# A Software System for Automatic Signed Italian Recognition

**Ignazio Infantino[1], Riccardo Rizzo[1], and Salvatore Gaglio[1,2]**

[1]Istituto di Calcolo e Reti ad Alte Prestazioni, Consiglio Nazionale delle Ricerche, sede di Palermo

Edif. 11, Viale delle Scienze, 90128, Palermo, Italy

[2]Dipartimento di Ingegneria Informatica, Università degli Studi di Palermo,

Edif. 6, Viale delle Scienze, 90128, Palermo, Italy

E-mail: infantino@pa.icar.cnr.it, ricrizzo@pa.icar.cnr.it, gaglio@unipa..it

**Abstract**

The paper shows a system for automatic recognition of Signed Italian sentences. The proposed system is based on a multi-level architecture that models and manages the knowledge involved in the recognition process in a simple and robust way, integrating a common sense engine in order to deal with sentences in their context. In this architecture, the higher abstraction level introduces a semantic control and an analysis of the correctness of a sentence given a sequence of previously recognized signs. Experimentations are presented using a set of signs from the Italian Sign Language (LIS) and a sentence template useful for domotic applications, and show a high recognition rate that encourages to investigate on larger set of sign and more general contexts.

## 1. INTRODUCTION

In the recent years various approaches dealing with different sign languages have been proposed: many of the works are based on American Sign Language (ASL), but there are some works on different languages as the Chinese (Ma et al., 2000; Wang et al., 2000), German (Bauer et al.,2000), Netherlands (Grobel & Assan, 1997), Taiwanese Language (Liang & Ouhyoung, 1998), and English language (Sweeney & Downton, 1997).

From the computer engineering point of view, the works about Sign Language Recognition has been focused on two main tasks: the single sign recognition and the continuous sign recognition. The first one deals with the reconstruction and classification of a determinate configuration of arms, hands, head, and body from a static snapshot, or a movement that represents a single word. A large part of the works on this problem uses neural networks techniques with various approaches and configurations (Hamilton & Micheli-Tzanakou, 1994; Kim et al., 1996; Yang et al., 2002; Wilson & Anspach, 1993). Other works are based on the recognition of a hand posture from a static image using the appearance approach or model based approach ( Cui & Weng, 2000; Triesh & von der Malsburg, 1996; Waldrom & Kim, 1995). The second research issue, i.e. the analysis of the sequence of signs, concerns the analysis of complex movements as continuous signals in space and time. Many of the works (see for example Vogler & Metaxas, 2001) use statistical approach based on Hidden Markov Models (HMMs), exploiting the assumption of the whole movement of a sign is decomposable in simpler components: chereme, i.e. the analogous of the phoneme in the sign language (Sweeney & Downton, 1997). Both single and continuous sign recognition need to detect and tracking spatial position (2D or 3D) of hands, arms, head of the signer. Often, this problem is easy resolved using marked gloves or gloves with special sensors. For example in Vogler & Metaxas, 2001, a set of magnetic markers were used in order to acquire the movement of hand and fingers. Recently, the non-manual gestures components are treated as a part of the sign language and processed together with the manual signs (see for example Grobel & Assan, 1997).

In this paper, we propose a inexpensive system that does not require special equipments or acquisition apparatus, but assure good performance using a sort of semantic context in a continuous sign language recognition task. The set of the admissible words, for a given detected sign, is disambiguated using the local context of the sentence. In order to perform that, we integrate a common sense engine in the process.

We explain the various functionalities implemented as a generic framework based on a cognitive architecture that allows to model and manage the knowledge of the recognition process in its wholeness. At the top of this architecture is the linguistic level that introduces the semantic context and allows the analysis of the correctness of a sentence given a sequence of recognized signs. Due to the predefined sentence templates that are used, the recognition is limited to a standardized gesture corresponding to words from the Italian Sign Language.

The paper is organized in five sections. After the introduction, the Sign Language features are highlighted to understand the implementation problems to approach. Section 2 deals with the various modules that compose the proposed system. Section 4 deals with a complete example of a sentence recognition. Finally, conclusions and discussion are reported.

