Segmentation and classification of hand gestures for man-machine communication

F García-Ugalde[1], D Gatica-Pérez[2] and V García-Garduño[3]

[1,3]Depto. de Ing. Eléctrica-DEP, FI-UNAM, Apdo. Postal 70-256
04510 México, D.F., MEXICO

[2]Human Interface Technology Laboratory, University of Washington,
Box 352142, Seattle, WA 98195, USA

garcia@verona.fi-p.unam.mx

ABSTRACT

Man-machine communication in a natural way, it means without cumbersome gloves, is still an open problem. Keeping in mind the need to develop some friendly tools for helping people with disabilities to use the computer as a support tool for training into reinforced methods for learning to read or any other application. In this work we have addressed the problem of communication with a computer using some recognition of very basic hand gestures. From an engineering point of view our system is based on a video camera which captures image sequences and in a first time a segmentation of hand gestures is developed in order to provide information for its posterior classification and recognition. For classifying the segmented fields named **e** of gestures, for instance *hand # 1* and *hand # 2,* see figures 5a and 6a, we have proceed first to obtain a binary version of these segmented fields comparing them with a threshold, so rendering the classification faster, then based on the Radon transform (Lim, 1990), a computation of the projected sum of the binary intensity of gestures has been done at directions 0° and 90°, see figures 1 and 2. For reducing the number of data to be processed a wavelet decomposition of the projected sum of the binary intensity for each orientation (0° and 90°) has been done using Daubechies filters: d4 (Daubechies, 1988). This projected and wavelet decomposed information has been used for classifying the gestures: training our system with our dictionary and computing the correlation coefficient between the wavelet coefficients corresponding to *trained sequences* and others captured and computed in continuous operation, the computer is able to recognize the very simple gestures. The region segmentation has been done using a dense motion vector field as the main information then each region is matched to a four-parameter motion model (Gatica-Pérez et al, 1997). Based on Markov Random Fields the segmentation model detects moving parts of the human body with different apparent displacement such as the hands (García-Ugalde et al, 1997). The motion vector field has been estimated by a Baaziz pel-recursive method (Baaziz, 1991) and considered together with others sources of information such as intensity contours, intensity values and non-compensated pixels as inputs of the Markov Random Field model. The maximum a posteriori criterion (MAP) is used for the optimization of the solution, and performed with a deterministic method: *iterated conditional modes* (ICM). The complete segmentation algorithm includes initializing, region numbering and labeling, parameter estimation of the motion model in each region, and optimization of the segmentation field. So our probabilistic approach takes into account the fact that an exact displacement field does not exist (errors usually occur at or around motion boundaries), and that better results can be attained if an indicator of the quality of the vector field is known, this indicator is obtained from the non-compensated pixels as well as the intensity contours (García-Ugalde et al, 1997).

1. INTRODUCTION

The problem of developing friendly interfaces for man-machine communication is still an open problem. However these interfaces become more and more necessary for helping people with some kind of disabilities to communicate with the computer. For instance, there exist some methods used for permitting handicapped people to learn to read: these methods consist essentially in defining one hand gesture for each of the 26

letters of the alphabet and use these hand gestures at the same time as a therapist shows the letter sign and makes the assocaited sound, this permits to deliver more information and then as a result people learning to read could learn more quickly because they assocaite to each letter three parameters: the sign of the letter, the sound and the hand gesture.

Our system is thinking as a support for the therapist, it has to recognize the hand gesture associated with each letter and then by synthetic speech generate the associated sound. So in these conditions people with disabilities could training without the constant presence of the therapist, they just need a computer with a video camera in front of them and when they are practicing to read, they are doing the hand gestures associated with letters and the computer is generating the sounds, making a close loop.

