articulators could be obtained from the estimated body pose. Indeed the generation software GeneALS uses a predefined skeleton (Fig. 4.60) which have to meet numerous anatomical and linguistic constraints according to the description in ZeBeDee. However we have mentioned that ZeBeDee describes what is meaningful in the sign without describing each single body part configuration, the generation takes into account a comfort measurement allowing to locate other body parts. Thus the same body pose through several generations may lead to different results being all of them understandable. In this case other articulators position are not relevant even though some of them (which are not annotated in ZeBeDee), such as shoulder position, convey linguistic information (Sec. 2.3.3). In addition we cannot obtain the position of all the bones from our tracking algorithm though other skeleton based tracking could approximate the 3D pose estimation. This is out of the scope of this PhD thesis.

An important point to take into account concerns the dependency of our measurements on the skeleton dimensions. We have to be very careful about features related to the skeleton shape and dimensions since they might not correspond to various signers morphology. Thus we prefer annotating relative measurements concerning motion and location.



Figure 4.61: Sign [BUY] in LSF. Linguistically both hands are used for performing the sign but only one hand moves. Source IVT [Girod 1997]

We propose to add information that can also be extracted from video. Data obtained after generation correspond to the 3D coordinates of hands (G_{h_1} and G_{h_2}) and head G_h in the skeleton axis. In our case we do not have the 3D coordinates but only the projection on the plane $X \perp Y$, thus annotation in ReZeBeDee might not concern depth since this cannot be obtained from our tracking algorithm for a mono-view video. Here we propose to extend the ZeBeDee description with the following features

- the number of moving hands
- the movement direction, and
- the relative position of hands with respect to key positions on the body.

The **number of moving hands** correspond to the number of hands (Static, 1H or 2H) that have change their location during the performance of signs. Our definition differs from what one-hand and two-hand signs represent from a linguistic point of view. Indeed from a linguistic point of view a two-hand sign uses both hands to express the meaning however both hands do not necessary move. For example in the sign [BUY] in LSF one hand do not move while the other does, this is illustrated in Figure 4.61. Although it corresponds to a two-hand signs only one hand moves. In this case the annotation corresponds to 1H. We propose to add this feature since that can be easily extracted from video, see following section.

The **movement direction** of hands between successive key postures corresponds to a range of directions seen during several random generations. If the movement direction is a dependency parameter, e.g. for the directional verb [SEND] (see description in Appendix B), the range will be very wide unlike signs where the direction is constant, then the range will be tiny and will depends on the comfort measurement. Indeed in the description the movement is defined as UP, a perfect movement cannot be performed by humans because of the kinematic chain. Here we only consider trajectory direction since we only have the position for each key posture so path cannot be considered. Figure 4.62 shows the label given to each quadrant direction. The example shows that hand moves horizontally the given label would correspond to II and III since it corresponds to a limit in quadrants.

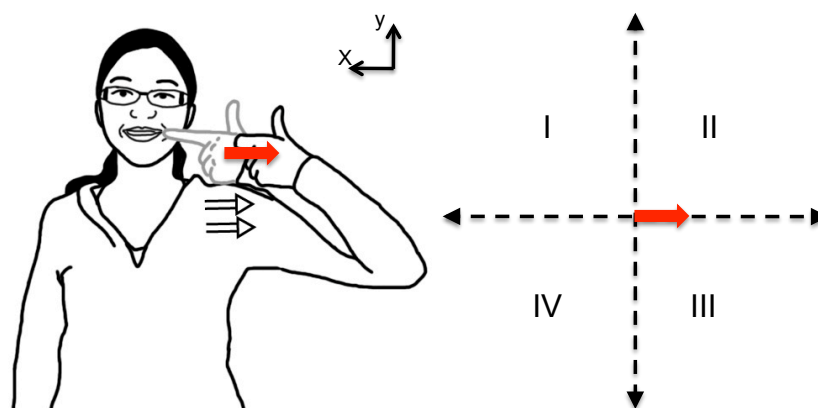


Figure 4.62: Movement direction quadrants.

The **relative position** consists on dividing the space on various sectors according to specific positions on the body. For each key posture each hand has a label corresponding to the sector in which it has been placed. Sectors and the associated labels are shown in Figure 4.63. Sectors limits have been determined using key positions on the skeleton which can be spotted in the image using anthropometrics. For example the limit of sector CENTER corresponds to the shoulder coordinates for lateral and to the CHIN for top. Limits between sectors correspond to the superposing of neighbouring sectors. When the hand is located in this intermediate sector it is labelled with both neighbouring sectors. Key positions correspond to the clavicle, shoulders, thorax and chin as illustrated in Figure 4.63. Locative signs as [BUILDING] (see Appendix B) which can be placed anywhere in the signing space might have several labels unlike signs with a constant location which can only be placed in one or in neighbouring sectors.

These features are automatically added to the initial ZeBeDee annotation file as a new xml tag at the beginning of the xml file, e.g. <NumOfHands> which are easily parsed. Now the ReZeBeDee file have features that are straight forward obtained from video. Adding all the possible labels seen during numerous generations handles sign variability and context-dependency. This respects the original file which can be used either for generation or recognition reaching the same goal as the original annotation, then the same description can describe the same sign in various contexts either for generation or for recognition. Also since the extension to ReZeBeDee is automatically performed from the generations this do not involve additional work to annotators.

Other features can be added like symmetry between hands, e.g. sagittal, central and alternated [Lefebvre-Albaret 2010]. What is important from these features and that has to be respected when adding further annotation is that they are not dependent to the coordinates in the skeleton axis. These new features are used in addition to what can be obtained directly from ZeBeDee (number of transitions and movement structure when defined), only the number of hands is obtained from the annotation and not from the

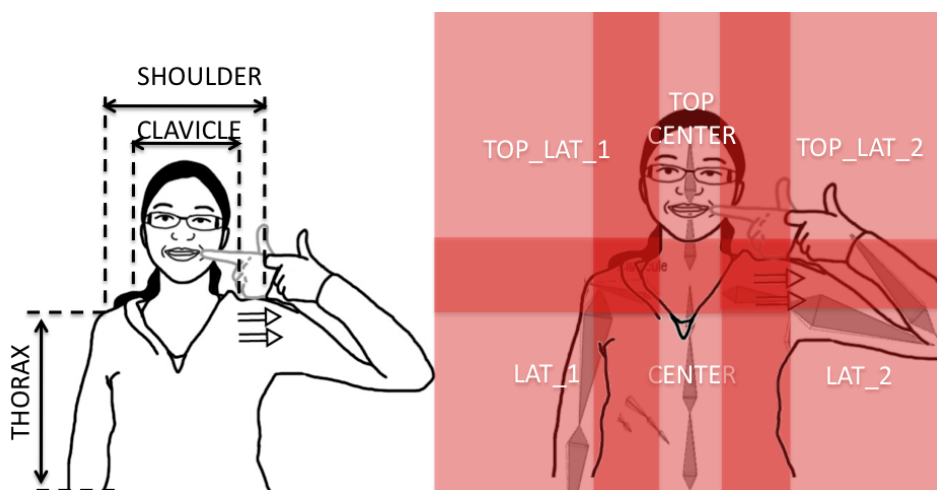


Figure 4.63: Signing space sectors. Hand is labelled according to the sector and its neighbours.

movement structure. Once we have the new features in the description we have to extract the same features from video using image processing techniques.

4.6.2.2 Visual features from image processing

Here we detail how the same features that have been extracted from synthetic data, can be defined from image processing techniques. These features are straight forward obtained from the tracking result and the temporal segmentation result. Features are extracted from video sequences as follows:

- **Number of moving hands** can be directly obtained from our motion classification step performed during the temporal segmentation (Sec. 4.5.1). The classes chosen in the classification correspond to static sign (S), one hand moving (1H) or two hands moving (2H) and are equivalent to what is annotated from synthetic data. For these features no further processing is required since they have already been performed for temporal segmentation.
- **Movement direction** is straight forward obtained from the position of hand at each key posture. Key postures have been defined using our segmentation approach (Sec. 4.5). It corresponds to the direction of the vector from the position of one key posture to the following. Intermediate frames are not considered to obtain similar results that the one obtained from generation. All signs for which this direction is a possibility will be proposed. For example in the sign [BUILDING], movement direction quadrants are I and II since it goes UP, thus this sign will be proposed for any direction between 0 and 180 degrees, other signs are rejected.
- **Relative position** requires the division of the space in the same way as it has been

Table 4.10: Feature classification results

| Gloss | Ground truth | |
|--------------|--------------|-------|
| | Nb. H | Traj. |
| Shoulder bag | 1 | A |
| Deaf | 1 | A |
| We | 1 | C |
| Give | 2 | C+C |

done for the annotation. For simplicity the key positions used in the annotation are manually selected in the first frame. Then for any following frame this is adjusted from the head position. Thus we are able to divide the space into several sectors, the same sectors used with the skeleton in the automatic generation. Although we are aware that shoulders and trunk motion play an important role in SL, only manual features are considered in ZeBeDee and in the generation .

The features extracted here correspond to the predicate that has to be verified. All the signs verifying all the predicates are proposed to the annotator. This method allows to introduce additional information in the description, adapted for recognition purposes. The proposed extended version of ZeBeDee is called in this PhD thesis as ReZeBeDee and is evaluated for pointing out the advantages and limitations of our approach.

4.6.3 Experiments and results

We have performed some experiments using ZeBeDee and ReZeBeDee. Experiments have been performed on the French Dicta-Sign corpora where vocabulary remains completely free. Glosses have been manually segmented and annotated. Table 4.10 shows some glosses with the number of hands and the kind of trajectory for 1*T*. Because of the novelty of our approach it is difficult to perform a comparison to any related work. However we show in this section some encouraging results.

4.6.3.1 Annotation from ZeBeDee

A selection of 95 signs with different number of transitions, number of hands and kind of trajectory (specified in ZeBeDee) is used to perform the experiment. Our experiment considers signs belonging to the 1*T* class which corresponds to 50% of signs in the selection. Table 4.11 shows the features extracted for some tested signs, number of hands -*column: Nb. H*- and kind of trajectory -*Column: Traj*- with and without statistics improvement (see Table 4.9). Notice that the performance of signs [SHOULDER BAG] and [DEAF] in different context do not lead to the same extracted features result because of the co-articulation effect, see Section 2.3.4. Indeed without considering statistics, possible trajectories combination between strong and weak hand shown in table 4.9, the results

are influenced by the context and do not correspond to the ground truth, see Table 4.10.

Figure 4.64(a) shows the sign [DEAF] in French Sign Language (LSF). It corresponds to $1H$ and an *Arc* movement. Figure 4.64(b) shows the performance of the same sign in a different context, this time left hand moves straight. In this context signer prepares the following sign which corresponds to a sign performed with two hands. In this case, the classification is improved using statistics over the movement structure. In fact a movement $A + S$ is hardly performed by a human and since one hand is moving to prepare the following sign the faster way of going from one point to another is through a straight S movement. Therefore the S is deleted.



Figure 4.64: Sign [DEAF] in French Sign Language in different context

Table 4.11: Feature classification results

| Gloss | Without statistics | | With statistics | |
|--------------|--------------------|-------|-----------------|-------|
| | Nb. H | Traj. | Nb. H | Traj. |
| Shoulder bag | 1 | A | 1 | A |
| Shoulder bag | 2 | A+S | 1 | A |
| Deaf | 1 | A | 1 | A |
| Deaf | 2 | A+S | 1 | A |
| We | 1 | C | 1 | C |
| Give | 2 | C+C | 2 | C+C |

Using the extracted features to query the database of descriptions in ZeBeDee we are able to propose the potential glosses to the annotator. The number of proposed glosses for some signs is shown in table 4.12. Figure 4.65 shows the sign [WE/US] in LSF with the potential glosses sorted alphabetically.

This results are promising and show that the selected features are discriminant though only about 10% of signs in the description can be processed this way. In order to extend the vocabulary size we introduce the features obtained from synthetic data and perform the classification from ReZeBeDee.

Figure 4.65: Sign *we/us* in FSL showing the potential glosses

Table 4.12: Number of potential glosses

| Gloss | Nb. of proposed glosses |
|--------------|-------------------------|
| Shoulder bag | 20 |
| Deaf | 20 |
| We/Us | 6 |
| Give | 8 |

4.6.3.2 Annotation form ReZeBeDee

Here we want to show some results from the generation for the extension of ZeBeDee. We have used the GeneASL software for generating several times (around 100 times) some signs in the database for the automatic extension of their description. The first problem faced concerns the time required for generating a sign which can be of about 5 minutes for some complex signs, in average is about 2 minute. This is because during the generation, signs are also randomly generated several times (about 10) and the one with the best comfort score is proposed. Thus to obtain 100 generations of one sign the software is computing 1000 generations. However this is not considered a problem since this off-line processing is done only once per sign.

We have selected about 30 signs for performing the generation. Most representative results are discussed here for the three features previously described; number of moving hands, movement direction and relative position.