## 2. FEATURES OF SIGN LANGUAGES

The interest in sign languages born with the Stokoe's work (Stokoe, 1978) that dealt with American Sign Language (ASL). The author discovered in ASL an organization similar to the common language, where a combination of simple sounds (phoneme) is used to create a very large number of words: in ASL a combination of simple gestures are used to generate a large number of signs with very different meanings. According with the analysis proposed by Stokoe a sign can be decomposed in three

parameters: the place in the space here the hands execute the sign (TAB); the configuration of the hands when they executes the sign (DEZ); the movement used to execute the sign (SIG). Another important parameter highlighted by Stokoe is the orientation of the palm, because some signs has the same DEZ, TAB and SIG but differs only for the palm orientation. It is to mention that in the set of signs of the ASL it is possible to identify the so-called Minimal Pairs: a couple of signs that differs only for a minimal variation of a single parameter. These non-manual components are very important in all the sign languages studied, because they convey an extra-linguistic information that is essential for the single sign recognition and also for continuous sign recognition.

In Italian Sign Language (LIS) the communication between the deaf and the other people the labial reading is frequently used since in the Italian language there is a very tight connection between the written and spoken language. The labial component is used in sign recognition when two sign are similar (Volterra, 1997).

## 3. SYSTEM OVERVIEW

The problem of sign translation must necessarily consider several aspects, known in literature as Recognition of Continuous Sign Language: the sign can begin and end in any instant of an observed sequence, since a temporal restriction in the execution of a sign does not exist; different signs have a variable duration, or the same sign can be executed with different duration; the transition from a sign to another is not exactly identified; a sign depends from the previous sign and the next sign (*coarticulation problem*); the begin and end of the single phrase is difficult to identify, the number of the signs in a phrase is not fixed.

All the cited problems make the recognition of continuous sign language a complex problem and a global solution is difficult to find. The aim of this work is to provide an architecture and a methodology to find a solution in the case of a single sentence. First of all we defined a subset of LIS sign that could be used in a home automation context. Such set is composed of verbs showing the key action to be performed, substantives, adjectives and time adverbs The system is based on a extensible words vocabulary, large enough to create a set of commands to give orders to a domestic robot or at a domotic automation system.

The proposed framework works in three main steps: sensorial input is processed in order to obtain a features vector by standard image processing algorithms; a Self Organizing Map (SOM neural network) is used to classified the single sign and to provide a list of probable meanings; a common sense engine finally choose the right meaning depending of the whole sentence context. An overview of the whole system architecture is depicted in figure 1. The system acquires a video sequence from a single video camera placed in frontal position with respect to the Signer.

The recognition is based on the space position of the signer's hands with respect to her head (Bowden et al., 2004; Charanyapan & Marble, 1992; Vogler & Metaxas, 2001). The main functionalities implemented are: the extraction and coding of the head and hands movement of the signer; the segmentation of the video input in single signs; the recognition of a single sign; the reconstruction of the whole sentence; the semantic analysis of the reconstructed sentence.

The incoming video is processed by the Movement detection Subsystem (MDS) that is the main component of the Pre-Processing Module (PM). This subsystem segments the video into several parts each of them representing a single sign, and generates a vector v. This vector is the input of the Sign Recognition subsystem that is in the Pattern Classification Module (PCM). This subsystem provides a list of the possible meanings for a single detected sign.

The last subsystem of the Pattern Classification Module collects the lists of meanings and reconstructs a set of sentences that are sent to the Sentence Recognition subsystem in Reasoning Module. This subsystem uses a common sense inferential software to check the meaning of the sentence. In the following subsections the implementation details of each single area are reported.