From an engineering point of view, for this system works the computer needs to recognize the hand gesture: segmenting and classifing it. The problem of segmenting moving regions was traditionally studied for coding video, accurate motion vector field estimation is crucial in this application as well as the segmentation and can be seen as interdependent problems, because one is needed to obtain the other with accuracy (estimation-segmentation ambiguity). Thus, motion-based segmentation is crucial for extracting high level information from the time-varying intensity of a sequence and for improving the motion measurement process (García-Garduño, 1995), (François, 1991), it represents a qualitative change from a local motion description to a regional one. In this paper, we present an algorithm to segment image sequences that begins with a Baaziz (1991) pel-recursive estimation of a motion vector field. Later, we model the image sequence using Markov Random Fields and pursue the optimization of the segmentation problem by a Bayesian estimation criterion (MAP) performed with a deterministic method: *iterated conditional modes* (ICM). Our probabilistic approach takes into account the fact that an exact displacement field does not exist (errors usually occur at or around motion boundaries), and that better results can be attained if an indicator of the quality of the vector field is known, this indicator is obtained from the non-compensated pixels as well as the contours. The classification of the segmented areas has been developed by computing for the *trained sequences* and others, a correlation coefficient of wavelet coefficients, of the projected sum of the intensity (at orientations 0^0 and 90^0) of the segmentation field (in its binary version).

2. GESTURE CLASSIFICATION

For classifying the segmented fields **e** of gestures, for instance *hand # 1* and *hand # 2* we have proceed first to obtain a binary version of these fields by comparing with a threshold, so rendering the classification faster. Then based on the Radon transform (Lim, 1990)

$$p(t) = \int_{u=-\infty}^{\infty} \mathbf{e}\left(t_1, t_2\right) \Big|_{t_1 = t\cos\theta - u\sin\theta,\ t_2 = t\sin\theta + u\cos\theta} du \qquad (1)$$

a computation of the projected sum of the binary intensity has been done at orientations: 0^0 and 90^0 see figures 1 and 2 for *hand # 1* and *hand # 2* respectively. Then for reducing the number of data to be processed a wavelet decomposition of the projected sum of the intensity for each orientation (0^0 and 90^0) has been done using Daubechies filters: d4 (Daubechies, 1988). This projected and wavelet decomposed information has been used for classifying the gestures, training our system with our dictionary and computing the correlation coefficient between the wavelet coefficients corresponding to *trained sequences* and others obtained in continuous operation of the system.

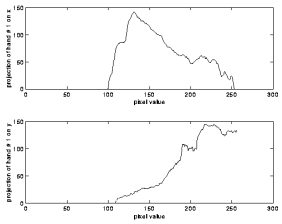

Figure 1. *Projected sum of the intensity of the binary version of the segmented field* **e** *of hand # 1 at* 0^0 *and* 90^0 *respectively.*

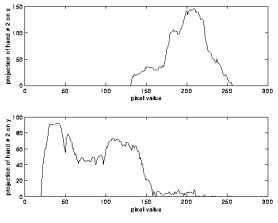

Figure 2. *Projected sum of the intensity of the binary version of the segmentation field* **e** *of hand # 2 at* 0^0 *and* 90^0 *respectively*.

3. THE SEGMENTATION ALGORITHM

From the point of view of image processing in presence of pure divergent motion, simple spatial clustering techniques for motion-based segmentation do not work well (García-Garduño, 1995). In this case, a model Θ_M of both the motion and the structure of the regions in the scene has to be introduced. Thus, the goal of the segmentation process is to assign each pixel in the image to one out of several regions, depending on the accuracy between each estimated motion vector and the assumed model. Each region is characterized by a motion parameter vector. The obtained regions can then be associated to different regions of the same object, or to different objects in the scene. The proposed segmentation algorithm is based on Markov Random Fields and estimation theory, using the *maximum a posteriori* criterion as optimality principle. Such a combined approach provides a common framework in which we can introduce information sources of distinct nature, model their interactions, and incorporate expected properties on the solution. Markov Random Fields modeling is appropriate for motion segmentation: we have a dense displacement vector field as the main information for separating an image into regions, but as we have discussed, motion information is not always correct, especially at movement discontinuities; in this case we may also include other data sources: intensity gray values, non-compensated pixels and intensity contours, as additional observations to improve the final solution. Furthermore, we add physical properties to the model: (a) a motion model Θ_M for each region in the scene to be segmented, (b) spatial continuity for the segmentation, (c) presence of motion boundaries only when strong intensity changes occur, and (d) expected geometrical shapes for the region boundaries. According to MRF theory, we will represent each information source as an *observation field* and each expected result as a *label field.* In this case, observations are:

- the estimated horizontal and vertical components of the motion field $\mathbf{d}_x$ and $\mathbf{d}_y$
- the binary non-compensated pixel field $\mathbf{p}$. As we mentioned earlier, it can be considered as a simplified way of removing motion outliers, for it represents a way of switching between displacement and more reliable information (intensity values) when the motion field is not accurately estimated
- the image intensity gray values field $\mathbf{i}$
- the binary intensity contour field $\mathbf{g}$ that favors the coincidence of motion boundaries and strong spatial gradients: 0 means no contour; 1 means contour.

On the other hand, desired *label* fields are:

- the desired segmentation label field $\mathbf{e}$ which has associated a four-parameter simplified linear motion model $\Theta_{MLS} = (t_x, t_y, k, \theta)$, that can describe combined translational, rotational, and divergent motions of planar surfaces parallel to the image plane (García-Garduño, 1995)

$$\begin{bmatrix} \mathrm{d_x} \\ \mathrm{d_y} \end{bmatrix} = \begin{bmatrix} \mathrm{t_x} \\ \mathrm{t_y} \end{bmatrix} + \begin{bmatrix} \mathrm{k} & -\theta \\ \theta & \mathrm{k} \end{bmatrix} \begin{bmatrix} \mathrm{x} - \mathrm{x_g} \\ \mathrm{y} - \mathrm{y_g} \end{bmatrix} \tag{2}$$

where (x_g, y_g) is the center of gravity of each surface.

- To improve the segmentation process we introduce an auxiliary binary motion discontinuity line field $\mathbf{l}$ along with the segmentation label field: motion boundaries (0 means no motion discontinuity; 1 means motion discontinuity).

In figure 3 we show the interaction model of observations, labels and physical assumptions.

Assuming *N* pixels in the image we formulate the motion-based segmentation as an estimation problem: simultaneously find the *label* fields $(\hat{\mathbf{e}}, \hat{\mathbf{l}})$ that maximize the *a posteriori* probability density function (**pdf**) of the labels, given the *observed* data :

$$(\hat{\mathbf{e}}, \hat{\mathbf{l}}) = \arg max_{\mathbf{e},\mathbf{l}} p(\mathbf{e}, \mathbf{l} | \mathbf{d}_x, \mathbf{d}_y, \mathbf{i}, \mathbf{p}, \mathbf{g}) \tag{3}$$

Reversing the problem using the Bayes rule, the last equation can be expressed as

$$(\hat{\mathbf{e}}, \hat{\mathbf{l}}) = \arg max_{\mathbf{e},\mathbf{l}} p(\mathbf{d}_x, \mathbf{d}_y, \mathbf{i} | \mathbf{e}, \mathbf{l}, \mathbf{p}, \mathbf{g}) p(\mathbf{e}, \mathbf{l} | \mathbf{p}, \mathbf{g}) \tag{4}$$

In (Gatica-Pérez et al, 1997) we have shown that maximizing the *a posteriori* **pdf** is equivalent to minimize a so-called *energy function* $U(e, l, d_x, d_y, i, p, g)$ which has the form

$$U(\mathbf{e}, \mathbf{l}, \mathbf{d}_x, \mathbf{d}_y, \mathbf{i}, \mathbf{p}, \mathbf{g}) = \alpha\ U_{\mathbf{d}}(\mathbf{d}_x, \mathbf{d}_y, \mathbf{e}, \mathbf{p}) + \beta\ U_{\mathbf{i}}(\mathbf{i}, \mathbf{e}, \mathbf{p}) + \gamma\ U_{\mathbf{e}}(\mathbf{e}, \mathbf{l}) + \kappa\, U_{\mathbf{l}}(\mathbf{l}, \mathbf{g}) \tag{5}$$

where α, β, γ and κ are weighting terms, all these energy terms has been also defined in (Gatica-Pérez et al, 1997).

