EURASIP Journal on Applied Signal Processing 2004:6, 861–870

An Algorithm for Motion Parameter Direct Estimate

Roberto Caldelli
Dipartimento di Elettronica e Telecomunicazioni, Università di Firenze, Via S. Marta 3, 50139 Firenze, Italy
Email: caldelli@lci.det.unifi.it

Franco Bartolini
Dipartimento di Elettronica e Telecomunicazioni, Università di Firenze, Via S. Marta 3, 50139 Firenze, Italy
Email: barto@lci.det.unifi.it

Vittorio Romagnoli
Dipartimento di Elettronica e Telecomunicazioni, Università di Firenze, Via S. Marta 3, 50139 Firenze, Italy
Email: romagnoli@lci.det.unifi.it

Received 29 January 2003; Revised 5 September 2003

Motion estimation in image sequences is undoubtedly one of the most studied research fields, given that motion estimation is a basic tool for disparate applications, ranging from video coding to pattern recognition. In this paper a new methodology which, by minimizing a specific potential function, directly determines for each image pixel the motion parameters of the object the pixel belongs to is presented. The approach is based on Markov random fields modelling, acting on a first-order neighborhood of each point and on a simple motion model that accounts for rotations and translations. Experimental results both on synthetic (noiseless and noisy) and real world sequences have been carried out and they demonstrate the good performance of the adopted technique. Furthermore a quantitative and qualitative comparison with other well-known approaches has confirmed the goodness of the proposed methodology.

Keywords and phrases: motion parameter estimation, MAP criterion, Markov random fields, iterated conditional mode, motion models.

1. INTRODUCTION

Estimation of motion fields and their segmentation are still an important task to be solved; in disparate applications ranging from pattern recognition to image sequence analysis, passing through object tracking and video coding, determining trajectories and positions of objects composing the scene is mandatory, and much effort has been spent in researching and devising a robust solution to adequately and satisfactory address this problem. Though for human visual system (HVS), motion recognition is effortless, the same thing cannot be assessed for computer-aided estimation. This is mainly due to the complex relationship existing between the movements of objects in a 3D scene and the apparent motion of brightness pattern in a sequence of 2D projections of the scene. Information about depth is lost and what appears as motion in the image plane can actually be determined by other phenomena, such as changes in scene illumination and shadowing effects. Furthermore, motion recognition is also hard to obtain because of some application hurdles, as the aperture problem [1] and regions occlusion; and although many algorithms and valuable approaches have been developed, this issue cannot be considered as completely investigated yet [2, 3, 4].

Different are the approaches to motion estimation task. One of the most well-known consists of representing motion fields by assigning independent motion vectors to each image pixel (*dense motion fields*) [5, 6]. Velocity vector estimate is generally performed by searching for the vector field, minimizing a predefined functional. As proposed in the basic paper by Horn and Schunck [1], this functional is composed by two contributions, the former weighs for the deviation from constancy of brightness intensity and the latter is used to impose a smoothness binding due to spatial correlation; the field which minimizes the functional is assumed to be the solution. Other techniques also impose the smoothness constraint in order to obtain an additional relationship to solve the underconstrained optic flow problem [7, 8]. In [9] the regularization of the velocity field, determined by a primary coarse least squares (LS) estimation, is achieved through a weighted vector median filtering operation. Motion estimation can also be performed through a Bayesian approach [6, 10] in which an inference framework is adopted to calculate the probability of a motion hypothesis given image data. In literature, some other algorithms use parametric motion models (e.g., [11]) to represent transformations by modelling relations between two successive images; in particular, the motion of a specific region is determined through an adopted model that, depending on its complexity, will be described by a different number of parameters (e.g., six parameters for affine motion model, eight parameters for perspective projection model [12]).

In this paper an algorithm which, by using a parametric motion model, deals with the direct estimation of model parameters is presented. This is the main characteristic of the proposed method, distinguishing it from other common approaches, that first estimate motion vectors and then evaluate motion parameters fitting the estimated vectors. Such a two-step approach poses problems from the point of view of segmentation, that should precede vectors aggregation, but should also benefit from knowledge of motion parameters. On the contrary, our technique directly obtains, for each image pixel, a parameter set describing the motion of the object the pixel belongs to; this information can then be successfully used for motion-based segmentation. Starting from two frames of an image sequence, the parameters describing the adopted motion model are computed for each image pixel through an iterative minimization of an ad hoc functional. The extracted motion parameters can be used for many higher-level analysis tasks beyond the already mentioned motion-based object segmentation, as for example, for reducing the motion description burden in coding operation (video coding), for describing the behavior of moving objects (event detection), for estimating the 3D structure of the surrounding world, and so on.

The remainder of this paper is organized as follows. In Section 2 the adopted motion model is introduced, and in Section 3 some theoretical arguments, which are important for work understanding, are discussed; in Section 4 the choice of the to be minimized functional is motivated and in Section 5 some experimental results both on synthetic and on real sequences are presented, finally Section 6 draws the conclusions.

2. CHOICE OF THE MOTION MODEL

Parametric motion models are introduced in many video processing applications. In most of these, they are used to efficiently analyse the moving objects that are present in a sequence. Motion can be described by adopting different models (translational, affine, projective linear, and so on) which have at their disposal a diverse number of parameters (degrees of freedom (DOF)); the greater this number the more complex the motion that can be represented. In this application, attention has been focused on the *affine* model which can be described as

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}, \tag{1}$$

where the parameters a, b, c, d, e, and f represent the 6 DOF, x and y are the coordinates of pixel initial position, and dx and dy are the components of its spatial displacement. In particular, the parameters e and f also take into account transformations (e.g., scaling and rotation) occurring with respect to a point (x_c, y_c) different from the image center, and their expressions are reported as follows:

$$\begin{aligned} e &= dx_0 - a \cdot x_c - b \cdot y_c, \\ f &= dy_0 - c \cdot x_c - d \cdot y_c, \end{aligned} \tag{2}$$

where dx_0 and dy_0 are, respectively, the initial horizontal and vertical displacement of the object with respect to the image center. With this model, transformations such as translations, rotations, and anisotropic scaling can be represented; geometric manipulations like projections (8 DOF) are not contemplated.

