Gibbs random field model based 3D motion estimation from video sequences

A. Aydin Alatan, Levent Onural

Gibbs Random Field Model based 3-D Motion Estimation from Video Sequences

A. Aydın Alatan and Levent Onural

Electrical-Electronics Engineering Department
Bilkent University 06533, Bilkent Ankara
e-mail: alatan@ee.bilkent.edu.tr ve onural@ee.bilkent.edu.tr

ABSTRACT

In contrast to previous global 3-D motion concept, a Gibbs random field based method, which models local interactions between motion parameters defined at each point on the object, is proposed. An energy function which gives the joint probability distribution of motion vectors, is constructed. The energy function is minimized in order to find the most likely motion vector set. Some convergence problems, due to ill-posedness of the problem, are overcome by using the concept of hierarchical rigidity. In hierarchical rigidity, the objects are assumed to be almost rigid in the coarsest level and this rigidness is weakened at each level until the finest level is reached. The propagation of motion information between levels, is encouraged. At the finest level, each point have a motion vector associated with it and the interaction between these vectors are described by the energy function. The minimization of the energy function is achieved by using hierarchical rigidity, without trapping into a local minimum. The results are promising.

Keywords: 3-D motion estimation, Gibbs random field modelling, hierarchical rigidity, non-rigidity.

1. INTRODUCTION

3-D motion estimation from a video sequence remains as a challenging problem. Although, the ill-posedness of 2-D motion estimation is regularized[1] , in three dimensions, the same problem is far from to be solved. The complexity of three dimensional motion is much higher with respect to its 2-D counterpart. 3-D motion is usually modelled by 5 or 6 degrees of freedom, whereas only two parameters are sufficient to define 2-D motion. Hence some assumptions has to be made on the moving object, in order to be able to get a unique solution. The assumption of a rigid moving object, results with assigning the same motion parameters for every point on the object and the uniqueness of a solution is guaranteed[2,3]. However for many cases, the rigidity assumption is invalid and hence limits the performance of the system.

Another problem in 3-D motion estimation is the segmentation of the moving object from the scene. For example, without segmentation, finding correspondence points is difficult, and a robust solution is almost impossible. However, segmentation of moving objects is also another problem,

0-8194-1638-X/94/$6.00

which has not been completely solved. It usually needs some iterative motion estimation - segmentation steps, performed one after the other[4]. Nevertheless, in such approaches, the accumulation of the estimation error, after motion estimation step, decreases performance of segmentation.

We propose a new approach to the well-known 3-D motion estimation problem. Our basic idea is to formulate the problem in such a way that, all the apriori information about the motion can be inserted into a cost function. Similar approaches were used successfully in image segmentation[5] and restoration[6] , and also 2-D optical flow determination[7,8] . The cost function is also equivalent to the energy function of a Gibbs distribution, which is written by defining some local interactions between its neighbors. These local interactions permit looser relations between neighboring parameters, on contrast to the rigidity assumption, which is ultimately tight. Therefore, such an approach may achieve non-rigid motion estimation[9], as well as rigid body motion estimation, which is simply a subset of non-rigid estimation. In this approach every point on the object can be assumed as a rigid object by itself, which has some interactions between its neighbors.

On the other hand, using such an approach, the segmentation of moving objects can be achieved synchronously with estimation of motion parameters, by properly defining some new parameters in the cost function. These new parameters enable gathering similiar parameters together, without disturbing the discontinuities, existing between moving bodies.

However, in order to be able to write such a cost function, motion parameters should be defined at each point on the object. By increasing the number of variables, the cost function becomes difficult to minimize. Non-uniqueness of motion at every point and adding some extra fields for segmentation, makes the the cost function non-convex. The computation time for getting the global minimum becomes excessively high. At that point, a hierarchical approach solves this problem. Initially, the object can assumed to be rigid and afterwards the rigidity can be weakened hierarchically, by decreasing the size of the rigid portions of the object. After each level, the solution at one level will serve as an initial estimate for the next level. Such an approach can be successful for escaping deep local minima, which is expected from our cost function.

