Segmentation of moving human body parts by a modified MAP-MRF algorithm

F. García-Ugalde
T.A. Furness III

J. Savage-Carmona
D. Gatica-Pérez
V. García-Garduño

Human Interface Technology Laboratory
University of Washington, Box 352142
Seattle, WA 98195

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

Abstract

Using a dense motion vector field as the main information we develop a region segmentation algorithm in which each region is matched to a four-parameter motion model. Based on Markov Random Fields the segmentation model detects moving parts of the human body with different apparent displacement such as the hands. The motion vector field has been estimated by a Baaziz pel-recursive method 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). Results on segmenting and classifying real sequences are shown and based on a roughly defined directional dictionary one application is pursuing for using the segmented regions as commands for a virtual robot. The classification is based on the correlation coefficient (between the trained sequences and others) of wavelet coefficients, of the projected sum of the intensity of the segmentation field (in its binary version).

1 Introduction

The problem of segmenting moving regions was traditionally studied for coding video, in this work we are interested in a different application oriented to allow communication with a virtual robot using a very simple directional gesture dictionary. The idea behind this very simple dictionary is to permit us to communicate with the robot without cumbersome gloves and to consider this kind of communication as complementary of speech communication [13] into a multimodal interface. 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 [4], [6], [7], [11], 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 [1] 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) [2]. 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 0^0 and 90^0) of the segmentation field (in its binary version).

2 Pel-recursive motion estimation

The mainly use of pel-recursive displacement estimation since it was proposed by Netravali and Robbins [12], 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. [3] 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 by

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

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 $i-1$ iterations
- $\hat{\mathbf{d}}^{\mathbf{i}}$ is the final estimation for each pixel, (or after i iterations)
- DFD is the displaced-frame-difference (reconstruction error)

$$DFD(\mathbf{z}, \mathbf{d}) = i(\mathbf{z}, t) - i(\mathbf{z} - \mathbf{d}, t - k\Delta t) \quad (2)$$

- 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 $t - k\Delta t$

$$\begin{aligned} i_x^j &= i_x(\mathbf{z}_j - \hat{\mathbf{d}}^{\mathbf{i-1}}, t - k\Delta t) \\ i_y^j &= i_y(\mathbf{z}_j - \hat{\mathbf{d}}^{\mathbf{i-1}}, t - k\Delta t) \end{aligned} \quad (3)$$

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

Figure 1: Sequence *hand # 1.* (a) Frame 1. (b) Frame 4.

Pel-recursive algorithms based on linear estimation includes local context information so its motion fields