To reduce the computational burden, it has been decided to concentrate solely on the case of roto translations, so the model is simplified and is based just on three parameters; (1) can be rewritten as

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \cos\theta - 1 & -\sin\theta \\ \sin\theta & \cos\theta - 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}. \tag{3}$$

The terms in the matrix in (1) are not independent anymore and the motion analysis will be demanded only to estimate the parameters θ, e, and f. The parameter θ takes into account rotations, and, as stated before, the parameters e and f include both the translational motion component (respectively, horizontal and vertical) and the rotation with respect to a point different from the image center. For the sake of clarity, in the following, a reference system centered in the middle of the image with x-axis directed to right and y-axis directed to top will be assumed. Moreover a clockwise rotation will be considered as negative (these issues are important to adequately understand the experimental results presented in Section 5).

3. MARKOV RANDOM FIELDS AND MAP ESTIMATION

Markov random fields (MRF) are often used in many image processing applications like motion detection and estimation. By simply making a direct multidimensional extension of a 1D Markov process, the definition of an MRF can be derived [13], here after the main characteristics of MRFs are outlined.

Let Λ be a sampling grid in R^N, $\eta(\mathbf{n})$ is a neighborhood of $\mathbf{n} \in \Lambda$, such that $\mathbf{n} \notin \eta(\mathbf{n})$ and $\mathbf{n} \in \eta(\mathbf{l}) \Leftrightarrow \mathbf{l} \in \eta(\mathbf{n})$. For example, a first-order bidimensional neighborhood consists of the closest *top, bottom, left*, and *right* neighbors of $\mathbf{n}$ (see Figure 1).

Let Π be a neighborhood system, that is, a collection of neighborhoods of all $\mathbf{n} \in \Lambda$; a random field Υ over Λ is a multidimensional random process such that each site $\mathbf{n} \in \Lambda$ is assigned a random variable whose $\nu \in \Gamma$ is an occurrence.

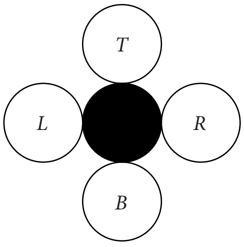


Figure 1: First-order bidimensional neighborhood.

A random field Υ with the following properties:

$$\begin{aligned} &P(\Upsilon = \nu) > 0, \quad \forall \nu \in \Gamma, \\ &P(\Upsilon_n = \nu_n \mid \Upsilon_l = \nu_l,\ \forall \mathbf{l} \neq \mathbf{n}) \\ &\quad = P(\Upsilon_n = \nu_n \mid \Upsilon_l = \nu_l,\ \forall \mathbf{l} \in \eta(\mathbf{n})), \\ &\qquad \forall \mathbf{n} \in \Lambda,\ \forall \nu \in \Gamma, \end{aligned} \tag{4}$$

where P is a probability measure, is called an MRF with state space Γ. Roughly speaking, in (4) it is asserted that the probability that the field assumes a certain value ν_n in the location $\mathbf{n}$, depending on all the other elements of the field, is the same probability of getting that value, depending only on the elements belonging to $\eta(\mathbf{n})$. To exploit MRFs characteristics in a practical way, we need to refer to the *Hammersley-Clifford theorem* which allows to set a relationship between MRFs and *Gibbs distributions*, by linking MRFs properties to distribution parameters by means of a potential function V. This theorem states that Υ is an MRF on Λ with respect to Π if and only if its probability distribution is a Gibbs distribution with respect to Λ and Π. A Gibbs distribution, with respect to Λ and Π, is a probability measure φ on Γ such that

$$\varphi(\nu) = \frac{1}{Z} e^{-U(\nu)/T}, \tag{5}$$

where the constants Z and T are called the *partition function* and *temperature*, respectively, and the *energy function* U is of the form

$$U(\nu) = \sum_{c \in C} V(\nu, c). \tag{6}$$

The term $V(\nu, c)$ is called *potential function* and depends only on the value of ν at sites that belong to the clique c. With clique c is intended a subset of Λ, defined over Λ with respect to Π, such that either c consists of a single site or every pair of sites in c are neighbors, according to η. The set of all cliques is denoted by C. Examples of two-element spatial cliques $\{\mathbf{n}, \mathbf{l}\}$ with respect to the first-order neighborhood of Figure 1 are two immediate horizontal and vertical neighbors.

3.1. MAP criterion

In order to estimate an unknown MRF realization, based on some observations, the maximum a posteriori probability (MAP) criterion is often used. In the sequel, the MAP approach is briefly described.

Let Y be a random field of observations and let Υ be a random field that it has to be estimated based on Y. Let y, ν be their respective realizations. For example, y could be the difference between two images, while ν could be a field of motion detection labels. In order to compute ν based on y, the MAP criterion can be used as follows:

$$\begin{aligned} \hat{\nu} &= \arg\max_{\nu} P(\Upsilon = \nu \mid y) \\ &= \arg\max_{\nu} \frac{P(\Upsilon = y \mid \nu) P(\Upsilon = \nu)}{P(Y = y)}, \end{aligned} \tag{7}$$

where $\max_\nu P(\Upsilon = \nu \mid y)$ denotes the MAP $P(\Upsilon = \nu \mid y)$ with respect to ν and arg denotes the argument $\hat{\nu}$ of this maximum such that $P(\Upsilon = \hat{\nu} \mid y) \geq P(\Upsilon = \nu \mid y)$ for any ν. In (7), by applying Bayes theorem, the final expression can be derived; moreover (7) can be simplified by not considering $P(Y = y)$ because it does not depend on ν.