In the next section, the definition of 3-D motion parameters will be given. In section 3, the motion parameters will be modelled as random variables, whose distribution is given by a Gibbs random field. After definition of some new random fields, which are necessary for segmentation purposes, the proposed energy function of the Gibbs distribution will be explained. Section 4 is devoted for the hierarchical rigidity concept and simulation results will be given after that section. The conclusion of the results can be found at section 6.

2. 3-D MOTION PARAMETERS

3-D motion of a point on a object can be defined as below:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} 1 & w_z & -w_y \\ -w_z & 1 & w_x \\ w_y & -w_x & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}$$

In the equation above, w_x, w_y, w_z, are the rotation parameters around the axes, x, y, z respectively. The T_x, T_y, T_z are the translation parameters of the object along the axes. According to the equation, the point (X, Y, Z) in the 3-D space, moved to the point (X', Y', Z') after making rotations and translations. Although, this model is correct for all points on a rigid object, we define this motion at each point on the object.

The projection of the object points into the image plane, is achieved by orthogonal projection[3], which is the direct mapping of X and Y coordinates of the object to the x and y coordinates of the image plane respectively. Although, orthogonal projection is less realistic than perspective, it is simple to use. Nevertheless, all our formulations can be easily converted into perspective projection without creating any problems.

After orthographic projection of the points in 3-D space into the image plane, the motion of the points in the image can be written in terms of image plane coordinates, which gives

$$\begin{aligned} x'(t+\Delta t) &= x(t) + (w_z y(t) - w_y Z(t) + T_x) \\ y'(t+\Delta t) &= y(t) + (-w_z x(t) + w_x Z(t) + T_y) \end{aligned}$$

In this equation, (x, y) is the coordinate of a point in the image, which moves to another coordinate (x', y') after 3-D motion in the scene. It can be easily observed that the translation along z axis dropped, due to the properties of the orthographic projection. It should be noted that in this work, the depth information, $Z(x, y)$, is assumed to be known.

The aim is to find the unknown motion parameters $\vec{w} = (w_x, w_y, w_z, t_x, t_y)$, which are defined at each point on the projection of the object(s) in the image plane.

3. GIBBS RANDOM FIELD MODELLING

The motion parameters can be modelled as random variables. $\vec{w}(x, y)$ is a vector random variable and it is defined on the lattice Λ_m, on which image intensities $I(x, y)$ are also defined. Each element of $\vec{w}(x, y)$ can take values from a finite set, which also determines the range and the resolution of the motion parameters.

The joint probability distribution of $\vec{w}$ can be written in terms of a Gibbs distribution, which is generally defined as

$$P(W = w) \doteq \frac{1}{K} e^{-U(w)/T}$$

$$K = \sum_{w} e^{-U(w)/T}$$

$$U(w) = \sum_{c} V_c(w)$$

In the above equation, the joint probability is given as an exponential distribution, where the energy (or cost) function, $U(w)$ is the sum of some clique functions, $V_c(w)$. These clique functions represent the interactions between neighboring variables and reflects the a priori knowledge to the distribution. K is simply a normalization constant.

There are three important properties of the motion vector fields and these properties must determine the clique and hence the energy function. The first of all, correct motion parameters must match the same point of an object on two consecutive frames, by using the intensity values. This property is independent of the neighbors and is also ill-posed. For many of the practical scenes, the 3-D motion parameters of local object points are similar (not necessarily to be equal) and this property regularizes the ill-posedness of the first property The discontinuity of the 3-D motion parameters must be modelled as well as the continuity, since it is probable to observe objects in the scene, moving with different motion parameters, passing each other and generating motion boundaries. In order to be able to segment objects, this third property needs some extra fields, apart from motion and intensity fields. As an extra property, it should be noted that, some unpredictable and model failure points must be discriminated from the others. The points on the uncovered background of a moving object are useless for the estimation of motion and can be assumed as unpredictable points. Similarly, the 3-D motion of some object points can not be modelled by our motion definition and therefore must be dropped during estimation. This property also needs a new field to be modelled.