3.1 Global optimization using iterated conditional modes method

To overcome the great computational cost required by simulated annealing, the global optimization of the solution is reached by using an iterative deterministic relaxation procedure: a modified *iterated conditional modes* (ICM) method based on an instability table (François, 1991). ICM methods minimize the local energy ΔU_s in each pixel $s(x, y)$ of the image. Our minimization scheme considers two phases in each iteration (García-Garduño, 1995): one for the optimization of the segmentation field through minimizing

$$U_1 = \alpha\ U_{\mathbf{d}}(\mathbf{d}_x, \mathbf{d}_y, \mathbf{e}, \mathbf{p}) + \beta\ U_{\mathbf{i}}(\mathbf{i}, \mathbf{e}, \mathbf{p}) + \gamma\ \ U_{\mathbf{e}}(\mathbf{e}, \mathbf{l}) \tag{6}$$

and the other for the optimization of the motion discontinuity line field, minimizing

$$U_2 = \gamma\ \ U_{\mathbf{e}}(\mathbf{e}, \mathbf{l}) + \kappa\ U_{\mathbf{l}}(\mathbf{l}, \mathbf{g}) \tag{7}$$

The term $U_{\mathbf{e}}(\mathbf{e}, \mathbf{l})$ represents a link term between the two stages of the optimization general process.

