Agent virtuel signeur - Aide à la communication pour personnes sourdes

Philippe Gorce, Nasser Rezzoug, Alexis Heloir, Sylvie Gibet, Nicolas Courty, Jean-François Kamp, Franck Multon, Catherine Pelachaud

Agent virtuel signeur
Aide à la communication pour personnes sourdes

Alexis Héloir, Sylvie Gibet
Nicolas Courty, Jean-François Kamp
VALORIA–SAMSARA
Université de Bretagne Sud
sylvie.gibet@univ-ubs.fr

Franck Multon
LPBEM
Université de de Rennes II
franck.multon@uhb.fr

Catherine Pelachaud
LINC-PARAGRAPHE
Université de Paris 8
pelachaud@univ-paris8.fr

Philippe Gorce, Nasser Rezzoug
LESP
Université de Toulon
gorce@univ-tln.fr

Abstract

Cet article présente un système permettant de capturer et de structurer une base de connaissance de gestes de la Langue des Signes Française (LSF), et de l'utiliser pour la génération de mouvement d'humanoide virtuel. On s'intéresse ici plus spécifiquement à la qualité expressive des gestes (quel style de geste : fluide, tendu, énervé), et à ses représentations sémantiques. L'application envisagée est celle d'un humanoide signeur capable de générer un ensemble de gestes de la LSF. Nous avons constitué une base d'information semi-structurée de gestes capturés comprenant les mouvements corporels, les gestes des bras et des mains ainsi que les mimiques faciales. L'analyse des signaux de cette base d'information nous permet d'extraire des caractéristiques propres à la variabilité des gestes, à la fois d'un point de vue sémantique (segmentation suivant la structure des gestes de la LSF) et d'un point de vue du style d'exécution. Ces caractéristiques sont intégrées dans des modéles de synthèse évalués qualitativement par des personnes sourdes expertes de la LSF, suivant des critères d'intelligibilité et de réalisme des animations produites.

1 Introduction

Les technologies de l'information et de la communication envahissent notre espace au quotidien en proposant une multitude de nouveaux services, facilitant ainsi la consultation et la production d'information pour leurs utilisateurs. Cependant, de telles avancées techniques ne sauraient être perçues comme un réel progrès pour une société, si elles excluent une partie de sa population. Un individu peut ressentir des difficultés d'accès à l'information pour diverses raisons : économiques, culturelles, linguistiques ou physiques. C'est le cas en particulier des personnes sourdes qui peuvent éprouver des difficultés d'ordre linguistique pour dialoguer avec des applications informatiques.

Nos recherches tentent d'apporter des éléments de réponse à ce problème : permettre à la machine d'émettre des informations suivant différentes modalités de communication (texte, image, son, geste). En particulier elles s'intéressent à l'extension des nouveaux modes de communication à travers la conception d'humanoides virtuels capables de commmuniquer en langue des signes.

Si la Langue des Signes Française (LSF) est aujourd'hui reconnue comme une langue à part entière [16, 17], elle a souffert en France d'un bannissement institutionnel de près d'un siècle auquel les sourds paient encore un lourd tribut. Aujourd'hui, on voit apparaître de plus en plus de logiciels à même de favoriser l'accès des sourds à l'information en la présentant dans leur langue naturelle et au moyen d'outils de visualisation pertinents. Les applications qui en découlent concernent entre autres la communication à distance en temps réel (visiophonie, vidéoconférence), l'enseignement assisté par ordinateur (dictionnaires informatisés, tutoriaux de langues des signes) et la diffusion d'émissions bilingues, signées et orales. Par ailleurs, les nouvelles modalités de communication, exploitées dans les systèmes de télécommunication récents, intègrent des agents artificiels capables d'améliorer les capacités d'interaction avec leurs utilisateurs. Ces agents peuvent jouer notamment le rôle d'assistants dans des applications interactives multimédia.

