Annotation automatique des configurations manuelles de la Langue des Signes Française à partir de données capturées

Lucie Naert, Caroline Larboulette, Sylvie Gibet

Journées Françaises d'Informatique Graphique, Rennes,
23 - 27 Octobre 2017

Annotation automatique des configurations manuelles de la Langue des Signes Française à partir de données capturées

Lucie Naert, Caroline Larboulette et Sylvie Gibet
Laboratoire IRISA, Université Bretagne Sud
Campus de Tohannic, Vannes, France
lucie.naert, caroline.larboulette, sylvie.gibet@univ-ubs.fr

Résumé

La capture de mouvement permet d'aboutir à des corpus précis et complets d'énoncés en Langue des Signes Française (LSF). Ceux-ci peuvent ensuite servir de base aussi bien à une étude linguistique qu'à de la synthèse de nouveaux énoncés en LSF sur un agent virtuel. L'annotation de ces données à différents niveaux est une étape nécessaire mais qui peut s'avérer fastidieuse si réalisée manuellement. Cet article s'attache à automatiser l'annotation des données capturées sur l'un de ces niveaux : celui des configurations manuelles de la LSF. La méthode d'annotation se décompose en deux étapes : une première étape de segmentation utilise les variations des distances entre les articulations de chaque main pour séparer les données de mouvement en segments de deux types, configuration manuelle ou transition, tandis que la deuxième étape s'attache à reconnaître, à l'aide d'un algorithme de classification supervisé, les configurations manuelles exécutées durant les phases segmentées correspondantes.

Mots clefs : Segmentation automatique, Annotation automatique, Langue des Signes Française (LSF), Capture de mouvement, Apprentissage automatique supervisé

1. Introduction

La Langue des Signes Française (LSF) est une langue visuelle et gestuelle utilisée par une partie importante de la communauté sourde de France. Cependant, les informations en LSF sont rares et passent souvent par l'utilisation de la vidéo. Or, cette dernière ne permet pas de préserver l'anonymat du signeur et les opérations d'édition y sont très complexes. Les avatars signeurs en LSF constituent ainsi un domaine de recherche très prometteur puisque leur utilisation permet de pallier en partie les limitations de la vidéo.

Nos travaux sur les avatars de LSF gravitent autour de la synthèse de signes "basée données", c'est-à-dire en utilisant les données de mouvement de signeurs humains. Des énoncés en LSF sont ainsi enregistrés par capture de mouvement (*Motion Capture*) avant d'être traités pour être utilisés ensuite dans un moteur de synthèse. Un de ces traitements consiste à annoter précisément la base des mouvements capturés sur les différents canaux de communication. Cette annotation, faite manuellement par des sourds, experts en annotation de données de LSF, est une étape laborieuse et qui est parfois l'occasion d'erreurs ou d'approximations.

La plupart des travaux existants sur la segmentation automatique des données de langues des signes partent de séquences vidéos et cherchent à segmenter à un niveau **signe** [YS06, LADG08, KPBB02]. Une segmentation au même niveau, utilisant les caractéristiques cinématiques des poignets, a été réalisée sur des données de *Motion Capture* dans [NLG17]. Ces segmentations ne prennent pas en considération l'aspect multicanal des langues des signes et aboutissent à une segmentation de haut-niveau, très dépendante du contexte de l'énoncé segmenté et ainsi difficilement réutilisable dans des contextes différents ou pour synthétiser de nouveaux énoncés. Une segmentation de plus bas niveau doit faciliter la composition de signes dans des contextes variés. En 1960, Stokoe [Sto60] donne une structure phonologique aux signes en définissant trois composantes permettant de décrire l'ensemble des signes : le **mouvement**, l'**emplacement** et la **configuration des mains**. À ces trois composantes initiales, deux composantes sont ensuite ajoutées : l'**orientation des mains** [Bat78] et les **expressions faciales**. Utiliser cette structure phonologique comme base de la segmentation est pertinent : les segments sont d'un niveau plus fin et conservent une valeur linguistique. Ainsi, Réverdy et al. [RGLM16] s'attachent à segmenter et à annoter les expressions faciales de la LSF à partir de données de *Motion Capture*. Héloir et al. [HGMC05], quant à eux, font une analyse en composantes principales (PCA) pour segmenter les configurations manuelles de l'alphabet dactylologique de la LSF.