3.1 Proposed Energy Function

Taking into account all the ideas above, an energy function is proposed for the estimation of 3-D motion parameters from consecutive two frames and depth field.

$$\begin{aligned}
\mathcal{U}(\vec{w}, l, s \mid I_t, I_{t+1}, Z(x,y)) &= \sum_x \sum_y \left[(1 - s(x,y))(I_t(x,y) - I_{t+1}(x',y'))^2 + s(x,y)T_s\right] \\
&+ \lambda_{mot} \left[\sum_x \sum_y \sum_{c(x,y)} ||\vec{w}(x,y) - \vec{w}(x_c,y_c)||^2 (1 - l(x_c,y_c))\right] \\
&+ \lambda_s \left[\sum_x \sum_y \sum_{c(x,y)} \phi\left(s(x,y), s(x_c,y_c)\right)(1 - l(x_c,y_c))\right] \\
&+ \lambda_l \left[\sum_x \sum_y L(x,y)\right] \\
&+ \lambda_{disc} \left[\sum_x \sum_y \sum_{c(x,y)} \frac{l(x_c,y_c)}{1 + (I_t(x,y) - I_t(x_c,y_c))^2}\right] \\
&+ \theta(x,y,Z,\vec{w}(x,y))
\end{aligned}$$

The first term in the energy function achieves matching of intensities between consecutive frames, I_t and I_{t-1}. (x,y) and (x',y') are the coordinates of any moving point between the consecutive frames. However this should be performed for only the predictable points in the image, which are marked by the binary random field, $s(x,y)$. The points, for which $s(x,y)$ equals to one, is assumed to be unpredictable or model failure. For the unpredictable and model failure points, there is a constant penalty, T_s, which is also a threshold for being predictable or not. The second term stands for the continuity of the neighboring motion parameters. It is defined on some cliques, $c(x,y)$, which are basically four neighbors of the point (x,y) on the lattice Λ_m. For the expected motion fields, L_2 norm of the neighbor motion vectors must be as small as possible. However, between object or motion boundaries, this norm can be high and this must not give a penalty to the energy function. $l(x,y)$

is a binary random field and prevents getting high penalties along motion boundaries. $l(x,y)$ field is defined on a lattice, λ_l, which is dual of Λ_m and both of them are shown in Figure 1.

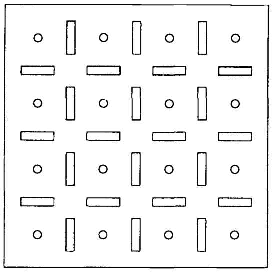

Figure 1. 'o' : lattice points for Λ_m
'[]': lattice points for Λ_l

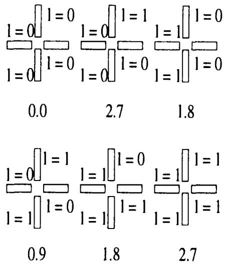

Figure 2. Penalties that are defined for $l(.,.)$ field.

The next term, supports the idea that unpredictable or model failure points, must be in groups in the image and the function ϕ gives some extra penalty if two neighboring $s(.,.)$ values are not equal. The function ϕ is simply an XOR operator in boolean algebra for the binary field $s(.,.)$. If a penalty is not given for setting $l(x,y)$ to one, then all values on this discontinuity field must be one, which will make the second term zero for all values of $\vec{w}(x,y)$. These penalties, which are represented by $L(.,.)$ function, must depend on the expected local structure of the discontinuity field[10] and shown in Figure 2. The fifth term, uses spatial discontinuity information, in order to locate the motion discontinuities. It encourages the locations for which $l(x,y)$ equals to one and spatial discontinuity is high. The very last term, gives a high penalty for the motion vectors which points a coordinate in 3-D space, which is occupied by another object point.