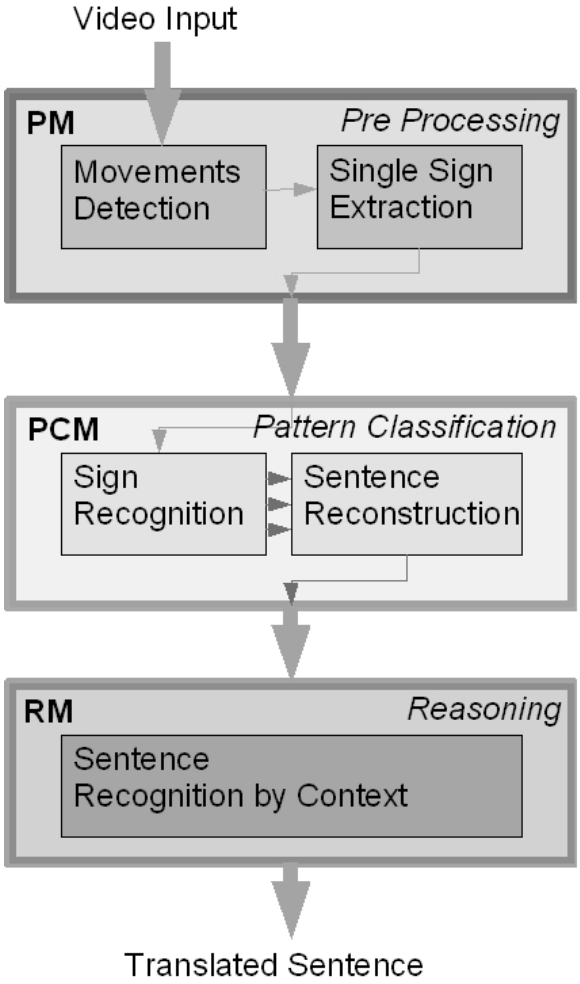



Figure 1: An overview of the proposed system.

The system is composed by the following processing blocks: Pre-Processing Module (PM), the Pattern Classification Module (PCM), the Reasoning Module (RM).

### SIGN DETECTION ON IMAGE

The first step of the process is the pre-processing phase by

the module PM. It aims to identify all the relevant image features (Bauer et al., 1990): we separated the skin areas of the image by using a colour based segmentation (see figure 2). The localization of head and hands is obtained by an algorithm based on detection of connected components. Only the three more extended regions are considered using the following criteria: head has greater area; hands are the others two regions if their extension is up to 10% of head area: in this way, the system is able to ignore non relevant regions; left and right hands are distinguished by their position on image. Each region is described by the following parameters: coordinates of the centroid ; width and height max ; width and height at the coordinates of the centroid; region area. The system follows the variation of these parameters, even if the occlusions happen.

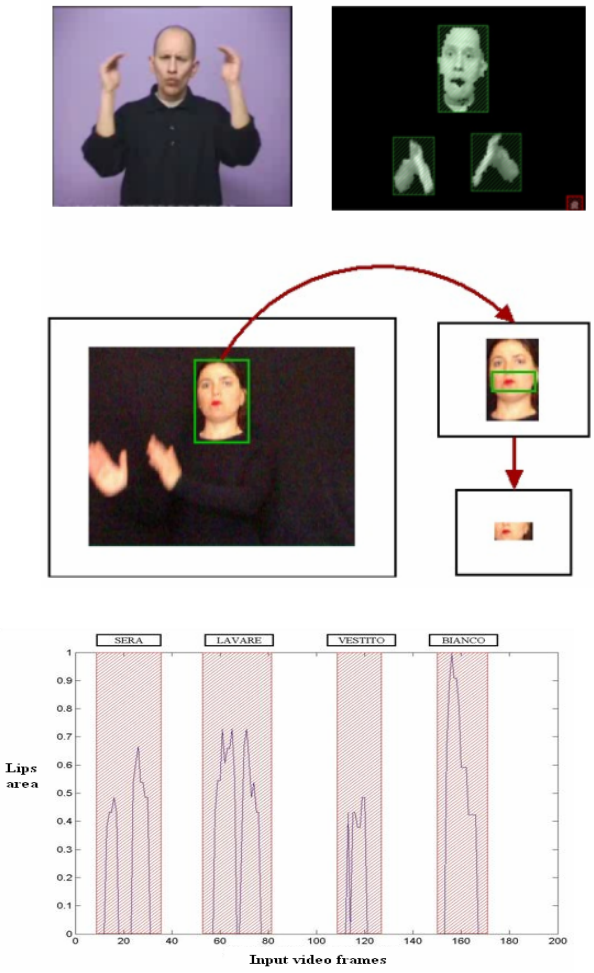



Figure 2: Pre-processing steps and results.

The pre-processing phase detect skin areas on the image and extract hand movements (first row). The motion of lips (middle row) is used to insulate single signs in a sentence (bottom row).

The data are collected for each frame of the sequence, and the whole movement corresponding to a sign is coded by a vector of real numbers. From the analysis of a huge number of sign videos the maximum sign duration observed is about 1300 ms. According to a given frame rate of 15 frames per second, a typical value even for a low-cost system, we choose n=10 a value that allows to capture the movements details. Having a fixed value of n allows obtaining a feature vector of 40 elements.

