High level models for sign language analysis by a vision system

Patrice DALLE
IRIT – Paul Sabatier University
118 route de Narbonne – 31062 Toulouse cedex 9 - France
E-mail: dalle@irit.fr

Abstract

Sign language processing is often performed by processing each individual sign and most of existing sign language learning systems focus on lexical level. Such approaches rely on an exhaustive description of the signs and do not take in account the spatial structure of the sentence. We present a high level model of sign language that uses the construction of the signing space as a representation of both (part of) the meaning and the realization of a sentence. We propose a computational model of this construction and explain how it can be attached to a sign language grammar model to help analysis of sign language utterances and to link lexical level to higher levels. We describe the architecture of an image analysis system that performs sign language analysis by means of a prediction/verification approach. A graphical representation can be used to explain sentence construction.

1. INTRODUCTION

As other languages, sign language relies upon several grammatical levels, namely lexical, syntactical, and semantical levels. However, most of the existing researches focus on the lexical level and moreover, on standard signs i.e, the ones that can be found in dictionaries, and not on iconic utterances (classifier or "proforms", transferts structures, …). On the other hand, iconic structures are widely used in spontaneous sign language so it seems appropriate to take them in account in automatic sign language processing systems.

The meaning of a sign language production can be recovered by considering the construction of the signing space (Cuxac 1999) (Cuxac, 2000): during this production, the signer uses this space to position the entities that are evoked in the sentence and to materialize their semantic relationships, so that the resulting construction can be considered as a representation of the meaning of the discourse.

We propose a computational representation of this organization, and describe how this representation can be used to help automatic interpretation of sign language by an image processing system, and how graphical representation can help sign language understanding and learning.

Most of previous works on sign language analysis focused on isolated sign translation by means of a finite set of parameters and values, from the Liddel and Johnson phonological description (Vogler 1998) or the Stokoe description system (Ouhyoung 1998). Datagloves are often used as input devices. Some works focus on increasing the recognition rate by using some additional knowledge on the signed sentence structure: statistics on consecutive pairs of signs (stochastic grammars) (Hienz 1999) or (Ouhyoung 1996), constraints on the structure of the sentence (Pentland 1995). Nevertheless these approaches do not take in account the spatial structure of the signed sentence. The resulting systems are only able to deal with sentences considered as a simple succession of isolated signs, eventually coarticulated. More complex aspects of sign language such as sign space utilization or classifiers have not been studied yet in vision-based sign language analysis, but some issues where brought out in recent works on sign language generation (Bossard 2003) (Huenerfauth 2004).

Our approach focuses on the fact that introducing knowledge about sign language syntax and grammar will allow a vision system to achieve the image analysis of the sequence and, thus, avoid us to systematically use complex reconstruction of gestures. Instead of direct sign recognition, we make much of identifying the structure of the sentence in terms of entities and relationships, which may be sufficient in a reduced-context application. This allows us to use a general model of sign language grammar and syntax. Hence, starting from an high level hypothesis about what is going to be said in the sign language sentence, this model let us compute a set of low level visual events that have to occur in order to validate the hypothesis. While verifying the fact that something has happened is simpler than detecting it, our approach permits the use of rather simple image processing mechanisms in the verification phase and reserves explicit reconstruction of gestures for the cases where prediction becomes impossible.

2. OVERVIEW OF OUR APPROACH

In order to analyse FSL utterances using a single video camera and simple image processing, we need to integrate a fair amount of knowledge (i) about FSL grammar and syntax for prediction and consistency checking of the interpretation but also (ii) about image processing for querying the low-level verification module.

The system integrates this knowledge in a multi-level architecture that is divided in three main subsystems:
1. The first subsystem consists in a representation of the interpretation of the discourse through a modeling of the signing space. During processing, the coherence of signing space instantiation is controlled by a set of possible behaviors resulting from the structure of the language and from a semantic modeling of the entities in the discourse.

2. The second subsystem is a knowledge representation system based on description logic formalism. The base contains some knowledge about FSL grammar and syntax that makes it able to describe high level events that occurred in signing space in terms of low level sequences of events on body components.
3. The last subsystem performs image processing; it integrates knowledge about the features it must analyze so as to choose the appropriate measurement on the data for the verification process.

Next sections describe the main aspects of the linguistic model and the verification process.

