Mobil phone camera Recognition Sign Language using a segmented multidecision filter adapted to the Parallel virtual machine (PVM)

A Marwa Elbouz[1,2], Ayman Alfalou[1], Habib Hamam[2]
*1- Laboratory L@BISEN ISEN-Brest, 20 rue cruirasse Bretagne, CS 42807,
29228, Brest cedex 2, France
marwa.el-bouz@isen.fr, ayman.al-falou@isen.fr*

*2- EMAT Laboratory, University of Moncton, 165 Massey av., E3V 2S8, Edmundston
(N.-B.), Canada
Habib.Hamam@umoncton.ca*

Abstract

In our society based on communication, the barrier between the spoken and gestural languages remains a main problem for the deaf-mute. To make this communication possible, we propose developing an interface using mobile phone camera with a module of recognition allowing the recognition and the restitution of the sign language in the form of synthesized words. This module is based on a spectral comparison between the image to recognize and a correlation multidecision segmented filter resulting from images reference coming from a predetermined database: correlation technique. The latter has been used for a long time to recognize a target image. However an adaptation of this technique is necessary to take into account particular characteristics of gestures (for which the target image to recognize evolves in time) and to be able to send it with a high transmission rate. Thus to obtain a reliable decision, the module of recognition must carry out a very large number of correlations, which slows down the translation process. In order to reduce this processing time, we propose implementing our recognition module by using multipost technique "PVM". This technique is based on the distribution of calculation on various stations in parallel. This technique enables us to distribute the required computing effort on various parallel stations using (PVM) a Parallel Virtual Machine. It showed significant gains in computing time but not sufficiently for a real-time translation (mobile phone communication). To further reduce processing time, we propose adding a Comprehension Pre-selection module to the initial recognition module. The former is based on the division of images (resulting from the gesture target) in various zones, in order to reduce and eliminate all the improbable gestures, according to the positions of hands. Thus, we reduce the needed identification time.

Key words: Sign Language, Parallel Virtual Machine, Optics, Correlation.

1. Introduction

To retranscribe the sign language, we purpose to develop an interface enabling the sign language recognition and to present the recognized signs in speech or synthesized word forms. Information is then transmitted by a mobile phone [1,2,3,4] in order to make communication possible between a deaf-mute and people not familiar with the sign language. For this purpose, our system (figure 1) must include:

- A camera in front of which the deaf-mute will place his hands and will make the desired gesture.

- A module of gesture comprehension (pre-selection) based on the identification of the both hands in space in order to eliminate improbable gestures in the references database.

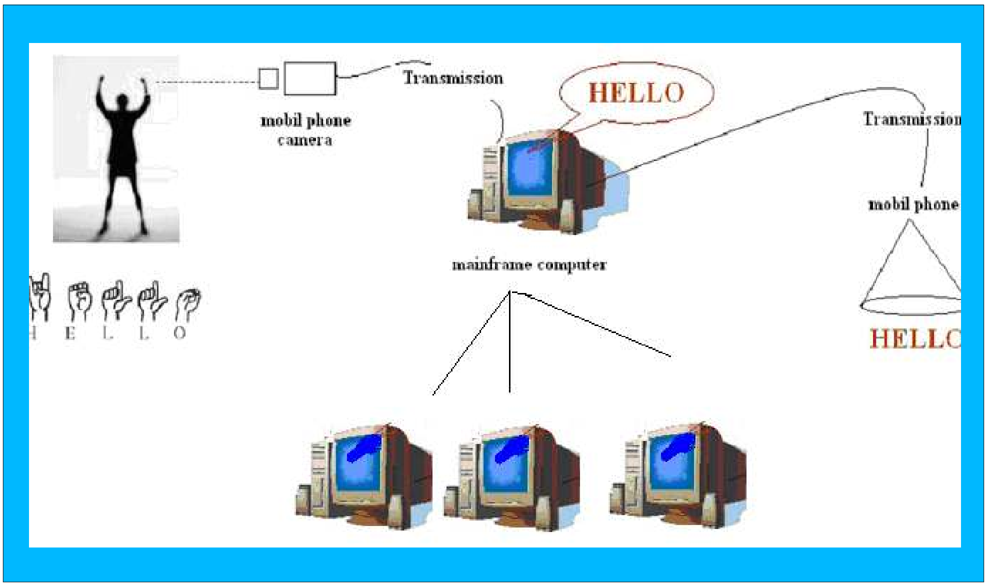


Figure 1. The synoptic diagram establishing the recognition module on several machines

-A module of correlation to recognize a target gesture. This module compares the target gesture with a limited set of gestures which are likely reference gestures selected by the Pre-selection comprehension module.

