Vision-based Sign Language processing using a predictive approach and linguistic knowledge

Boris Lenseigne

IRIT-TCI
Paul Sabatier University
Toulouse, France

Patrice Dalle

IRIT-TCI
Paul Sabatier University
Toulouse; France

Abstract

While the study of Sign languages provides a wide a field of applications for computer vision systems, the analysis of such gestures often leads to complex $3D$ reconstruction or to ambiguities. In this paper, we describe the architecture of an image analysis system that performs sign language analysis by means of a prediction/verification approach. The system integrates a model of sign language structure and uses it during analysis for predicting visual events so that simple $2D$ features can be used to determine whether the image corroborates the prediction or not.

1 Introduction

In the field of communication gesture analysis, the main difference between gestual interfaces and sign language processing, is the linguistic aspect of the sign language production, which means that a language model can be used to help image analysis. In this case, additional knowledge comes from the syntax of the language and, in a reduced context, from semantics. In this paper, we show how this kind of knowledge can be represented and used in a vision-based sign language analysis system in order to make some application “possible”. We first present some previous works in the field of linguistic representation in computer-based sign language analysis. Then, we give an overview of the system and describe the way it runs, before to consider in details each part of the model. Our approach takes place simultanneously in two fields of application :

1. linguistic studies, where image processing results provided by model-based image processing can be compared to the linguist’s interpretation ;

2. sign language interpretation for translation or in order to answer a question in sign language.

1.1 Previous work

Most of previous works on sign language linguistic focused on isolated sign description by the mean of a finite set of parameters and values. Resulting transcription systems have been used for machine translation by C. Vogler and D. Metaxas [3] that uses the Liddel and Johnson phonological description or in [9] that uses the Stokoe description system for sign recognition using datagloves. Some other works focus on increasing the recognition rate by using some additionnal knowledge on the signed sentence structure, which is done by using statistics on consecutive pairs of signs (so-called stochastic grammars) such as in [6] or [8], or by adding constraints on the structure of the sentence [10]. But none of them really takes in account the spatial structure of the signed sentence. Those 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[1] have not been studied yet in vision-based sign language analysis, but some issues where brought out in recent works on sign language generation [7][1].

1.2 Our approach

Our approach is focused on the fact that introducing knowledge about sign language syntax and grammar will make the analysis of the sequence possible and avoid us to systematicaly use complex reconstruction of gestures. Instead of direct sign recognition, we focus on identifying the structure of the sentence in terms of entities and relationships, which is generally sufficient in a reduced-context application. This allows us to use a general model of sign language grammar and syntax. So that, 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

[1] classifiers are gestures that are used to remainder entities that where previously created. Generally, classifier movement is directly relied to some physical aspect of the referenced entity.

visual events that have to occur in order to validate the hypothesis. While verifying that something has happened is simpler than detecting it, this approach will permit the use of rather simple image processing in the verification phase and reserve explicit reconstruction of gestures for the cases where prediction becomes impossible.

2 Overview of the system

Our system analyses french Sign language (FSL) gestures based on the fact that those gestures follow the rules of the grammar of this language. In order to make it possible to perform this task using a single video camera and simple image processing, we need to integrate plenty of knowledge about FSL grammar and syntax for prediction and consistency checking of the interpretation and about image processing for querying the low-level verification module.

2.1 Architecture of the system

Our system integrates these knowledges 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 modelling of the signing space[2]. During processing, the coherence of signing space instantiation is controlled by a set of possible behaviours resulting from the structure of the language and from a semantic modelling of the entities in the discourse (fig. 1 (A)).

2. The second subsystem is a knowledge representation system based on description logic formalism. The base contains 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 (fig. 1 (B)).

3. The last subsystem performs image processing, it integrates knowledge about the features it must analyse so that it can choose the appropriate measurement on the data for the verification process (fig. 1 (C)).

2.2 Prediction/verification cycle

A normal analysis cycle begins with an hypothesis on the sence of the sentence in terms of a signing space modification with given parameters. The linguistinc model make it then possible to infer a sequence of gestures (ie. components description ordered in time) from it and the image processing module chooses the right operator, in that given context, to verify the predicted values for each valued property in the predicted description. Verification results for each property are finally merged to produce the final answer in a tree state decision process that allows *indeterminated* values to be produced.