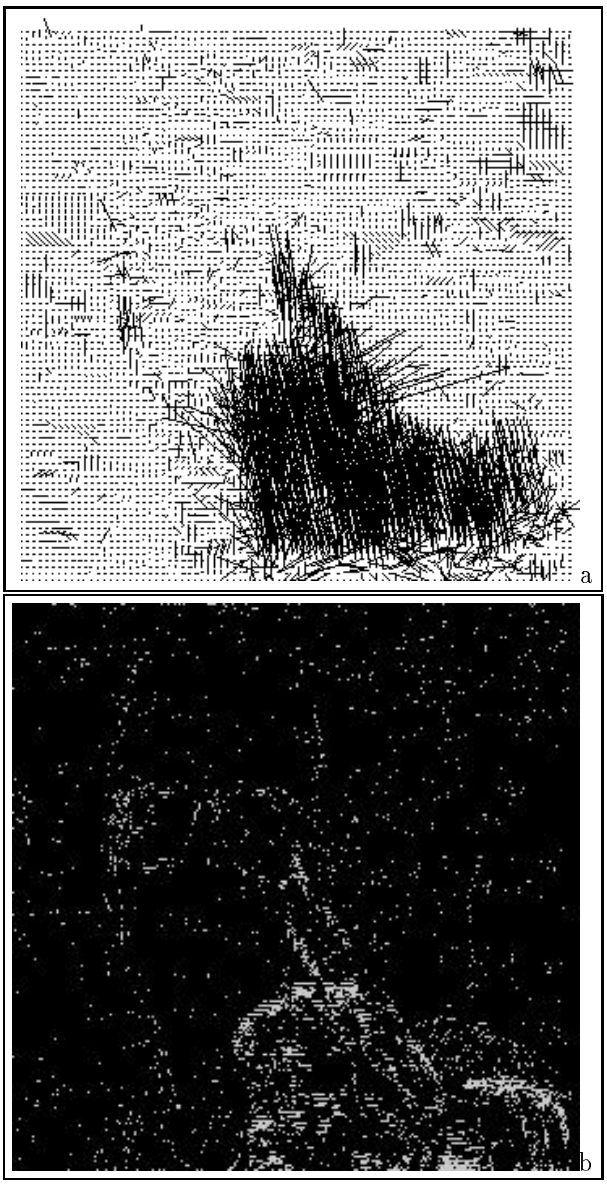

Figure 2: (a) Motion field obtained with the Baaziz method using *hand # 1*, frames 1 and 2. (b) Non-compensated pixel binary image (0 = compensated pixel, 1 = non-compensated pixel).

Figure 3: Sequence *hand # 2*. (a) Frame 1. (b) Frame 9.

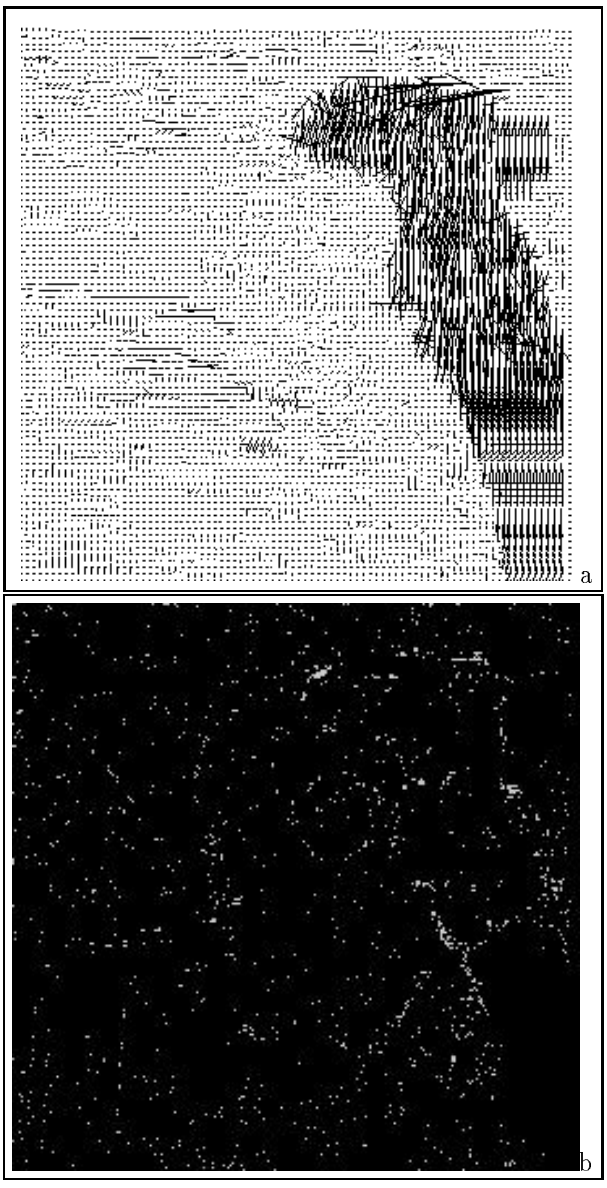

Figure 4: (a) Motion field obtained with the Baaziz method using *hand # 2*, frames 6 and 7. (b) Non-compensated pixel binary image (0 = compensated pixel, 1 = non-compensated pixel).

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 1a and 1b show frames 1 and 4 of the test sequence *hand # 1*. In figures 2a and 2b 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 3a and 3b show frames 1 and 9 of the test sequence *hand # 2*, in which the hand is moving from top to bottom, in figures 4a and 4b we present respectively the motion field and the non-compensated pixel image, obtained using frames 6 and 7 and the Baaziz method.

3 The segmentation algorithm

In presence of pure divergent motion, simple spatial clustering techniques for motion-based segmentation do not work well [7]. 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 [7]

$$\begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} + \begin{bmatrix} k & -\theta \\ \theta & k \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} \quad (4)$$

  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 5 we show the interaction model of observations, labels and physical assumptions.