For the number of moving hands results are very interesting. The set of 30 signs has been generated to obtain the number of times each sign has been detected as a static sign, 1H or 2H. First of all we have noticed that within the total number of generations the same sign can be detected as zero, one or two hands moving. Only for about 57.5% of signs, all their generations have done exactly the same results concerning the number of moving hands. The remaining signs have some generation classified as zero, one or two hands. Table 4.13 shows some signs with the percentage associated to each class in the number of hands. Notice that same sign can be classified in various number of hands classes which is not possible since in the description it is well defined when hands moves. For example if only one hand moves the other hand is not in the description. The correct number of hands is given by the highest score.

Table 4.13: Number of hands percentage

| Gloss | No. of Gen | Zero | 1H | 2H |
|------------|------------|-------|-------|------|
| Error | 85 | 18.82 | 78.82 | 2.35 |
| Be thirsty | 100 | 12 | 88 | 0 |
| Good | 100 | 19 | 81 | 0 |
| School | 100 | 0 | 23 | 77 |

Table 4.14: Movement direction results

| SIGNE | Movement direction | | | | | | | |
|------------------|--------------------|-----------|------------|-----------|------------|-----------|------------|-----------|
| | I | | II | | III | | IV | |
| | Right Hand | Left Hand | Right hand | Left Hand | Right Hand | Left Hand | Right Hand | Left Hand |
| avoir_faim.S1669 | 15 | 0 | 22 | 0 | 29 | 0 | 34 | 0 |
| avoir_soif.S801 | 29 | 0 | 19 | 0 | 20 | 0 | 20 | 0 |
| beau.S1350 | 25 | 0 | 7 | 0 | 12 | 0 | 56 | 0 |
| blanc.S410 | 39 | 0 | 10 | 0 | 5 | 0 | 46 | 0 |
| difficile.S1378 | 7 | 0 | 46 | 0 | 38 | 0 | 5 | 0 |
| ecole.S308 | 11 | 9 | 4 | 6 | 39 | 36 | 46 | 26 |
| ecouter.S1126 | 13.69863 | 0 | 82.191781 | 0 | 2.739726 | 0 | 1.369863 | 0 |
| ouvert.S894 | 15 | 5 | 15 | 7 | 33 | 50 | 37 | 38 |

Concerning the movement direction some results are shown in Table 4.14. Cells in green correspond to the correction by an expert in SL. We can see that different generation give various directions that cannot be done. Here we notice that for the sign [AVOIR FAIM] which means [BE HUNGRY] only 40% of the generation correspond to what has been selected by the annotator. The best score corresponds to the sign [LISTENING] or [ECOUTER] in French about 82%. The last feature proposed concerns the relative position of hand with respect to the body (Sec 4.6.2.1). Table 4.15 shows some signs and the percentage of generation for the label *TOP_CENTER* and *TOP_LAT_1*. Notice that for the sign [WHITE] or [BLANC] in French 99% of the generation gives the good results for the end location of hand. These example for the generation shows that the verification and correction of the proposed annotation has to be systematically performed. This has to be investigated deeply.

The same example shown previously [WE/US] this time allow to propose less glosses to the annotator. In fact the gloss [FACE] and [EUROPE] are performed in the *TOP_CENTER* and *TOP_LAT_1* sectors which allow to reject them from the potential glosses. Also the sign [SPAIN] is performed in the *LAT_2* and the quadrant I and II since the circle is going up while [WE/US] movement direction is II and IV, see Figure 4.66. Finally we propose a list of only 3 glosses from about 150 glosses.

The evaluation of the annotation using ReZeBeDee has to be investigate more in detail with the whole database of 1600 signs, here we point out how the new description can be used for reducing the list of potential glosses.

Table 4.15: Relative position results

| SIGNE | TOP_CENTER | | | | TOP_LAT_I | | | |
|------------------|------------|-----------|------------|-----------|------------|-----------|------------|-----------|
| | Begin | | End | | Begin | | End | |
| | Right Hand | Left Hand | Right Hand | Left Hand | Right Hand | Left Hand | Right Hand | Left Hand |
| avoir_faim.S1669 | 9 | 0 | 2 | 0 | 11 | 0 | 7 | 0 |
| avoir_soif.S801 | 74 | 0 | 65 | 0 | 14 | 0 | 20 | 0 |
| blanc.S410 | 49 | 0 | 0 | 0 | 49 | 0 | 99 | 0 |

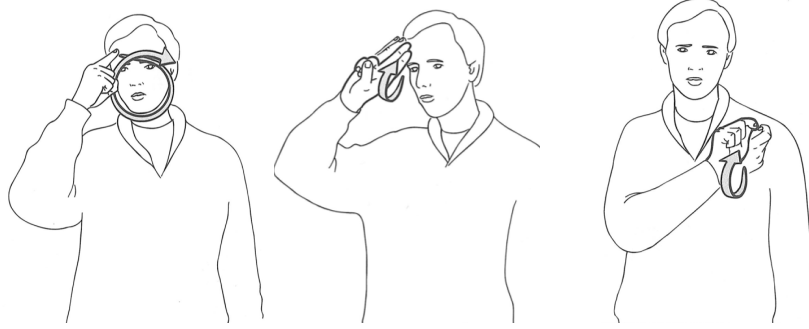


Figure 4.66: Sign [FACE], [EUROPE] and [SPAIN] respectively.

4.6.4 Conclusion

In this section we have presented the proposed approach for recognising potential glosses. A novel approach using a linguistic description of signs and SL generation is proposed. After the extraction of low level features using the approaches previously described, now we intend to introduce linguistic knowledge for recognising glosses. The problem faced is the way in which such an information will be introduced avoiding the use of training data. This is the main concerns in this PhD thesis since we have to avoid the use of annotated data for performing the annotation. For these reason we have proposed using a linguistic description of signs, ZeBeDee.

The first method for recognising glosses uses the description of signs in ZeBeDee. For this we extract features from video that are already described in ZeBeDee. The problem faced concerns the consistency of the annotation. Since ZeBeDee has been designed for SL generation, the same characteristic can be described in several way without penalising the generation results. Thus, querying the sign database become impossible. For this we proposed to use only few features that are consistently annotated such as the temporal structure, the number of moving hands and the path. However this reduces the size of our vocabulary. Indeed many signs do not have any defined trajectory since what is important do not correspond to the trajectory but about going from one point to another.

Although image processing techniques are able to classify hand shape, a hand shape ZeBeDee filter is difficult to develop because the same hand configuration can be described in several ways. It is the same problem for signs described in terms of a position. These problems are addressed by introducing new features that can be easily extracted from video.

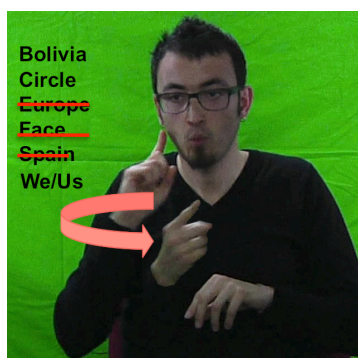


Figure 4.67: Sign *we/us* in FSL showing the potential glosses

We have presented an extended description of signs called ReZeBeDee where Re stands for recognition. For avoiding using large amount of annotated data we proposed using synthetic data which is obtained using the automatic SL generation. Indeed ZeBeDee can be used for piloting a signing avatar. In this way we can generate numerous times the same sign. By giving random values to the dependency parameter we are able to have several performances considering signs variability.

From the generated data we can extract visual features for adding them to the description. Same features can be extracted from video and be used for filtering glosses in the database. In this way we automatically build a description of signs adapted to the visual modality of image processing techniques. Also using synthetic data avoid any training though no statistical model in terms of occurrences could be built from this data.

Experiments have shown promising results and could be investigated in future work. This approach can be used to annotate any kind of gestures or SL described in ZeBeDee since all the other features correspond to low level features extracted from video.

Conclusions and perspectives

Résumé : Conclusion et perspectives

Ici nous présentons nos contributions, le respect des spécifications et nos perspectives. Plusieurs approches ont été abordées dans cette thèse afin de réussir nos objectifs.

Nous avons proposé un modèle de peau robuste aux changements de luminosité et spécifique au signeur dans la vidéo. Nous avons adapté le modèle des mains et de la tête ainsi que l'algorithme de suivi pour les différents membres. En effet la forme et la dynamique du mouvement sont très différents entre les mains et la tête. De plus nous avons introduit une nouvelle fonction de pénalisation permettant de gérer les occultations entre les objets. Nous avons aussi introduit une nouvelle méthode de segmentation de la main. Pour ça nous proposons une méthode de détection d'occultations afin d'utiliser la méthode adaptée au cas où la main se trouve devant le visage. Cette méthode utilise les contours et l'apparence afin d'extraire précisément la région de la main.

La segmentation temporelle des signes qui a été proposée utilise les caractéristiques de mouvement pour la segmentation et la forme de la main pour la correction de la sur-segmentation. Contrairement à d'autres approches dans la littérature nous n'utilisons pas de capteurs permettant d'extraire de manière précise la configuration mais nous proposons d'utiliser la silhouette ce qui permet d'extraire des caractéristiques géométriques. Ceci permet de proposer des limites potentielles à l'annotateur, qui à son tour sélectionne les limites qui correspondent au début ou à la fin d'un signe. De plus notre approche nous permet de déterminer la structure temporelle des signes ce qui n'a pas été réalisé auparavant dans la littérature.

Finalement nous avons proposé une méthode de reconnaissance des signes avec un système de représentation lexicale des signes adapté à la reconnaissance. Il s'agit de la première description contenant des caractéristiques visuelles explicitement décrites. Ces caractéristiques sont extraites à partir de données synthétiques générées automatiquement à l'aide d'un système de génération automatique de la LS.

Les méthodes issues de nos travaux de recherche nous permettent d'assister l'annotation et d'éviter de faire de l'apprentissage avec des données qui biaiserait nos résultats. De plus ceci rend l'annotation reproductible et moins dépendante de l'expérience et des connaissances de l'annotateur. Quand l'interaction de l'annotation est nécessaire, nous proposons plusieurs possibilités issues de mesures objectives afin d'éviter une influence

importante de l’annotateur. Finalement nous avons intégré nos contributions dans le système distribué proposé en [Collet 2010] afin qu’ils soient accessibles par les chercheurs en France et à l’étranger. Les contributions ici présentées sont le résultat d’une étude détaillée de la problématique et des spécifications à respecter imposées par notre application.

Nous avons toute au long de cette thèse travaillé avec une seule caméra afin de rendre nos algorithmes accessible à des linguistes et informaticiens. Afin de gérer la manque de profondeur nous utilisons plusieurs caractéristiques de mouvement et de forme. Dans notre problématique nous avons mentionné n’est pas forcément une contrainte tant que la qualité des résultats sont robustes et nos algorithmes n’utilisent pas des données d’apprentissage.

Nos contributions permettent de résoudre la problématique dans chaque problème abordé. Dans le cas du suivi des composantes corporelles notre méthode est robuste aux occultations sans contraindre la dynamique du mouvement ou la vitesse de réalisation des signes. De plus elle est robuste à la similarité de couleur entre les différents membres et à la manque d’information de profondeur. Pour la segmentation de la main nous avons proposé une approche permettant d’identifier les pixels appartenant à la main et au visage même quand la forme de la main change pendant l’occultation.

En ce qui concerne la segmentation temporelle, nous avons introduit des mesures objective permettant de caractériser les bords des signes dans un discours en LS. La représentation des signes adapté pour la reconnaissance respecte la grande variabilité des signes et la dépendance au contexte des signes. Finalement nous proposons une liste des signes potentiels.

Dans le future plusieurs chemins sont envisagés que ça soit d’un point de vue informatique ou linguistique. Par exemple, pour le suivi des composantes corporelles nous ne sommes pas en mesure de déterminer quelle est la position de la main droite et de la main gauche. De plus notre méthode n’est utilisable que quand un signeur a des habits à manches longues. Dans le cas de la segmentation des mains quand elles se trouvent devant le visage, les résultats de segmentation présentent des artefacts et des trous qui peuvent être supprimés.

En ce qui concerne la segmentation temporelle, nous n’utilisons que peu de mesures afin de caractériser les bord des signes. Même si ça a montré de bons résultats, l’utilisation d’autres caractéristiques permettrait améliorer les résultats de segmentation. De plus nous pourrions identifier les segments en tant que signes ou transitions en utilisant par exemple l’étude des mouvements balistiques.

Pour la reconnaissance nous avons introduit que de caractéristique de mouvement alors que la configuration ou la forme de la main nous permettrait de réduire la liste des signes potentiels proposés à l’annotateur.

Nous sommes convaincus que notre approches peut rester complètement libre et non contraint par le contexte ou le vocabulaire dans un contexte de reconnaissance de la LS.

In this chapter we present our main conclusions about our contributions and their originality. Also we discuss how they address the problem statement and they respect the specifications. Further work to be done for improving our contributions is described in the next section.

5.1 Conclusion

First of all we will summarise our contributions and we will argue how they have, on one's hand, solved the problem statement and on the other's hand respected the specification.