As said in section 2, the Italian signers often use lips movements during sign execution, especially when communicate with other people, and they don't feel as limiting to do it when use the system. This can be successfully exploited to solve the sentence segmentation problem. The pauses between lips movements reproducing the word of corresponding sign allow us to segment the sentence in fast and robust way (see figure2). This segmentation can be refined using an empirical consideration: usually an absence of signal can be considered a pause if it is about 500 ms. Considering frames rate of 15 fps this allows us to ignore signal interruptions that are shorter than 8 frames.

## SIGN REPRESENTATION

The aim of the Pattern Classification Module (PCM) is to work as a bridge connecting the perception to the symbolic processing, and to label the incoming sign representation. Using the signs representation as a vector it is possible to transform the comparison operation in a metrics measurement, and clustering allows to easily labelling the incoming pattern.

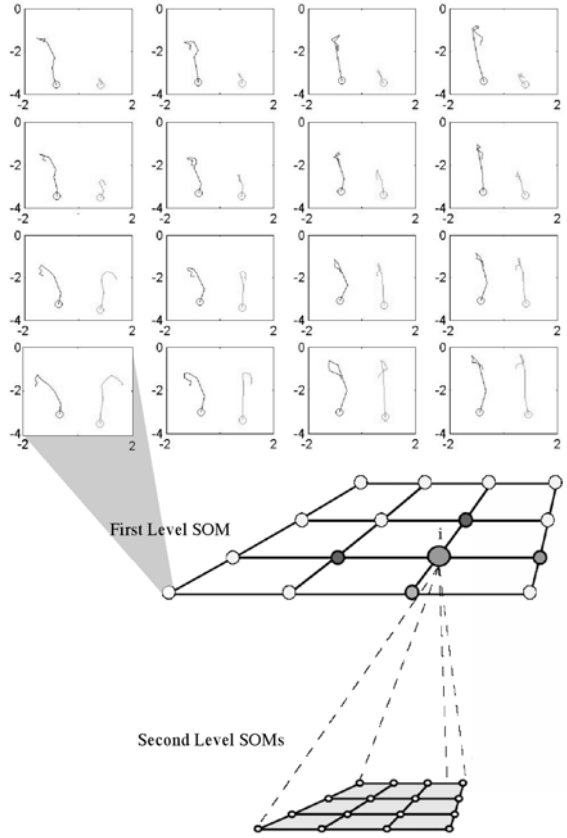
In our system the sentence recognition is made by the linguistic area, so the meaning of the sign is decided at the end of the processing chain, when the sentence is analyzed as a whole. To obtain this it is necessary to have a ranked list of possible meanings for each sign instead of a single answer (label) from the sign recognition block. This ranked list constitutes a search space for the Reasoning Module, and it is obtained exploring the representation space to find, inside the training set, similar signs. A topological map, obtained using a SOM neural network (Kohonen, 1997), supports the exploration of the representation space.

Using SOM, if an arbitrary pattern is mapped onto a unit, all points in a neighbourhood of it are mapped either to itself or to one of the units in the neighbourhood. This property is highlighted in the upper section of figure 3: the input patterns of our SOM are vectors that represent hands movements, and the weight vectors approximating these trajectories are visualized. Notice that the movements that involve only the right hand are mapped on the upper left corner of the map.

A SOM clustering system is usually obtained training the SOM with a set of patterns, then labelling the SOM units. This labelling can be done using the labels associated with the training patterns (in our case the meaning of the sign) using a voting criteria. Usually patterns that are neighbourhood in space have the same label or different label that can have a semantic relationship. If this property exists it is transmitted to the topological mapping. In our application due to the presence of the so-called *minimal pairs* (a sort of false friends for sign language), similar patterns can have very different label. Moreover for a single sign we can have few examples because for the user is a waste of time to repeat many and many times the same sign.

Many training of SOM with different topologies and different learning parameters were performed and a good compromise was obtained using a 4 X 4 topology. This allows to have a gross grain separation of the patterns, then, in order to obtain a finer separation of the signs inside the clusters, another layer of SOM networks was used (Miikkulainen, 1990). In second layer there is a SOM

network for each unit of the first layer; i.e. the second layer is made by 16 SOM network of the same 4 by 4 topology (see figure 3). The hierarchical SOM classifier uses a total of 272 neurons, but is more efficient of a single layer SOM network. The SOM multilayer classifier associates a ranked list of labels to the input pattern (the single sign detected); this ranked list came from the labels




associated to the training patterns.