Fields $\mathbf{e}$, $\mathbf{d}_x$, $\mathbf{d}_y$, $\mathbf{p}$ and $\mathbf{i}$ are defined over a lattice S of pixel sites s, and fields $\mathbf{l}$ and $\mathbf{g}$ are defined over a lattice S_l of line (interpixel) sites s_l (figure 6). We will assume that there are 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 :

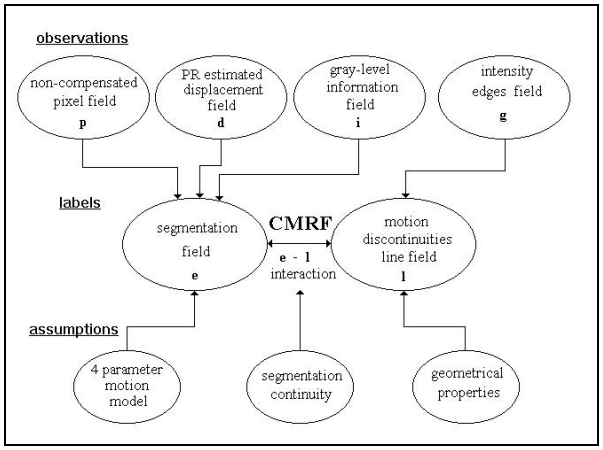

Figure 5: Interaction model for the motion-based segmentation algorithm.

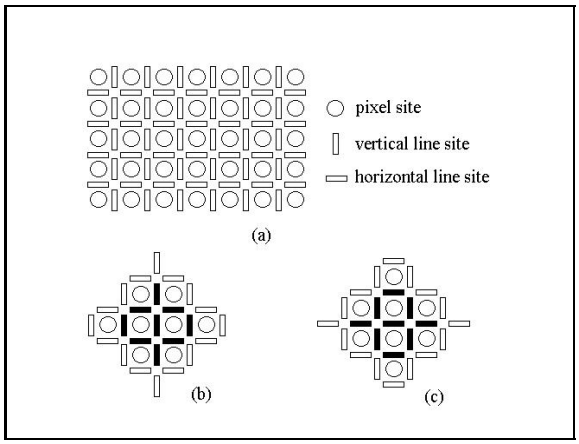

Figure 6: (a) Definition of pixel sites s and line sites s_l. MRF neighborhood system including pixel and line sites: (b) vertical; (c) horizontal.

$$(\hat{\mathbf{e}}, \hat{\mathbf{l}}) = argmax_{\mathbf{e},\mathbf{l}} \quad p(\mathbf{e}, \mathbf{l}/\mathbf{d}_x, \mathbf{d}_y, \mathbf{i}, \mathbf{p}, \mathbf{g}) \quad (5)$$

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

$$(\hat{\mathbf{e}}, \hat{\mathbf{l}}) = argmax_{\mathbf{e},\mathbf{l}} \quad p(\mathbf{d}_x, \mathbf{d}_y, \mathbf{i}/\mathbf{e}, \mathbf{l}, \mathbf{p}, \mathbf{g}) p(\mathbf{e}, \mathbf{l}/\mathbf{p}, \mathbf{g}) \quad (6)$$

In [8] we have shown that maximizing the a posteriori **pdf** is equivalent to minimize a so-called *energy function* $U(\mathbf{e}, \mathbf{l}, \mathbf{d}_x, \mathbf{d}_y, \mathbf{i}, \mathbf{p}, \mathbf{g})$ which has the form

$$U(\mathbf{e}, \mathbf{l}, \mathbf{d}_x, \mathbf{d}_y, \mathbf{i}, \mathbf{p}, \mathbf{g}) = \alpha U_{\mathbf{d}}(\mathbf{d}_x, \mathbf{d}_y, \mathbf{e}, \mathbf{p}) + \beta U_{\mathbf{i}}(\mathbf{i}, \mathbf{e}, \mathbf{p}) + \gamma U_{\mathbf{e}}(\mathbf{e}, \mathbf{l}) + \kappa U_{\mathbf{l}}(\mathbf{l}, \mathbf{g}) \quad (7)$$

where α, β, γ y κ are weighting terms, all these energy terms has been also defined in [8].

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 [6]. 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 [7]: one for the optimization of the segmentation field through minimizing

$$U_1 = \alpha U_{\mathbf{d}}(\mathbf{d}_x, \mathbf{d}_y, \mathbf{e}, \mathbf{p}) + \beta U_{\mathbf{i}}(\mathbf{i}, \mathbf{e}, \mathbf{p}) + \gamma U_{\mathbf{e}}(\mathbf{e}, \mathbf{l}) \quad (8)$$

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

$$U_2 = \gamma U_{\mathbf{e}}(\mathbf{e}, \mathbf{l}) + \kappa U_{\mathbf{l}}(\mathbf{l}, \mathbf{g}) \quad (9)$$

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

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 (Figure 7).