Dans cet article, nous nous intéressons plus généralement à l'annotation automatique du canal des configurations manuelles. La figure 1 montre un exemple d'annotation sous le logiciel ELAN : le canal des configurations manuelles de la main droite y est encadré. Ces configurations correspondent à des postures discriminantes et, parfois, signi-

fiantes des mains (voir exemples de la figure 3). Leur nombre et leur nature changent en fonction des sources. Cuxac liste 39 configurations [Cux00] † pour la LSF tandis que Boutora en recense 77 [Bou06]. Dans cet article, nous utilisons les 33 configurations manuelles définies dans les annotations du corpus *Sign3D* [GLH*16].

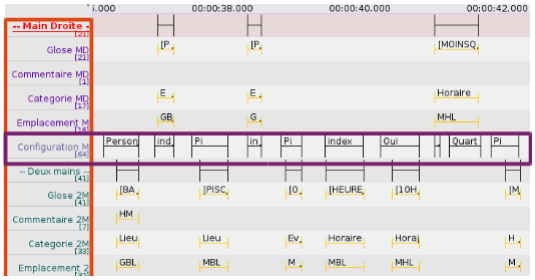

Figure 1: *Annotation sous le logiciel ELAN. Le cadre vertical orange montre une partie des canaux choisis pour représenter la LSF dans le projet* Sign3D *[GLH*16]. Le cadre horizontal violet délimite les annotations des configurations de la main droite.*

La méthode d'annotation automatique des configurations manuelles présentée ici comprend deux étapes. L'étape de segmentation permet de distinguer les séquences de *configuration manuelle* des séquences de *transition* entre deux configurations manuelles. L'étape de reconnaissance s'attache à déterminer la nature précise de la configuration pour chaque frame des segments de type *configuration manuelle* déterminés à la phase précédente. Un segment est ensuite étiqueté selon le label présent majoritairement sur celui-ci (voir figure 2).

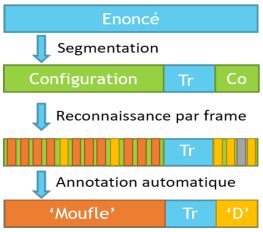

Figure 2: *Vue d'ensemble.*

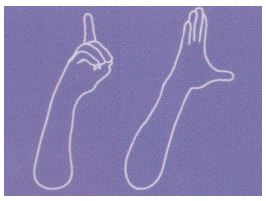

Figure 3: *Les configurations manuelles 'D' et 'Moufle'.*

Dans la suite de cet article, la section 2 décrit l'acquisition des données. La section 3 s'attache à la phase de segmentation des séquences de mouvement tandis que la section 4 détaille la méthode et les résultats obtenus en reconnaissance de configuration manuelle frame à frame. La section 5 décrit l'utilisation consécutive de la segmentation et de la reconnaissance conduisant à l'annotation automatique.

2. Acquisition des données

2.1. Corpus

Les données étudiées dans cet article sont issues d'un corpus rassemblant plusieurs énoncés signés en LSF. Ce corpus a été constitué par *Motion Capture* dans le cadre du projet *Sign3D* [GLH*16] destiné à la synthèse de LSF sur un humain virtuel. Il est composé de 8 séquences de mouvements contenant chacune plusieurs énoncés ou des signes isolés. Les séquences ont été réalisées par un seul signeur. L'acquisition des données est multimodale : des marqueurs ont été placés sur l'ensemble du corps du sujet pour capturer en simultané les mouvements réalisés sur tous les canaux de communication (visage, main, buste, regard). Le corpus n'est donc pas limité à l'étude des configurations manuelles. La figure 4 montre notre avatar de LSF animé par *Motion Capture*.