Figure 3. SOM multilayer structure.

The bottom part of the figure shows the SOM multilayer structure. The second level SOM will be trained using the training patterns of the gray units of the first level. In the upper section the weights of the neurons are visualized as the corresponding movement. In each square there is the movement of the hands, for example it can be noticed that signs with only right hand are in the upper-left corner.

## COMMON SENSE CONTEXT ANALYSIS

The Reasoning Module Area is responsible for the syntactic control of the sequence of the signs, and generates the most probable sentence referring a given context. This module is based on OpenCyc[1], the open source version of the Cyc technology that implements a complete general knowledge base and commonsense reasoning engine. OpenCyc has been used as the basis of a wide variety of intelligent applications and expert systems. The principal functionalities of the module are: verification of the semantic correctness of a complete sentence; search of a single error in the sentence and suggestion of possible substitutes; correction of the error by evaluation of the most probable sentence using all the possible substitutions. Moreover the system is able to suggest next sign (or a set of possible signs) if an incomplete sentence is given. In the case of man-machine interaction, i.e. signer and recognition system are involved in a simple structured discussion (command/ request/ question+ answer+ …), the general context is defined but every recognized sentence generates a current context useful for process the next sentence. The context is introduced in Cyc defining a *microtheory*, i.e. a constant denoting assertions which are grouped together because they share a set of assumptions. It is accessible by querying to the Cyc Server about the truth of a sentence (formula), which may or may not contain an undefined sign (free variable). If the formula contains variables, then Cyc server supplies bindings to those variables which make the formula true (correct sentence); otherwise, it simply answers whether the formula is true. In the following experimental part examples of query are reported explaining the various functionalities of semantic control.

## 4. EXPERIMENTATIONS

The system is mainly written in Java code: it recalls the Preprocessing Module implemented using Matlab, that also includes the video acquisition capabilities; a specific Java class of the Reasoning Module manages the queries and the answers to/from the OpenCyc engine. The SOM multi-layer classifier was implemented using the Matlab SOM Toolbox.

The video sequences used for the training and the recognition has been acquired with a digital camera using a resolution of 320*240 at 15 fps (Pentax Optio 330GS). Moreover, we have mainly used some low resolution videos from a free database of sign ("Dizionario Italiano dei Segni", DIZLIS[2]). The implemented system has optimal performance if the signer is in front of the camera and the background is uniformed colored. Figure 4 shows the implemented graphical user interface that allows to manage all the computation.

The performances of the Pattern Classification Module are due to the Sign Recognition subsystem and the Sentence segmentation. The Sign Recognition subsystem is based on the SOM classifier and performances are determined by the training set that constitutes the vocabulary of the system. The larger vocabulary tested is made by 40 signs and there were added few minimal pairs. Each sign was repeated 4 times in order to obtain a set of 160 video fragments. The data extraction procedure generates a matrix of 160 rows and 40 columns. The training set is small compared with the dimension of the input space but to have more video samples is a problem because the user should repeat a single sign a lot of times. To artificially add more vectors to the training set we replicated the same representing vectors 4 times adding Gaussian noise with zero means and 0.1 variance. Using this method a matrix of 640 rows and 40 columns was obtained. Each row was labelled with a reference to the original video segment.

We submitted to the system 80 videos that where not part of the training set of the system and take the first five

[1] Cycorp, Inc. OpenCyc, http://www.cyc.com.

[2] http://www.dizlis.it

labels obtained from the Pattern Classification Module: for 50 videos the correct label was the first one, for the remaining videos the second label was checked and it was found correct in 17 cases, and so on. For two videos the correct label was not on the first five choices and these videos are considered not classified.