5.1.1 Summary: Our Contributions

Numerous approaches have been studied in this work to achieve our goal. Signs are characterised by their motion as well as the hand shape which are used to perform temporal segmentation and split sequences into signs or transition between signs, in continuous SL. An approach to robustly track hands and head even during occlusion has been proposed 4.3 and a hand segmentation method to extract hand when this overlaps the face 4.4. A temporal segmentation method is presented 4.5 to separate sign from transitions and to determine the temporal structure of signs. Finally we have proposed an extended representation of signs from ZeBeDee adapted for SL recognition. All this methods are used for querying a sign description database for proposing a list of potential glosses.

In our work instead of training the system using the performance of people, we synthesize data to extract features, this has the advantage of considering only what matters in the realisation of a sign. In this way the training is implicitly carried out without biasing it to a person background and way of signing. What is important here is the set of possibilities that can be stocked in our representation to be as general as possible. Our main contributions are the following :

- **Skin model** : we proposed a simple and fast specific skin model robust to illumination changes and adapted for each signer in the video.
- **Tracking** : we have adapted the model for hands and head as well as the tracking algorithm because of the shape and dynamic difference between these objects. In addition we have proposed a penalisation function wisely defined to robustly handle inter-objects occlusions.
- **Head segmentation** : we presented an occlusion detection function for determining whether and adapted algorithm is required. When hand occludes the face we use a classification approach using contours and appearance. This method allows to extract in a very accurate way hands region for further processing.

- **Temporal segmentation** : a new approach using motion for segmenting sequences and hands shape for correcting over-segmentation has been proposed. Unlike other methods here hands shape is considered without any device but using its silhouette. We propose limits to the annotator that can correspond to borders. In addition this approach is able to determine the temporal structure of sign which has not been addressed in the literature.
- **Gloss recognition** : For this we have proposed an adapted representation of signs for SL recognition. This is the first description having visual characteristics in it. Also we have proposed using synthetic data from automatic SL generation methods. Unlike trained approaches this considers the variability of signs.

The proposed methods allow us to (i) assist the annotation avoiding any learning step, so that our results are not dependent on the training data, (ii) make the annotation reproducible and less dependent from annotators knowledge and experience. This is achieved by avoiding as possible human interaction. When human interaction is required, e.g. for selecting the corresponding gloss, we propose several solutions from objective measurements. In this way the final annotation remains somehow unbiased to annotators knowledge, and (iii) make available the annotation algorithms to a large scientific community. Our algorithms are implemented in an automatic annotation architecture [Collet 2010] so that it is available for any person with access to this architecture. These contributions have been developed considering the problem statement and the specifications to be met by our work.

5.1.2 Problem statement and specifications

In the problem statement (Sec. 1.4) we have defined the problems faced during this PhD thesis. Here we discuss how our contributions have successfully addressed the problem statement. We have effectively considered the complexity of the language with only a mono-camera since this is widely used by linguists. Our approaches robustly deal with the lack of depth information using various features.

In our problem statement we mentioned that the execution time is not a constraint and that it is preferable to avoid using any learning step though other solutions could be computationally expensive. This has been respected in our approach since we do not use any training step and the execution time is not greater than the time required for manually annotate. In addition the annotation becomes less dependent from annotator and reproducible. Thus though the spent time is equivalent the quality of the annotation is better.

Our contributions have successfully dealt with the problems mentioned in the problem statement.

- **tracking**: our approach handles hand over face occlusion successfully without restricting hand dynamics or performance speed. It deals correctly with the similarity of colour and the lack of depth information.

- **hand segmentation:** the proposed method is able to classify pixels as belonging to the face or to the hand even if the hand shape changes during occlusion.
- **temporal segmentation:** we propose a method making the temporal segmentation approach reliable on objective measurement. This is achieved by proposing limits to annotators who select the ones that better correspond to signs borders. This make the segmentation less subjective to annotators experience and interpretation of the language.
- **sign representation:** we have proposed using a linguistic representation of signs that respects the high sign variability and the context-dependency of signs. In addition we have propose an extension of this model for describing sign for SL recognition.
- **gloss recognition:** We propose a method in which annotator can select the corresponding gloss from a list of potential signs. Disambiguating homosigns has not been addressed in this work since we require other linguistic models considering semantics and grammar.

In conclusion we have proposed robust features extraction algorithms, for continuous SL processing, to obtain representative characteristics from a performance. Also an extended sign representation for SL or gesture recognition is presented for filtering the glosses from the sign representation using the extracted features. This makes an advancement on the state-of-the-art.

5.2 Perspectives

In this PhD thesis we have addressed several works from computer vision and automatic SL processing. We have introduced a novel approach that has to be investigated more in detail. Numerous tests have to be performed at each stage in our procedure.

We have presented a novel approach which involves numerous processing techniques corresponding to a high amount of work in this PhD thesis. For this reason it would be suitable to improve each stage with other features.

In future work several ideas come to our minds either from image processing technique, e.g., improving feature extraction from video, or from SL processing. For example for

- **Tracking :** we are able to determine the position of hands and head, however we do not know which hand position corresponds to right or left hand. Indeed hand filters can be exchanged during the tracking. Also our method is constrained to long-sleeves clothes. This can be addressed tracking elbow with an adapted filter.
- **Hands segmentation :** the main drawback are the holes and artefacts in the final result. This could be improved by adding a verification step in which we can verify that all the contours exist in the image since artefacts correspond to one contour in the template and the other in the image. Several methods could be used for

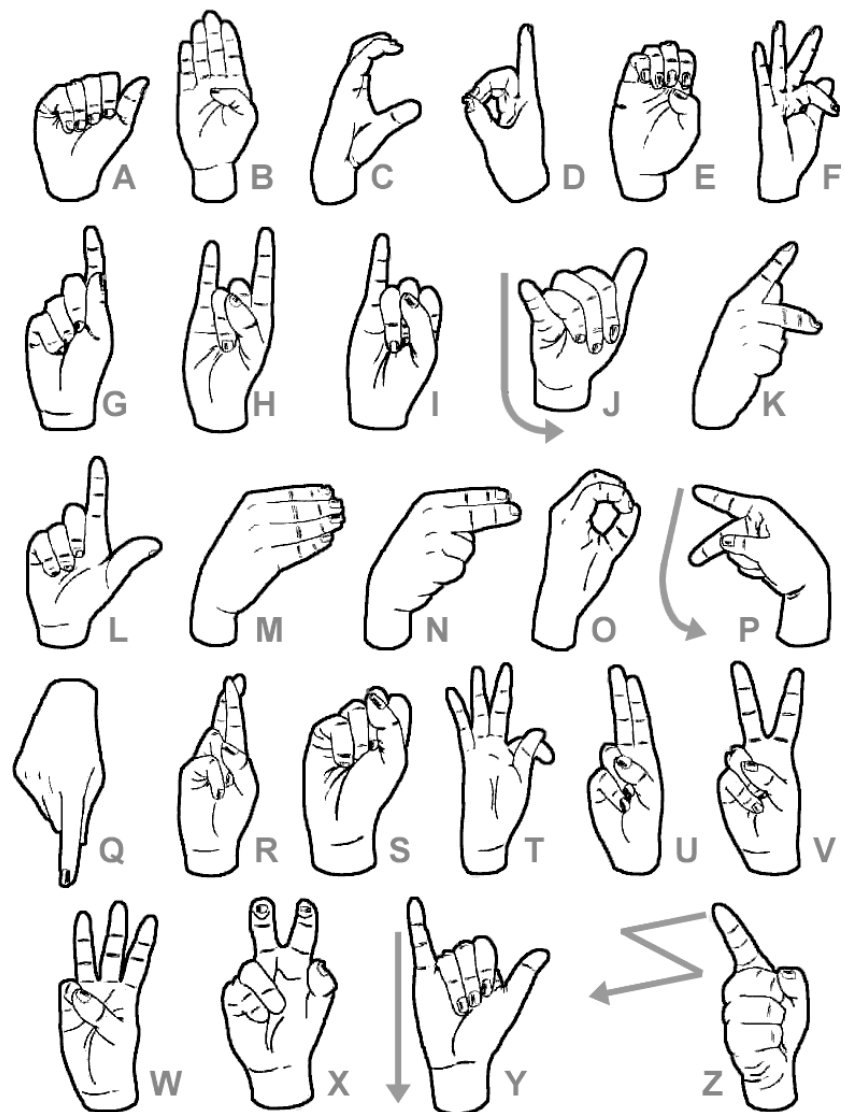
improving the segmentation result.

- **Temporal segmentation** : we only use two geometrical features for hand shape comparison. Although this has shown good results, adding other features might improve segmentation. Also we could label segments as *Signs* by studying the motion of transitions (ballistic movements).
- **Gloss recognition** we could add more features to the description, e.g. kind of symmetry. Adding hands shape to the ReZeBeDee representation might allow us to disambiguate signs with similar motion features but different hand shape. This approach will not be able to distinguish homosigns, for this we need the introduction of other features either from a linguistic level or from other non-manual features. Indeed non-manual feature for disambiguating signs could make of our approach, a system for recognition of freely SL performance, so far it remains focused on annotation because only a list of potential glosses is proposed.

We are convinced that this approach could remain completely free and unconstrained for recognising any SL in any context. We are very interesting in continuing our work.

Finger spelling

Here is the alphabet corresponding to the hand configuration associated.



D'après Albert Taboot

Zededee XML example

Description of sign [BUILDING] in LSF. the lime DEP corresponds to external dependences concerning the context.

SEQUENCE "immeuble"

DEP loc = @ABST + <FWD | medium>

DEP height = large

DEP foundation = medium

KEY_POSTURE(0){

KEEP :

For \$h=s,w

#L_closed(\$h)

#R_closed(\$h)

#M_closed(\$h)

#wrench(\$h, small)

Orient palm(\$h) _|_ UP

End

Orient NRM!palm(w) along -NRM!palm(s)

HERE:

For \$h=s,w

Place @T_TIP(\$h) at [loc] - <DIR!index(\$h,3) | [foundation]>

End }

TRANSITION(10) { Accel 1 For \$h =s,w Arc @PA(\$h) : <0,0,0> End }

KEY_POSTURE(5){ HERE: For \$h=s,w Place @T_TIP(\$h) at [loc] - <DIR!index(\$h,3) | [foundation]> + <UP | [height]> End }

End "immeuble"

```

SEQUENCE "choc"
<language=LSF>
<numvidlimsi="218">
<refdico="2-124-7">
<described_by="Nadège, Flora">
<sens="choc, être choqué">
KEY_POSTURE(0)
KEEP:
For $h=s,w
#all4_claw($h)
#T_lateral($h)
#all4_spread($h)
Orient NRM!palm($h) along BWD
End
Orient DIR!palm(w) along UP+LAT
Orient DIR!palm(s) along UP-LAT
HERE:
Place @M_TIP(s) at @ABST + <LAT | tiny>
Place @M_TIP(w) at @ABST - <LAT | tiny>
TRANSITION(10)
Accel 1
KEY_POSTURE(5)
HERE:
Place @M_TIP(s) at @ST + <LAT | tiny>
Place @M_TIP(w) at @ST - <LAT | tiny>
End "choc"

```

```

SEQUENCE "envoyer"
<language=LSF>
<numvidlimsi="324">
<refdico="2-52-5">
<described_by="Flora, Nadège">
DEP target = @SH(s) + <FWD | large>
ALIAS _line -> <@ST, [target]>
ALIAS _start -> @ST + <FWD | medium>
KEY_POSTURE(0)
HERE:
#all4_contact(s)
Place @T_TIP(s) at @M_TIP(s)
Place @PA(s) at _start
Orient DIR!palm(s) along -LAT
TRANSITION(10)
Accel 1
KEY_POSTURE(5)
HERE:
#all4_extended(s)
#all4_spread(s)
#T_straight(s)
Place @PA(s) at _start + <_line | 2*medium>
Orient NRM!palm(s) along _line
End "envoyer"

```


Bibliography

- [Adamo-Villani 2004] N. Adamo-Villani, J. Doublestein and Z. Martin. *The MathSigner: an interactive learning tool for American sign language*. In Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on, pages 713 – 716, july 2004. (Cited on page 73.)
- [Adamo-Villani 2008] N. Adamo-Villani. *3D rendering of american sign language finger-Spelling: a comparative study of two animation techniques*. International Journal of Human and Social Sciences, vol. 3, no. 4, 2008. (Cited on pages 24, 73 and 75.)
- [Aerts 2004] S. Aerts, B. Braem, K. Van Mulders and L. De Weerd. *Searching SignWriting signs*. In LREC 2004: 4th International Conference on Language Resources and Evaluation (RPSL Workshop),, pages 79–81, Lisbon, Portugal, 2004. (Cited on page 42.)
- [Ahmad 1997] 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, pages 345–352, 1997. (Cited on page 56.)
- [Al-Rousan 2009] M. Al-Rousan, K. Assaleh and A. Tala’a. *Video-based signer-independent Arabic sign language recognition using hidden Markov models*. Applied Soft Computing, vol. 9, no. 3, pages 990–999, 2009. (Cited on page 68.)
- [Alon 2006] J. Alon. *Spatiotemporal gesture segmentation*. Rapport technique, Boston University Computer Science Department, 2006. (Cited on page 68.)
- [Assan 1998] M. Assan and K. Grobel. *Video-based sign language recognition using hidden markov models*. Gesture and Sign Language in Human-Computer Interaction, pages 97–109, 1998. (Cited on page 68.)
- [Awad 2006] G. Awad, J. Han and A. Sutherland. *A unified system for segmentation and tracking of face and hands in sign language recognition*. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, volume 1, pages 239–242. IEEE, 2006. (Cited on page 55.)
- [Battison 1978] R. Battison. *Lexical Borrowing in American Sign Language*. 1978. (Cited on pages 39, 41 and 145.)
- [Bauer 2002] B. Bauer and K. Karl-Friedrich. *Towards an automatic sign language recognition system using subunits*. Gesture and Sign Language in Human-Computer Interaction, pages 123–173, 2002. (Cited on pages 47 and 63.)
- [Bébian 1825] A. Bébian. *Mimographie, ou essai d’écriture mimique propre à régulariser le langage des sourds-muets*. L. Colas, 1825. (Cited on page 41.)
- [Bernier 2009] O. Bernier, P. Cheung-Mon-Chan and A. Bouguet. *Fast nonparametric belief propagation for real-time stereo articulated body tracking*. Computer Vision and Image Understanding, vol. 113, no. 1, pages 29–47, 2009. (Cited on page 53.)