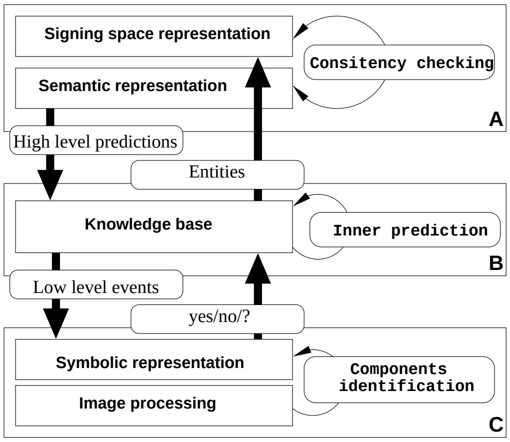

Figure 1: General overview of system architecture and communications between different subsystems during the prediction/verification procedure. Higher level module *(A)* uses a representation of signing space and of the sense of the discourse for semantic prediction and consistency checking of the results provided by the intermediate subsystem *(B)*. This subsystem uses knowledge about the FSL grammar to infer low level events on components from events predicted above. Finally, the last module *(C)* processes images to determine whether or not they corroborate the predicted events.

Depending on the results of the verification phase, the system will validate the courent hypothesis, reject it and formulate a new one (eventually by taking in account additionnal informations achieved by reconstruction) or choose an alternative strategy to solve indeterminations.

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

3 Modelling signing space

The model of the FSL we use is based on the iconicity theory from C. Cuxac [5][4]. This theory points out the fact that there can be found a direct correspondance between the sense of the sentence (in terms of entities and relations) and the way signing space is used.

3.1 Signing space utilization

If we consider the semantic of the sentence as a set of entities and relationships relying those entities, the signing space will be filled with those entities in order to satisfy the desired relationships. The sentence is realized by putting the different entities in place in the space surrounding the

[2] Signing space is the space surrounding the signer where gestures are performed

signer so that their respective location is relied to the semantic relationships among these items.

In our application, that means that one can perform an analysis of the sequence in order to determine the global organisation of the discourse without taking in account the lexicon in a at this level of analysis.

3.2 Signing space representation

Signing space is used at the same time as a representation of the sense of the sentence, and as a direct representation of the way the sentence is signed. Thus the model consists in two parts: a $3D$ representation of the spatial structure of signing space and a object-based representation of the underlying semantics.

Geometric representation of signing space: Signing space is represented by a cube surrounding the signer, regularly divided into *Site*(s)[3]. Each location may contain a single *Entity*, each of them having a *Referent*. A *Referent* is a semantic notion that can be found in the discourse. Once it has been placed in signing space, it becomes an *Entity* and has a role in the sentence. In this model, building a sentence in sign language consists in creating a set of *Entities* in *SigningSpace*. Figure (fig 2) gives an example of an instantiated signing space.

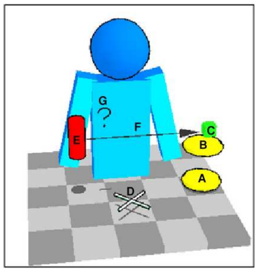

Figure 2: An example of construction of signing space that corresponds to the FSL question (signs order has been respected): *"In the city of Toulouse (A), in the movie theatre called Utopia (B), the movie that plays (C), on Thursday February 26th at 9.30 pm (D), the one (E) who made it (F), who is it (G) ?"*. In this figure, one can see that, the sentence is realized by putting the different entities in place in the space surrounding the signer and that their respective place is relied to the semantic relationships among these items.

Underlying sense representation: The representation of the sense contained in current signing space instantiation is represented in terms of *Entities*(s) whose *Referent* can have successively different *function* during the construction of the sentence (*locative, agent, . . .*). Each kind of *Referent* has a predefined subset of possible *Function*(s). A set of rules maintains the consistency of the representation by verifying that enough and coherent informations have been provided when a request for creating an entity is passed to this module. The figure (fig. 3) gives an overview of he global architecture of the subsystem in UML notation standard.

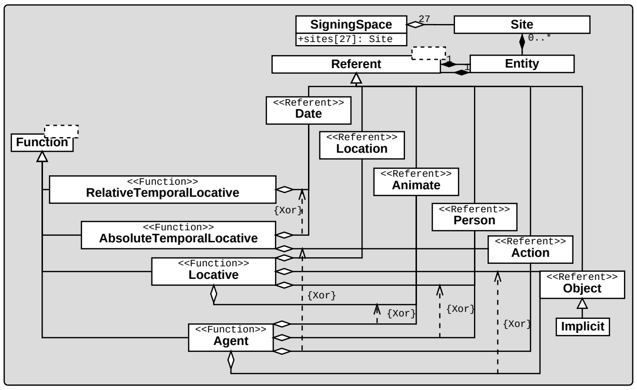

Figure 3: UML class diagram from the semantic representation of the *SigningSpace*. *SigningSpace* is regularity divided into *Sites*. Each *Site* can contain a single *Entity* whose *Referent* can have several *Function*(s) during the realisation of the sequence.