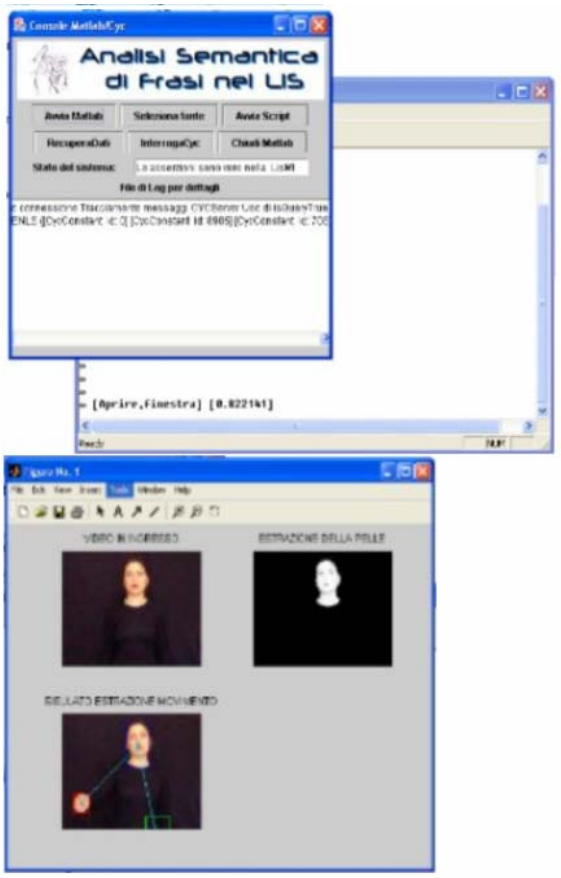



Figure 4: The Graphical User Interface of the system

The sign segmentation subsystem is based on the lips movement and uses a threshold on the width of the lips area of segment the sentence.
As mentioned before, we used OpenCyc capabilities to implement the reasoning module: we defined a microtheory called LisMt that is a specialization of AgentMt. The other modules are able to interact with the Cyc server by a suitable Java application that is executed in the Matlab workspace.
The defined a microteory LisMt includes 40 signs, and represents the current context. The first pre-processing step eliminates adjectives that usually are not relevant to investigate the correctness of the sentence. In the following an example of processed sentence is reported. The true signs are "ROBOT LIBRO BIANCO PRENDERE" (robot take white book).
In the example the adjective bianco (white) is not considered in the first evaluation: it will be considered if no possible correct sentence is founded, and in this case the system will search for a possible substitute as it will be shown in the next example. The syntactic analysis and the consistence with the local context are obtained generating the following query:

(#$and(#$relationAllExists #$performedBy #$prendere #$robot)
(#$relationAllExists #$objectActedOn #$prendere #$libro))

The Cyc server checks if
- #$robot is an actor
- #$prendere is a verb
- #$libro is a object

the sentence is compatible with LisMt and returns TRUE and the sentence is accepted. Now, we describe a sentence with an error: robot cucinare vestito [robot dress cooks]. The query generated is

(#$and (#$relationAllExists #$performedBy #$cucinare #$robot)
(#$relationAllExists #$objectActedOn #$cucinare #$vestito))

and Cyc returns FALSE, because there is not the context validation:
- #$robot is an actor
- #$cucinare is a verb
- #$vestito is a object
- but vestito is not a #$Food and cannot be cooked

From the analysis of the SOM maps we see that vestito was the first recognized sign, but pasta – the correct one – was the second with a similar rank index in the ranked list of possible meanings. The sign cucinare was recognized with a consistent difference from the second most probable. Then, the system tries to investigate the possible substitutes of vestito. A query with a free variable is generated:

```
CycList error_query= cycAccess.makeCycList(" (#$relationAllExists
#$objectActedOn #$cucinare ?X)");
CycList substitutes_list= cycAccess.askWithVariable( error_query, new
CycVariable("?X"),mt);
```

The returned list is {pane, pasta}, i.e. all the objects of the current context that can be cooked. Pasta is selected because is included in the SOM candidates. The whole process, from video acquisition to sentence recognition, takes less than 1 second plus the movement duration (typically ~4 seconds for a complete sentence): the delay is mainly dues to the wait of the answer from OpenCyc.

| | |
|---|---|
| Correct segmented sentences | 76 (95.5%) |
| Correct translated sentences | 66 (82.5%) |
| Erroneous translated sentences | 10 (12.5%) |

Table 1. Experimental results.

A test set of 80 videos of sentences has been used to check the system performances. We have obtained the 95% of correct segmented sentences (76 of 80). Moreover, 66 of this 76 sentences has been correctly translated, obtained a final success rate of 82,5%.

## 5. CONCLUSIONS

We have proposed a complete framework for sign language recognition that integrates a common sense engine in order to deal with sentences. The proposed architecture allows modelling and managing the knowledge of the recognition process in a simple and robust way. Moreover, the introduction of the semantic context resolves the problem of the analysis and validation of a sentence.