Nous présentons dans cet article un projet de recherche

qui vise à concevoir l'animation d'humanoides virtuels autonomes et réalistes, capables de générer des gestes de la Langue des Signes Française. L'accent est mis sur la qualité expressive des gestes, et sur le lien entre ses représentations symboliques (quel style de geste et d'expression : fluide, raide, tendu, etc.) et biomécaniques (quels paramètres physiques pour une exécution donnée). Ce projet a permis à quatre équipes de recherche pluridisciplinaires, spécialisées en sciences du mouvement et en informatique (animation par ordinateur, agents conversationnels) de collaborer. Cette collaboration n'a pu se faire qu'avec la participation en amont et en aval du projet de personnes sourdes capables de signer. Les contributions des équipes de recherche sont complémentaires. Elles concernent :

- la constitution d'une base de donnée de gestes intégrant les gestes des bras, des mains et les expressions faciales,
- le développement d'outils pour l'édition du mouvement capturé, l'annotation, la segmentation et la visualisation de l'humanoide virtuel,
- la synthèse de mouvements corporels et faciaux.

Après un état de l'art (section 2) et une présentation générale du projet (section 3), les sections suivantes détaillent plus précisément la constitution d'une base d'information multimodale structurée (section 4), la mise en oeuvre d'outils interactifs pour le traitement et la segmentation des mouvements capturés (section 5), le développement de modèles de synthèse de mouvement (section 6), et les perspectives qu'offrent un tel projet (section 7).

2 Etat de l'art

Un certain nombre d'approches visent à concevoir des agents conversationnels capables de générer des gestes expressifs qui accompagnent la parole [1, 19, 18, 5]. Ils permettent la synthèse d'énoncés multimodaux pour des comportements verbaux et non verbaux.

Un nombre important d'études sont dédiées à la synthèse de gestes de la langue des signes. Ces systèmes intègrent des techniques inspirées des études sur le langage naturel, laissant l'animation elle-même en arrière-plan. Stokoe, pionnier en la matière, a proposé une description de la langue des signes américaine (ASL) en terme d'unités de mouvement sémantiques appelées chérèmes, et un système de transcription écrit basé sur la combinaison de ces chérèmes [23, 22]. La notation originale consiste en un nombre limité de symboles (chérèmes), distribués en trois classes, chacune représentantun un paramètre formationnel d'un signe : l'emplacement du signe (TAB), la forme de la main (DEZ) et le mouvement (SIG).

Parmi les applications informatiques, Lee et Kunii [10]ont développé un logiciel qui traduit du langage naturel en langue des signes en utilisant un ensemble fini de formes de la main et des expressions faciales préenregistrées pour générer des gestes de l'ASL. Sagawa et al. Sagawa and al [21] ont développé un système de traduction de la langue des signes japonaise en texte et vice versa. Losson [12] a proposé un système de description grammaticale des gestes relativement complet, s'appuyant sur la description linguistique de Liddell et Johnson [11]. Lebourque et al. [9, 2] développent un langage de spécification de gestes naturels, fondé sur une description qualitative de haut niveau de la commande gestuelle. Le langage développé s'appuie sur une analyse structurelle des gestes de la Langue des Signes Française (LSF), et sur une représentation discrète de l'espace autour du personnage virtuel. Un certain nombre de primitives de base sont identifiées, qui correspondent aux principaux mouvements, orientations et configurations de la main. Ces primitives sont combinées pour former des unités motrices atomiques appelées gestèmes, elles-mêmes assemblées en séquence et en parallèle pour constituer des gestes élémentaires et des gestes coordonnés. La composition de ces différents éléments permet d'obtenir des gestes allant de simples mouvements de désignation à des gestes complexes. Les expressions construites sont ensuite traduites en données quantitatives contrôlant l'animation des membres supérieurs d'un personnage virtuel.

Plus récemment, le projet européen VISICAST a cherché à élaborer un ensemble d'outils pour la communication en langue des signes [4]. Dans le contexte de ce projet, un langage de description SiGML basé sur les langages textuels Extended Markup Language, a été développé. Ce langage s'appuie sur la notation HamNoSys [20] qui permet de décrire précisément les gestes des langues des signes. Un traducteur permet de passer du codage HamNoSys vers SiGML.