4. THE POTENTIAL FUNCTION

According to (7) and just reporting this general case to the case of motion parameter estimate in an image sequence, the best-fitting parameter set for each point $(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})_{\text{opt}}$ can be obtained based on the MAP criterion. This is made evident in (8) where $(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})$ is the parameter set realization of the random field $(\boldsymbol{\Theta}, \mathbf{E}, \mathbf{F})$ and g_{t+dt} is the image at time $t + dt$ (realization of G_{t+dt}) and g_t is the image at time t:

$$\begin{aligned} &(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})_{\text{opt}} \\ &\quad = \arg\max_{(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})} P((\boldsymbol{\Theta}, \mathbf{E}, \mathbf{F}) = (\boldsymbol{\theta}, \mathbf{e}, \mathbf{f}) \mid G_{t+dt} = g_{t+dt};\ G_t = g_t). \end{aligned} \tag{8}$$

The expression to be maximized can be rewritten, also in this case, as

$$\begin{aligned} &P((\boldsymbol{\Theta}, \mathbf{E}, \mathbf{F}) = (\boldsymbol{\theta}, \mathbf{e}, \mathbf{f}) \mid G_{t+dt} = g_{t+dt};\ G_t = g_t) \\ &\quad = P(G_{t+dt} = g_{t+dt} \mid (\boldsymbol{\Theta}, \mathbf{E}, \mathbf{F}) = (\boldsymbol{\theta}, \mathbf{e}, \mathbf{f});\ G_t = g_t) \\ &\qquad \cdot P((\boldsymbol{\Theta}, \mathbf{E}, \mathbf{F}) = (\boldsymbol{\theta}, \mathbf{e}, \mathbf{f});\ G_t = g_t). \end{aligned} \tag{9}$$

The two terms of the product, in the right member, represent, respectively, two contributions: the first one accounts for the probability to have the image g_{t+dt} given the parameter values $(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})$ and the previous image g_t, the second one accounts for the a priori *probability* by considering all the information available about the field $(\boldsymbol{\Theta}, \mathbf{E}, \mathbf{F})$ and the image G_t.

In the light of this consideration, this maximization has been achieved by defining a potential function W_{TOT}, itself composed by two terms and directly depending on the motion parameters, in such a way that the optimal set will be chosen in correspondence of the minimum of this potential function,

$$\begin{aligned} (\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})_{\text{opt}} &= \arg\min_{(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})} W_{\text{TOT}} \\ &= \arg\min_{(\boldsymbol{\theta}, \mathbf{e}, \mathbf{f})} \sum_{(x,y) \in \Re} W_{(x,y)}, \end{aligned} \tag{10}$$

where $\Re$ represents the whole image. The assumption to deal with MRFs [13] permits to consider the motion of a generic point as depending on the motion of the other points belonging to its neighborhood. In the proposed approach for each pixel (x,y), only its four neighbors of first order (T, B, R, and L) (this set will be indicated with the notation $N_{(x,y)}$) have been deemed as relevant. The potential $W_{(x,y)}$ can be expressed as evidenced in (11) to better highlight the meaning of its composing terms:

$$W_{(x,y)} = \alpha \cdot A_{(x,y)} + B_{(x,y)}. \tag{11}$$

The term $A_{(x,y)}$ is defined as

$$A_{(x,y)} = \left| G_t(x,y) - G_{t+dt}(x+dx, y+dy) \right| \tag{12}$$

and it takes into account the goodness of matching between the brightness $G_t(x,y)$ of the pixel (x,y) at time t and the corresponding brightness $G_{t+dt}(x+dx, y+dy)$ in the successive frame in the location $(x+dx, y+dy)$; if dx and dy have been correctly estimated, the value of $A_{(x,y)}$ will be very low. On the other side, the term $B_{(x,y)}$ gives a contribution to the potential function from the point of view of motion field smoothness (see (13))

$$B_{(x,y)} = \sum_{(\tilde{x},\tilde{y}) \in N_{(x,y)}} V_c\big((x,y),(\tilde{x},\tilde{y})\big),$$
$$V_c\big((x,y),(\tilde{x},\tilde{y})\big) = \begin{cases} 0 & \text{if } (\theta,e,f)_{(x,y)} = (\theta,e,f)_{(\tilde{x},\tilde{y})}, \\ \gamma & \text{otherwise,} \end{cases} \tag{13}$$

with γ >0. $B_{(x,y)}$ will be low if the parameters under judgement are homogeneous with their neighbors. Lastly, in the definition of the potential function W_{TOT}, there is the factor α which allows to balance the two effects, frame matching and field smoothness. During the optimal parameter search, from a computational point of view, to exhaustively test all the possible values for each pixel results to be prohibitive. Therefore a deterministic relaxation is adopted to obtain a succession of estimated fields, bringing in a suboptimal solution but with reduced convergence time. The method used to sequentially visit all the points of the image and to update their values is the iterated conditional mode (ICM) [14, 15, 16]. At this point, we analyze in detail how the computing and the updating of the potential take place. We suppose that this computing and updating be on the generic point (x,y) which has got the parameter set $(\theta_t, e_t, f_t)_{(x,y)}$, and we test the candidate parameters $(\theta_c, e_c, f_c)_{(x,y)}$ by calculating $W_{(x,y)}$ (the new potential value on the considered point) and the four values $W_{(\tilde{x},\tilde{y})}$, for all $(\tilde{x},\tilde{y}) \in N_{(x,y)}$ (potentials of the four points near to (x,y)); these last ones are checked because albeit only the parameter set referred to (x,y) is modified, also the $B_{(\tilde{x},\tilde{y})}$ terms are affected. The so far best fitting set $(\theta_t, e_t, f_t)_{(x,y)}$ will be substituted by the candidate set $(\theta_c, e_c, f_c)_{(x,y)}$ if the relation expressed in (14) is verified:

$$\left(W_{(x,y)} + \sum_{(\tilde{x},\tilde{y}) \in N_{(x,y)}} W_{(\tilde{x},\tilde{y})} \right)_{(\theta_c,e_c,f_c)_{(x,y)}} < \left(W_{(x,y)} + \sum_{(\tilde{x},\tilde{y}) \in N_{(x,y)}} W_{(\tilde{x},\tilde{y})} \right)_{(\theta_t,e_t,f_t)_{(x,y)}}, \tag{14}$$

otherwise the set $(\theta_c, e_c, f_c)_{(x,y)}$ will be rejected. The parameter 3D space has to be investigated, and by depending on the parameter search step, the computational complexity will be differently onerous. Finally the optimum set, which minimizes the addition of the five potentials, related to the point and to $N_{(x,y)}$, will be obtained. The parameter field gets stable after 7–8 complete iterations, and variations are not recorded anymore.

4.1. The macropixel approach

One of the crucial problems in dealing with dense fields is to obtain homogeneous motion regions; ideally the proposed estimation approach should yield to the recognition of rigid moving objects characterized by the same motion parameters, but this does not happen because a specific motion, in some particular object areas, could be adequately represented, for example, by a uniform rotation or by a smoothly variable translation, without any relevant difference in the potential function evaluation. To avoid this, a multiresolution approach can be used; blocks of pixels (named *macropixel*), forming a 4 × 4 or 2 × 2 window, are constrained to move with the same parameters, thus resulting in a superior motion field homogeneity. On the other side, loss of resolution is a drawback from moving object detection point of view, in fact the boundaries of these could appear enlarged with respect to their real size. A good tradeoff between these two aspects has been achieved by adopting the macropixel arrangement (macropixel size has been set to 2 × 2) just for the first two or three iterations, then resolution is augmented again to the single pixel level; doing so a primary raw estimation is obtained which is successively refined in the subsequent steps.

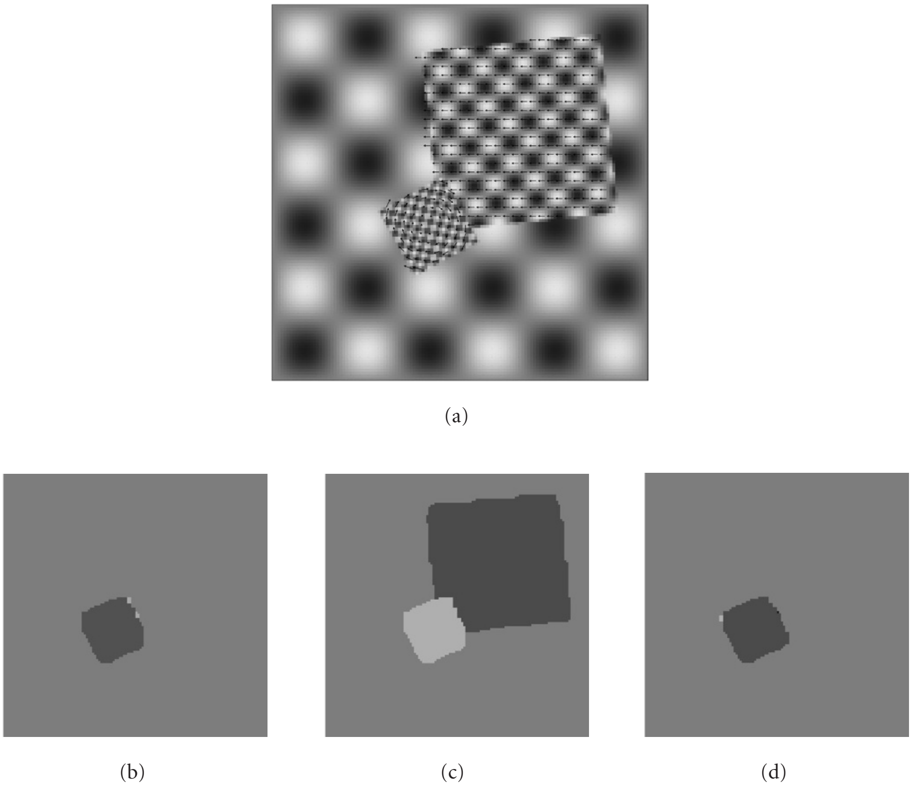
5. EXPERIMENTAL RESULTS

The proposed approach has been tested both on synthetic sequences, with and without added noise, and on real world sequences; and some experimental results confirming the good performance of the method are presented in this section.

5.1. Testing on synthetic sequences

In the synthetic sequence (see Figure 2a), there are two textured squares of different size moving on a slightly textured background. The big square has got only a translational motion towards left direction by 1 pel/frame and the small one rotates clockwise around its center by 5 deg/frame.