3.2 Minimizing Energy Function

The energy function must be minimized in order to find the maximum a priori estimate of the $\vec{w}(.,.)$, $l(.,.)$ and $s(.,.)$ fields. There are excessive number of unknowns and there exist methods to solve such optimization problems. Simulated Annealing[11] (SA) is one of the most popular among them. It is computationally very costly and converges with probability to global minimum in infinite time. However, in finite time and with some suboptimal temperature schedules, it gives reasonable results. Gibbs Sampler[6] (GS) is another version of simulated annealing, but it does not give any improvements, when the computation time is considered. Iterated Conditional Modes[12] (ICM) is another popular optimization method, which has "hard decision" nature on contrast to annealing methods. It converges to comparable results with annealing based methods, if the initial estimates are successful. Both SA and ICM are used during simulations, for the minimization of the energy function.

4. HIERARCHICAL RIGIDITY

The uniqueness of the 3-D motion parameters is investigated in the literature[2,3]. Having more than one parameter set for any motion in 3-D space, usually makes an associated energy function divergent. When each point in the image has a 3-D motion vector associated with it, the uniqueness of the motion is no more guaranteed even if there are local interactions between them. A solution to this problem is imposing some constraints to the problem, which will help solution to converge; , in other words regularizing the ill-posed problem. This constraint can be defined as making the neighboring pixels exactly equal at some predefined neighborhood. Instead of defining the motion vectors at each point on the image, they can be defined, in a subsample version of the lattice Λ_m. This will have an effective result of assigning the same motion vector to some number of image points, which are defined on the finest level of hierarchy. Therefore, an image can be assumed to be consisting of some rigid rectangular patches with different sizes at each level. In the coarsest level, there are large rigid bodies, whereas going down through the levels, we finally reach to the finest level, on which every image point has one motion vector associated with it. At each level of hierarchy, the interaction between neighbor motion vectors still exist.

The advantage of such an approach is to guarantee the uniqueness of the motion parameters at the coarsest level (easy convergence) and propagating this result through the levels, without trapping into a local minima. At the finest level, we still reach to our initial "weakly rigid" model, in which the motion vectors are weakly connected by a similarity term in the energy function. Each level create some motion vector set, which will serve as an initial estimate in the next level and makes the energy function converge easily.

Some simulations are carried on, in order to test the performance of such a motion estimator and the results are presented in the next section.

5. SIMULATION RESULTS

Although, the formulation is valid for non-rigid motion, all the simulations are carried on an artificial sequence, which has rigid motion. The depth information is assumed to be known.

The experiments are carried in two different stages. In the first stage, the proposed energy function is minimized by using only the finest level and in the second one, hierarchical rigidity is used during minimization.

SA is used for the minimization of the energy function for the finest level with a suboptimal "staircase" temperature schedule. The temperature is kept constant for some number of iterations, until the number of accepted perturbations converges to a steady value. Afterwards the temperature is decreased by some amount and the same procedure is applied iteratively until temperature is cold enough. For coarser levels, ICM instead of SA is used for minimization, since it is easy to get a convergence in the coarser levels, even with ICM.

The values of the constants, λ's, in the energy function is found by experimentation and tabulated in table 1. Although there is no successful method in the literature to estimate these parameters in such a complex energy function, as a rule of thumb, it can be stated that, the best values are the one's that normalize (or equalize) each energy term with respect to the others. This result can also be justified intuitively by stating that each piece of the a priori information has the same importance.

λ_{mot}	λ_s	λ_l	λ_{disc}
0.01	0.5	0.5	20

Table 1. λ values

The result of minimization of the energy term without the hierarchical rigidity concept is given in Figure 3. The histograms of the motion parameters are given in Figure 4. Although during SA, the temperature is lowered very slowly, even SA trapped into a local minima, resulting from the ill-posedness of the 3-D motion. The quality of the reconstructed image is good, however the result is not correct. The energy function is trapped into suboptimal solution, whose energy is more than (but small enough) the optimal solution.