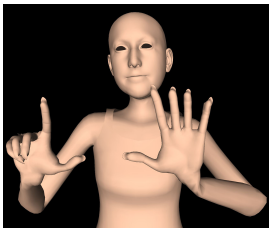

Figure 4: *Avatar de LSF réalisant les configurations manuelles '2' et '5'.*

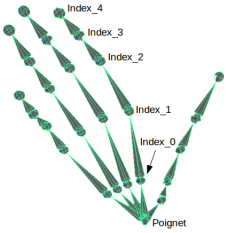

Figure 5: *Modèle du squelette de la main utilisé pour le calcul des distances.*

2.2. Données d'étude : les distances entre articulations

Le modèle des mains utilisé dans cette étude est montré dans la figure 5. D'après ce modèle, chaque main comporte 26 articulations (5 par doigt et 1 pour le poignet). Contrairement aux positions et orientations des articulations qui varient en fonction du repère choisi, les distances euclidiennes entre deux articulations sont invariantes quel que soit le référentiel. Nous avons donc choisi d'utiliser ces distances, au nombre de 325, pour cette étude. Les distances entre deux articulations consécutives (comme les articulations "Index_1" et "Index_2" de la figure 5), correspondant biologiquement à des os, varient naturellement très peu et ne contiennent pas d'information. Ces distances, dites *rigides*, n'ont pas été utilisées dans la segmentation mais l'ont été dans certaines expérimentations de l'étape de reconnaissance pour étudier l'impact du choix des distances.

3. Segmentation

Pendant la réalisation d'un énoncé en LSF, le signeur alterne entre des périodes stables, où la configuration manuelle varie très peu, et des périodes de transitions entre deux configurations. L'étape de segmentation consiste donc à séparer les énoncés en segments de deux types : *configuration manuelle* ou *transition*. Les segments de type *configuration manuelle* sont ensuite étiquetés en fonction de la nature de la configuration dans une étape de reconnaissance (section 4).

Pour réaliser la segmentation, la variation des distances entre les différentes articulations de la main est considérée. L'équation (1) montre le calcul de la variation des distances pour la frame n et pour la main droite (MD). $d(ij)$ désigne la distance entre l'articulation i et l'articulation j. Il s'agit donc de faire une somme de la variation des distances

†. Il précise toutefois qu'il ne s'agit pas d'un "inventaire fermé".

entre deux frames consécutives. La figure 6 permet de visualiser cette variation ainsi que la segmentation manuelle correspondante. Les pics de la courbe de variation (excepté le pic (1) ‡) ont lieu entre deux segments de type *configuration manuelle*, confirmant que la variation des distances est plus importante lors des segments de transitions que lors des périodes désignées par les annotateurs comme étant des configurations manuelles. La segmentation automatique repose ainsi sur l'utilisation d'un seuil. Si la variation dépasse ce seuil, un segment *transition* sera détecté tandis que, si la variation est inférieure à ce seuil, le segment sera considéré comme un segment *configuration manuelle*.

$$VarDist_{n,MD} = \sum_{i \in MD} \sum_{\substack{j \in MD \\ d(ij) \neq rigide}} |d(ij)_n - d(ij)_{n-1}| \quad (1)$$

Pour évaluer les performances de notre segmentation en fonction du seuil choisi, le *Simple Matching Coefficient* (SMC) a été utilisé [SM]. Il permet de mesurer la similarité entre deux ensembles, ici la segmentation faite à la main contre la segmentation réalisée automatiquement. Le SMC se calcule en faisant le rapport du nombre de frames de recouvrement entre les deux segmentations sur le nombre de frames total. La figure 7 montre la variation du SMC de l'ensemble du corpus en fonction du seuil choisi. La valeur de similarité maximale (SMC = 80%) est atteinte pour un seuil d'une valeur de 9 cm. La figure 6 montre le résultat de la segmentation sur un des 25 énoncés du corpus pour ce seuil.

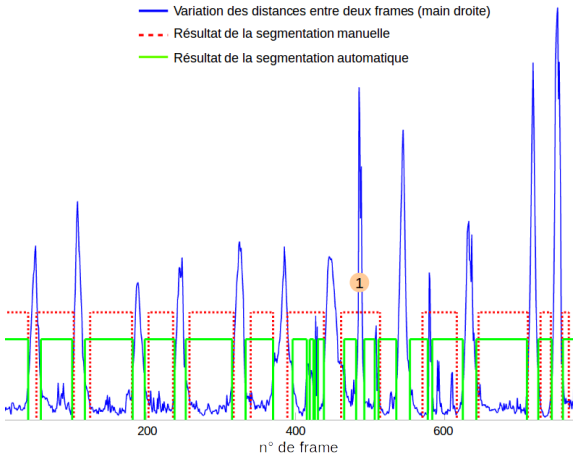

Figure 6: *Résultat de la segmentation sur un énoncé pour un seuil de 9 cm. Les porte rouges en pointillés et vertes, délimitent les segments de type configuration manuelle pour la segmentation manuelle et automatique respectivement.*