In Figure 2b the estimated values of the parameter θ are depicted; it can be noted that the rotating square is exactly and homogeneously recognized (dark gray states for


(a)

(b) (c) (d)

Figure 2: Synthetic sequence: (a) a frame with the superimposed ideal motion vector field, (b) the estimated motion parameters θ, (c) e, and (d) f.

negative values, clear gray for positive); contributions on the big square, that has no rotational components, have not been rightly revealed. On the contrary, the big square horizontal motion is correctly detected through the parameter e as illustrated in Figure 2c; in this picture and also in Figure 2d, for the parameter f, it appears that the values over the small square are not zero although its motion has not any translational component: these are due to the fact that this object rotates around a point which is not the center of the image and this gives origin to two translational components in the model, as described in (2). In Table 1 the mean absolute error (MAE) between the true displacements and the estimated ones, computed both through the proposed method and through the well-known Horn and Schunck (H&S) technique [1], is proposed. This algorithm has been running with the parameter that balances the two-component terms in the functional set at 1 and the number of iterations set at 128 (this has been maintained also for real world sequences). Errors have been computed on the whole image, in the interior and on the boundaries of the moving objects; two cases, perfect data and data with noise addition (Gaussian noise with $\sigma^2 = 20$), have been taken into account. Errors related to the proposed method are widely lower than those obtained with the H&S method, especially in the interior of the moving objects, thanks to the adoption of the model-based approach.

Table 1: MAE between ideal displacements and estimates computed through the proposed and H&S methods with perfect and noisy ($\sigma^2 = 20$) data.

		MAE		
		Overall	Interior	Contours
Perfect data	Proposed	0.029	0.001	0.251
	H&S	0.058	0.024	0.324
Noisy data	Proposed	0.042	0.003	0.346
	H&S	0.156	0.134	0.329

5.2. *Testing on real sequences*

In this subsection experimental tests carried out on three different real world sequences are proposed.

5.2.1. *Carphone*

The first sequence examined is *Carphone*. The same frames (QCIF format), numbers 168 and 171, considered in [17] have been processed to make a possible comparison with some numerical results presented in that paper.

In Figure 3a, the estimated motion vector field has been superimposed to the frame 171; the vectors over the head of the man and over his left shoulder are quite accurate, but regions that are visible through the car window, on the right

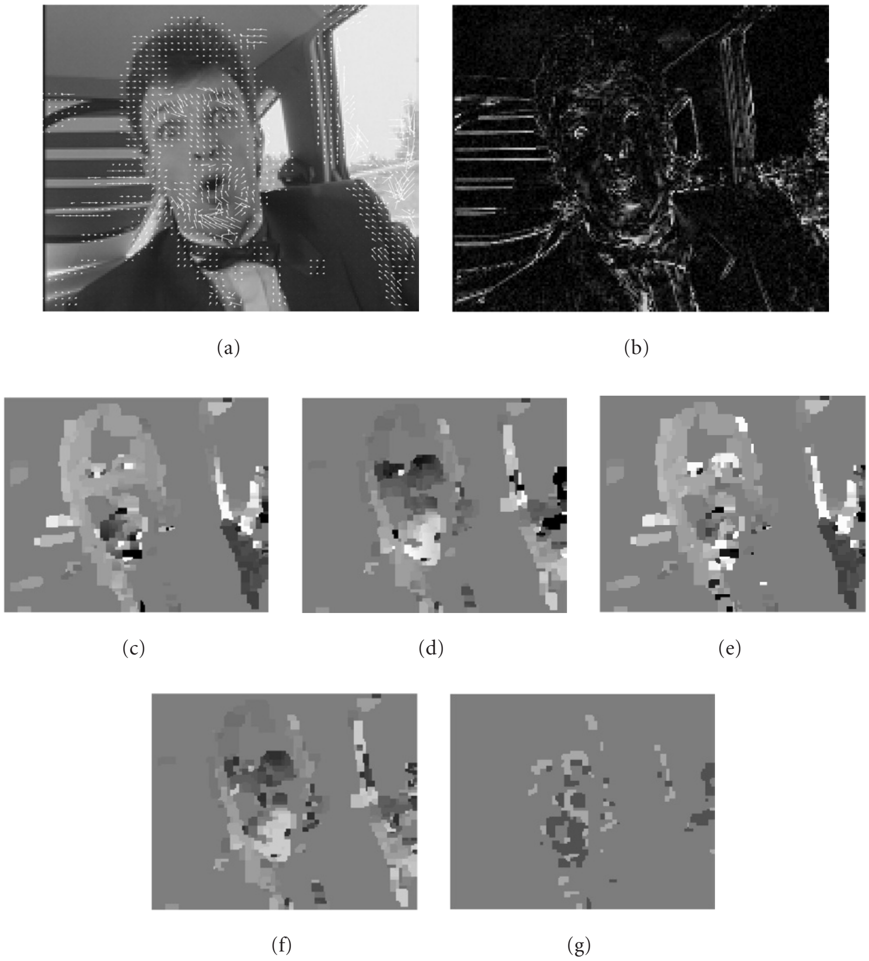


Figure 3: Real world sequence (Carphone): (a) frame 171 with the superimposed motion field estimated through the proposed method; (b) pixel-per-pixel squared difference between frame 171 and its motion compensated version; estimates obtained by means of the proposed method: the displacements (c) dx and (d) dy, the motion parameters (e) e, (f) f, and (g) θ.

