MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
13-18

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

The study of sign languages provides a wide field of applications for computer vision systems, but the analysis of such gestures often leads to complex $3D$ reconstructions or to ambiguities. In this paper, we describe the architecture of an image analysis system that performs sign language analysis by using a prediction/verification approach. The system integrates a model of the sign language's structure and uses it during analysis to predict visual events enabling simple $2D$ features to be used to determine whether the images corroborate 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 that 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 applications "possible". Our approach is related to two fields of application: linguistic studies, where results provided by model-based image processing can be compared to a linguistical interpretation; sign language interpretation for automatic translation or in order to answer a question.

1.1 Previous work

Most of previous works on sign language linguistics focused on the description of an isolated sign using a finite set of parameters and values. Resulting transcription systems have been used for automatic translation as in [11], this system uses the Liddel and Johnson phonological description, or in [9], in this one datagloves are combined with the Stokoe description system. To be able to take in account the variations between several realisations of the same sign, another system, [2], uses a qualitative description of the signs based on a linguistic feature vector for vison-based recognition of british sign language. Other works focus on increasing the recognition rate by using additionnal knowledge on the signed sentence structure. This is done by using statistics on consecutive pairs of signs (so-called stochastic grammars) [5] [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 succession of isolated signs, eventually coarticulated. More complex aspects of sign language such as the utilization of the signing space 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 [6][1].

1.2 Our approach

Instead of focusing on direct sign recognition, we rather try to identifying the structure of the sentence in terms of entities and relationships which can be done by observing the way the signer expresses those entities in the space surrounding him (the signing space). Such a representation is generally sufficient in a reduced context. This approach allows us to use a general model of the grammmar and the syntax of sign language. Therefore, starting with a high level hypothesis on what is going to be said, this model lets us compute a set of low level visual events that have to occur in order to validate the hypothesis. As verifying that something has happened is simpler than detecting it, we will be able to use rather simple image processing in the verification phase with this approach and reserve explicit reconstruction of gestures for the cases where prediction becomes impossible.

[1] classifiers are gestures that are used to reference entities whose movement or configuration is directly relied to some physical aspects of the referenced entity.

2 Overview of the system

Our system analyzes French Sign language (FSL) gestures based on the fact that those gestures follow this language's grammatical rules. In order to perform this task using a single video camera and simple image processing, we need to integrate plenty of knowledge in our system : on FSL grammar and syntax for prediction and interpretations verification; on image processing for image-level verification module queries.

2.1 System architecture

Our system integrates these knowledges in a multi-level architecture that is divided in three main subsystems:

1. The first subsystem represents the interpretation of the discourse based on a signing space[2] modelling. During processing, the coherence of the 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 represents knowledge about FSL grammar and syntax using a description logic formalism. This subsystem is used to describe high level events that occurred in the signing space in terms of low-level sequences of events involving 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 measurements 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 meaning of the sentence, done on a signing space modification with given parameters. Then the linguistic model enables us to infer a sequence of gestures (ie. the description of components ordered in time). Finally the image processing module chooses an operator, in that given context, to verify the predicted values for each valued property in the predicted description. The verification results for each property are finally merged to produce the final answer in a three state decision process that allows *indeterminated* values to be produced.

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

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

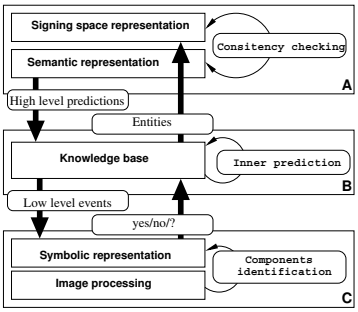

Figure 1: General overview of system architecture and communications between different subsystems during the prediction/verification procedure. Higher level module *(A)* uses a representation of signing space and of the sense of the discourse for semantic prediction and consistency checking of the results 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.

The next sections will describe the main aspects of the linguistical model and the verification process. Further details about that model can be found in [7].

3 Modelling the signing space

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

3.1 Using the signing space

If we consider the semantics of the sentence as a set of entities and relationships linking them, the sentence is realized by putting the different entities in place in the space surrounding the signer so that their respective location is linked 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 to represent both the meaning of the sentence, and the way the sentence is signed. Thus the model consists in two parts: a $3D$ representation of the spatial structure of the signing space and a object-based representation of the underlying semantics.

The geometric representation the of signing space: The 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 the 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 the *SigningSpace*. Figure (fig 2) gives an example of an instantiated signing space.