4. Reconnaissance des configurations manuelles frame à frame

Pour reconnaître les configurations manuelles, une méthode d'apprentissage supervisé de classification multiclasse

‡. Ce pic correspond à un contact entre la main droite et le bras gauche dans le signe [HEURE]. Il peut être la conséquence d'un mouvement induit par ce contact ou de bruit causé par ce contact pendant la capture (e.g. occultation d'un marqueur).

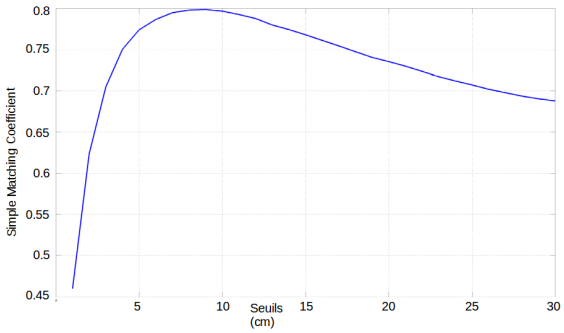

Figure 7: *Mesure de similarité sur l'ensemble du corpus en fonction du seuil choisi. Le maximum (SMC = 80%) est atteint pour un seuil d'une valeur de 9 cm.*

a été mise en place. Plus précisément, nous avons choisi d'appliquer une stratégie de classification de type One-versus-all [RK04] en utilisant un algorithme de régression logistique. Le corpus totalise 33 configurations manuelles différentes qui sont donc les 33 classes de notre classification. Les données d'apprentissage sont les distances euclidiennes entre chaque articulation de chaque main (on verra, par la suite, que nous avons également testé la reconnaissance sur un sous-ensemble de ces distances). Ces distances ont été calculées sur chaque frame annotée manuellement comme présentant une configuration manuelle de la LSF. On recense ainsi 27460 exemples $\times$ 325 distances. L'ensemble d'entraînement (*training set*) est constitué de 23533 de ces exemples (86% des données). Les 14% restants sont dédiés à l'ensemble de test (*test set*). Celui-ci correspond en pratique à l'une des huit séquences de mouvements capturées. Toutes les configurations manuelles présentes dans l'ensemble de test le sont également dans l'ensemble d'entraînement mais l'ensemble de test ne recouvre que 19 des 33 configurations manuelles du corpus.

La reconnaissance frame à frame a été appliquée sur l'ensemble de test. La précision de la reconnaissance en utilisant les 325 distances – distances *rigides* incluses – est de 83,6%. Lorsque seules les distances avec les plus fortes variations (qui sont souvent les distances entre les extrémités des doigts et une autre articulation) sont conservées, la précision est de 86,1%. Le temps d'entraînement est naturellement plus court lorsque moins de distances sont étudiées. Le choix des distances est un problème de *features selection* qu'une étude complémentaire nous permettra d'approfondir.

Dans la matrice de confusion de la figure 8, on peut voir, entre autres, des confusions entre :
Les configurations 'N' et 'U'. Il ne s'agit pas d'une erreur : il est normal que ces deux configurations soient confondues car elles sont identiques du point de vue de la posture de la main, seule l'orientation du poignet est différente.
Les configurations 'O' et 'C'. Cette forte confusion est caractéristique d'une confusion due à la *Motion Capture* : le contact entre le pouce et les autres doigts pour la configuration du 'O' peut être mal détecté ou reconstruit imparfaitement, entraînant la reconnaissance d'un 'C'.
Les configurations 'X fermé' et 'X ouvert'. Ces configurations sont également très proches. Les annotateurs eux-

même ont parfois du mal à les différencier. Des erreurs dans le corpus d'entraînement ont été relevées ce qui peut expliquer les erreurs de la reconnaissance.