3 Vue générale du projet

L'architecture de notre système est représentée de manière schématique par la figure 1. Toute la chaîne d'analyse/synthèse des gestes est matérialisée par trois modules fonctionnels. Le module *Analyse* est chargé de l'acquisition des mouvements et des expressions faciales, de l'identification de paramètres biomécaniques et de la constitution d'une base d'information de gestes élémentaires de la LSF. Le module *Synthèse* gère le contrôle et la synthèse du mouvement des humanoïdes virtuels. Le module *Description et spécification des gestes* permet de définir les gestes que l'on souhaite rejouer ou synthétiser à partir d'une représentation qualitative de haut niveau caractérisant les diverses modalités.

Nous présentons dans cet article les protocoles expérimentaux, les techniques et méthodes utilisées pour concevoir la base d'information structurée multimodale, comprenant un ensemble limité de gestes de la LSF réalisés

avec des variations de styles et de dynamiques. Cette base d'information est utilisée pour le développement des recherches autour de l'analyse et de la synthèse du mouvement d'humanoides virtuels, basées sur la capture de mouvement humain. Nous présentons également les outils d'édition, de visualisation et d'annotation de gestes de la LSF développés dans le cadre du projet. Enfin, nous présentons les résultats en terme de génération du mouvement d'humanoides virtuels.

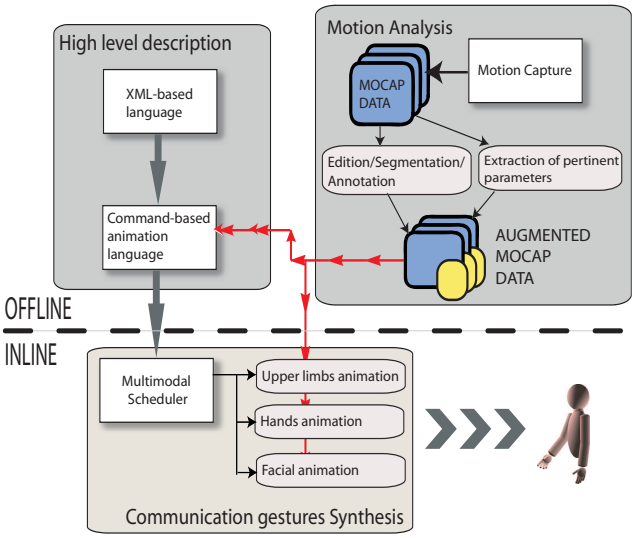

FIG. **1. Architecture fonctionnelle du projet**

4 Constitution d'une base d'information structurée multimodale

Plusieurs séquences d'enregistrements ont été réalisées avec un système VICON MX (produit d'Oxford Metrics) du laboratoire LPBEM (Université de Rennes 2). Ce système, composé de 12 caméras infrarouges cadencées à 110 Hz a été utilisé pour capturer les déplacements 3D de 24 marqueurs réflexifs placés sur des emplacements anatomiques du corps, comme le montre la figure 2. Le VICON a été également utilisé pour l'acquisition d'expressions faciales, à l'aide de marqueurs plus petits placés en des points MPGEG4 spécifiques. Les mouvements de mains ont été capturés par une paire de gants de données du laboratoire LESP (université de Toulon et du Var). Ces données manuelles ont été synchronisées avec les données corporelles, après reconstruction et ré-échantillonnage à 60 Hz. Plusieurs corpus ont été mis en oeuvre à partir du choix de gestes spécifiques de la LSF. Un sujet sourd a participé à ces enregistrements. Le premier corpus est constitué de 22 phrases relatives à l'énoncé de bulletins météorologiques, avec un nombre limité de signes et des séquences d'enregistrement limitées à 60 s. Le second corpus est constitué de 10 phrases traitant également de météo, réalisées avec différentes dynamiques et émotions. Le troisième corpus contient 8 séquences de messages relatifs à des incidents pouvant survenir dans une gare SNCF. Enfin plusieurs séquences supplémentaires de signes isolés ont été réalisées. L'une correspondant à l'alphabet dactylologique a permis de calibrer et de segmenter les mouvements de mains. Une autre séquence correspond à la signature d'un ensemble de villes françaises. Après avoir modélisé le squelette, un processus de suivi automatique permet d'étiqueter les trajectoires cartésiennes utilisées pour animer le personnage virtuel. A partir de ces informations de mouvement brutes, nous avons ensuite fusionné les différentes sources d'information en un format BVH (Bio-Vision Hierarchical data). Les données faciales sont également traitées et traduites dans le format FAP (Facial Animation Parameters) de la norme MPEG4.