This is only one part of the sign language [5]. Indeed, the movements of the arms and those of the mouth are also important elements to be considered and will be treated in future articles.

Thanks to the recent progresses in the field of real time correlators, based on adapted filtering, it is now possible to implement complex architectures for image processing [6,7]. In this article, we focus on a numerical implementation, using initially only one station (then we planned a numerical implementation on DSP, or FPGA processors) to design our correlator based on segmented multicorrelation filters elementary operator [8,9,10]. Indeed, this correlation filter, originally designed for an optical implementation [11], showed a high capacity of discrimination [12] and enabled multidecision [9,13] to identify, for an example, an object with various angles of sight.

The main idea of this article is to associate this complex architecture with a broad data base of gestures which makes it possible to identify the whole set of the signs and thus to translate the language [14]. The major problem of this technology is that it requires a high processor performance and a large memory which goes largely beyond what is offered by current computers. Indeed, a correlation can be carried out with a computer in a reasonable time. In our case, in order to carry out the recognition of the gesture, we need hundreds of successive correlations almost running in real time. For this reason, we opted for reducing the reference gestures (by choosing the likely ones) and a multi-station technique saving time while keeping the processing flexibility to the hybrid correlation. We will see the reason of choosing a master/slave architecture through the use of **PVM** (**P**arallel **V**irtual **M**achine) software [7]. In this article, we limit ourselves to validating the principle of this new architecture adapted to the multi-station use for the recognition of gestures.

2. Correlation by adapted filter

The correlation, as a technique of patter recognition, is well adapted to two-dimensional signals: images. The advantages of this technique are, on the one hand, the high capacity of discrimination, and on the other hand, the possibility of treating the image as a whole (parallelism). The principle is simple (figure 2) and consists in comparing the object to be recognized (the "**?**" target), generally being in a scene limited in space, with predetermined objects (reference **R**$_i$). To carry out this comparison, the spectrum of the target (**S**), obtained after carrying out a **F**ourier **T**ransformation (**TF**), is multiplied by the inverse complex of the reference spectrum (**H**$_i$), which we refer to as the adapted filter. Then, a second Fourier Transform provides the product of correlation (the Correlation plane: target/reference). In this plane, applying a threshold enables making a decision.