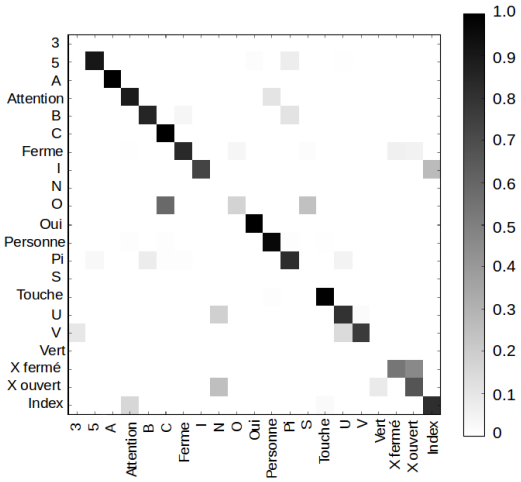

Figure 8: *Matrice de confusion entre les configurations manuelles pour l'ensemble de test.*

5. Annotation automatique des configurations manuelles

Il s'agit maintenant d'annoter automatiquement les configurations manuelles d'une séquence de mouvements continue en appliquant :
1. la segmentation pour séparer les phases de configurations stables des phases de transition (section 3), puis
2. la reconnaissance frame à frame sur les segments de type *configuration manuelle* (section 4), et enfin,
3. en déterminant la classe majoritaire dans chaque segment.

Les résultats de l'annotation automatique d'un énoncé de LSF (ne faisant pas partie de l'ensemble d'entraînement pour l'étape de reconnaissance) sont visibles sur la figure 9 pour un seuil de segmentation fixé ici à 7 cm (seuil obtenant les meilleurs résultats pour cet énoncé particulier). Les labels attribués sont cohérents avec les labels de l'annotation réalisée à la main par des experts (les 11 labels de l'annotation manuelle de cet exemple sont détectés correctement) bien qu'une différence soit visible (le label 'Touche' détecté automatiquement n'apparaît pas dans l'annotation manuelle). Le taux de recouvrement (SMC) au niveau de l'annotation pour la séquence en question est de 68,2% sachant que le taux de recouvrement de la segmentation seule est de 71,5%. La précision de la reconnaissance est donc de 95,3% sur cet exemple. L'augmentation significative obtenue par rapport à la précision de 86,1% de la reconnaissance frame à frame (section 4) prouve l'intérêt d'utiliser le label majoritaire sur les segments de type *configuration manuelle*.

6. Conclusion, discussion et travaux futurs

Cet article présente une technique d'annotation automatique de configurations manuelles de la LSF à partir de données capturées. L'étape de segmentation donne un taux de recouvrement avec la segmentation manuelle et pour l'ensemble du corpus pouvant aller jusqu'à 80%. La précision

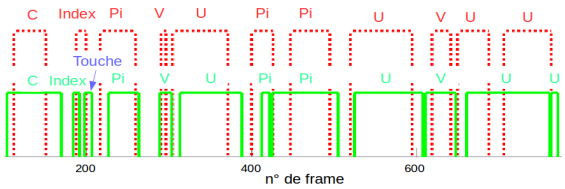

Figure 9: *Résultat de l'annotation automatique sur un énoncé. Les labels indiquent la nature des configurations manuelles telles qu'annotées manuellement (en rouge) ou trouvées automatiquement (en vert).*

de la reconnaissance par frame est de 86% mais elle peut être grandement améliorée avec la détermination de la classe majoritaire par segment. Nous aboutissons ainsi à un premier système d'annotation automatique des configurations manuelles performant. Les différences relevées entre l'annotation réalisée à la main et l'annotation automatique peuvent être dues à : (i) des erreurs d'annotation manuelle : les annotateurs peuvent confondre deux configurations proches ou estimer de façon approximative les limites temporelles de début et de fin de la configuration ; (ii) du bruit enregistré lors de la *Motion Capture* ; (iii) des erreurs lors de la reconstruction du squelette ou de la modélisation de la main.