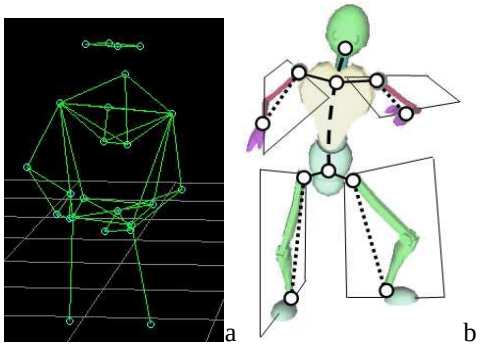

FIG. **2. (a) Marqueurs réflexifs placés sur des emplacements anatomiques (b) Représentation normalisée du squelette**

5 Traitement et segmentation des données

5.1 Outils interactifs pour l'édition, la visualisation et l'annotation

Les méthodes d'analyse et de segmentation du geste permettent d'enrichir les bases de données de mouvements capturés en introduisant des éléments structurels propres aux gestes de la LSF. Les données enrichies contribuent à faciliter la spécification du mouvement pour la synthèse. Les outils de segmentation développés et mis en œuvre au sein de notre projet sont présentés dans cette section. Ces outils sont classés en deux catégories : segmentation manuelle pour l'annotation et segmentation automatique.

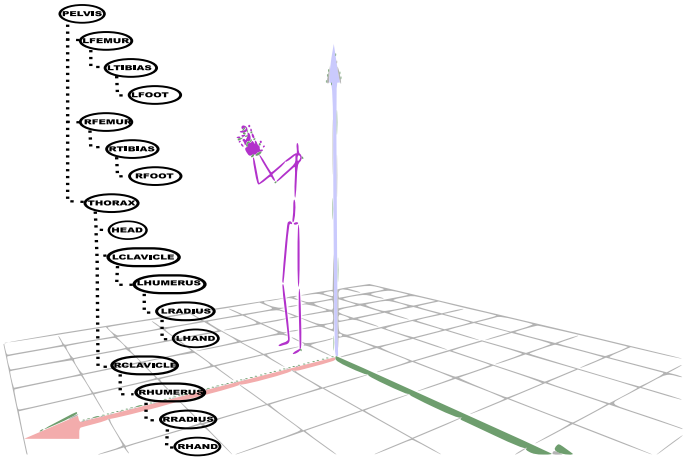

FIG. **3. Description hiérarchique de la morphologie d'un squelette d'animation selon le format BVH**

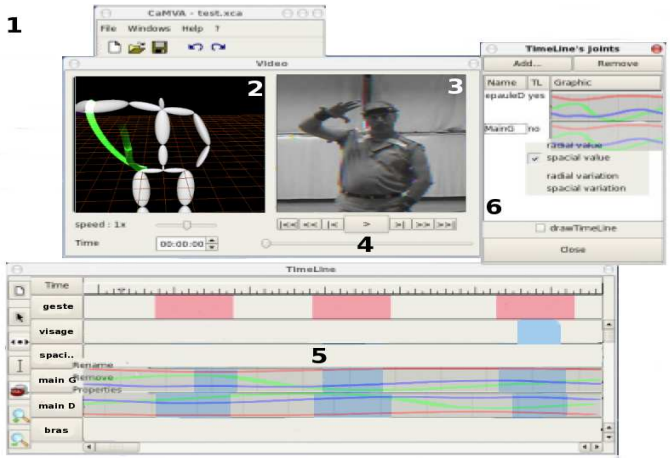

FIG. **4. Interface de visualisation et d'annotation**

5.2 Outils interactifs pour la visualisation, la segmentation manuelle et l'annotation

Un ensemble d'outils logiciels permet de représenter et de visualiser en 3D le personnage virtuel animé. La représentation hiérarchique du personnage est représentée sur la figure 3. L'animation du personnage reconstitué est synchronisée avec la vidéo de l'acteur enregistrée pendant la séquence de capture. Il est possible également de visualiser, éditer et annoter les trajectoires des signaux numériques dans des pistes temporelles préalablement spécifiées, comme le montre la figure 4. L'utilisation de cette représentation offre en outre la possibilité d'enrichir les données capturées de marqueurs temporels délimitant différentes phases de mouvements à partir d'un processus de segmentation manuelle. Les données mouvement sont enrichies d'un ensemble de descripteurs et d'annotations permettant de caractériser la structure des mouvements capturés ainsi que la segmentation des différents canaux multimodaux.