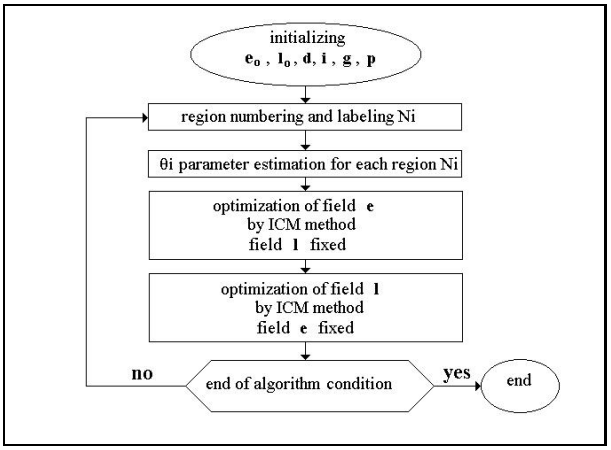

Figure 7: General diagram of the proposed motion-based segmentation algorithm.

4 Segmentation Results

Results obtained on the test sequences *hand # 1* and *hand # 2* for segmenting the moving parts are presented in figures 8 and 9 respectively. The segmentation fields obtained using the proposed algorithm are shown in figures 8a and 9a. A superposition of *hand # 1* and *hand # 2* with their respective segmentation can be seen in figures 8b and 9b. 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 (for each case: *hand # 1* and *hand # 2*) of the segmentation algorithm, the tiny regions remaining in the background could be fused in posterior iterations.

No further processing has been done on the segmentation frontiers. The motion vector fields obtained with the Baaziz method, figures 2a and 4a, are somewhat homogeneous and properly adjusted to the moving areas. The non-compensated pixels shown in figures 2b and 4b, 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. When the segmentation of the *hands* has been completed, we have defined a very simple dictionary of movements for commanding a virtual robot in a two dimensional space. In a first phase this dictionary will permit the displacement of the virtual robot depending on the directional movement of the *hands* and in a second phase the interpretation of some signs or gestures of the *hands* will permit to realize more complicated actions.

5 Gesture Classification

For classifying the segmented fields $\mathbf{e}$ of *hand # 1* and *hand # 2* we have proceed first to obtain a binary version of these fields comparing them with a threshold, so rendering the classification faster. Then based on the Radon transform [9]

$$p(t) = \int_{u=-\infty}^{\infty} \mathbf{e}(t_1, t_2) \mid_{t_1 = t\cos\theta - u\sin\theta, t_2 = t\sin\theta + u\cos\theta} du \quad (10)$$

a computation of the projected sum of the intensity has been done at 0^0 and 90^0 and are shown in figures 10 and 13 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

Figure 8: MAP motion-based segmentation algorithm. (a) Segmentation field **e** of *hand # 1*. (b) Superposition of frame 3 of *hand # 1* and segmentation field **e**.

Figure 9: MAP motion-based segmentation algorithm. (a) Segmentation field **e** of *hand # 2*. (b) Superposition of frame 7 of *hand # 2* and segmentation field **e**.

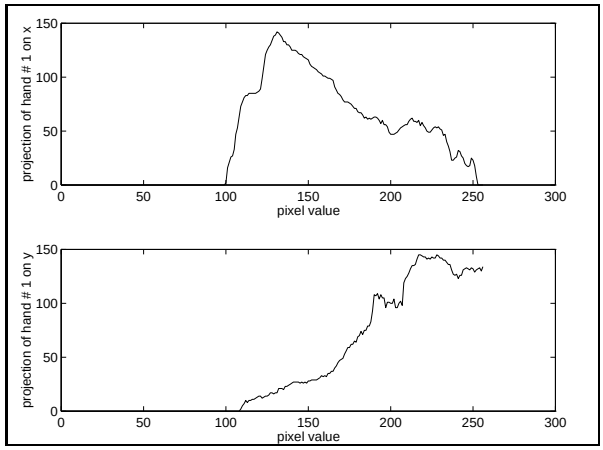

Figure 10: Projected sum of the intensity of the binary version of the segmentation field **e** of *hand # 1* at 0^0 and 90^0 respectively.