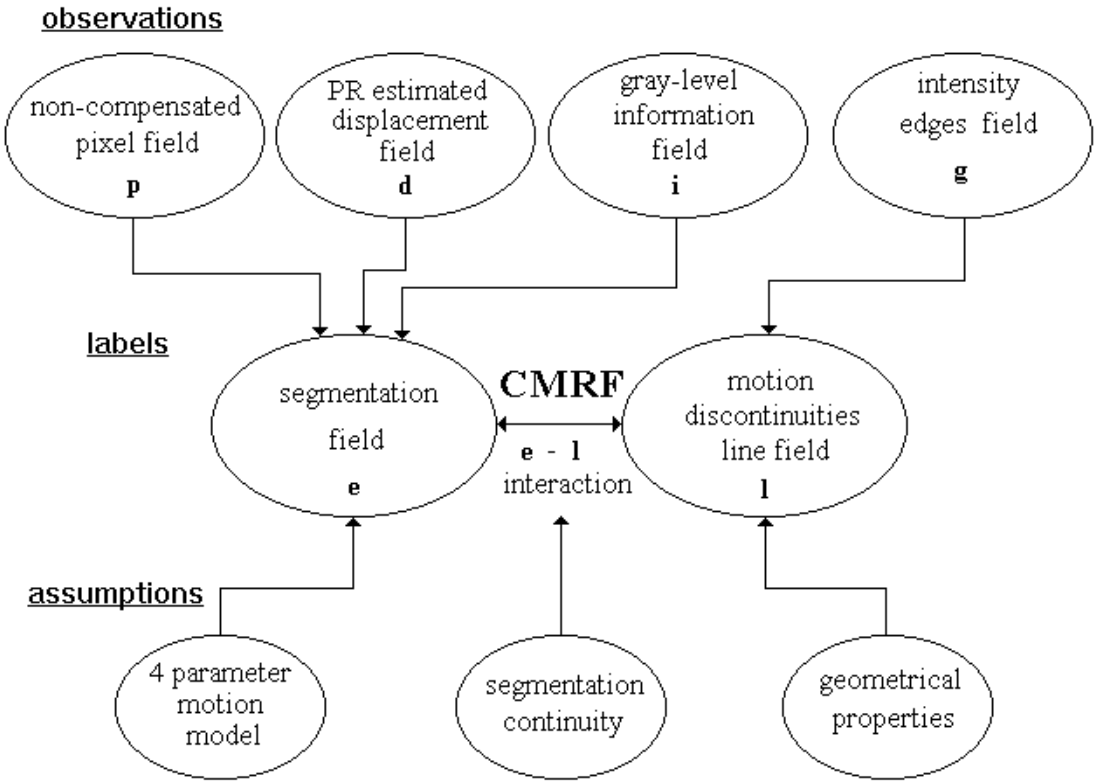

Figure 3. *Interaction model for the motion-based segmentation algorithm.*

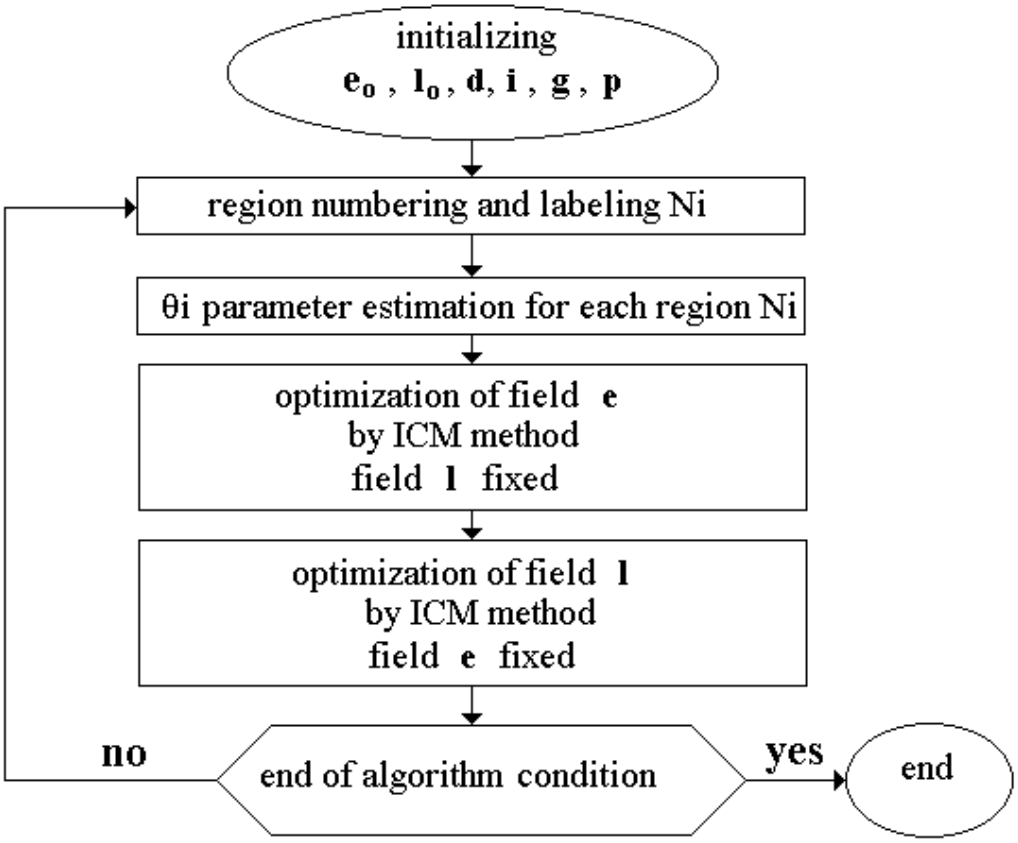

Figure 4. *General diagram of the proposed motion-based segmentation algorithm.*

3.2 The complete motion-based segmentation algorithm

The complete motion-based segmentation algorithm includes four stages (a) initializing, (b) numbering and labeling of each region in the image, (c) motion model parameter estimation in each region, and (d) optimization of the label fields. These steps are repeated until the method reaches the maximum number of iterations allowed, or until the segmentation becomes stable. An advantage of our algorithm is that the number of regions in the image is not fixed through the segmentation process, see figure 4.