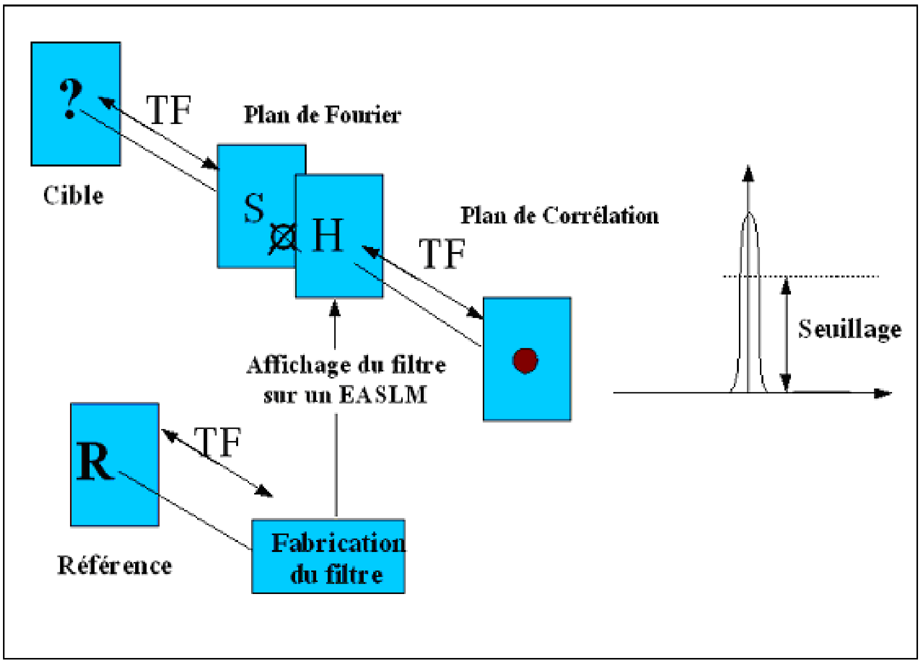


Figure 2. General diagrams of a correlator by adapted filtering

As we have just seen before and owing to the fact that, in a gesture, the target image evolves/moves in time, it is necessary to use a large number of reference images. As a consequence it is necessary to carry out a large number of comparisons to make a decision. This results in hugely increasing the time required for making a right decision since it is necessary to compare the target image with all the reference images. To overcome this problem, we propose a first solution, which consists in using the technique of multidecision correlation [8,9]. This technique is based on only one comparison between the image to recognize (target image) and a filter of correlation regrouping information from several reference images. The technique of merging information originating from various references within only one filter referred to as a segmented filter, already showed its high performance in other applications [8,9].

3. Suggested correlation module

However, to obtain a system working in real time, this computing reduction is still insufficient. Moreover, making a decision by only considering the shapes of both hands does not ensure robustness and accuracy system. Indeed, a sign is defined by the shapes of the hands but also by the positions of the two hands (in space). This makes recognition difficult. For this reason, we propose to add to the first recognition module another interpretation module which is concerned by reducing the number of correlations to be made.

This second module is founded on a number of space-time rules limiting gesture. Indeed, some space areas are specific to some signs [5] (figure 3 - a,b). With a second camera (figure 3-c), we calculate the positions of the two hands.

This enables optimizing the processing. According to the initial positions of the hands, we can exclude some signs, thus limiting the number of comparisons to be carried out.

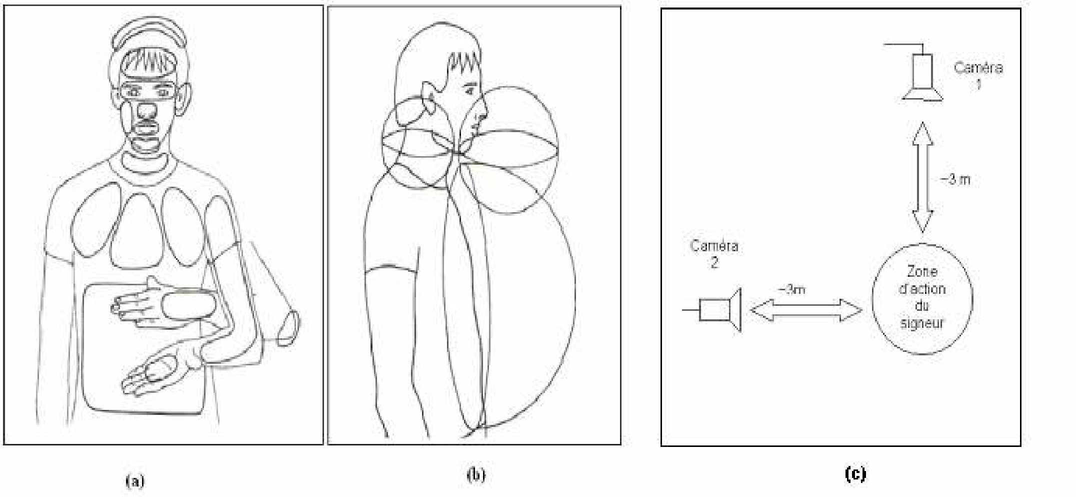


Figure 3. (a-b) Areas used by the person making signs, (c) device and configuration to be used to detect the signs.

Thus not all the images of the gestures in the video sequence are investigated, but only some potential images referred to as "**key images**". This results in a faster processing and requires a data base composed of fewer reference images. Indeed, after having chosen these key images the

Master sends them one by one to different slave stations to carry out their identification tasks (figure 4). If a slave identifies the image with an element of its data base, it sends the results to the Master, and so on, for all the slaves (figure 4).