- [Bertsekas 2002] Dimitri P. Bertsekas and John N. Tsitsiklis. Introduction to probability. Athena Scientific, 2002. (Cited on page 86.)
- [Birchfield 1998] S. Birchfield. *Elliptical head tracking using intensity gradients and color histograms*. In Proc. CVPR, pages 232–237, 1998. (Cited on pages 48 and 51.)
- [Boutora 2008] L. Boutora. *Fondements historiques et implications théoriques d’une phonologie en langue des signes: étude de la perception catégorielle des configurations manuelles en LSF et réflexion sur la transcription des langues des signes*. PhD thesis, Université Paris 8, 2008. (Cited on page 13.)
- [Boutora 2011] L. Boutora and A. Braffort. DEfi Geste Langue des Signes. Corpus DEGELS1. Corpus ID oai:crdo.fr:crdo000767 : Video en LSF, informateur A, 2011. (Cited on pages 27, 29, 137 and 140.)
- [Bowden 2004] R. Bowden, D. Windridge, T. Kadir, A. Zisserman and M. Brady. *A linguistic feature vector for the visual interpretation of sign language*. Computer Vision-ECCV 2004, pages 390–401, 2004. (Cited on pages 67, 70 and 71.)
- [Bowyer 1999] K. Bowyer, C. Kranenburg and S. Dougherty. *Edge detector evaluation using empirical ROC curves*. In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., volume 1. IEEE, 1999. (Cited on page 48.)
- [Bradski 2002] G.R. Bradski. *Real time face and object tracking as a component of a perceptual user interface*. In Applications of Computer Vision, 1998. WACV’98. Proceedings., Fourth IEEE Workshop on, pages 214–219. IEEE, 2002. (Cited on page 51.)
- [Braffort 2001] A. Braffort, C. Cuxac, A. Choisier, C. Collet, P. Dalle, I. Fusellier, R. Gherbi, G. Jausions, G. Jirou, F. Lejeune et al. *Projet LS-COLIN. Quel outil de notation pour quelle analyse de la LS*. Journées Recherches sur la langue des signes. UTM, Le Mirail, Toulouse, 2001. (Cited on pages 27 and 29.)
- [Braffort 2004] 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 4th International Conference on Language Resources and Evaluation - LREC 2004, volume 1, pages 201–203, Lisbon, Portugal, May 2004. (Cited on pages 14, 21 and 33.)
- [Braffort 2005] A. Braffort, B. Bossard, J. Segouat, L. Bolot and Lejeune F.(2005). *Modélisation des relations spatiales en langue des signes française*. Proceedings of traitement Automatique de la Langue des Signes. CNRS, ATALA, 2005. (Cited on page 68.)
- [Braffort 2008] A. Braffort and P. Dalle. *Sign language applications: preliminary modeling*. Universal access in the information society, vol. 6, no. 4, pages 393–404, 2008. (Cited on pages x and 45.)
- [Braffort 2010] A. Braffort, L. Bolot, E. Chételat-Pelé, A. Choisier, M. Delorme, M. Filhol, J. Segouat, C. Verrecchia, F. Badin and N. Devos. *Sign Language corpora for analysis, processing and evaluation*. Seventh conference on International Language Resources and Evaluation (LREC10), 2010. (Cited on pages xi, 73 and 74.)

- [Braffort 2011] A. Braffort, L. Bolot and J. Segouat. *Virtual signer coarticulation in Octopus, a Sign Language generation platform*. In GW 2011: The 9th International Gesture Workshop, 2011. (Cited on pages 74 and 77.)
- [Braffort 2012] Annelies Braffort and Leïla Boutora. *Défi d'annotation DEGELS2012 : la segmentation*. In JEP-TALN-RECITAL 2012, pages 1–8, June 2012. (Cited on pages xi, 60 and 61.)
- [Brashear 2005] H. Brashear, T. Starner, P. Lukowicz and H. Junker. *Using multiple sensors for mobile sign language recognition*. In Wearable Computers, 2003. Proceedings. Seventh IEEE International Symposium on, pages 45–52. IEEE, 2005. (Cited on page 24.)
- [Brentari 1998] D. Brentari. A prosodic model of sign language phonology. The MIT Press, 1998. (Cited on pages 39 and 42.)
- [Brentari 2006] D. Brentari. *Effects of language modality on word segmentation: An experimental study of phonological factors in a sign language*. Papers in laboratory phonology, vol. 8, pages 155–164, 2006. (Cited on page 60.)
- [Brentari 2011] D. Brentari. *Handshape in sign language phonology*. Companion to phonology, pages 195–222, 2011. (Cited on page 39.)
- [Brooks 1997] R.A. Brooks. *The intelligent room project*. In Cognitive Technology, 1997. 'Humanizing the Information Age'. Proceedings., Second International Conference on, pages 271–278. IEEE, 1997. (Cited on page 52.)
- [Brugaille 2006] JL Brugaille, J. Dalle and MP Kellerhals. *«Une expérience d'utilisation de forme graphique dans la scolarité des enfants sourds: Méthodes de travail et premières observations»*, colloque Syntaxe, interprétation, lexicque des langues signées. Université Lille, vol. 3, 2006. (Cited on page 42.)
- [Bruno 2002] B. Bruno. *Problèmes posés par la reconnaissance de gestes en Langue des Signes*. In Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues, 2002. (Cited on page 67.)
- [Bungerot 2008] J. Bungerot, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way and L. Van Zijl. *The ATIS sign language corpus*. 2008. (Cited on pages 28 and 31.)
- [Chen 2002] Y. Chen, W. Gao, Z. Wang, C. Yang and D. Jiang. *Text to avatar in multimodal human computer interface*. Asia-Pacific Human Computer Interface (APCHI2002), vol. 2, pages 636–643, 2002. (Cited on pages 73 and 74.)
- [Chen 2003] F.S. Chen, C.M. Fu and C.L. Huang. *Hand gesture recognition using a real-time tracking method and hidden Markov models*. Image and Vision Computing, vol. 21, no. 8, pages 745–758, 2003. (Cited on page 48.)
- [Chételat-Pelé 2008a] É. Chételat-Pelé, A. Braffort and J. Véronis. *Sign Language Corpus Annotation: Toward a New Methodology*. In 6th International Conference on Language Resources and Evaluation. Marrakech. Morocco, 2008. (Cited on page 31.)

- [Chételat-Pelé 2008b] Emilie Chételat-Pelé, Annelies Braffort and Jean Véronis. *Annotation of Non Manual Gestures: Eyebrow movement description*. In 3rd Workshop on the Representation and Processing of Sign Languages, pages 28–32, 2008. (Cited on page 31.)
- [Cheung 2005] K. Cheung, S. Baker and T. Kanade. *Shape-from-silhouette across time part ii: Applications to human modeling and markerless motion tracking*. International Journal of Computer Vision, vol. 63, no. 3, pages 225–245, 2005. (Cited on page 50.)
- [Collet 2010] C. Collet, M. Gonzalez and F. Milachon. *Distributed System Architecture for Assisted Annotation of Video Corpora*. International workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC), Valletta, Malte, pages 49–52, Mai 2010. (Cited on pages 23, 33, 34, 164 and 166.)
- [Companys 2004] M. Companys, F. Tourmez and Y. Delaporte. Dictionnaire 1200 signes: français-lsf. M. Companys, 2004. (Cited on pages x and 41.)
- [Coogan 2006] T. Coogan and A. Sutherland. *Transformation invariance in hand shape recognition*. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, volume 3, pages 485–488. IEEE, 2006. (Cited on page 58.)
- [Cooper 2007] H. Cooper and R. Bowden. *Large lexicon detection of sign language*. Human–Computer Interaction, pages 88–97, 2007. (Cited on pages 33, 70 and 71.)
- [Cooper 2011] H. Cooper, B. Holt and R. Bowden. *Sign Language Recognition*. Visual Analysis of Humans, pages 539–562, 2011. (Cited on pages 13, 21, 66 and 68.)
- [Corina 1993] D.P. Corina. *To branch or not to branch: Underspecification in ASL hand-shape contours*. Current issues in ASL phonology, vol. 3, pages 63–95, 1993. (Cited on page 39.)
- [Costa 2003] A. C. R. Costa and G.P. Dimuro. *SignWriting and SWML: Paving the way to sign language processing*. Traitement Automatique des Langues de Signes, Workshop on Minority Languages,, June 2003. (Cited on page 42.)
- [Cox 2002] S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt and S. Abbott. *TESSA, a system to aid communication with deaf people*. In Proceedings of the fifth international ACM conference on Assistive technologies, pages 205–212. ACM, 2002. (Cited on pages 24 and 73.)
- [Cui 1995] Y. Cui, D.L. Swets and J.J. Weng. *Learning-based hand sign recognition using SHOSLIF-M*. In Computer Vision, 1995. Proceedings., Fifth International Conference on, pages 631–636. IEEE, 1995. (Cited on page 55.)
- [Cui 2000] Y. Cui and J. Weng. *Appearance-based hand sign recognition from intensity image sequences*. Computer Vision and Image Understanding, vol. 78, no. 2, pages 157–176, 2000. (Cited on pages 48 and 55.)
- [Cuxac 2000] C. Cuxac. Langue des signes française, les voies de l’iconicité, volume 15–16. Ophrys, 2000. (Cited on pages 10, 11, 13, 15 and 16.)

- [Dalal 2005] N. Dalal and B. Triggs. *Histograms of oriented gradients for human detection*. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. Ieee, 2005. (Cited on page 48.)
- [Dalle-Nazébi 2006] S. Dalle-Nazébi. *Chercheurs, Sourds et Langue des Signes. Le travail d’ÀŠun objet et de repères linguistiques*. PhD thesis, Thèse de Sociologie, Université Toulouse 2, 2006. (Cited on page 13.)
- [Darrell 1993] T. Darrell and A. Pentland. *Space-time gestures*. In Computer Vision and Pattern Recognition, 1993. Proceedings CVPR’93., 1993 IEEE Computer Society Conference on, pages 335–340. IEEE, 1993. (Cited on page 68.)
- [Dat Nguyen 2011] T. Dat Nguyen and S. Ranganath. *Facial expressions in American sign language: Tracking and recognition*. Pattern Recognition, 2011. (Cited on page 33.)
- [Davis 1994] J. Davis and M. Shah. *Recognizing hand gestures*. Computer Vision&ECCV’94, pages 331–340, 1994. (Cited on page 55.)
- [Delamarre 1999] Q. Delamarre and O. Faugeras. *3D articulated models and multi-view tracking with silhouettes*. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 2, pages 716–721. Ieee, 1999. (Cited on page 52.)
- [Delamarre 2001] Q. Delamarre and O. Faugeras. *3D articulated models and multiview tracking with physical forces*. Computer Vision and Image Understanding, vol. 81, no. 3, pages 328–357, 2001. (Cited on page 52.)
- [Delorme 2009] M. Delorme, M. Filhol and A. Braffort. *Animation Generation Process for Sign Language Synthesis*. In Advances in Computer-Human Interactions, 2009. ACHI ’09. Second International Conferences on, pages 386–390, feb. 2009. (Cited on page 75.)
- [Delorme 2011] Maxime Delorme. *Modélisation du squelette pour la génération réaliste de postures de la langue des signes française*. PhD thesis, Université Paris-Sud 11, 2011. (Cited on pages xiv, 44, 76, 77, 151 and 152.)
- [Deutscher 2000] J. Deutscher, A. Blake and I. Reid. *Articulated body motion capture by annealed particle filtering*. In Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, volume 2, pages 126–133. IEEE, 2000. (Cited on pages 48, 51, 53, 95 and 96.)
- [Devos 2011] N. Devos, F. Badin and É. Chételat-Pelé. *Les pointages en Langue des Signes: une méthodologie d’ÀŠannotation*. DEGELS, 2011. (Cited on page 29.)
- [Diamanti 2008] O. Diamanti and P. Maragos. *Geodesic active regions for segmentation and tracking of human gestures in sign language videos*. In Image Processing, 2008. IICIP 2008. 15th IEEE International Conference on, pages 1096–1099. IEEE, 2008. (Cited on page 56.)
- [Ding 2009] L. Ding and A.M. Martinez. *Modelling and recognition of the linguistic components in American Sign Language*. Image and vision computing, vol. 27, no. 12, pages 1826–1844, 2009. (Cited on page 52.)