3. MODELING THE SIGNING SPACE

3.1 SIGNING SPACE MODEL

In the FSL, entities are evoked through signs and located in the signing space so that their relative positions will correspond to spatial relationships between those entities in the real world. Temporal relationships are evoked through entities that are located on "time lines". Binary actions are evoked through directional verbs and more complex ones by grammatical structures called "transfers" (Cuxac 1999). The different kinds of entities depend on the kinds of relationships in which each entity may be involved: *dates* can be involved in temporal relationships, *places* in spatial relationships; *animates* can perform an action or be located relative to another entity, *actions* can be referenced as a moment in time or as one of the protagonists of an action. The specificities of the FSL grammar require to consider some additional kind of entities: one needs to make a distinction between entities that whenever involved in a complex action are evoked by the signer taking their role (*persons*1) and the entities that cannot be evoked this way (*objects*). Finally, due to the temporal ordering of the signs, one needs to take in account the case of actions that are evoked before one of their protagonists; the type of this entity is *implicit*.

3.2 SIGNING SPACE REPRESENTATION

The symbolic representation of the signing space consists of a volume surrounding the signer, regularly divided into Sites. Each location may contain a single Entity, each Entity having a Referent. A Referent is a semantic notion that can be found in the discourse. Once it has been placed in the signing space, it becomes an Entity and has a role in the sentence. Hence, building a representation of a sign language sentence consists in creating a set of Entities in the SigningSpace. A graphical representation of the signing space can be built to explain how FSL uses the space as seen in figure 1.

The meaning contained in this signing space construction is represented in terms of Entities whose Referents can have successively different function(s) during the construction of the sentence (locative, agent, actions, …). A set of rules maintains the consistency of the representation by verifying that sufficient coherent information has been provided when one needs to create a new entity in the signing space. The global architecture of the model can be represented in UML notation standard.

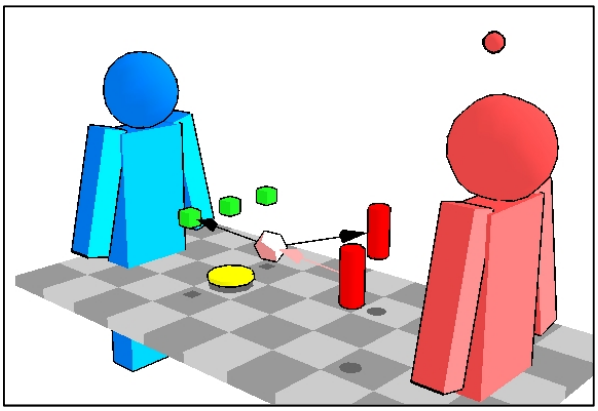

Figure 1: Signing space representation

4. A MODEL TO CONSTRUCT THE SIGNING SPACE

4.1 RULES OF THE SIGNING SPACE CONSTRUCTION

We need rules of FSL grammar to describe the signing space construction. As modifying the signing space only consists in creating new entities, our model focuses on the gestures that are used to create those entities. Without lexical knowledge, it is not possible to make a distinction between entities that are neither dates nor actions. So that creating such an entity relies on a generic mechanism.

Creating an entity of a given type relies on the following mechanisms:

• Creating a generic entity: entities are created and localized in the signing space by signs that can be performed either directly in the desired location or localized on the signer's body for lexical reasons. In the second case, the production of the sign is followed by an explicit designation of the desired location.

• Creating a date: in our reduced context, dates are explicitly evoked by standard signs, performed in a neutral location (if front of signer's chest) and located simultaneously on one of the time lines.

• Creating an action: binary actions are evoked through directional verbs, which implies some gestures that explicitly connect two locations containing entities in the signing space. For complex actions, "great iconicity" structures such as those where the signer plays the role of one of the action's protagonist have to be used. We have not yet study such complex actions. .

The formalization of that grammar relies on the fact that each of those mechanisms can be described by a gesture sequence.

4.2 DESCRIBING THE CONSTRUCTION OF THE SIGNING SPACE

A modification in the signing space is defined by the kind of the entity that is created and its localization. The behavior model attaches to each kind of entity a gesture sequence that describes the state of the components involved and the way they are synchronized.

The computational representation of that grammar relies on a description logic formalism and uses the CLASSIC

knowledge representation system (Brachman 1991). This system expresses the representation of FSL grammar as a set of hierarchically organized concepts. Concepts are structured objects, with roles (concepts of a given type) and associated with automatic inference mechanisms and user-defined propagation rules.