side of the image and near the chin of the man, contain some wrong nonhomogeneous vectors. In particular, the errors visible on the objects at the right extreme of the window are due to the fact that these objects were not present in the previous frame, thus confusing motion estimation. On the other side, the few not well-estimated vectors on the chin correspond to uniform grey-level regions of the face, where local motion estimation algorithms often encounter problems. In Figure 3b a pixel-per-pixel squared difference between frame 171 and its motion compensated version is depicted. A clear gray level means a high discrepancy between the two images; also in this picture significant errors are confirmed in the same areas as before. To better evaluate the obtained results, in Table 2 the value of prediction error (PE), computed with the proposed method, is compared to the data provided in [17], regarding the same sequence, and to H&S technique [1]: the proposed method performs better with respect to the other kind of approaches. In Figures 3c and 3d the computed displacements (dx and dy) are also depicted. Finally, in Figures 3e, 3f, and 3g the motion parameters, respectively, representing the horizontal and vertical translation, and the rotation, are presented. In particular, by observing Figure 3e, it can easily be noticed that the left-side movement of the left shoulder of the man is correctly recognized by the dark (negative) homogeneous region. The same shoulder has also a light up-side motion as evidenced by the bright region in Figure 3f in that location. The rotation parameter θ is zero almost everywhere, with the exception of some zones in

Table 2: PE for Carphone sequence (higher value means a better prediction). The results for the first three methods are taken from [17].

Kind of adopted approach	PE
Block-based prediction [17]	31.8 dB
Pixel-based prediction [17]	35.9 dB
Region-based prediction [17]	35.4 dB
Horn&Schunck	30.4 dB
Proposed method	36.7 dB

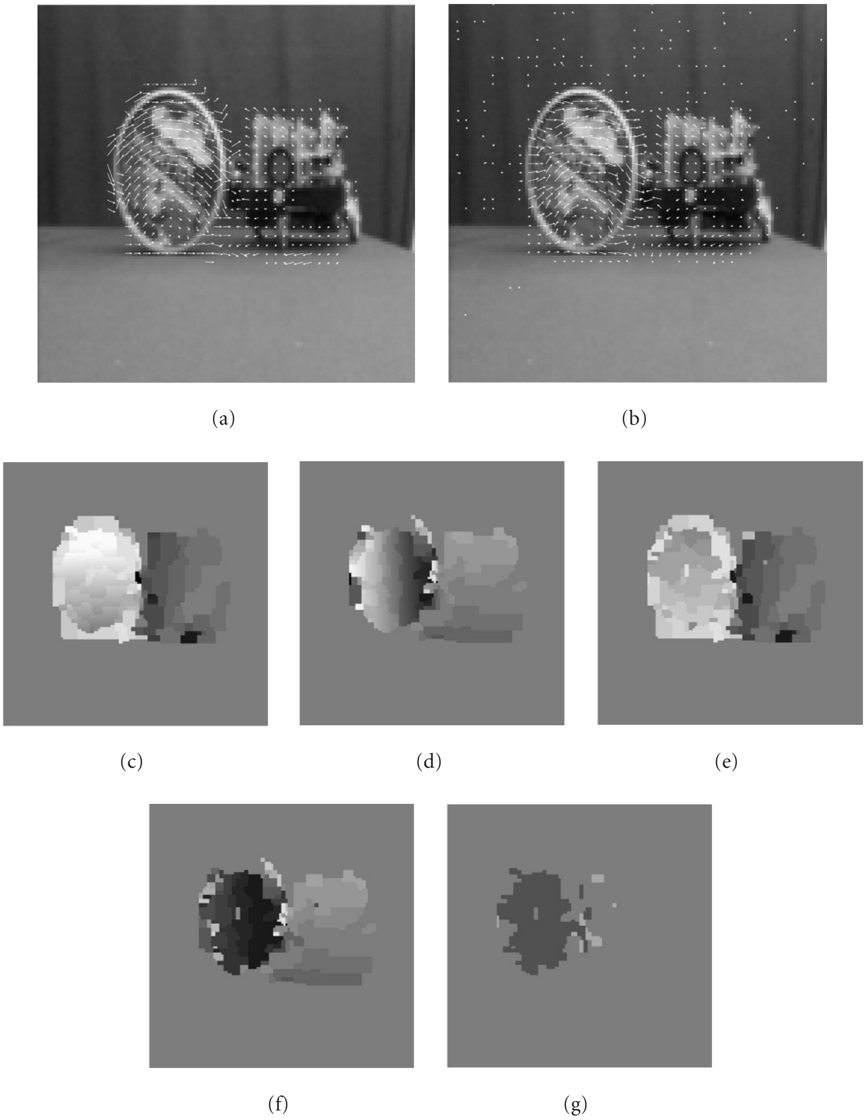


Figure 4: Real world sequence (Robox): frame 15 with the superimposed motion field estimated through (a) the proposed method and (b) the H&S approach; estimates by means of the proposed method: the displacements (c) dx and (d) dy, and the motion parameters (e) e, (f) f, and (g) θ.

correspondence of the mouth and of the nose where motion is quite complex, and small rotational components are detected by the algorithm.

5.2.2. *Robox*

Experimental tests carried out with sequence named *Robox* are illustrated in Figure 4 and discussed in the sequel; frames taken into consideration are numbers 15 and 17. This sequence is composed by two moving objects: a round box which rotates clockwise over a table and a small robot moving towards the camera. In Figures 4a and 4b, frame 15 of the sequence with the motion field superimposed, computed, respectively, by means of the proposed method and through the H&S technique, is pictured. It can be easily noted how the motion field is more properly and precisely detected in Figure 4a with respect to the other methodology, in particular, for the rotating object.