- [Downton 1992] A.C. Downton and H. Drouet. *Model-based image analysis for unconstrained human upper-body motion*. In Image Processing and its Applications, 1992., International Conference on, pages 274–277, April 1992. (Cited on page 52.)
- [Dreuw 2007] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi and H. Ney. *Speech recognition techniques for a sign language recognition system*. In Eighth Annual Conference of the International Speech Communication Association, 2007. (Cited on pages 28 and 67.)
- [Dreuw 2008a] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff and H. Ney. *Benchmark databases for video-based automatic sign language recognition*. In International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, pages 1115–1121, 2008. (Cited on page 28.)
- [Dreuw 2008b] P. Dreuw and H. Ney. *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, 2008. (Cited on pages 33 and 34.)
- [Dreuw 2008c] P. Dreuw, D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth and H. Ney. *Spoken language processing techniques for sign language recognition and translation*. Technology and Disability, vol. 20, no. 2, pages 121–133, 2008. (Cited on page 67.)
- [Drummond 2001] T. Drummond and R. Cipolla. *Real-time tracking of highly articulated structures in the presence of noisy measurements*. In Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, volume 2, pages 315–320. IEEE, 2001. (Cited on page 53.)
- [Duan-Sheng 2006] C. Duan-Sheng and L. Zheng-Kai. *A survey of skin color detection*. J]. Chinese Journal of Computers, vol. 29, no. 2, pages 194–207, 2006. (Cited on page 48.)
- [Duarte 2010] K. Duarte and S. Gibet. *Heterogeneous data sources for signed language analysis and synthesis*. In Proceedings of the 7th International Conference on Language Resources and Evaluation, pages 19–21, May 2010. (Cited on page 75.)
- [Dubot 2012] Rémi Dubot and Christophe Collet. *Improvements of the Distributed Architecture for Assisted Annotation of Video Corpora*. In 5th Workshop on the Representation and Processing of Sign Languages, LREC 2012, 2012. (Cited on page 36.)
- [Duetscher 2000] J. Duetscher, A. Blake and I. Reid. *Articulated body motion capture by annealed particle filtering*. In cvpr, page 2126. Published by the IEEE Computer Society, 2000. (Cited on pages x, 52 and 53.)
- [Efthimiou 2007] E. Efthimiou and S.E. Fotinea. *GSLC: creation and annotation of a Greek sign language corpus for HCI*. Universal Access in Human Computer Interaction. Coping with Diversity, pages 657–666, 2007. (Cited on page 31.)
- [Efthimiou 2009] E. Efthimiou, S.E. Fotinea, C. Vogler, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos and J. Segouat. *Sign language recognition*,

- generation, and modelling: a research effort with applications in deaf communication*. Universal Access in Human-Computer Interaction. Addressing Diversity, pages 21–30, 2009. (Cited on page 29.)
- [Eleni Efthimiou 2010] Thomas Hanke John Glauert Richard Bowden Annelies Braffort Christophe Collet Petros Maragos François Goudenove Eleni Efthimiou Stavroula-Evita Fotinea. *DICTA-SIGN: Sign Language Recognition, Generation and Modelling with application in Deaf Communication*. International workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC), Valleta, Malte, pages 80–83, Mai 2010. (Cited on page 4.)
- [Elliott 2000] R. Elliott, J. R. W. Glauert, J. R. Kennaway and I. Marshall. *The development of language processing support for the ViSiCAST project*. In Proceedings of the fourth international ACM conference on Assistive technologies, Assets '00, pages 101–108, New York, NY, USA, 2000. ACM. (Cited on pages xi, 74 and 75.)
- [Elliott 2004] R. Elliott, JRW Glauert, V. Jennings and JR Kennaway. *An overview of the SiGML notation and SiGMLSigning software system*. 2004. (Cited on page 42.)
- [Elliott 2010] R. Elliott, J. Bueno, R. Kennaway and J. Glauert. *Towards the integration of synthetic sl animation with avatars into corpus annotation tools*. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Valletta, Malta, 2010. (Cited on page 29.)
- [Elmezain 2009] M. Elmezain, A. Al-Hamadi, S.S. Pathan and B. Michaelis. *Spatio-temporal feature extraction-based hand gesture recognition for isolated American Sign Language and Arabic numbers*. In Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on, pages 254–259. IEEE, 2009. (Cited on page 27.)
- [Encrevé 2008] F. Encrevé. *Sourds et société française au XIXe siècle: 1830-1905*. PhD thesis, Université Paris 8- Vincennes Saint-Denis, 2008. (Cited on page 13.)
- [Erdem 2002] U. Erdem and S. Sclaroff. *Automatic detection of relevant head gestures in American Sign Language communication*. In International Conference on Pattern Recognition, volume 16, pages 460–463. Citeseer, 2002. (Cited on page 28.)
- [Fang 2002] G. Fang, W. Gao, X. Chen, C. Wang and J. Ma. *Signer-independent continuous sign language recognition based on SRN/HMM*. Gesture and sign language in human-computer interaction, pages 163–197, 2002. (Cited on page 63.)
- [Fang 2004] G. Fang, X. Gao, W. Gao and Y. Chen. *A novel approach to automatically extracting basic units from Chinese sign language*. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 4, pages 454–457. IEEE, 2004. (Cited on page 70.)
- [Filhol 2007] M. Filhol, A. Braffort and L. Bolot. *Signing avatar: Say hello to Elsi*. International Workshop on Gesture in Human-Computer Interaction and Simulation (GW), 2007. (Cited on page 73.)
- [Filhol 2008] M. Filhol. *Modèle descriptif des signes pour un traitement automatique des langues des signes*. PhD thesis, Université Paris 11, 2008. (Cited on pages x, 42, 43 and 44.)

- [Filhol 2009a] M. Filhol. *Internal report on Zebedee*. Rapport technique 2009-08, LIMSI-CNRS, 2009. (Cited on pages 43, 45 and 46.)
- [Filhol 2009b] M. Filhol. *Zebedee: a lexical description model for Sign language synthesis*. LIMSI report, 2009. (Cited on pages 71 and 75.)
- [Fillbrandt 2003] H. Fillbrandt, S. Akyol and K.F. Kraiss. *Extraction of 3D hand shape and posture from image sequences for sign language recognition*. In IEEE International Workshop on Analysis and Modeling of Faces and Gestures, volume 17, pages 181–186, 2003. (Cited on page 58.)
- [Fischer 2011] S. Fischer and Q. Gong. *Marked Hand Configurations in Asian Sign Languages*. Formational Units in Sign Languages, vol. 3, page 19, 2011. (Cited on page 39.)
- [Fleischer 1917] M. Fleischer. *Method of producing moving-picture cartoons*, October 9 1917. US Patent 1,242,674. (Cited on page 73.)
- [Flood 2002] C.M. Flood. *How Do Deaf and Hard of Hearing Students Experience Learning to Write Using Signwriting, a Way to Read and Write Signs?* PhD thesis, University of New Mexico, 2002. (Cited on page 42.)
- [Fotinea 2008] Stavroula-Evita Fotinea, Eleni Efthimiou, George Caridakis and Kostas Karpouzis. *A knowledge-based sign synthesis architecture*. Universal Access in the Information Society, vol. 6, pages 405–418, 2008. (Cited on page 75.)
- [Francke 2007] H. Francke, J. Ruiz-del Solar and R. Verschae. *Real-time hand gesture detection and recognition using boosted classifiers and active learning*. Advances in Image and Video Technology, pages 533–547, 2007. (Cited on page 58.)
- [Freeman 1995] W.T. Freeman and M. Roth. *Orientation histograms for hand gesture recognition*. In International Workshop on Automatic Face and Gesture Recognition, volume 12, pages 296–301, 1995. (Cited on page 48.)
- [Frey 1996] W. Frey, M. Zyda, R. Mcghee and B. Cockayne. *Off-the-shelf, real-time, human body motion capture for synthetic environments*. Computer Science Department, Naval Postgraduate School, USA, 1996. (Cited on page 24.)
- [Fritsch 2002] J. Fritsch, S. Lang, A. Kleinhagenbrock, GA Fink and G. Sagerer. *Improving adaptive skin color segmentation by incorporating results from face detection*. In Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on, pages 337–343. IEEE, 2002. (Cited on page 48.)
- [Gall 2007] J. Gall, J. Potthoff, C. Schnörr, B. Rosenhahn and H.P. Seidel. *Interacting and annealing particle filters: Mathematics and a recipe for applications*. Journal of Mathematical Imaging and Vision, vol. 28, no. 1, pages 1–18, 2007. (Cited on page 95.)
- [Gall 2009] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn and H.P. Seidel. *Motion capture using joint skeleton tracking and surface estimation*. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1746–1753. Ieee, 2009. (Cited on page 50.)

- [Gao 2004] W. Gao, G. Fang, D. Zhao and Y. Chen. *Transition movement models for large vocabulary continuous sign language recognition*. In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pages 553–558. IEEE, 2004. (Cited on page 64.)
- [Garcia 2011] B. Garcia, M.A. Sallandre, C. Schoder, M.T. L’Huillier *et al.* *Typologie des pointages en Langue des Signes Française (LSF) et problématiques de leur annotation*. Actes du 1er DEfi Geste Langue des Signes (DEGELS 2011), pages 109–121, 2011. (Cited on page 29.)
- [Gianni 2009] F. Gianni, C. Collet and P. Dalle. Robust tracking for processing of videos of communication’s gestures, pages 93–101. LNAI 5085. Springer, 2009. (Cited on pages 28, 48, 49, 51, 55, 91, 93, 96, 99, 105, 106 and 107.)
- [Gibet 2007] S. Gibet and P.F. Marteau. *Approximation of curvature and velocity using adaptive sampling representations-Application to hand gesture analysis*. In Gesture workshop, 2007. (Cited on page 63.)
- [Gibet 2008] Sylvie Gibet and Alexis Héloir. *Formalisme de description des gestes de la langue des signes française pour la génération du mouvement de signeurs virtuels*. TAL, Special issue on Modeling and processing of sign languages, vol. 3, no. 48, pages 111–145, 2008. (Cited on page 43.)
- [Gillot 1998] Dominique Gillot. *Le droit des Sourds*. page 75, June 1998. (Cited on page 4.)
- [Girod 1997] M. Girodet *et al.* *La langue des signes française, dictionnaire bilingue LSF/français, tomes 2 et 3*. Paris, éditions d’ŒIVT, 1997. (Cited on pages ix, xi, xiii, xiv, 11, 76, 142, 143 and 153.)
- [Gleicher 1998] M. Gleicher. *Retargeting motion to new characters*. In ACM SIGGRAPH 98, 1998. (Cited on page 75.)
- [Gonzalez 2010] 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. (Cited on page 29.)
- [Gonzalez 2012a] M. Gonzalez, C. Collet *et al.* *Segmentation semi-automatique de corpus vidéo en Langue des Signes*. 2012. (Cited on page 29.)
- [Gonzalez 2012b] Matilde Gonzalez. *Un système de segmentation automatique de gestes appliqué à la Langue des Signes*. In DEfi Geste Langue des Signes, JEP-TALN-RECITAL 2012, pages 93–98, 2012. (Cited on page 60.)
- [Grobel 1997] K. Grobel and M. Assan. *Isolated sign language recognition using hidden Markov models*. In IEEE Int. Conference on Systems, Man, and Cybernetics, volume 1, pages 162–167. IEEE, 1997. (Cited on page 67.)
- [Habibi 2004] N. Habibi, 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, pages 1086–1097, 2004. (Cited on pages 48, 50, 55, 87 and 111.)