Ce travail constitue une étude préliminaire qu'il est nécessaire d'approfondir. La conception d'un corpus dédié à l'étude de ces configurations est une première étape pour raffiner le processus présenté. Le corpus étudié est, en effet, insuffisant par rapport à nos objectifs de synthèse. De plus, il ne permet pas de tester la classification entre sujets puisqu'un seul signeur y a contribué. Pour pouvoir généraliser les résultats à plusieurs signeurs en prenant en compte leurs différences morphologiques, des opérations de normalisations devront être mises en place.
Différents types d'algorithmes de classification supervisée tels que les réseaux de neurones, les plus proches voisins (kNN) (utilisés, par exemple, pour la reconnaissance de gestes en flux dans les travaux de Dupont et Marteau [DM15]) ou les Machines à Vecteurs de Support (SVM) seront étudiés pour tester leur efficacité dans l'étape de reconnaissance frame à frame. Pour ce qui est de la segmentation, il pourrait être intéressant d'employer l'analyse en composantes principales (PCA) développée dans [HGMC05] afin de rendre cette étape plus robuste et moins dépendante d'un seuil fixé empiriquement.
La question de l'évaluation des résultats de l'annotation automatique mérite également d'être approfondie. Nous avons fait ici le choix d'utiliser la mesure de similarité de type *Simple Matching Coefficient* qui mesure le recouvrement entre la segmentation déterminée par des spécialistes et notre segmentation calculée automatiquement. Cependant, les annotateurs humains sont imprécis et peuvent faire des erreurs. L'utilisation des annotations faites à la main comme vérité terrain constitue t-elle vraiment un choix pertinent ?

7. Remerciements

Les observations et travaux relatés dans cet article sont basés sur les données de *Motion Capture* et les annotations du projet *Sign3D* [GLH*16].

Références

[Bat78] BATTISON R. : *Lexical borrowing in American sign language.* ERIC, 1978.

[Bou06] BOUTORA L. : Vers un inventaire ordonné des configurations manuelles de la langue des signes française. In *Journées d'Études sur la Parole (JEP)* (2006).

[Cux00] CUXAC C. : *La langue des signes française (LSF) : les voies de l'iconocité*. Ophrys, 2000.

[DM15] DUPONT M., MARTEAU P.-F. : Coarse-dtw for sparse time series alignment. In *International Workshop on Advanced Analytics and Learning on Temporal Data* (2015).

[GLH*16] GIBET S., LEFEBVRE-ALBARET F., HAMON L., BRUN R., TURKI A. : Interactive editing in French sign language dedicated to virtual signers : requirements and challenges. *Universal Access in the Information Society. Vol. 15*, Num. 4 (2016), 525–539.

[HGMC05] HELOIR A., GIBET S., MULTON F., COURTY N. : Captured motion data processing for real time synthesis of sign language. In *Motion in Games* (2005).

[KPBB02] KIM J.-B., PARK K.-H., BANG W.-C., BIEN Z. : Continuous korean sign language recognition using gesture segmentation and hidden markov model. In *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems* (2002).

[LADG08] LEFEBVRE-ALBARET F., DALLE P., GIANNI F. : Toward a computer-aided sign segmentation. In *Language Resources and Evaluation Conference (LREC). European Language Resources Association* (2008).

[NLG17] NAERT L., LARBOULETTE C., GIBET S. : Coarticulation analysis for sign language synthesis. In *International Conference on Universal Access in Human-Computer Interaction* (2017).

[RGLM16] REVERDY C., GIBET S., LARBOULETTE C., MARTEAU P.-F. : Un système de synthèse et d'annotation automatique à partir de données capturées pour l'animation faciale expressive en lsf. In *Journées Françaises d'Informatique Graphique (AFIG)* (2016).

[RK04] RIFKIN R., KLAUTAU A. : In defense of one-vs-all classification. *Journal of machine learning research. Vol. 5*, Num. Jan (2004), 101–141.

[SM] SOKAL R., MICHENER C. : *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin.

[Sto60] STOKOE W. C. : Sign language structure : An outline of the visual communication systems of the american deaf. *Studies in Linguistics, Occasional Papers. Vol. 8* (1960).

[YS06] YANG R., SARKAR S. : Detecting coarticulation in sign language using conditional random fields. In *Proceedings of the 18th International Conference on Pattern Recognition* (2006).