4 Computing Visual events (VE) from High level events (HLE)

Communication between signing space representation module and the second subsystem is limited to messages concerning the creation of *Entities*. So that the role of this subsystem is only, being given an event in signing space, to infer from it signer's gestures in signing space. This subsystem is implemented as a knowledge representation system that uses description logic formalism and CLASSIC knowledge representation system [2] which allows representation and inference on complex structured objects.

4.1 Knowledge base architecture

In FSL, it was shown that meaningful events that produced changes in signing space could be described in terms of sequences of events that implies properties of several components of the signer's body such as hands motion and position, gaze, chest movements, facial expression. . . Those sequences are used in the computation of a set of low level visual events.

Components and parameters: At this level, we distinguish four components in our model: *Hands, Body*, and *Gaze*. *Hand*(s) are separated in "dominating hand" (*DH*, and "dominated hand" (*dh*). Components have a set of parameters which have several predefined symbolic values.

[3] Terms written using a *slanted* font are elements of the model.

During instantiation mechanism, all or part of the parameters will have their value fixed. An object describing a component whose parameters are (eventually partially) instantiated is called *ComponentState*.

HLE description: HLE descriptions are based on the sequences of gestures that are used in FSL for the corresponding signing space construction. In our knowledge base, they are represented as a set of *ComponentState*. Additional behaviour comes directly from object hierarchy and predefined values in HLE definitions.

Inference mechanism: Inference uses both Classic value propagation through concept hierarchy and rules firing when appropriate. It allows partial objects instantiation so that only a few informations are needed in order to complete the description of each *ComponentState* object.

5 Image analysis subsystem

This part of the system relies on an operator description that consider them as high level entities that gives a three-state answer for the verification of each valued property of a component. Each of them is attached to tree sets of functions :

1. *measurement functions* which opers on properties of the datas in the images ;

2. *test functions* that produces the three-states answer based on the result of the *measurement function* ;

3. *application functions* that determine wether the operator can be used in that given context or not.

5.1 Verification process

This part of the system receives queries from the upper knowledge base as lists of VE. Its goal is, for each VE, to answer if data in the images are in contradiction with the prediction or not. Each operator is attached to a VE it can verify, so tht for each requested VE, the system checks *application functions* to choose the right operator and, if an operator was choosen, applies *measurement* and *test functions* in order to perform the verification. If no operator could be choosen, the system lefts the value for that givn VE ‘indeterminated and continues the analysis.

When all VEs where checked, a global answer is send to the intermediate system which take it in account for further prediction or requests for additionnal informations, obtained by reconstruction to correct the current hypothesis.

6 Conclusion

The use of a detailled linguistic model is a strong guideline to permit sign language image sequences analysis that avoids complex motion reconstruction. This paper has shown the main aspects of such an application that can be used to answer simple queries made in sign language without taking in account the lexicon. Furthermore, this kind of system can be used for an evaluation purpose of the language model so that it provides some formal approach for sign language analysis.

References

[1] M. Jardino B. Bossard, A. Braffort. Some issues in sign language processing. In 5^{th} *International Workshop On Gesture And Sign Language Based Human-Computer Interaction*, Genova, Italy, April 15-17 2003.

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

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

[4] C. Cuxac. French sign language: proposition of a structural explanation by iconicity. In Springer: Berlin, editor, *Lecture Notes in Artificial Intelligence : Procs 3rd Gesture Workshop'99 on Gesture and Sign-Language in Human-Computer Interaction*, pages 165–184, Gif-sur-Yvette, France, march 17-19 1999. A. Braffort, R. Gherbi , S. Gibet , J. Richardson, D. Teil.

[5] C. Cuxac. *La langue des Signes française. Les voies de l'iconicité.* ISBN 2-7080-0952-4. Faits de langue, Orphys, Paris, 2000.

[6] K.F. Kraiss H. Hienz, B. Bauer. Hmm-based continuous sign language recognition using stochastic grammars. In Springer: Berlin, editor, *Lecture Notes in Artificial Intelligence : Procs 3^{rd} Gesture Workshop'99 on Gesture and Sign-Language in Human-Computer Interaction*, pages 165–184, Gif-sur-Yvette, France, march 17-19 1999. A. Braffort, R. Gherbi , S. Gibet , J. Richardson, D. Teil.

[7] M. Huenerfauth. 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)*, pages 24–31, Lisbon Portugal, 30 May 2004.

[8] M. Ouhyoung R.H. Liang. 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*, pages 59–66, Hongkong, June 1996.

[9] M. Ouhyoung R.H. Liang. A real-time continuous gesture recognition system for sign language. In 3^{rd} *International conference on automatic face and gesture recognition*, pages 558–565, Nara, Japan, 1998.

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