4. SEGMENTATION RESULTS

Results obtained on the test sequences *hand # 1* and *hand # 2* for segmenting the moving parts are presented in figures 5 and 6 respectively. The segmentation fields obtained using the proposed algorithm are shown in figures 5a and 6a. A superposition of *hand # 1* and *hand # 2* with their respective segmentation regions can be seen in figures 5b and 6b. From the results it can be observed that the *hands* in motion have been well segmented from the rest of the scene. This result is qualitatively correct and reached only after 2 iterations of the segmentation algorithm (for each case: *hand # 1* and *hand # 2*), the tiny regions remaining in the background could be fused in posterior iterations.

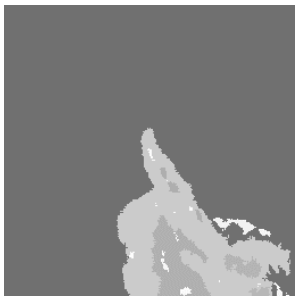

Figure 5.a. *MAP motion-based segmentation algorithm. Segmentation field* **e** *of hand # 1.*

Figure 5.b. *Superposition of frame 3 of hand # 1 and segmentation field* **e.**

Figure 6.a. *MAP motion-based segmentation algorithm. Segmentation field* **e** *of hand # 2.*

Figure 6.b. *Superposition of frame 7 of hand # 2 and segmentation field* **e**.

No further processing has been done on the segmentation frontiers. The motion vector fields obtained with the Baaziz method, figures 8a and 10a, are somewhat homogeneous and properly adjusted to the moving areas. The non-compensated pixels shown in figures 8b and 10b, represent respectively only 4.65 % and 1.99 %. A very important input to the segmentation algorithm is the initial binary segmentation regions, this initialization was obtained by thresholding the difference between, respectively frames 1 and 2 for *hand # 1* and frames 6 and 7 for *hand # 2*, and passing the resulting difference through a median filter of window 3x3. Keeping in mind the use of segmented regions for helping people with some kind of disability to communicate with a computer, when the segmentation of the *hands* has been completed, we have defined a very simple dictionary of gestures which are used to reinforce methods designed for learning to read.

5. PEL-RECURSIVE MOTION ESTIMATION

The mainly use of pel-recursive displacement estimation since it was proposed by Netravali and Robbins (1979), has been on predictive motion-compensated image sequence coding. A dense motion field based on the spatio-temporal varying intensity of a sequence is produced, such a field can be considered as a low level information source. For its computing, in this work we have selected the Baaziz method which consist on two main steps: in the first one it uses the method proposed by Biemond et al. (1987) and in a second step for reducing the number of non-compensated pixels the Walker and Rao method is used (on non-compensated sites only). We have applied it towards a higher level representation of an image sequence: a dense field will constitute the main clue to guide the pixel fusion process into regions of similar motion. Displacement is computed along the scan direction according to a prediction-updating scheme until convergence is obtained. One proper criteria for convergence is the recursive minimization of the reconstruction error; this minimization can also be iterative. Thus, in the Wiener-based algorithm, the displacement is estimated for each pixel until the DFD has been minimized using the equation

$$\hat{\mathbf{d}}^{\mathbf{i}} = \hat{\mathbf{d}}^{\mathbf{i-1}} - \begin{pmatrix} \sum_{j=1}^{N_n} \left(i_x^j\right)^2 + \mu & \sum_{j=1}^{N_n} i_x^j i_y^j \\ \sum_{j=1}^{N_n} i_x^j i_y^j & \sum_{j=1}^{N_n} \left(i_y^j\right)^2 + \mu \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_{j=1}^{N_n} i_x^j \cdot \mathrm{DFD}\left(\mathbf{z}_j, \hat{\mathbf{d}}^{\mathbf{i-1}}\right) \\ \sum_{j=1}^{N_n} i_y^j \cdot \mathrm{DFD}\left(\mathbf{z}_j, \hat{\mathbf{d}}^{\mathbf{i-1}}\right) \end{pmatrix} \tag{8}$$

where

- $i(x,y,t)$ is the intensity of each pixel of the sequence
- $\mathbf{z} = (x,y)$ is the position of each pixel
- $\mathbf{d}(x,y,t) = (d_x(x,y,t), d_y(x,y,t))$ is the displacement vector of each pixel in the interval $(t - k\Delta t, t)$
- $\hat{\mathbf{d}}^{\mathbf{i-1}}$ is the initial displacement estimation (prediction) for each pixel. If estimation is iterative, it represents the displacement after $\mathbf{i}-1$ iterations
- $\hat{\mathbf{d}}^{\mathbf{i}}$ is the final estimation for each pixel, (or after $\mathbf{i}$ iterations)