- [Hager 2002] G.D. Hager and P.N. Belhumeur. *Efficient region tracking with parametric models of geometry and illumination*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 10, pages 1025–1039, 2002. (Cited on page 51.)
- [Hamada 2002] 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. (Cited on page 55.)
- [Hamada 2004] Y. Hamada, N. Shimada and Y. Shirai. *Hand shape estimation under complex backgrounds for sign language recognition*. In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pages 589–594. IEEE, 2004. (Cited on page 58.)
- [Han 2009] J. Han, G. Awad and A. Sutherland. *Modelling and segmenting subunits for sign language recognition based on hand motion analysis*. Pattern Recognition Letters, vol. 30, no. 6, pages 623–633, 2009. (Cited on pages 63 and 70.)
- [Hanke 2002] T. Hanke. *Hamnosys in a sign language generation context*. In Progress in sign language research (International Studies on Sign Language and Communication of the Deaf, editeurs, Rolf Schulmeister and Heimo Reinitzer, volume 40, pages 249–264, 2002. (Cited on page 75.)
- [Hanke 2008] T. Hanke and J. Storz. *iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography*. In Proc. of 6th International Conference on Language Resources and Evaluation, LREC 2008, pages W25–64–W25–67, Marrakesh, May 2008. (Cited on page 33.)
- [Hanke 2010] T. Hanke, L. König, S. Wagner and S. Matthes. *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, pages 106–110, 2010. (Cited on pages 23, 27 and 29.)
- [Hasanuzzaman 2004] M. Hasanuzzaman, V. Ampornaramveth, T. Zhang, MA Bhuiyan, Y. Shirai and H. Ueno. *Real-time vision-based gesture recognition for human robot interaction*. In Robotics and Biomimetics, 2004. ROBIO 2004. IEEE International Conference on, pages 413–418. IEEE, 2004. (Cited on page 27.)
- [Havasi 2005] L. Havasi and H.M. Szabó. *A motion capture system for sign language synthesis: overview and related issues*. In Computer as a Tool, 2005. EUROCON 2005. The International Conference on, volume 1, pages 445–448. IEEE, 2005. (Cited on page 24.)
- [Heap 1995] T. Heap and F. Samaria. *Real-time hand tracking and gesture recognition using smart snakes*. Proc. Interface to Human and Virtual Worlds, Montpellier, France, 1995. (Cited on page 50.)
- [Heloir 2006a] A. Heloir, S. Gibet, N. Courty and F. Multon. *Alignement temporel de séquences gestuelles communicatives*. Groupe de Travail Animation et Simulation, 2006. (Cited on page 68.)
- [Heloir 2006b] Alexis Heloir, Sylvie Gibet, Franck Multon and Nicolas Courty. *Captured Motion Data Processing for Real Time Synthesis of Sign Language*. In Gesture in

- Human Computer Interaction and Simulation. Springer Berlin Heidelberg, 2006. (Cited on page 73.)
- [Hernandez-Rebollar 2002] J.L. Hernandez-Rebollar, R.W. Lindeman and N. Kyriakopoulos. *A multi-class pattern recognition system for practical finger spelling translation*. In Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on, pages 185–190. IEEE, 2002. (Cited on page 24.)
- [Hienz 1999] H. Hienz, B. Bauer and K.F. Kraiss. *HMM-based continuous sign language recognition using stochastic grammars*. Gesture-Based Communication in Human-Computer Interaction, pages 185–196, 1999. (Cited on page 47.)
- [Holden 2001] E.J. Holden and R. Owens. *Visual sign language recognition*. Multi-Image Analysis, pages 270–287, 2001. (Cited on pages 47 and 50.)
- [Holden 2005] E.J. Holden, G. Lee and R. Owens. *Australian sign language recognition*. Machine Vision and Applications, vol. 16, no. 5, pages 312–320, 2005. (Cited on pages 50 and 56.)
- [Horain 2002] P. Horain and M. Bomb. *3D model based gesture acquisition using a single camera*. In Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on, pages 158–162, 2002. (Cited on page 52.)
- [Howe 1999] N. Howe, M. Leventon and W. Freeman. *Bayesian reconstruction of 3d human motion from single-camera video*. In Neural Information Processing Systems, volume 1999, page 1. Cambridge, MA, 1999. (Cited on page 53.)
- [Howe 2008] L.W. Howe, F. Wong and A. Chekima. *Comparison of hand segmentation methodologies for Hand gesture recognition*. In Information Technology, 2008. IT-Sim 2008. International Symposium on, volume 2, pages 1–7. IEEE, 2008. (Cited on page 55.)
- [Hrůz 2008] M. Hrůz, P. Campr and M. Železný. *Semi-automatic annotation of sign language corpora*. In LREC 3rd Workshop on the Representation and Processing of Sign Languages Construction and Exploitation of Sign Language Corpora, numéro 3, pages 78–81, 2008. (Cited on page 33.)
- [Huang 1998] C.L. Huang and W.Y. Huang. *Sign language recognition using model-based tracking and a 3D Hopfield neural network*. Machine vision and applications, vol. 10, no. 5, pages 292–307, 1998. (Cited on page 68.)
- [Huang 2001] C.L. Huang and S.H. Jeng. *A model-based hand gesture recognition system*. Machine vision and applications, vol. 12, no. 5, pages 243–258, 2001. (Cited on page 48.)
- [Imagawa 1998] K. Imagawa, S. Lu and S. Igi. *Color-based hands tracking system for sign language recognition*. In 3th IEEE International Conference on Automatic Face and Gesture Recognition, pages 462–467. IEEE, 1998. (Cited on page 50.)
- [Isard 1998] M. Isard and A. Blake. *Condensation-conditional density propagation for visual tracking*. International journal of computer vision, vol. 29, no. 1, pages 5–28, 1998. (Cited on pages xi, 51 and 93.)

- [Jang 2002] D.S. Jang, S.W. Jang and H.I. Choi. *2D human body tracking with structural Kalman filter*. Pattern Recognition, vol. 35, no. 10, pages 2041–2049, 2002. (Cited on page 51.)
- [Johnson 2010] R.E. Johnson and S.K. Liddell. *Toward a phonetic representation of signs: Sequentiality and contrast*. Sign Language Studies, vol. 11, no. 2, pages 241–274, 2010. (Cited on page 42.)
- [Johnson 2011] R.E. Johnson and S.K. Liddell. *A segmental framework for representing signs phonetically*. Sign Language Studies, vol. 11, no. 3, pages 408–463, 2011. (Cited on page 42.)
- [Ju 1996] S.X. Ju, M.J. Black and Y. Yacoob. *Cardboard people: A parameterized model of articulated image motion*. In Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on, pages 38–44. IEEE, 1996. (Cited on page 50.)
- [Kadir 2004] T. Kadir, R. Bowden, E.J. Ong and A. Zisserman. *Minimal training, large lexicon, unconstrained sign language recognition*. In Proc. BMVC, 2004. (Cited on pages 47, 70 and 71.)
- [Kakumanu 2007] P. Kakumanu, S. Makrogiannis and N. Bourbakis. *A survey of skin-color modeling and detection methods*. Pattern recognition, vol. 40, no. 3, pages 1106–1122, 2007. (Cited on page 48.)
- [Karpouzis 2007] K. Karpouzis, G. Caridakis, S.-E. Fotinea and E. Efthimiou. *Educational resources and implementation of a Greek sign language synthesis architecture*. Computers and Education, vol. 49, no. 1, pages 54 – 74, 2007. (Cited on page 75.)
- [Kelly 2009] D. Kelly, J. McDonald and C. Markham. *Recognizing spatiotemporal gestures and movement epenthesis in sign language*. In Machine Vision and Image Processing Conference, 2009. IMVIP’09. 13th International, pages 145–150. IEEE, 2009. (Cited on page 64.)
- [Kennaway 2003] R. Kennaway. *Experience with and requirements for a gesture description language for synthetic animation*. In Gesture Workshop, pages 300–311, 2003. (Cited on page 75.)
- [Kennaway 2007] J. R. Kennaway, J. R. W. Glauert and I. Zwitserlood. *Providing signed content on the Internet by synthesized animation*. ACM Trans. Comput.-Hum. Interact., vol. 14, no. 3, September 2007. (Cited on page 75.)
- [Keskin 2011] C. Keskin, F. Kirac, Y.E. Kara and L. Akarun. *Real time hand pose estimation using depth sensors*. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pages 1228–1234. IEEE, 2011. (Cited on page 27.)
- [Khan 2011] S. Khan, D. Bailey and G.S. Gupta. *Delayed absolute difference (DAD) signatures of dynamic features for sign language segmentation*. In Automation, Robotics and Applications (ICARA), 2011 5th International Conference on, pages 109–114. IEEE, 2011. (Cited on pages 61 and 62.)

- [Kim 1996] J.S. Kim, W. Jang and Z. Bien. *A dynamic gesture recognition system for the Korean sign language (KSL)*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 26, no. 2, pages 354–359, 1996. (Cited on page 24.)
- [Kim 2005] J.H. Kim, D.G. Kim, J.H. Shin, S.W. Lee and K.S. Hong. *Hand gesture recognition system using fuzzy algorithm and RDBMS for post PC*. Fuzzy Systems and Knowledge Discovery, pages 487–487, 2005. (Cited on page 24.)
- [Kim 2009] S.H. Kim, H.J. Yang and K.S. Ng. *Temporal sign language analysis based on DTW and incremental model*. In Communications (MICC), 2009 IEEE 9th Malaysia International Conference on, pages 586–591. IEEE, 2009. (Cited on page 68.)
- [Kipp 2001] M. Kipp. *Anvil - A Generic Annotation Tool for Multimodal Dialogue*. In Proc. of 7th European Conference on Speech Communication and Technology (Eurospeech), pages 1367–1370, 2001. (Cited on page 33.)
- [Kipp 2011] Michael Kipp, Alexis Heloir and Quan Nguyen. *Sign Language Avatars: Animation and Comprehensibility*. In Hannes Vilhjálmsón, Stefan Kopp, Stacy Marsella and Kristinn Thórisson, editors, Intelligent Virtual Agents, volume 6895 of *Lecture Notes in Computer Science*, pages 113–126. Springer Berlin / Heidelberg, 2011. (Cited on page 75.)
- [Kiruluta 2002] A. Kiruluta, M. Eizenman and S. Pasupathy. *Predictive head movement tracking using a Kalman filter*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 27, no. 2, pages 326–331, 2002. (Cited on page 51.)
- [Kitagawa 1996] G. Kitagawa. *Monte Carlo filter and smoother for non-Gaussian non-linear state space models*. Journal of computational and graphical statistics, pages 1–25, 1996. (Cited on page 94.)
- [Klima 1980] E. Klima and U. Bellugi. *The sign of language*. The Sign of Language, 1980. (Cited on page 41.)
- [Koizumi 2002] A. Koizumi, H. Sagawa and M. Takeuchi. *An annotated japanese sign language corpus*. In Proc. Int’l Conf. Language Resources and Evaluation, volume 3, pages 927–930, 2002. (Cited on page 31.)
- [Kong 2008] WW Kong and S. Ranganath. *Automatic hand trajectory segmentation and phoneme transcription for sign language*. In Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference on, pages 1–6. IEEE, 2008. (Cited on pages 63 and 70.)
- [Kovac 2003] J. Kovac, P. Peer and F. Solina. *Human skin color clustering for face detection*. In EUROCON International Conference on Computer as a Tool, volume 2, pages 144–148, 2003. (Cited on page 85.)
- [Kulkarni 2010] V.S. Kulkarni and SD Lokhande. *Appearance Based Recognition of American Sign Language Using Gesture Segmentation*. International Journal on Computer Science and Engineering, vol. 2, no. 03, pages 560–565, 2010. (Cited on page 49.)

- [Lee 1999] H.J. Lee and J.H. Chung. *Hand gesture recognition using orientation histogram*. In TENCON 99. Proceedings of the IEEE Region 10 Conference, volume 2, pages 1355–1358. IEEE, 1999. (Cited on page 48.)
- [Lee 2002] M.W. Lee, I. Cohen and S.K. Jung. *Particle filter with analytical inference for human body tracking*. In Motion and Video Computing, 2002. Proceedings. Workshop on, pages 159–165. IEEE, 2002. (Cited on page 54.)
- [Lee 2009] Y.H. Lee and C.Y. Tsai. *Taiwan sign language (TSL) recognition based on 3D data and neural networks*. Expert Systems with Applications, vol. 36, no. 2, pages 1123–1128, 2009. (Cited on page 24.)
- [Lefebvre-Albaret 2008] F. Lefebvre-Albaret and P. Dalle. *Une approche de segmentation de la Langue des Signes Française*. Traitement Automatique des Langues Naturelles, 2008. (Cited on pages 33, 62 and 64.)
- [Lefebvre-Albaret 2009] F. Lefebvre-Albaret and P. Dalle. *Body posture estimation in a sign language video*. In Proc of The 8th International Gesture Workshop, Feb 2009. (Cited on pages 28 and 50.)
- [Lefebvre-Albaret 2010] F. Lefebvre-Albaret. *Traitement automatique de vidéos en LSF, modélisation et exploitation des contraintes phonologiques du mouvement*. Phd thesis, University of Toulouse, October 2010. (Cited on pages 48, 50, 51, 52, 55, 93, 96, 105, 106, 107, 108 and 154.)
- [Lefebvre-Albaret 2012] F. Lefebvre-Albaret and J. Segouat. *Influence de la segmentation temporelle sur la caractérisation de signes*. Defi Geste Langue des Signes, JEP-TALN-RECITAL, pages 73–83, June 2012. (Cited on page 60.)
- [Lenseigne 2005] B. Lenseigne and P. Dalle. *Modélisation de l'espace discursif pour l'analyse de la langue des signes*. Inorkshop on Traitement Automatique des Langues des Signes, TALN, 2005. (Cited on pages ix, 14 and 15.)
- [Liang 1998] R.H. Liang and M. Ouhyoung. *A real-time continuous gesture recognition system for sign language*. In Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pages 558–567. IEEE, 1998. (Cited on pages xi and 63.)
- [Lichtenauer 2008a] J. Lichtenauer, E. Hendriks and M. Reinders. *Learning to recognize a sign from a single example*. In Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, pages 1–6. IEEE, 2008. (Cited on pages 70 and 71.)
- [Lichtenauer 2008b] J.F. Lichtenauer, E.A. Hendriks and M.J. Reinders. *Sign language recognition by combining statistical DTW and independent classification*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 11, pages 2040–2046, 2008. (Cited on page 68.)
- [Liddell 1984] S.K. Liddell. *THINK and BELIEVE: sequentiality in American Sign Language*. Language, pages 372–399, 1984. (Cited on pages 32 and 42.)
- [Liddell 1989] S.K. Liddell and R.E. Johnson. *American sign language: The phonological base*. Linguistics of American Sign Language, vol. 64, pages 195–277, 1989. (Cited on pages 42 and 43.)