If the hierarchical rigidity approach is used instead, the results are shown in Figures 5,7,9 and the histograms are illustrated in Figure 6,8,10. At the coarsest level, rectangular blocks of size 8x8 were used. Even with ICM, the motion parameters tend to converge into correct values, as it can be observed from Figure 6. At finer levels, the convergence continues and quality of the reconstructed image improves and SNR_{peak} increases. The final solution is not optimal, but represent the correct motion which is observed from the histograms.

The segmentation of the moving object is also successful as it can be observed from the line field. Without the line field, at the boundary between moving object and static background, blurring of the motion field is observed. However, the transition must be sharp and this is achieved by the help segmentation. Another important segmentation is the discrimination of unpredictable and model failure points. This is important especially for video coding, since these points must not be compensated by motion vectors.

6. CONCLUSIONS

In this paper, a new approach to 3-D motion estimation problem is presented. The main idea is consideration and utilitization of all available information about 3-D motion, in order to solve this ill-posed problem. In contrast to the other methods, motion is not considered globally on the object. The local interactions are defined, which will model the overall motion of the object. However, such an approach has a drawback of converging into another solution, which has approximately equal energy with the correct solution. This result does not approve the modelling success of the energy function.

However, this problem, which is due to ill-posedness, can be tackled by imposing the concept of hierarchical rigidity into the energy function. Previous works show that, some number of correspondence points guarantees a unique solution for the rigid 3-D motion[2,3]. Obviously, this result also eases convergence to the correct solution. By forcing some number of points to use the same motion vector parameter, the global concept of the motion is also taken into account. In the coarser levels, the uniqueness of the 3-D motion is also guaranteed by that means. Simulation results approves all our discussions.

Our scheme is a new approach to 3-D motion estimation. Currently, its only drawback seems to be the computation time. In the future studies, this problem will tried to be examined. The estimation of depth must also be considered in a 3-D motion estimation method and by adding some new terms into the energy function, the unknown structure field will be tried to estimated.

7. REFERENCES

1. B.K.P.Horn and B.G.Schunk "Determining Optical Flow", Artificial Intelligence , vol.17, pp 185-203, 1981.

2. J. Weng, N.Ahuja and T.S. Huang, "Optimal Motion and Structure Estimation", IEEE Trans. on PAMI, vol.15, no.9, pp.864-884, Sep. 1993.

3. X. Hu and N. Ahuja, "Motion Estimation under Orthographic Projection", IEEE Trans. on Robotics and Automation, vol.7, no.6, pp. 848-853, Dec. 1991.

4. D.W. Murray and B.F. Buxton, "Scene segmentation from visual motion using global optimization", IEEE Trans. on PAMI, vol.9, pp.220-228, 1987.

5. M.M.Chang, M.I.Sezan and A.M.Tekalp, "A Bayesian Framework For Combined Motion Estimation and Scene Segmentation in Image Sequences",in IEEE ICASSP Proc., (Adeliade, Austr.), May 1994.

6. D. Geman and S. Geman, "Stochsatic Relaxation, Gibbs Distribution, and Bayesian Restoration of Images", IEEE Trans. on PAMI, vol.6 , no.6 , pp.721-741 , Nov. 1984.

7. J. Konrad and E. Dubois, "Bayesian Estimation of Motion Vector Fields", IEEE Trans. on PAMI, vol.14, no.9, pp.910-927, Sep. 1992.

8. F. Heitz and P. Bouthemy, "Multimodal Estimation of Discontinuous Optical Flow using Markov Random Fields", IEEE Trans. on PAMI, vol.15, no.2, pp.1217-1232, December 1993.

9. T.S.Huang, "Modelling, Analysis and Visualization of Nonrigid Object Motion", in Int. Conf. on Pat. Recog., pp.361-364, (Atlantic City, NJ), 1990.

10. R.Buschmann, "Joint Multigrid estimation of 2-D motion and object boundaries using boundary patterns", in SPIE's Symp. on VCIP, pp.106-118, 1993.