In each of these cases, each slave will return to the Master the outputs of its treatment. The Master will reorganize all the information which are returned to him in order to give the final result, that is to say: to return the word or the sentence which will have been played in front of the camera (gesture). To reach this solution, it would have been necessary to analyse the sign language in order to optimise its interpretation.

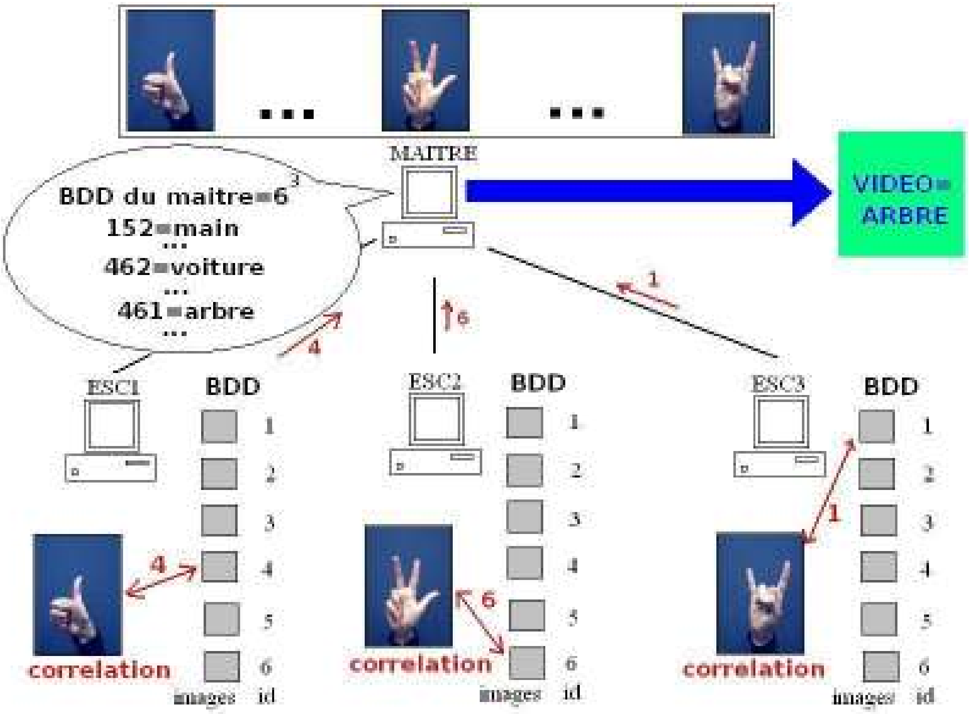


Figure 4. architecture proposed using PVM

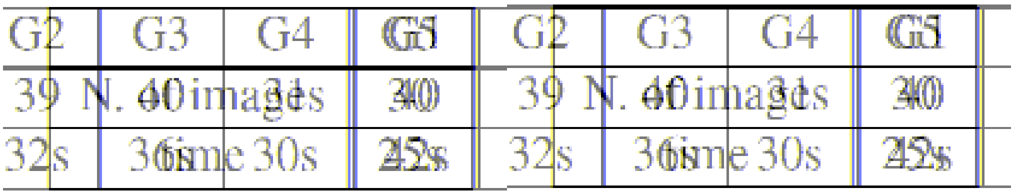
3.1. Simulations Results of the recognition module

In order to validate the principle of our architecture, we will carry out an algorithm of image processing using correlation on several machines which work in parallel (PVM). The aim is to interpret gestures into the French sign language (LSF). Because of the heaviness of calculations and the required important memory capacity, in this article, we will dispatch the treatments on several machines: "**one"** Master and "**four**" slaves. It will be easy to adapt it to a larger data-processing park in the future.