5.3 segmentation automatique des données

Les méthodes de réduction de dimension d'espace tels que l'analyse en composantes principales (PCA) peuvent conduire à une segmentation proche de la représentation sémantique faite par un observateur humain. En suivant la même démarche, nous appliquons cette technique aux mouvements de mains [3]. Nous considérons que ces mouvements peuvent être représentés par des suites de tenues et de changements de configuration caractérisant des transitions nonlinéaires dans les mouvements. D'un point de vue algorithmique, nous exploitons le fait que lors des transitions entre segments de mouvement, l'erreur de reconstruction induite par la projection de la posture considérée sur un hyperplan de dimension fixée doit augmenter rapidement (cf figure 5). Ces résultats de segmentation manuelle sont utilisés pour segmenter globalement les mouvements, l'hypothèse étant faite que les zones de transition constituent des points d'articulation du discours en langue des signes.

6 Synthèse de gestes et d'expressions faciales

6.1 Synthèse de gestes

En collaboration avec le projet SIAMES de l'IRISA, le LPBEM a développé un environnement d'animation temps réel pour les humanoïdes appelé MKM[1] qui est le produit de deux thèses co-encadrées [13, 6]. Cet environnement est composé d'une librairie dédiée à l'animation temps réel à partir de données issues de captures de mouvements [15, 14] et d'un outil pour générer des fichiers de gestes indpéndants de la morphologie [8]. Grâce à cette représentation il est possible de passer automatiquement d'un mouvement capturé sur une personne à un personnage synthétique ayant des dimensions différentes. Cependant, pour assurer que cette opération s'effectue sans dégrader le mouvement initial, il est nécessaire de préserver les contraintes qui lui sont intrinsèquement liées. Par exemple, en langue des signes, si la main doit toucher la bouche pour préserver le sens de la phrase, le système doit être capable d'assurer que cette contrainte est effectivement vérifiée sur le squelette synthétique. L'outil que nous avons développé

[1] Voir www.irisa.fr/siames/MKM

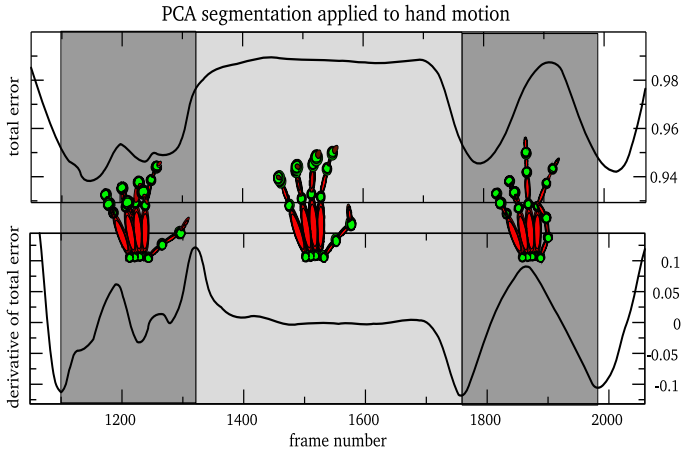

FIG. **5. Segmentation automatique s'appuyant sur la PCA : la détection intervient lorque la dérivée de l'erreur de reconstruction dépasse trois déviations de la moyenne**

FIG. **6. Personnage virtuel signeur du système MKM**

permet d'éditer ce type de contraintes. Plus généralement, il permet de spécifier des contraintes de distance entre deux points du squelette, entre un point du squelette et un point de l'environnement, des contraintes d'orientation d'un segment dans l'espace, de vitesse…Ces contraintes sont ensuite résolues dans l'environnement temps réel par un algorithme optimisé de cinématique et de cinétique inverse [7].