- [Lippmann 1987] R. Lippmann. *An introduction to computing with neural nets*. ASSP Magazine, IEEE, vol. 4, no. 2, pages 4–22, 1987. (Cited on page 67.)
- [Lombardo 2010] V. Lombardo, F. Nunnari and R. Damiano. *A Virtual Interpreter for the Italian Sign*. In Intelligent Virtual Agents: 10th International Conference, IVA 2010, page 201, 2010. (Cited on page 75.)
- [Lombardo 2011] V. Lombardo, F. Nunnari and R. Damiano. *The ATLAS Interpreter of the Italian Sign Language*. In International Workshop on Sign Language Translation and Avatar Technology (SLTAT) International Workshop on Sign Language Translation and Avatar Technology (SLTAT) International Workshop on Sign Language Translation and Avatar Technology (SLTAT), 2011. (Cited on page 77.)
- [Losson 2000] O. Losson. *Modélisation de geste communicatif et réalisation d’un signeur virtuel de phases en langue des signes française*. PhD thesis, Université de Lille, 2000. (Cited on page 43.)
- [LS-COLIN 2002] LS-COLIN. *13:08 - 15:15 le 11 septembre 2001 par Nasredine Chab* http://corpusdelaparoie.in2p3.fr/spip.php?article30&ldf_id=oai:crdo.vjf.cnrs.fr:crdo-FSL-CUC020_SOUND, 2002. (Cited on pages ix, 16, 17, 29, 107, 114, 137, 138 and 140.)
- [Lu 2003] Shan Lu, D. Metaxas, D. Samaras and J. Oliensis. *Using multiple cues for hand tracking and model refinement*. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 2, pages II – 443–50 vol.2, June 2003. (Cited on pages x, 48, 50 and 51.)
- [Lu 2009] P. Lu and M. Huenerfauth. *Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects*. In Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility, pages 83–90. ACM, 2009. (Cited on pages 24 and 75.)
- [Lu 2010a] P. Lu. *Modeling animations of American Sign Language verbs through motion-capture of native ASL signers*. ACM SIGACCESS Accessibility and Computing, no. 96, pages 41–45, 2010. (Cited on page 24.)
- [Lu 2010b] P. Lu and M. Huenerfauth. *Collecting a motion-capture corpus of American Sign Language for data-driven generation research*. In Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, pages 89–97. Association for Computational Linguistics, 2010. (Cited on page 24.)
- [MacCormick 2000a] J. MacCormick. *Probabilistic models and stochastic algorithms for visual tracking*. PhD thesis, PhD thesis, University of Oxford, 2000. (Cited on page 95.)
- [MacCormick 2000b] J. MacCormick and A. Blake. *A probabilistic exclusion principle for tracking multiple objects*. International Journal of Computer Vision, vol. 39, no. 1, pages 57–71, 2000. (Cited on pages 92 and 100.)
- [MacCormick 2000c] J. MacCormick and M. Isard. *Partitioned sampling, articulated objects, and interface-quality hand tracking*. Computer Vision—ECCV 2000, pages 3–19, 2000. (Cited on page 95.)

- [MacLaughlin 1997] D. MacLaughlin. *The structure of determiner phrases: Evidence from American Sign Language*. PhD thesis, Boston University, 1997. (Cited on page 14.)
- [Mak 2011] Joe Mak and Tang Gladys. Movement types, repetition, and feature organization in hong kong sign language. *formational units in sign language*, volume 3 of *Sign Language Topology*. Nijmegen/Berlin: Ishara Press/Mouton de Gruyter, 2011. (Cited on page 41.)
- [Mandel 1981] M.A. Mandel. *Phonotactics and morphophonology in american sign language*, volume 1981. University of California, Berkeley, 1981. (Cited on page 39.)
- [Marshall 2003] Ian Marshall and Éva Sáfár. *A prototype text to British Sign Language (BSL) translation system*. In Proc. of the 41st Annual Meeting on Asso. for Computational Linguistics. Vol 2, ACL '03, pages 113–116, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. (Cited on page 75.)
- [Martin 2004] D.R. Martin, C.C. Fowlkes and J. Malik. *Learning to detect natural image boundaries using local brightness, color, and texture cues*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 26, no. 5, pages 530–549, 2004. (Cited on page 48.)
- [Matthes 2010] Silke Matthes, Thomas Hanke, Jakob Storz, Eleni Efthimiou, Nassia Dimiou, Panagiotis Karioris, Annelies Braffort, Annick Choisier, Julia Pelhate and Eva Safar. *Elicitation Tasks and Materials designed for Dicta-Sign's Multilingual Corpus*. International workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC), Valletta, Malte, pages 158–163, May 2010. (Cited on page 23.)
- [Maung 2009] T.H.H. Maung. *Real-time hand tracking and gesture recognition system using neural networks*. World Academy of Science, Engineering and Technology, vol. 50, pages 466–470, 2009. (Cited on page 48.)
- [Meyer 1992] K. Meyer, H.L. Applewhite and F.A. Biocca. *A survey of position trackers*. In Presence: Teleoperators and Virtual Environments (ISSN 1054-7460), vol. 1, no. 2, p. 173–200., volume 1, pages 173–200, 1992. (Cited on page 24.)
- [Micilotta 2004] A. Micilotta and R. Bowden. *View-based location and tracking of body parts for visual interaction*. In Proc. of British Machine Vision Conference, volume 2, pages 849–858, 2004. (Cited on pages 48, 50, 51, 55, 93 and 96.)
- [Mikic 2001] I. Mikic, M. Trivedi, E. Hunter and P. Cosman. *Articulated body posture estimation from multi-camera voxel data*. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, pages I–455. IEEE, 2001. (Cited on page 53.)
- [Millet 2012] Agnès Millet and Isabelle Estève. *Segmenter et annoter le discours d'un locuteur de LSF : permanence formelle et variabilité fonctionnelle des unités*. DEfi Geste Langue des Signes, pages 57–72, June 2012. (Cited on page 60.)
- [Mlouka 2011] M.B. Mlouka, J. Dalle and P. Dalle. *DEGELS2011: Analyses d'Annotations pour la reconnaissance*. 2011. (Cited on page 29.)

- [Moeslund 1999] T.B. Moeslund. *Computer vision-based human motion capture—a survey*. University of Aalborg Technical Report LIA, vol. 99, no. 02, 1999. (Cited on pages ix and 24.)
- [Munib 2007] Q. Munib, M. Habeeb, B. Takruri and H.A. Al-Malik. *American sign language (ASL) recognition based on Hough transform and neural networks*. Expert Systems with Applications, vol. 32, no. 1, pages 24–37, 2007. (Cited on page 68.)
- [Myers 1980] C. Myers, L. Rabiner and A. Rosenberg. *Performance tradeoffs in dynamic time warping algorithms for isolated word recognition*. Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 28, no. 6, pages 623–635, 1980. (Cited on page 67.)
- [Nayak 2009] S. Nayak, S. Sarkar and B. Loeding. *Automated extraction of signs from continuous sign language sentences using iterated conditional modes*. IEEE Conference CVPR, pages 2583–2590, June 2009. (Cited on page 64.)
- [Neidle 2001] C. Neidle, S. Sclaroff and V. Athitsos. *SignStream: A tool for linguistic and computer vision research on visual-gestural language data*. Behavior Research Methods, vol. 33, no. 3, pages 311–320, 2001. (Cited on page 33.)
- [Newkirk 1998] D. Newkirk. *On the temporal segmentation of movement in American Sign Language*. Sign language & linguistics, vol. 1, no. 2, pages 173–211, 1998. (Cited on page 42.)
- [Ong 2004] E.J. Ong and R. Bowden. *A boosted classifier tree for hand shape detection*. In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pages 889–894. IEEE, 2004. (Cited on pages 49 and 58.)
- [Ong 2005a] S.C.W. Ong and S. Ranganath. *Automatic sign language analysis: a survey and the future beyond lexical meaning*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pages 873–891, 2005. (Cited on pages 13, 47 and 66.)
- [Ong 2005b] S.C.W. Ong and S. Ranganath. *Automatic sign language analysis: A survey and the future beyond lexical meaning*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, no. 6, pages 873–891, 2005. (Cited on pages 21, 23 and 59.)
- [O’Rourke 1980] J. O’Rourke and N.I. Badler. *Model-based image analysis of human motion using constraint propagation*. IEEE TRANS. PATTERN ANALY. AND MACH. INTELLIG., vol. 2, no. 6, pages 522–536, 1980. (Cited on page 52.)
- [Paulraj 2010] M.P. Paulraj, S. Yaacob, M.S. bin Zanar Azalan and R. Palaniappan. *A phoneme based sign language recognition system using skin color segmentation*. In Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium on, pages 1–5. IEEE, 2010. (Cited on page 70.)
- [Perlmutter 1992] D.M. Perlmutter. *Sonority and syllable structure in American Sign Language*. Linguistic Inquiry, vol. 23, no. 3, pages 407–442, 1992. (Cited on page 42.)

- [Pezeshkpour 1999] F. Pezeshkpour, I. Marshall, R. Elliott and JA Bangham. *Development of a legible deaf-signing virtual human*. In Multimedia Computing and Systems, 1999. IEEE International Conference on, volume 1, pages 333–338. IEEE, 1999. (Cited on page 73.)
- [Phan 2009] D. Phan, T. Nguyen and T. Bui. *A 3D conversational agent for presenting digital information for deaf people*. Agent Computing and Multi-Agent Systems, pages 319–328, 2009. (Cited on page 73.)
- [Phung 2005] S.L. Phung, A. Bouzerdoun Sr and D. Chai Sr. *Skin segmentation using color pixel classification: analysis and comparison*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, no. 1, pages 148–154, 2005. (Cited on page 48.)
- [Piater 2010] J. Piater, T. Hoyoux and W. Du. *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, pages 22–23. Citeseer, 2010. (Cited on page 51.)
- [Pitsikalis 2010] V. Pitsikalis, S. Theodorakis and P. Maragos. *Data-driven sub-units and modeling structure for continuous sign language recognition with multiple cues*. In LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, volume 1, 2010. (Cited on page 64.)
- [Pitsikalis 2011] V. Pitsikalis, S. Theodorakis, C. Vogler and P. Maragos. *Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition*. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, pages 1–6. IEEE, 2011. (Cited on page 70.)
- [Prillwitz 1989] S. Prillwitz, R. Leven, H. Zienert, T. Hamke and J. Henning. *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introductory Guide*. International Studies on Sign Language and Communication of the Deaf. Sigmun, vol. 5, page 40, 1989. (Cited on pages 4, 29 and 42.)
- [Rabiner 2010] L. Rabiner and R. Schafer. *Theory and applications of digital speech processing*. Prentice Hall Press, 2010. (Cited on page 67.)
- [Raducanu 2006] B. Raducanu and Y. Vitrià. *A Robust Particle Filter-Based Face Tracker Using Combination of Color and Geometric Information*. Image Analysis and Recognition, pages 922–933, 2006. (Cited on page 52.)
- [Ramamoorthy 2003] A. Ramamoorthy, N. Vaswani, S. Chaudhury and S. Banerjee. *Recognition of dynamic hand gestures*. Pattern Recognition, vol. 36, pages 2069–2081, 2003. (Cited on pages 55 and 111.)
- [Roberts 2004] T. Roberts, S. McKenna and I. Ricketts. *Human pose estimation using learnt probabilistic region similarities and partial configurations*. Computer Vision-ECCV 2004, pages 291–303, 2004. (Cited on page 50.)
- [Sagawa 2000a] H. Sagawa and M. Takeuchi. *A method for recognizing a sequence of sign language words represented in a japanese sign language sentence*. In Automatic

- Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pages 434–439. IEEE, 2000. (Cited on page 33.)
- [Sagawa 2000b] H. Sagawa and M. Takeuchi. *A method for recognizing a sequence of sign language words represented in a japanese sign language sentence*. In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pages 434–439. IEEE, 2000. (Cited on page 63.)
- [Sallandre 2003] M.-A. Sallandre. *Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d’une grammaire de l’iconicité*. PhD thesis, Université Paris 8, 2003. (Cited on page 16.)
- [Sandler 1989] W. Sandler. Phonological representation of the sign: Linearity and non-linearity in american sign language, volume 32. Foris Pubns USA, 1989. (Cited on page 42.)
- [Sandler 2012] W. Sandler. *The Phonological Organization of Sign Languages*. Language and Linguistics Compass, vol. 6, no. 3, pages 162–182, 2012. (Cited on pages x, 14, 38, 39 and 40.)
- [Segouat 2010] J. Segouat and A. Braffort. *Toward Modeling Sign Language Coarticulation*. volume Gesture in Embodied Communication and Human-Computer Interaction, 2010. (Cited on pages 18 and 74.)
- [Shakhnarovich 2003] G. Shakhnarovich, P. Viola and T. Darrell. *Fast pose estimation with parameter-sensitive hashing*. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pages 750–757. IEEE, 2003. (Cited on page 48.)
- [Shi 1994] J. Shi and C. Tomasi. *Good features to track*. In Computer Vision and Pattern Recognition, 1994. Proceedings CVPR’94., 1994 IEEE Computer Society Conference on, pages 593–600. IEEE, 1994. (Cited on page 47.)
- [Shiosaki 2008] T. Shiosaki, T. Matsuo, Y. Shirai and N. Shimada. *Motion Segmentation using Hand Movement and Hand Shape for Sign Language Recognition*. 2008. (Cited on page 57.)
- [Sidenbladh 2000] H. Sidenbladh, M. Black and D. Fleet. *Stochastic tracking of 3D human figures using 2D image motion*. Computer Vision—ECCV 2000, pages 702–718, 2000. (Cited on page 53.)
- [Simon 1908] Th. Simon and A. Binet. *Étude sur l’art d’enseigner la parole aux sourds-muets*. L’année psychologique, vol. 15, no. 1, pages 373–396, 1908. (Cited on page 12.)
- [Smith 2007] Paul Smith, Niels da Vitoria Lobo and Mubarak Shah. *Resolving hand over face occlusion*. Image and Vision Computing, vol. 25, no. 9, pages 1432 – 1448, 2007. (Cited on pages x, 56 and 59.)
- [Soontranon 2005] N. Soontranon, S. Aramvith and TH Chalidabhongse. *Improved face and hand tracking for sign language recognition*. In Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on, volume 2, pages 141–146, 2005. (Cited on page 50.)

- [Starner 1995a] T. Starner and A. Pentland. *Real-time american sign language recognition from video using hidden markov models*. In Computer Vision, 1995. Proceedings., International Symposium on, pages 265–270. IEEE, 1995. (Cited on page 47.)
- [Starner 1995b] T.E. Starner. *Visual Recognition of American Sign Language Using Hidden Markov Models*. Rapport technique, DTIC Document, 1995. (Cited on page 68.)
- [Stein 2006] D. Stein, J. Bungeroth and H. Ney. *Morpho-syntax based statistical methods for sign language translation*. In 11th Annual conference of the European Association for Machine Translation, Oslo, Norway, pages 223–231, 2006. (Cited on page 27.)
- [Stenger 2001] B. Stenger, P.R.S. Mendonça and R. Cipolla. *Model-based 3D tracking of an articulated hand*. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 2, pages II–310. IEEE, 2001. (Cited on pages 51 and 52.)
- [Stokoe 1960] W. Stokoe. *Sign Language structure: an outline of the visual communication system of the American Deaf*. Studies in linguistics, occasional papers, vol. 8, 1960. (Cited on pages 39 and 41.)
- [Stokoe 1976] William C. Stokoe, Dorothy C. Casterline and Carl G. Croneberg. A dictionary of american sign language on linguistic principles. Linstok Press (Silver Spring, Md.), 1976. (Cited on pages 15 and 41.)
- [Stokoe 1980] WC Stokoe. *Sign language structure*. Annual Review of Anthropology, vol. 9, no. 1, pages 365–390, 1980. (Cited on page 32.)
- [Suszczańska 2002] N. Suszczańska, P. Szmal and J. Francik. *Translating Polish Texts into Sign Language in the TGT System*. In 20th IASTED International Multi-Conference. Applied Informatics AI, pages 282–287, 2002. (Cited on page 75.)
- [Sutherland 1996] A. Sutherland. *Real-time video-based recognition of sign language gestures using guided template matching*. In Proceedings of Gesture Workshop on Progress in Gestural Interaction, pages 31–38. Springer-Verlag, 1996. (Cited on page 47.)
- [Sutton 1995] V.J. Sutton. Lessons in sign writing. Center Sutton Movement Writing, 1995. (Cited on page 4.)
- [Tanibata 2002] N. Tanibata, N. Shimada and Y. Shirai. *Extraction of hand features for recognition of sign language words*. In The 15th International Conference on Vision Interface, pages 391–398, 2002. (Cited on pages xi, 49, 51, 52, 56 and 57.)
- [Theodorakis 2009] S. Theodorakis, A. Katsamanis and P. Maragos. *Product-HMMs for automatic sign language recognition*. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, volume 00, pages 1601–1604. IEEE Computer Society, 2009. (Cited on page 68.)
- [Thompson 2006] R. Thompson, K. Emmorey and R. Kluender. *The relationship between eye gaze and verb agreement in American Sign Language: an eye-tracking study*.

- Natural Language & Linguistic Theory, vol. 24, no. 2, pages 571–604, 2006. (Cited on page 14.)
- [Vamplew 1998] P. Vamplew and A. Adams. *Recognition of sign language gestures using neural networks*. Australian Journal of Intelligent Information Processing Systems, vol. 5, no. 2, pages 94–102, 1998. (Cited on page 68.)
- [Vezhnevets 2003] V. Vezhnevets, V. Sazonov and A. Andreeva. *A survey on pixel-based skin color detection techniques*. In Proc. Graphicon, volume 3. Moscow, Russia, 2003. (Cited on page 48.)
- [Viola 2002] P. Viola and M. Jones. *Robust real-time object detection*. International Journal of Computer Vision, vol. 57, no. 2, pages 137–154, 2002. (Cited on pages 50, 51, 97 and 98.)
- [Vlasic 2007] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik and J. Popović. *Practical motion capture in everyday surroundings*. In ACM Transactions on Graphics (TOG), volume 26, page 35. ACM, 2007. (Cited on page 24.)
- [Vogler 1997] C. Vogler and D. Metaxas. *Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods*. In Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on, volume 1, pages 156–161. IEEE, 1997. (Cited on pages 53 and 70.)
- [Vogler 1999a] C. Vogler and D. Metaxas. *Parallel hidden markov models for american sign language recognition*. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 1, pages 116–122. IEEE, 1999. (Cited on pages 69, 70 and 71.)
- [Vogler 1999b] C. Vogler and D. Metaxas. *Toward scalability in ASL recognition: Breaking down signs into phonemes*. Gesture-Based Communication in Human-Computer Interaction, pages 211–224, 1999. (Cited on pages 61 and 70.)
- [Vogler 2001] C. Vogler and D. Metaxas. *A framework for recognizing the simultaneous aspects of american sign language*. Computer Vision and Image Understanding, vol. 81, no. 3, pages 358–384, 2001. (Cited on pages 17, 33 and 67.)
- [Vogler 2003] C.P. Vogler. *American sign language recognition: reducing the complexity of the task with phoneme-based modeling and parallel hidden markov models*. 2003. (Cited on page 68.)
- [Vogler 2004] C. Vogler and D. Metaxas. *Handshapes and movements: Multiple-channel american sign language recognition*. Gesture-Based Communication in Human-Computer Interaction, pages 431–432, 2004. (Cited on pages 24 and 28.)
- [Vogler 2008] C. Vogler and S. Goldenstein. *Facial movement analysis in ASL*. Universal Access in the Information Society, vol. 6, no. 4, pages 363–374, 2008. (Cited on page 28.)
- [Von Agris 2008] U. Von Agris, J. Zieren, U. Canzler, B. Bauer and K.F. Kraiss. *Recent developments in visual sign language recognition*. Universal Access in the Infor-

- mation Society, vol. 6, no. 4, pages 323–362, 2008. (Cited on pages xi, 21, 56, 57, 58, 59, 66, 68, 69 and 70.)
- [Wang 1991] L.-C.T. Wang and C.C. Chen. *A combined optimization method for solving the inverse kinematics problems of mechanical manipulators*. Robotics and Automation, IEEE Transactions on, vol. 7, no. 4, pages 489–499, aug 1991. (Cited on page 76.)
- [Wang 2009] R.Y. Wang and J. Popović. *Real-time hand-tracking with a color glove*. In ACM Transactions on Graphics (TOG), volume 28, page 63. ACM, 2009. (Cited on pages x, 47, 48 and 53.)
- [Welch 2002] G. Welch and E. Foxlin. *Motion tracking: No silver bullet, but a respectable arsenal*. Computer Graphics and Applications, IEEE, vol. 22, no. 6, pages 24–38, 2002. (Cited on page 24.)
- [Wells 1999] M. Wells, F. Pezeshkpour, I. Marshall, M. Tutt and JA Bangham. *Simon: an innovative approach to signing on television*. In Proc. Int. Broadcasting Convention, volume 146, 1999. (Cited on page 73.)
- [Wittenburg 2006] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann and H. Sloetjes. *ELAN: a professional framework for multimodality research*. In Proc. of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pages 1556–1559, 2006. (Cited on pages 33 and 140.)
- [Wu 2001] Ying Wu and T.S. Huang. *Hand modeling, analysis and recognition*. Signal Processing Magazine, IEEE, vol. 18, no. 3, pages 51–60, May 2001. (Cited on page 50.)
- [Wu 2004] Y. Wu and T.S. Huang. *Robust visual tracking by integrating multiple cues based on co-inference learning*. International Journal of Computer Vision, vol. 58, no. 1, pages 55–71, 2004. (Cited on page 48.)
- [Yang 2002] M.H. Yang, N. Ahuja and M. Tabb. *Extraction of 2d motion trajectories and its application to hand gesture recognition*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 8, pages 1061–1074, 2002. (Cited on page 69.)
- [Yang 2005] C. Yang, R. Duraiswami and L. Davis. *Fast multiple object tracking via a hierarchical particle filter*. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 1, pages 212–219. IEEE, 2005. (Cited on page 48.)
- [Yang 2006a] R. Yang and S. Sarkar. *Detecting coarticulation in sign language using conditional random fields*. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, volume 2, pages 108–112. IEEE, 2006. (Cited on page 63.)
- [Yang 2006b] 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, pages 635–642, 2006. (Cited on page 33.)

- [Yang 2007] R. Yang, S. Sarkar and B. Loeding. *Enhanced level building algorithm for the movement epenthesis problem in sign language recognition*. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007. (Cited on page 17.)
- [Yang 2010] R. Yang, S. Sarkar and B. Loeding. *Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 3, pages 462–477, 2010. (Cited on page 64.)
- [Yeasin 2000] M. Yeasin and S. Chaudhuri. *Visual understanding of dynamic hand gestures*. Pattern Recognition, vol. 33, no. 11, pages 1805–1817, 2000. (Cited on page 70.)
- [Yilmaz 2006] A. Yilmaz, O. Javed and M. Shah. *Object tracking: A survey*. Acm Computing Surveys, vol. 38, no. 4, page 13, 2006. (Cited on pages x, 47 and 49.)
- [YoungJoon 2010] Chai YoungJoon, Shin SeungHo, Chang Kyusik and Kim TaeYong. *Real-Time User Interface using Particle Filter with Integral Histogram*. IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pages 510–515, May 2010. (Cited on page 51.)
- [Yuan 2005] Q. Yuan, S. Sclaroff and V. Athitsos. *Automatic 2D hand tracking in video sequences*. In Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on, volume 1, pages 250–256. IEEE, 2005. (Cited on page 28.)
- [Zafrulla 2011] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton and P. Presti. *American sign language recognition with the kinect*. In Proceedings of the 13th international conference on multimodal interfaces, pages 279–286. ACM, 2011. (Cited on page 27.)
- [Zahedi 2006] M. Zahedi, D. Keysers and H. Ney. *Pronunciation clustering and modeling of variability for appearance-based sign language recognition*. Gesture in Human-Computer Interaction and Simulation, pages 68–79, 2006. (Cited on page 28.)
- [Zhao 2007] L. Zhao and J. Tao. *Fast Facial Feature Tracking with Multi-Cue Particle Filter*. In International Conference on Image and Vision Computing. Hamilton, New Zealand, pages 7–12, 2007. (Cited on page 52.)
- [Zhou 2004] H. Zhou, D.J. Lin and T.S. Huang. *Static hand gesture recognition based on local orientation histogram feature distribution model*. In Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on, pages 161–161. IEEE, 2004. (Cited on page 48.)
- [Zhu 2000] X. Zhu, J. Yang and A. Waibel. *Segmenting hands of arbitrary color*. In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pages 446–453. IEEE, 2000. (Cited on page 55.)
- [Zieren 2004] J. Zieren and K.F. Kraiss. *Non-intrusive sign language recognition for human-computer interaction*. In Proc. IFAC/IFIP/IFORS/IEA symposium on analysis, design and evaluation of human machine systems, 2004. (Cited on page 27.)