To carry out these tests, we started by filming some gestures. Then these films were divided into several images at the rate of "**5"** images by second (key images). Then, we labelled these various gestures and associated each of them with a sound file. For our study, we filmed "**4"** gestures with "**40"** images each. We thus obtained approximately "**160**" images that we exploited. After observing each of these gestures, we selected î6î images characterizing each gesture. These "**6**" images/references will be used to manufacture the various filters of correlation to proceed to the recognition of the various gestures. Thus, we created a reference data base with "**24"** images references/filters. By using a segmented filter gathering information coming from two ("**2**") images/references, we thus obtain "**12**" filters. In this study, we made the choice to place one gesture by slave. To test the behaviour of our architecture, we filmed a gesture randomly (i.e. **40** images which make it up). These "**40**" images are correlated with different data base on each slave (one gesture corresponds to one slave). After each treatment for one input image, each slave returns to the Master the image it thinks to have identified. The Master makes a decision according to the level of correlation needed, measured by using the PCE correlation criterion (peaks to correlation Energy) [15]. The identification decision consists in taking the image with the highest "**PCE"** Value. This image is identified by a number and the station (slave) from which it comes. At the end of the "**40"** correlations, if the Master has identified a maximum of images on station number "**n**", he makes the following decision: it is the gesture-N which is filmed in input and thus it will start the corresponding sound file with the gesture "**n**". Thus we succeeded in recognizing the gesture filmed in input with a rate of success of 87%.

3.2. Discussion

In fact, we did not succeed in identifying the gesture with **100%** of success. This is due essentially to the fact that the images of the same gesture differed according to the angle and the scale of the shot. To solve this problem, we increased the reference images in order to make our filters more robust with respect to these two problems. We note that to increase the recognition rate, it is necessary to increase the number of reference images in our data base to take into account the possible modifications of the images (rotation, scale). However this solution will considerably increase the identification time. Thus it is necessary to find a compromise between the number of reference images to incorporate into our data base and the desired processing time. In our example, to have a recognition without errors, it would have been necessary to multiply by "**2**" the number of reference filters to be used "**12 x 2 = 24**". After having validated our architecture, it is certain that our program is not completely optimized especially concerning the processing time. It would be necessary to use more slaves for example. Another correlation approach would perhaps be more time saving. With our approach and to have a **100%** recognition rate, we need today a recognition time of about "**35**"s for a video sequence of about "**40**" images with "**24**" filters (table 1).



	G2	G3	G4	G5
N. of images	39	40	31	40
time	32s	36s	30s	25s

Table 1. Needed time to recognize gestures with our method

4. Pre-selection Module with divided zones

To reduce time and the power necessary for the gesture recognition, we propose to add to our recognition module another one : a comprehension and Pre-selection module. The main role of this module is to reduce the number of reference gestures (ex: one gesture by machine) to be compared with the target gesture. This enables to benefiting from all the machines in order to recognize the gesture and thus reducing the computing time. This module is based on the target image division in various zones (take into account the characteristic of gesture) in order to identify the position of hands (figure 5). To do this, in this article, we used the movement detection to follow hands (To optimize this diction we investigating several methods see details in future paper).

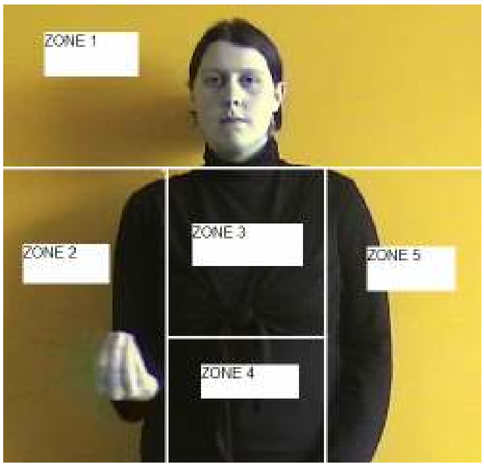


Figure 5. target gesture divided on several zones

Thus, we can eliminate all the improbable gestures without the hands positions. However it was necessary to classify our reference gestures according to their hands positions. With the gestures used higher, we succeeded in eliminating 1/3 of the needed gestures/references while reducing as much the recognition time presented table (1). however, to reduce “more and more” the number of gestures/references it is necessary to reduce the size of different zones more and more figure(**5**), while increasing the complexity of the comprehension module and its processing time. We are making tests on a very large number of real gesture in order to find a compromise between the number of probable gestures/reference to keep and the processing time.