11. S.Kirkpatrick, C.D. Gellat and Jr.M.P.Vecchi, "Optimization by Simulated Annealing", Science, pp.661-680, 1983.

12. J.Besag, "On the statistical analysis of dirty pictures", J.Roy.Statis. Soc., B48(3), 1886.

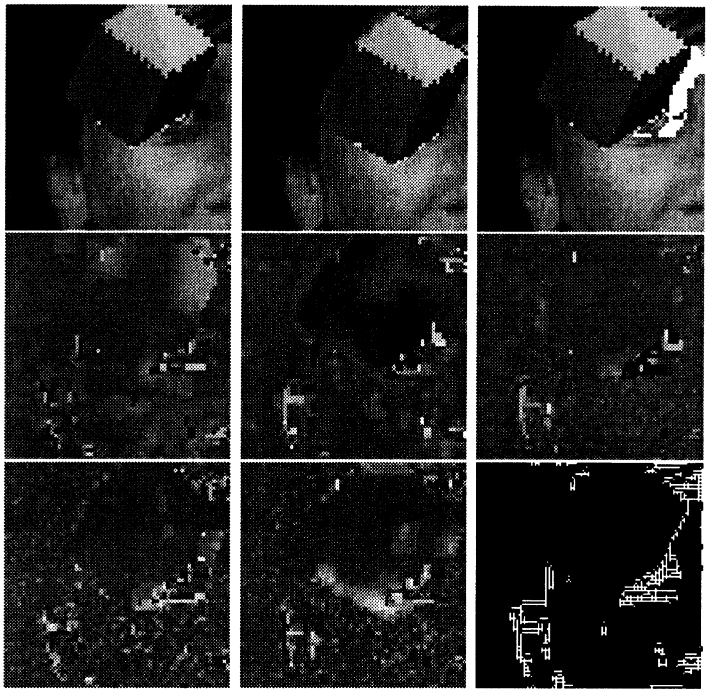

Figure 3. The first two frames of the artificial "cube" sequence. The reconstructed first frame after the estimation of 3-D motion parameters. PSNR is about 40dB.
The points of white intensity belongs to $s(.,.)$ field (uncovered background) and they are unpredictable. [top row, left to right]
Estimated w_x, w_y and w_z angle values of the motion. Dark intensities represent negative, whereas light ones positive motion values.
The values are mostly equal to zero. [middle row: left to right]
Estimated T_x and T_y values. The $l(,.,)$ field is able to locate motion discontinuities at some points. [bottom row: left to right]

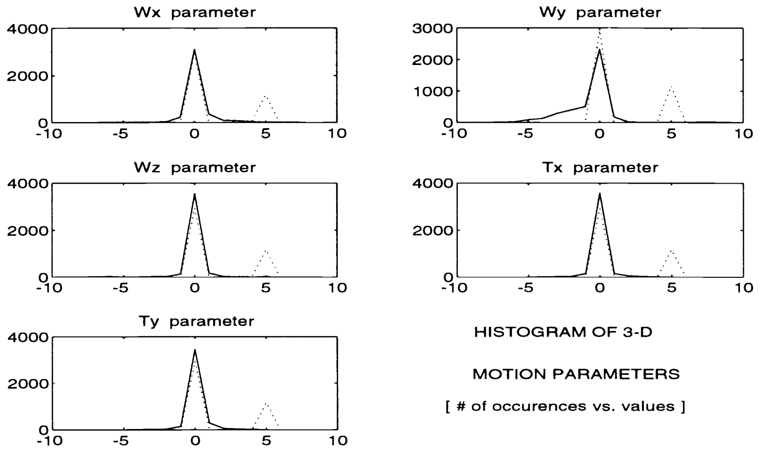

Figure 4. The histograms of the motion parameters in Figure 3. Solid lines are the estimated values, whereas the dotted lines give the correct solutions. These values are calculated by using only one (finest) motion level only and the results are far from being correct.