6.2 Synthèse d'expressions faciales

Le système Greta prend en entrée un texte que l'agent doit dire augmenté des fonctions communicatives (spécifiant comment le texte doit être dit). Ces informations sont spécifiées par des étiquettes du langage de représentation "Affective Presentation Markup Language" (APML). Ce langage suit le format XML. Le système Greta interprète tout d'abord les étiquettes APML et décide quel signal sera communiqué avec quelle modalité (regard, visage, tête) pour chaque fonction communicative. Chaque expression est synchronisée avec le texte qu'elle accompagne ; sa durée est donnée par les étiquettes APML qui entourent cette partie de texte. Les étiquettes APML sont ensuite instantiées et les fonctions communicatives traduites en signaux multimodaux. L'animation est obtenue en convertissant chaque signal du visage et du regard en paramètres définis par le standard MPEG-4.

6.3 Intégration

Quelques résultats liés à l'expérimentation menée à Rennes avec tous les partenaires sont donnés en figure 7. Dans cette démo, les mouvements sont automatiquement adaptés au personnage synthétique. On voit d'ailleurs, en figure 6, plusieurs personnages effectuant le même geste. Suite à la segmentation des séquences capturées lors de la dernière expérimentation, il sera possible d'ajouter les contraintes intrinsèques à chaque geste grâce aux outils développés dans MKM. Un travail en cours consiste à exporter les animations au format Mpeg4 et à fusionner le résultat avec les données liées au visage. A plus long terme, il est prévu de fusionner les deux outils (MKM et Greta) pour n'avoir qu'un seul environnement d'animation plutôt que de passer par fichiers.

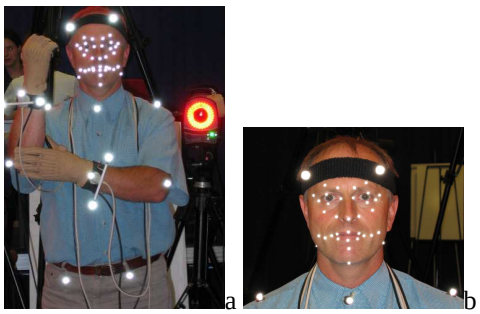

FIG. **7. (a) Capture du mouvement corporel du signeur (b) Capture des expressions faciales du signeur**

7 Conclusions et perspectives

Nous avons présenté dans cet article une plateforme expérimentale dédiée à l'enregistrement de gestes de la LSF, et à la génération du mouvement d'humanoides signeurs. Cette plateforme a permis de constituer plusieurs

corpus de données multimodales synchronisées, intégrant des mouvements corporels, des gestes des mains et des expressions faciales. Nous avons choisi un ensemble fini de signes, issu de deux contextes applicatifs différents : la présentation des bulletins météorologiques d'une part et la diffusion de messages d'incidents dans des gares SNCF d'autre part. La collaboration entre plusieurs équipes de recherche pluridisciplinaires a permis d'identifier et de résoudre un ensemble de problèmes techniques (emplacement des capteurs, prétraitement de données comme le filtrage, reconstruction 3D, ...). Nous avons également défini un ensemble de paramètres permettant de caractériser la modulation des gestes et des expressions faciales. Un ensemble d'outils d'analyse et de segmentation de mouvements a été également mis en place pour les mouvements corporels et manuels. La plateforme permet d'animer des personnages virtuels en utilisant des techniques variées : rejeu avec adaptation pour l'animation corporelle, interpolation de paramètres FAPS pour l'animation faciale. Quelques résultats de recherche montrent l'utilisation de mouvements capturés pour la synthèse de mouvements coarticulés et expressifs. Nous envisageons à court terme d'élaborer une base d'information de gestes, indexée et segmentée, qui pourra être utilisée pour effectuer des analyses statistiques sur les gestes de la LSF. Elle permettra également d'animer un ensemble varié d'humanoides signeurs, à partir d'un langage de spécification interactif.