5. Conclusions

This project relates to the assistance of people with special needs. The tests (in real time) carried out in our laboratory, by using gestures with a duration corresponding to forty images, show that by using these two pre-selection (comprehension) and recognition modules, we succeeded in reducing significantly the time necessary to make a decision. This is essentially due to the pre-selection (comprehension) module which pre-selected some possible gestures. This pre-selection was carried out by selecting key positions of both hands in one gesture. For example, this makes it possible to draw aside the gesture where the hands do not coincide with the reference positions.

In order to reduce the processing time to get a real time system, studies are being made to determine the hand position more rapidly according to specific areas (figure-3 and figure-5). Tests are also under way to validate the robustness of our approach with respect to false alarm and the absences of detection.

6. References

[1] H. Cooper, R. Bowden. **Sign Language Recognition Using Boosted Volumetric Features**. Proceedings of the MVA2007 : IAPR conference on Machine Vision Applications (**MVA 2007**). Page(s): 359-362, May 2007.

[2] J.-C. Terrillon, A. Pilpre;Y. Niwa, K. Yamamoto. **A realtime system for robust multiple face detection, tracking and hand posture recognition in color video sequences**. Proceedings of the 17th International Conference on Pattern Recognition, 2004 (**ICPR 2004**). Volume 3, Page(s): 302-305, Aug. 2004.

[3] R. Bowden, D.Windridge, T. Kadir, A. Zisserman, M. Brady **A Linguistic Feature Vector for the Visual Interpretation of Sign Language**, Proceedings of the 8th European Conference on Computer Vision, **ECCV04**, LNCS3022, Springer-Verlag, Volume 1, Page(s): 391- 401, (2004).

[4] Y. Liu,bY. Su, Y. Yang, F. Wang, M. Yuan, Z. Ren **A Facial Sketch Animation Generator for mobile cummunication**. Proceedings of the MVA2007 : IAPR conference on Machine Vision Applications (**MVA 2007**). Page(s): 532-535, May 2007.

[5] B. Moody. **La langue des signes 1**, *IVT Editions*, ISBN: 2-904641-17-3.

[6] F. T. S. Yu and S. Jutamulia. **Optical Pattern recognition**. *Cambridge university Press*, ISBN: 0-521-46517-6.

[7] D. Kranzlmuller, P. Kacsuk, J. Dongarra. **Recent Advances in Parallel Virtual Machine and Message Passing Interface 2004: Proceedings of the 11th European PVM/MPI (LNCS 3241)**. Edited by *Springer* 2004.

[8] A. Al Falou, G. Keryer, J.-L. de Bougrenet de la Tocnaye. **Optical Implementation of Segmented Composite Filtering**. *Applied Optics*, Vol. 38, Issue 29, pp. 6129-6135, 1999.

[9] A. Al Falou, M. El Bouz and H. Hamam. **Segmented phase only filter binarized with a new approach of error diffusion method**. *Journal of Optics A: Pure and Applied Optics*, Vol. 7, pp: 183-191, 2005.

[10] M. Farhat,A. Al Falou, H. Hamam. **Face Recognition by Optical Correlation using 3D Filter** Proceedings of the 2nd IEEE conferenc on Information and Communication Technologies (*ICTTA '06*. Volume: 1, page(s): 1568- 1572, ISBN: 0-7803-9521-2, April 2006.

[11] A.B. Vander Lugt. **Signal detection by complex spatial filtering**, *IEEE*, IT-10, pp: 139-145, 1964.

[12] A.V. Openheim. **The importance of phase signals**. In proceedings of the IEEE, V. 69, P. 529-541, 1981.

[13] G. Keryer, J. L. de Bougrenet, A. AL Falou. **Performance comparison of ferroelectric liquid-crystal-technologybased coherent optical multichannel correlators**. *Applied Optics*, Vol. 36, Page(s): 3043-3055, 1997.

[14] A. Al Falou. **A hybrid correlator for language sign recognition based on the used of a nonlinear segmented filter**. WSEAS transactions on electronics, Issue 3, Vol. 1, July 2004, pp. 547-550.

[15] D. Roberge and Y. Sheng. **Optical wavelet matched filter.** *Applied Optics*, Vol. 33, Issue 23, pp. 5287-5293, 1994.