DFD is the displaced-frame-difference (reconstruction error)

$$DFD(\mathbf{z},\mathbf{d}) = i(z,t) - i(z-d, t-k\Delta t) \tag{9}$$

- N_n represents the number of pixels in a small casual spatial neighborhood of each pixel
- i_x^j y i_y^j are the components of the intensity gradient vector ∇i on the displaced positions in frame $t - k\Delta t$

$$i_x^j = i_x(z_j - \hat{\mathbf{d}}^{\mathbf{i-1}}, t - k\Delta t)$$

$$i_y^j = i_y(z_j - \hat{\mathbf{d}}^{\mathbf{i-1}}, t - k\Delta t) \tag{10}$$

$\mu = \sigma_v^2 / \sigma_u^2$ is the ratio between linearizing error variance and actualization term variance respectively.

Figure 7.a. *Sequence hand # 1. Frame 1.*

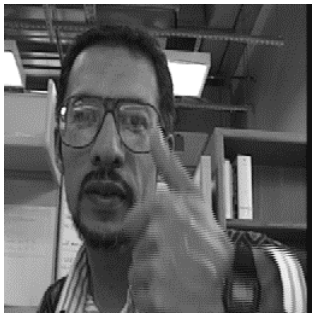

Figure 7.b. *Sequence hand # 1. Frame 4.*

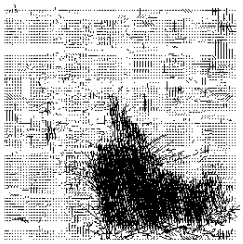

Figure 8.a. *Motion field obtained with the Baaziz method using hand # 1 , frames 1 and 2.*

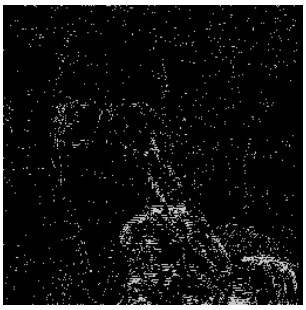

Figure 8.b. *Non-compensated pixel binary image using hand # 1 (0=compensated, 1=non-compensated)*

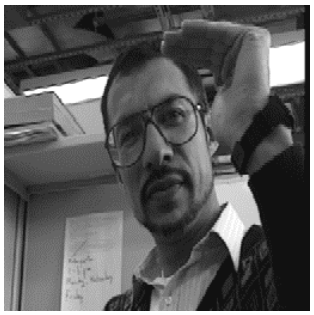

Figure 9.a. *Sequence hand # 2. Frame 1.*

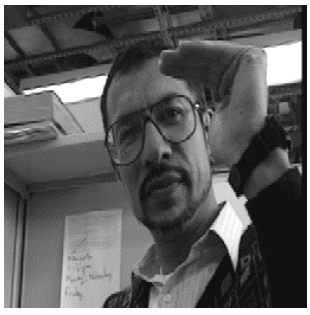

Figure 9.b. *Sequence hand # 2. Frame 9.*

Proc. 2nd Euro. Conf. Disability, Virtual Reality & Assoc. Tech., Skövde, Sweden, 1998

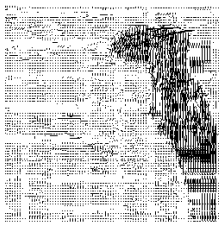

Figure 10.a. *Motion field obtained with the Baaziz method using hand # 2, frames 6 and 7.*

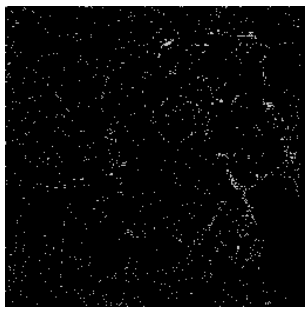

Figure 10.b. *Non-compensted pixel binary image using hand # 2 (0=compensated, 1=non-compensated).*

Pel-recursive algorithms based on linear estimation includes local context information so its motion fields are more immune to noise and quantitatively more accurate, but simple enough to compute. On those pixels in which the convergence criterion is not satisfied we had applied the Walker and Rao method and even if after this second step the pixel remains non-compensated (NC), this information will be useful during the segmentation process as a simple partial confidence measure of the motion estimation quality.