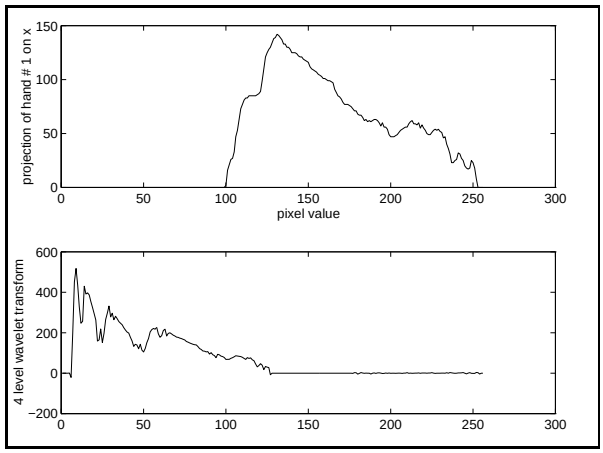

Figure 11: Projected sum of the intensity of the binary version of the segmentation field **e** of *hand # 1* at 0^0 and its 4 level wavelet transform respectively.

the intensity for each orientation (0^0 and 90^0) has been done using Daubechies filters: d4 [5]. The wavelet transform of $f(t)$ at scale 2^i is given by

$$WF_i(t) = f * \psi_i(t) = \int_{-\infty}^{\infty} f(\tau)\psi_i(t-\tau)d\tau \qquad (11)$$

Where from Mallat and Zhong [10] $\psi(t)$ is a wavelet function with average $\int \psi(t)dt = 0$ and $\psi_i(t) = 1/2^i\psi(t/2^i)$ is its dilation by a factor 2^i. The dyadic wavelet transform is the sequence of functions $WF[f()] = [WF_i(t)]icZ$. Figures 11 and 12 show the four level wavelet decomposition for *hand # 1* at 0^0 and 90^0 respectively. And figures 14 and 15 show the four level wavelet decomposition for *hand # 2* at 0^0 and 90^0 respectively.

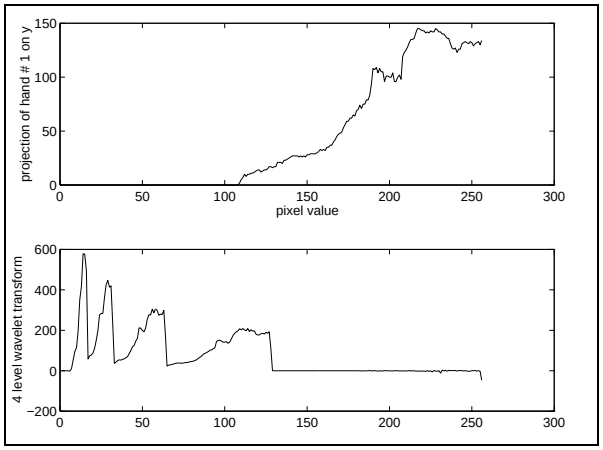

Figure 12: Projected sum of the intensity of the binary version of the segmentation field **e** of *hand # 1* at 90^0 and its 4 level wavelet transform respectively.

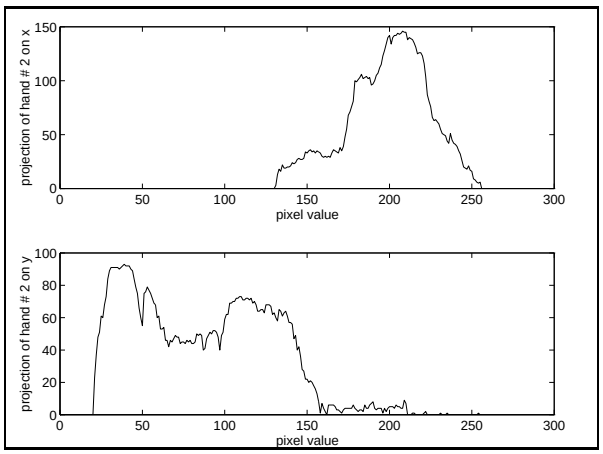

Figure 13: Projected sum of the intensity of the binary version of the segmentation field **e** of *hand # 2* at 0^0 and 90^0 respectively.

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.