The presented experiments show that the system maintains the recognition rate high when the set of sign grows, correcting erroneous recognized single sign using the context. Table 1 shows the experimental results using 80 videos of sentences using 40 signs. Experiments demonstrate the goodness of the proposed approach. Future research will deal with the extension of the number of signs, allowing to use the system in more general contexts.

## 6. REFERENCES


Bauer, B., Hienz, H., Kraiss, K.F. (2000). Video-based continuous sign language recognition using statistical methods. In proceedings of the Intl. Conference on Pattern Recognition, vol. 2, pp. 463-- 366.

Bowden, R., Windridge, D., Kadir T., Zisserman A, Brady M. (2004). A linguistic Feature Vector for the Visual Interpretation of sign language. In Tomas Pajdla, Jiri Matas (Eds), proc. of 8th European Conf. on Computer Vision, ECCV04, LNCS3022, Springer Verlag, vol. 1, pp. 391--401.

Charanyapan, C., Marble, A. (1992). Image processing system for interpreting motion in the American Sign Language. Journal of Biomedical Engineering, vol. 14(5), pp. 419--425.

Cui Y., Weng, J. (2000). Apparence based hand sign recognition from intensity image sequences. Computer Vision and Image Understanding, May 2000, pp. 175--176.

Erdem, M. U., Sclaroff, S. (2002). Automatic detection of relevant head gesture in American Sign Language communication. In proceedings of the International Conference on Pattern Recognition, pp. 460--463.

Grobel, K., Assan, M. (1997). Isolated sign language recognition using hidden Markov models. In proceedings of the IEEE Internationak Conference on System, Man and Cibernetics, vol. 1 (12-15), pp. 162--167.

Hamilton, J., Micheli-Tzanakou, E. (1994). Alopex neural network for manual alphabet recognition. In proceedings of the IEEE Conference on Engineering in Medicine and Biology: Engineering Advances: New Opportunities for Biomedical Engineers, pp. 1109--1110.

Liang, R. H., Ouhyoung, M. (1998). A real-time continuous gesture recognition system for sign language. In proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 558--567.

Kim, J.S., Jang, W., Bien, Z.N. (1996). A dynamic gesture recognition system for the Korean sign language (KSL). IEEE Transaction on System, Man, and Cybernetics: Part B, April 1996, pp. 354--359.

Kohonen, T. (1997). Self-Organizing Maps, Springer-Verlag, NY.

Ma J., Gao, W., Wang, C., Wu, J. (2000). A continuous Chinese sign language recognition system. In proceeding of the International Conf. on Automatic Face and Gesture Recognition, pp. 428--433.

Miikkulainen, R. (1990). Script recognition with hierarchical feature maps. Connection Science, Vol. 2, pp. 83--101.

Starner, T., Pentland, A. (1995). Real time American Sign Language recognition from video using hidden Markov models. In proceedings of Symposium on Computer Vision, pp. 265--270.

Stokoe, W. (1978). Sign Language, Silver Spring, Linstok Press.

Sweeney, G.J., Downton C, (1997). Sign language recognition using a cheremic architecture. In proceedings of the Sixth International Conference on Image Processing and Its Applications, vol. 2(14-17), pp. 483--486.

Triesch, J., von der Malsburg, C. (1996). Robust classification of hand postures against complex backgrounds. In proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 170-175.

Yang, M.H., Ahuja, N., Tabb, M. (2002). Extraction of 2d motion trajectories and its application to hand gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Aug 2002, (pp. 1061--1074).

Vogler, C., Metaxas, D. (2001). A framework of recognizing the simultaneous aspects of American Sign Language. Computer Vision and Image Understanding, n. 81, (pp. 358--384).

Volterra, V. (1987). La lingua italiana dei segni: La comunicazione visivo-gestuale dei sordi. Il Mulino, Bologna.

Waldrom, M.B., Kim, S. (1995). Isolated ASL sign recognition system for deaf persons. IEEE Trans. on Rehabilitation Engineering, vol. 3(3), pp. 261--271.

Wang, C., Gao W., Shan, S. (2002). An approach based on phonemes to large vocabulary Chinese sign language recognition. In proceedings of the International Conference on Automatic Face and Gesture Recognition, (pp. 393--398).

Wilson, E. J., Anspach, G. (1993). Applying neural network developments to sign language translation. In proceedings of the IEEE-SP Workshop on Neural Networks for Processing, (pp. 301--310).