Figures 7a and 7b show frames 1 and 4 of the test sequence *hand # 1*. In figures 8a and 8b we present respectively the motion field and the non-compensated pixel image, obtained using frames 1 and 2 and the Baaziz method, the hand is moving from bottom to top. Figures 9a and 9b show frames 1 and 9 of the test sequence *hand #* 2, in which the hand is moving from top to bottom, in figures 10a and 10b we present respectively the motion field and the non-compensated pixel image, obtained using frames 6 and 7 and the Baaziz method.

6. CONCLUSIONS

The results obtained for segmenting and classifing hand gestures as a way for helping people with disabilities to communicate eassily with a computer are very encouraging, for this very simple dictionary of gestures a 100% successful has been obtained.

Our system is very simple consisting only in a computer and a video camera attached to it. However our major drawback is time processing, we are not able to operate in real-time. Comparing our work with others developped previously (Starner and Pentland, 1995), (Quek et al,), we have a more general system able to recognize more diverse hand gestures but not working in real-time jet.

Acknowledgments: This work was supported in part by ``Universidad Nacional Autónoma de México" (UNAM) and ``Consejo Nacional de Ciencia y Tecnología" (CONACyT).

7. REFERENCES

[1] A.N. Netravali and J.D. Robbins (1979), ``Motion-compensated television coding: part I," *The Bell System Technical Journal*, Vol. 58, No. 3, pp. 631-670, March.

[2] D. Gatica-Pérez, F. García-Ugalde and V. García-Garduño (1997), ``Segmentation algorithm for image sequences from a pel-recursive motion field." *Proc. of the VCIP, SPIE*, Vol. 3024, pp. 1152-1163.

[3] E. François (1991), *Interpretation qualitative du mouvement a partir d'une séquence d'images*, Ph. D. Thesis, Université de Rennes I, France, June.

[4] F. García-Ugalde, J. Savage-Carmona, T. A. Furness III, D. Gatica-Pérez and V. García-Garduño (1997), "Segmentation of moving human body parts by a modified MAP-MRF algorithm," *Proc. of the VSMM, IEEE,* pp. 197-205.

[5] F. K. H. Quek, T. Mysliwiec and M. Zhao. "Finger mouse: A freehand pointing interface." *Vision Interfaces and* Systems *Lab. (VISLab), The University of Illinois at Chicago.*

[6] I. Daubechies (1988), ``Orthonormal bases of compactly supported wavelets," *Comm. Pure and Applied Mathematics*, Vol. 41, pp. 909-996.

[7] J. Biemond, L. Looijenga, D.E. Boekee and R. Plompen (1987), ``A pel-recursive Wiener-based displacement estimation algorithm." *Signal Processing*, Vol. 13, No. 4, pp. 399-412, December.

[8] J.S. Lim (1990), *Two dimensional signal and image processing*, Englewood Cliffs, NJ: Prentice Hall.

[9] N. Baaziz (1991), *Approches d'estimation et de compensation de mouvement multiresolutions pour le codage de séquences d'images*, Ph. D. Thesis, Université de Rennes I, France, October.

[10] T. Starner and A. Pentland (1995), "Real-time american sign language recognition from video using hidden Markov models." *MIT Media Lab. Perceptual Computing Section Technical Report* No. 375.

[11] V. García-Garduño (1995), *Une approche de compression orientée-objets par suivi de segmentation basée mouvement pour le codage de séquences d'images numériques*, Ph. D. Thesis, Université de Rennes I, France, May.