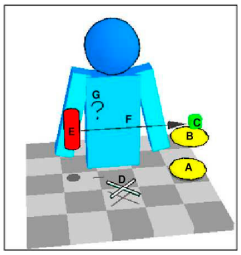

Figure 2: An example of the construction of the signing space that corresponds to the FSL question (the order of the signs has been respected): *"In the city of Toulouse (A), in the movie theatre called Utopia (B), the movie that plays (C), on Thursday February 26th at 9.30 pm (D), the one (E) who made 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.

The representation of underlying meaning: The representation of the meaning contained in the current signing space instantiation is represented in terms of *Entities*(s) whose *Referent*s can have successively different *function*s 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 given to this module. The figure (fig. 3) gives an overview of the global architecture of the subsystem in UML notation standard.

4 Computing Visual Events (VE) from High Level Events (HLE)

Communication between the signing space representation module and the second subsystem is limited to messages concerning the creation of *Entities*. The role of this subsystem is only, to infer signer's gestures description being given an event in the signing space. This subsystem is implemented as a knowledge representation system that uses

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

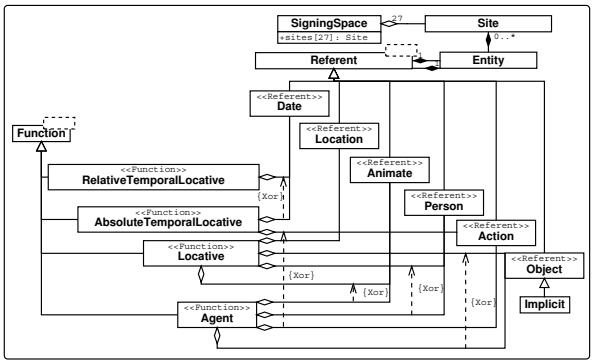

Figure 3: UML class diagram from the semantic representation of the *SigningSpace*. *SigningSpace* is regularity divided into *Sites*. Each *Site* can contain a single *Entity* whose *Referent* can have several *Function*(s) during the realisation of the sequence.

description logic formalism and CLASSIC's knowledge representation system [3] which allows representation and inference on complex structured objects.

4.1 Architecture of the knowledge base

In FSL, it has been shown that meaningful events that produce changes in the signing space can be described by event sequences where the properties of the components of the signer's body (eg. hands motion and position, gaze, chest movements...) will have particular values. Those sequences are used in the computation of low level visual events that will be searched in the sequence.

Components and parameters: We distinguish, for now, four components in our model: two *Hands*, *Body*, and *Gaze*. We distinguish the "dominating hand" (*DH*), and the "dominated hand" (*dh*). The components have a set of parameters which have several predefined symbolic values. During the instantiation mechanism, some of the parameters will have their value fixed. An object describing a component whose parameters are (eventually partially) instantiated is called a *ComponentState*.

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

Inference mechanism: Inference uses both CLASSIC's value propagation through concept hierarchy and rules fired when appropriate. It allows the manipulation of partial des-

critions 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 the operators as high level entities and gives a three-state answer for the verification of each valued property of a component. Each of the operators is attached to three sets of functions: *measurement functions* which oper on data properties in the images; *test functions* that produce the three-state answers based on the result of the *measurement function* and *application conditions* that determine whether 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 the data are in contradiction with the prediction. For each VE to verifiy, there are several operators that can perform this verification. For each requested VE, the system checks *application conditions* to choose the right operator and applies the *measurement* and *test functions* in order to perform the verification. If no operator could be choosen, the system lefts the value for that given VE "indeterminated" and continues the analysis.

When all VEs are checked, results are merged and a global answer is sent to the intermediate system which takes it in account for further prediction or requests for additionnal informations, obtained by reconstruction for instance, 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 avoiding 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] B. Bossard, A. Braffort, and M. Jardino. 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. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *Proc. 8^{th} European Conference on Computer Vision, ECCV04*, volume 1 of *Lecture Notes in Computer Science*, pages 391–401. Springer-Verlag, 2004.

[3] 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.

[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] H. Hienz, B. Bauer, and K.F. Kraiss. 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.

[6] 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.

[7] B. Lenseigne and P. Dalle. A computational model for sign language grammar. In *2^{nd} Language and technology Conference*, page (to be published), Poznam Poland, April 21-23 2005.

[8] R.H. Liang and M. Ouhyoung. 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] R.H. Liang and M. Ouhyoung. 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] T. Starner and A. Pentland. 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.

[11] C. Vogler and D. Metaxas. 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.