8 Remerciements

Nous remercions les personnes sourdes (en particulier Alain Cahut), ainsi que leurs interprètes qui ont permis les séances d'enregistrement de gestes de la LSF. Ces personnes ont apporté également leur expertise dans la construction et l'interprétation des corpus de gestes de la LSF. Les travaux présentés dans cet article sont financés d'une part par la région Bretagne (Réf. B/1042/2004/SIGNE) et d'autre part par le département STIC du CNRS (projet RobEA HuGEx).

Références

[1] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Horn, W. Becket, B. Douville, and S. P. M. Stone. Animated conversation : Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics, Annual Conference Series, ACM*, pages 413–420, 1994.

[2] S. Gibet and T. Lebourque. High level specification and animation of communicative gestures. In *Journal of Visual Languages and Computing*, volume 12, pages 657–687, 2001.

[3] A. Heloir, S. Gibet, F. Multon, and N. Courty. Captured motion data processing for real time synthesis of sign language. In *Lecture Notes in Artificial Intelligence, Vol.*, 2006.

[4] R. Kennaway. Experience with, and requirements for, a gesture description language for synthetic animation. In *5th Int Work. on Gesture and Sign-Language based Human-Computer Interaction*, Genova, Italy, 2003.

[5] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15 :39–S52, 2004.

[6] R. Kulpa. *Adaptation interactive et performante des mouvements d'humanoides synthétiques : aspects cinématiques, cinétiques et dynamiques*. PhD thesis, INSA Rennes, November 2005.

[7] R. Kulpa and F. Multon. fast inverse kinematics and kinetics solver for human-like figures. In *Proceedings of IEEE Humanoids*, pages 38–43, Tsukuba, Japan, december 2005.

[8] R. Kulpa, F. Multon, and B. Arnaldi. Morphology-independent representation of motions for interactive human-like animation. *Computer Graphics Forum, Eurographics 2005 special issue*, 24(3) :343–352, 2005.

[9] T. Lebourque and S. Gibet. High level specification and control of communication gestures : the GESSYCA system. In *Proc. of IEEE Computer Animation*, Geneva, May 1999.

[10] J. Lee and T. Kunii. Computer animated visual translation from natural language to sign language. *Journal of Visualization and Computer Animation*, 4(2) :63–78, 1993.

[11] S. Liddell and R. Jonhson. American sign language : the phonological base. In *Sign Language Studies 64*, pages 195–277, 1989.

[12] O. Losson and J. Vannobel. Sign language formal description and synthesis. *Int.Journal of Virtual Reality*, 3(4) :27–34, 1998.

[13] S. Ménardais. *Fusion et adaptation en temps réel de mouvements acquis pour l'animation d'humanoïdes synthétiques*. PhD thesis, École doctorale Matisse, 2003.

[14] S. Menardais, R. Kulpa, F. Multon, and B. Arnaldi. Synchronization for dynamic blending of motions. In *Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 325–336, Grenoble, France, August 2004.

[15] S. Ménardais, F. Multon, R. Kulpa, and B. Arnaldi. Motion blending for real-time animation while accounting for the environment. In *Computer Graphics International*, June 2004.

[16] B. Moody. International Visual Theatre, 1998.

[17] B. Moody. In I. V. Theatre, editor, *La langue des signes : Introduction à l'histoire et à la grammaire de la langue des signes*, 1998.

[18] C. Pelachaud. *Contextually Embodied Agents*, chapter Deformable Avatars, pages 250–263. Kluwer Publishers, 2001.

[19] I. Poggi and C. Pelachaud. *Performative facial expressions in animated faces*, chapter Embodied Conversational Agents, pages 155–188. MIT press, 2000.

[20] S. L. Prillwitz, R. Leven, H. Zienert, R. Zienert, T.Hanke, 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, 1989.

[21] H. Sagawa, M. Ohki, T. Sakiyama, E.Oohira, H. Ikeda, and H. Fujisawa. Pattern recognition and synthesis for a sign language translation system. *Journal of Visual Languages and Computing*, 7 :109–127, 1996.

[22] W. Stokoe, D. Casterline, and C. Croneberg. *A dictionary of American Sign Language on Linguistic principles*. Linstok Press, Silver Spring, 1978.

[23] W. C. Stokoe. *Semiotics and Human Sign Language*. Mouton, The Hague, 1972.