6 Conclusions

For this particular application and even in this preliminary phase of the project the results of motion estimation, segmentation and classification of the segmented fields are very encouraging. The segmentation of moving regions has been well realized with this MAP-MRF algorithm. The classification of *hands* gestures are correctly and quickly realized with the combination of projected sum of intensities at 0^0 and 90^0,

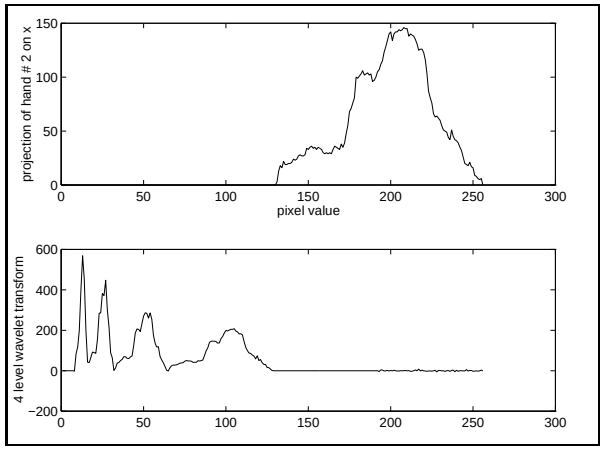

Figure 14: Projected sum of the intensity of the binary version of the segmentation field **e** of *hand # 2* at 0^0 and its 4 level wavelet transform respectively.

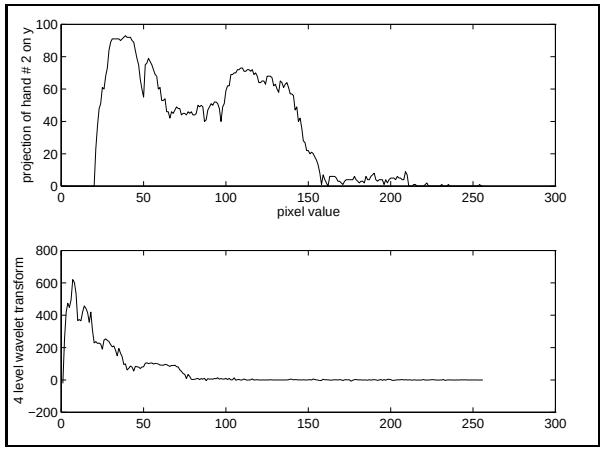

Figure 15: Projected sum of the intensity of the binary version of the segmentation field **e** of *hand # 2* at 90^0 and its 4 level wavelet transform respectively.

computing then the wavelet decomposition of these projections and finally the correlation coefficient between the wavelet coefficients corresponding to *trained sequences* and others. For this very simple dictionary we have obtained a 100 % successful hand gesture recognition system. Currently now we are integrating this processing to a general multimodal interface, in which others sources of information as speech will be taking into account for communicating with a virtual robot without cumbersome gloves.

Acknowledgments

This work was supported in part by "Universidad Nacional Autónoma de México" (UNAM) and "Consejo Nacional de Ciencia y Tecnología" (CONACyT) and realized during the sabbatical leave of the first author from 8/1/96 to 7/31/97 at Human Interface Technology Laboratory, University of Washington.

References

[1] N. Baaziz. *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 1991.

[2] J. Besag. "On the statistical analysis of dirty pictures." *Journal of the Royal Statistical Society B*, Vol. 48, No. 3, pp. 259-302, 1986.

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

[4] M. M. Chang, A. M. Tekalp and M. I. Sezan. "Motion-field segmentation using an adaptive MAP criterion," *Proc. of the ICASSP*, Vol. 5, pp. 33-36, 1993.

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

[6] E. Francois. *Interpretation qualitative du mouvement a partir d'une séquence d'images,* Ph. D. Thesis, Universite de Rennes I, France, June 1991.

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

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

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

[10] S. Mallat and S. Zhong. "Characterization of signals from multiscale edges," *Trans. Pattern Anal. Machine Intell.*, Vol. PAMI-14, No. 7, pp. 710-732, July 1992.

[11] D.W. Murray, and B.F. Buxton. "Scene segmentation from visual motion using global optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. PAMI-9, No. 2, pp. 220-228, March 1987.

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

[13] J. Savage-Carmona. *A hybrid system with symbolic AI and statistical methods for speech recognition,* Ph. D. Dissertation, University of Washington, July 1995.