On the basis of the description logic formalism, describing the creation of an entity consists in defining a set of concepts with specific constraints on some of their roles:
1. The concept representing the creation of an entity is called ACTS (ACtion Transforming Signing space). It is described by a location, a temporal interval and a gesture sequence.
2. Gesture sequences consist in a list of component descriptions associated with constraints on the values of the component roles.
3. Additional knowledge propagation rules concern vertical information propagation from an ACTS description to gestures defined in the corresponding sequence (e.g. the localization of the hand must be the same as the one of the entity). Horizontal information propagation mechanisms are used between different gesture descriptions in the same sequence (e.g. both hands must have the same location). Finally gestures synchronization rules are based on Allen's algebra operators.
This formalization leads to a global representation of the FSL grammar as a concept hierarchy associated with additional propagation rules sets (figure 2)

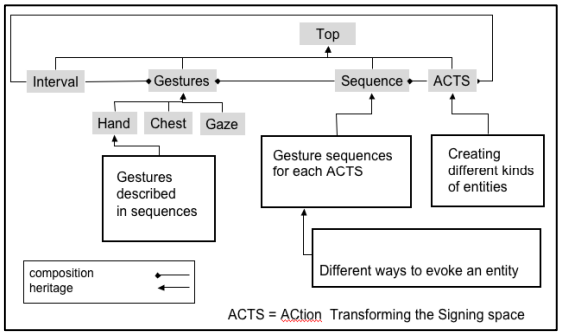

Figure 2 Concept hierarchy and inference mechanisms

For each kind of entity, there is a specialization of the ACT concept with a specific *GestureSequence*. This sequence can be derived depending on the different ways of creating an entity of that type. Gestures that can be found in *GestureSequences* are specializations of generic *Component* descriptions that include additional constraints on their roles.

5. IMAGE-BASED SIGN LANGUAGE ANALYSIS

The representation of the signing space can be linked to the meaning of the discourse by giving access to the relationships between entities that were evoked and referenced. On the other hand, the iconicity theory by (Cuxac 1999) provides a description of the grammar of the sign language in terms of gesture sequences that leads to creating a new entity in the signing space. As a result, this permits to link this representation to the gestures that were used to create the current signing space instantiation. Such a predictive model can be used for analysis of sign language sentences.

Using that model for sign language analysis leads to two classes of tools: (i) interactive tools intended for linguists to evaluate the model or for teachers to explain sign language, (ii) automatic analysis tools that can be used in many fields of application (linguistic analysis, automatic interpretation,).
An interactive tool has been developed in order to represent the construction of the signing space during the production of the utterance (fig. 3). This tool consists of a transcription software that allows to synchronously link the different steps of the construction of the signing space and the video sequence that is transcripted. This application was designed to evaluate the model with respect to several kinds of utterances and to determine how this model can be considered as a generic representation of sign language utterances.

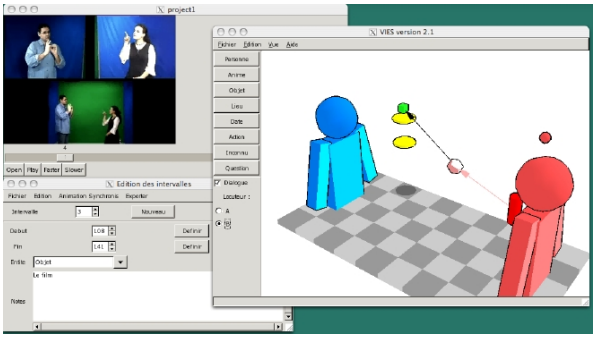

Figure 3 : interactive tool to build signing space

In the field of automatic analysis, using a single camera, it is not possible to build an exhaustive description of the gestures that are used. Therefore, automatic vision-based sign language analysis, the model of the signing space is used as a general representation of the structure of the sentence that simultaneously gives access to the meaning of the discourse.
The grammar of the sign language that can be attached to this construction allows the use of a prediction verification approach (Dalle 2005): from an hypothesis on the meaning of the discourse in terms of a signing space modification, it is possible to infer the gestures that were used to create the new entity in the signing space. Analyzing the utterance is then reduced to verify whenever the data corroborates this prediction or not.
Such an analysis can be performed without taking in account the lexicon, so that the gestures descriptions that can be used need to be less precise that the ones required for exhaustive sign recognition. This makes the analysis of low resolution images possible.

However, in a reduced context, the spatial structure of the sentence may be an interesting guideline to identify the signs as it can be done by only considering discriminative aspects of the signs. The behavior model infers a gesture sequence and asks the image processing module to verify it. The system describes each item of the gesture sequence in visual features. This reformulation is made in a

qualitative way. For instance it does not need an exact knowledge about hand shape, but only to know whether it is changing or not. Then, each of these features can be verified using simple 2D clues. For instance, to test hand shape properties, we only have to consider simple 2D shape properties as area or bounding box; to test if the signer looks at the location of the entity, we measure the dissymmetry of the face from the chest axis. Without this prediction process, in a bottom-up analysis, we should have to extract and recognize arm movement or hand configuration and so, to use more complicated measures as 3D tracking trajectories, shape descriptors, gaze direction or 3D face orientation.