In Figures 4c and 4d, the displacements dx and dy estimated by means of the proposed technique are presented; it is interesting to highlight that the box, which rotates around its contact point with the table, has dx's values increasing from the bottom to the top (e.g., whiter regions in Figure 4c) and also dy's values increasing from its center towards the right edge (e.g., darker regions with negative values) and towards the left edge (e.g., brighter regions with positive values). Similar considerations, regarding the rotating object, can be drawn by observing Figures 4e and 4f where the translation parameters e and f, that take into account the fact that the rotation is not occurring around the image center, are depicted. The other object (robot), that moves forward, has got values in displacement dx especially in the robox left side

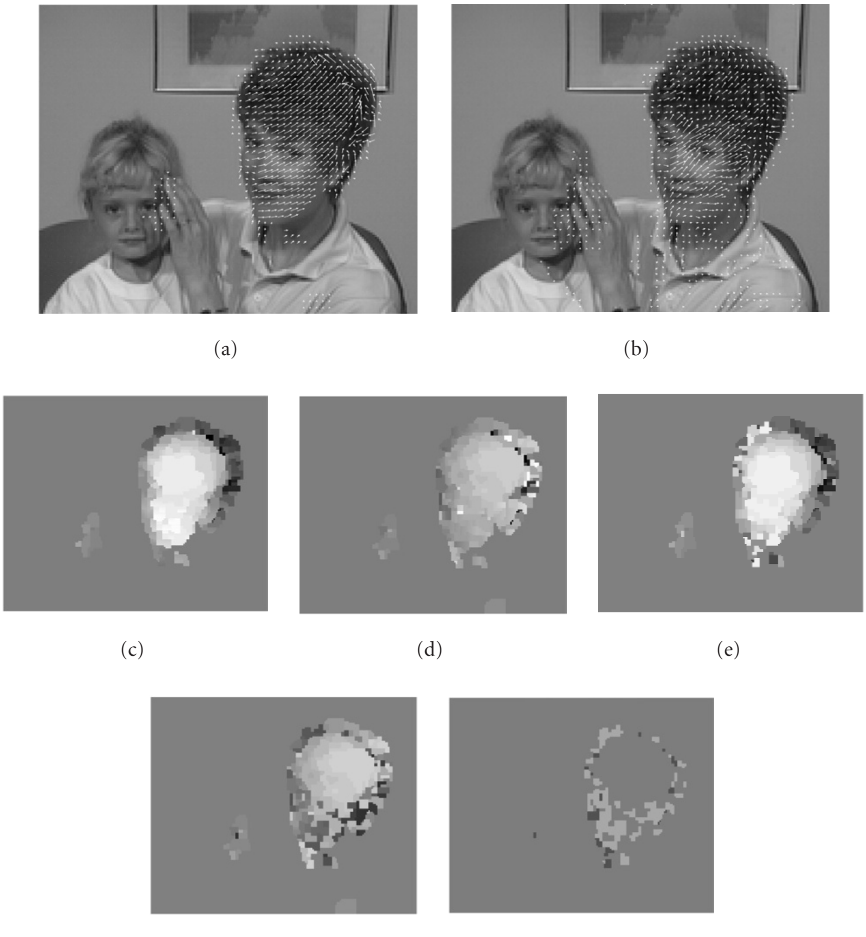


Figure 5: Real world sequence (M&D): frame 39 with the superimposed motion field estimated through (a) the proposed method and (b) the H&S approach; estimates by means of the proposed method: the displacements (c) dx and (d) dy, and the motion parameters (e) e, (f) f, and (g) θ.

(Figure 4c) and has got values in displacement dy increasing in magnitude going from its center towards the top and the bottom, thus resulting in correct description of a zooming effect. In Figure 4g the parameter θ is illustrated; only coefficients related to pure rotation (the box) are detected. As done before, also in this case, the PE has been computed and its value is reported in Table 3.

5.2.3. Mother&Daughter

Experimental tests carried out with a sequence called *Mother&Daughter* are presented in Figure 5 and debated hereafter.

In this video a mother caressing her daughter hair is depicted; the mother moves her head towards right and, in addition, slightly rotates up her neck; frames (QCIF format) that have been considered are numbers 38 and 40. In Figures 5a and 5b the motion vector field respectively estimated by the proposed methodology and the H&S approach are presented. It appears immediately that, in the first case, the field obtained is smoother and the vectors are very similar to each other; at the right end of the mother's head, the estimation is not so accurate and this is due to occlusions happened because of the rotation of her head. Furthermore, the global field appears more clean and does not show small vectors on the shoulders and on the breast of the mother, and on the daughter's head. As done before, in Figures 5c and 5d the values of the displacements dx and dy obtained with the proposed approach are presented. It is interesting to notice that pixels, belonging to the central part of the mother's face, which are in the 3D space closer to the camera, present a higher motion towards the right with respect to those back positioned. The head, in Figure 5c, appears as composed by different overlapped ovals, becoming darker while going from foreground to background, adequately explaining the movement in act. The backward part of the head is dark-colored and states that there is a motion towards the left side of the image as this region really has; in fact it is located behind the rotational axis of the head. The movement of the mother's hand is correctly detected as directed up and right as witnessed by regions brighter than the background

Table 3: PE for Robox sequence (higher value means a better prediction).

Kind of adopted approach	PE
Horn&Schunck	28.22 dB
Proposed method	38.19 dB

Table 4: PE for sequence *M&D* (higher value means a better prediction).

Kind of adopted approach	PE
Horn&Schunck	32.55 dB
Proposed method	38.34 dB

in Figures 5c and 5d. In Figures 5e, 5f, and 5g the estimated motion parameters are presented. Figures 5e and 5f look quite similar to Figures 5c and 5d already analyzed in detail. On the contrary, Figure 5g contains very interesting information because it clearly indicates that there is an object with an anticlockwise rotation (bright gray pixels) and its rotation center can easily be supposed to be in the middle of the circular region individuated. Also in this case the PE has been computed and its value is reported in Table 4.