The three different elements of such automatic tool (signing space representation, grammatical model, low level image processing) have been evaluated separately. It has been shown that in a reduced context, the prediction/verification approach was relevant and allowed to use simple 2D image processing operators instead of complex gesture reconstruction algorithms to performs the identification of the different kinds of entities that where used in the utterance.

6. CONCLUSION

In conclusion, this model is our first formalization of spatio-temporal structure of the signing space. Its purpose is to help sign language image analysis.
The main interests of this approach are:

- the use of a qualitative description of the gestures that can be easily identified with simple and robust image processing techniques,
- the use of a prediction / verification approach where only significant events have to be identified and that avoid an exhaustive reconstruction of the gestures,
- the descriptions used in that model provide a strong guideline for the design of those operators.

Implementation of the model and tools we have built help linguists to evaluate their linguistic model of sign language and teachers to explain FSL structures..
Further works concern:

- The extension of the model to dialog situation, with shared entities,
- The implementation of more complex transformations as “transfer structures”.

Finally, signing space representation could be used for the specification of a graphical form of sign language.

7. REFERENCES

Bossard, B., Braffort, A., Jardino, M. (2003). Some issues in sign language processing. In A. Camurri and G. Volpe (Eds.), Lecture Notes in Artificial Intelligence : 5th International Workshop on Gesture-Based Communication in Human-Computer Interaction, (pp. 15-17) Genova, Italy

Brachman, R.J. and al. (1991). Living with classic: When and how to use a kl-one-like language. In J. Sowa, (Eds.), Principles of Semantic Networks: Explorations in the representation of knowledge, (pp.401–456). Morgan-Kaufmann, San Mateo, California.

Cuxac, C. (1999). French sign language: proposition of a structural explanation by iconicity. In A. Braffort, R. Gherbi, S. Gibet, J. Richardson and D. Teil, (Eds.), Lecture Notes in Artificial Intelligence : Procs 3rd Gesture Workshop’99 on Gesture and Sign-Language in Human Computer Interaction (pp. 165-184) Gif-sur-Yvette, France, Springer: Berlin

Cuxac, C. (2000). La langue des signes française (LSF). Les voies de l’iconicité, Faits de Langues n° 15-16, Ophrys, Paris.

Dalle P., Lenseigne B. (2005). Vision-based sign language processing using a predictive approach and linguistic knowledge. In: IAPR conference on Machine Vision Applications – MVA (pp. 510-513). Tsukuba Science City, Japon,. IAPR.

Dalle P., Lenseigne B. (2005). Modélisation de l’espace discursive pour l’analyse de la langue des signes. In TALS 2005, atelier de TALN 2005. [on line] [http://tals.limsi.fr/actes/s7.pdf].

Hienz, H., Bauer, B., Kraiss, K.F. Hmm-based continuous sign language recognition using stochastic grammars. In A. Braffort, R. Gherbi, S. Gibet, J. Richardson and D. Teil, (Eds.), Lecture Notes in Artificial Intelligence : Procs 3rd Gesture Workshop’99 on Gesture and Sign-Language in Human Computer Interaction (pp. 185-196) Gif-sur-Yvette, France, Springer: Berlin.

Huenerfauth, M. (2004). Spatial representation of classifier predicates for machine translation into american sign language. In Workshop on Representation and Processing of Sign Language, 4th Internationnal Conference on Language Ressources and Evaluation (LREC 2004), (pp. 24–31), Lisbon Portugal.

Ouhyoung M. Liang. R.H.(1996). A sign language recognition system using hidden markov model and context sensitive search. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, (pp . 59–66), Hongkong

Ouhyoung, M., Liang, R.H. (1998) A real-time continuous gesture recognition system for sign language. In 3rd International conference on automatic face and gesture recognition, (pp. 558–565), Nara, Japan,

Pentland, A. , Starner, T. (1995). Real-time american sign language recognitio from video using hidden markov models. Technical Report TR-375, M.I.T Media Laboratory Perceptual Computing Section.

Vogler, C., Metaxas, D. (1998). Asl recognition based on a coupling between hmms and 3d motion analysis. In Proceedings of the International Conference on Computer Vision, (pp. 363–369), Mumbai, India.