6. CONCLUDING REMARKS

A new approach aiming at direct estimation of motion parameters in a sequence of images has been developed. The method is based on the minimization of a potential function which is composed by two basic components accounting for frame matching and smoothness binding, respectively. This potential has been derived by exploiting MAP criterion and MRF modelling. The technique has given positive results both with synthetic and with real world sequences. In particular, in addition to allow the direct estimation of motion parameters, the proposed technique shows excellent results also from the point of view of correct motion prediction (as demonstrated by the superior PE performances). This is due to fact that our approach constraint the estimated motion to adapt to a precise model, thus reducing the effects of noise. The main drawback of the algorithm, as for most of MRF-based techniques, is the high computational cost. To improve this aspect, to enhance the precision of parameter estimate, and to better handle large displacements, a multiresolution approach is under investigation. Work is also in progress to adapt the algorithm to deal with a more complex kind of motion (zooming objects) by introducing a more general motion model composed by a higher number of parameters.

REFERENCES

[1] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, 1981.

[2] J. Konrad and C. Stiller, “On Gibbs-Markov models for motion computation,” in *Video Compression for Multimedia Computing - Statistically Based and Biologically Inspired Techniques*, H. Li, S. Sun, and H. Derin, Eds., pp. 121–154, Kluwer Academic Publishers, Boston, Mass, USA, June 1997.

[3] A. M. Tekalp, *Digital Video Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1995.

[4] A. C. Bovik, *Handbook of Image & Video Processing*, Academic Press, New York, NY, USA, 2000.

[5] C. Stiller, “Object-based estimation of dense motion fields,” *IEEE Trans. Image Processing*, vol. 6, no. 2, pp. 234–250, 1997.

[6] J. Konrad and E. Dubois, “Bayesian estimation of motion vector fields,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, pp. 910–927, 1992.

[7] E. C. Hildreth, “Computations underlying the measurement of visual motion,” *Artificial Intelligence*, vol. 23, no. 3, pp. 309–354, 1984.

[8] H.-H. Nagel, “On the estimation of optical flow: Relations between different approaches and some new results,” *Artificial Intelligence*, vol. 33, no. 3, pp. 299–324, 1987.

[9] L. Alparone, M. Barni, F. Bartolini, and R. Caldelli, “Regularization of optic flow estimates by means of weighted vector median filtering,” *IEEE Trans. Image Processing*, vol. 8, no. 10, pp. 1462–1467, 1999.

[10] J. Konrad and E. Dubois, “Estimation of image motion fields: Bayesian formulation and stochastic solution,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1072–1075, April 1988.

[11] L. Lucchese, “A frequency domain technique based on energy radial projections for robust estimation of global 2D affine transformations,” *Computer Vision and Image Understanding*, vol. 81, no. 1, pp. 72–116, 2001.

[12] R. Y. Tsai and T. S. Huang, “Estimating three-dimensional motion parameters of a rigid planar patch,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1147–1152, 1981.

[13] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[14] J. Besag, “On the statistical analysis of dirty pictures,” *J. Roy. Statist. Soc. Ser. B*, vol. 48, no. 3, pp. 259–279, 1986.

[15] F. Heitz and P. Bouthemy, “Multimodal estimation of discontinuous optical flow using Markov random fields,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1217–1232, 1993.

[16] M. M. Chang, M. I. Sezan, and A. M. Tekalp, “An algorithm for simultaneous motion estimation and scene segmentation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, pp. V/221–V/224, Adelaide, Australia, May 1994.

[17] C. Stiller and J. Konrad, “Estimating motion in image sequences,” *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 70–91, 1999.

Roberto Caldelli was born in Figline Valdarno (Florence), Italy, in 1970. He graduated (cum laude) in electronic engineering from the University of Florence, in 1997, where he also received his Ph.D. degree in computer science and telecommunications engineering in 2001. He works now as a Postdoctoral Researcher with the Department of Electronics and Telecommunications at the University of Florence. He holds one Italian patent in the field of digital watermarking. His main research activities, witnessed by several publications, include digital image sequence processing, digital filtering, image and video digital watermarking, image processing applications for the cultural heritage field, and multimedia applications.

Franco Bartolini was born in Rome, Italy, in 1965. In 1991, he graduated (cum laude) in electronic engineering from the University of Florence, Florence, Italy. In November 1996, he received his Ph.D. degree in informatics and telecommunications from the University of Florence. Since November 2001, he has been an Assistant Professor at the University of Florence. His research interests include digital image sequence processing, still and moving image compression, nonlinear filtering techniques, image protection and authentication (watermarking), image processing applications for the cultural heritage field, signal compression by neural networks, and secure communication protocols. He has published more than 130 papers on these topics in international journals and conferences. He holds three Italian and one European patents in the field of digital watermarking. He is a Member of the Program Committee of the SPIE/IST Workshop on Security, Steganography, and Watermarking of Multimedia Contents, and Technical Program Cochair of the IEEE MMSP Workshop 2004. Dr. Bartolini is a Member of IEEE, SPIE, and IAPR.

Vittorio Romagnoli was born in Abbadia S. Salvatore (Siena), Italy, in 1976. In 1994 he got the High School degree in industrial electronic from the "I.T.I.S. Amedeo Avogadro" in Abbadia S. Salvatore. In February 2001 he graduated (cum laude) in electronic engineering from the University of Florence with a thesis on motion estimation in video sequences. From March 2001 to September 2002, he worked in a software company in Florence, where he developed java application on Linux platform and performed relational databases. Since October 2002, he has been working for a company, near Siena, operating in automation field, in particular, dealing with programmable logic controllers and industrial robots.