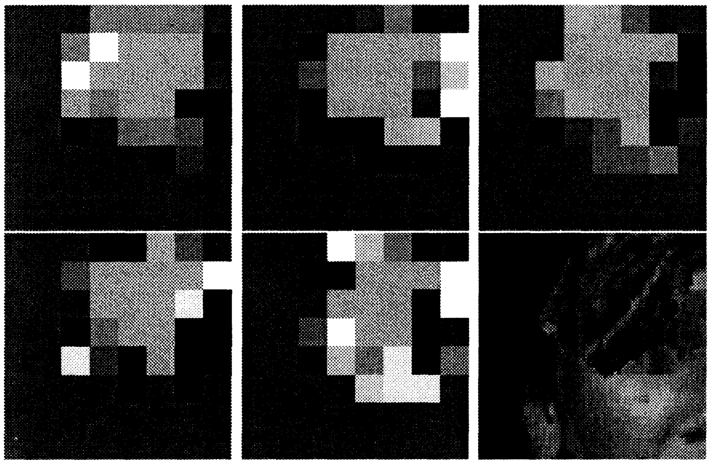

Figure 5. Estimated w_x, w_y and w_z parameter values of the motion. [up row: left to right]
Estimated T_x and T_y values. The reconstructed first frame after the estimation of 3-D motion parameters.
PSNR is 33dB. in the reconstructed image.
The cube is textured with the same motion parameters. The block size is 8x8 for this case.

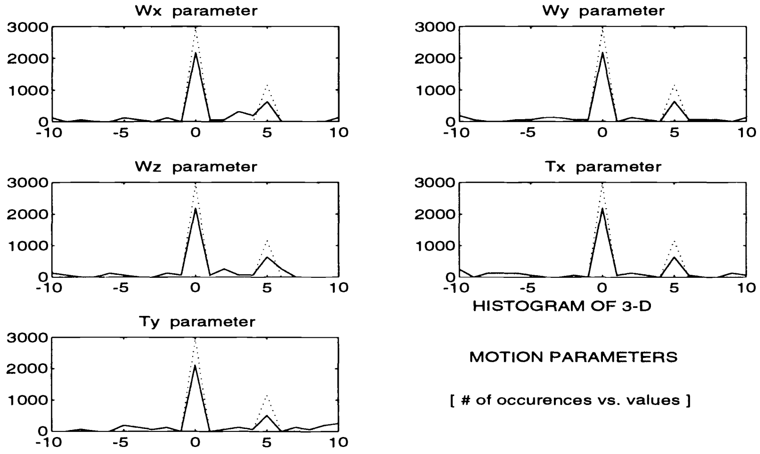

Figure 6. The histogram of the motion parameters in Figure 5. Solid lines are the estimated values,
whereas the dotted lines give the correct solution.
The convergence of motion parameters are much more better than Figures 3 and 4.

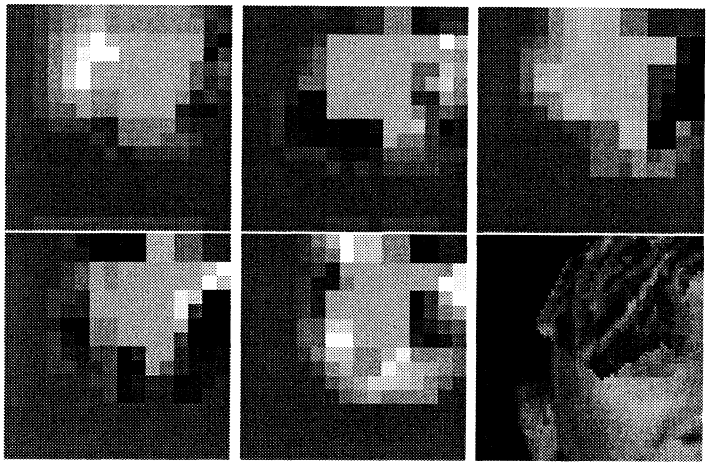

Figure 7. Estimated w_x, w_y and w_z parameter values of the motion. [up row: left to right] Estimated T_x and T_y values. The reconstructed first frame after the estimation of 3-D motion parameters. The results of Figure 5 is used as an initial estimate for this step. PSNR is 34dB in the reconstructed image. The block size is 4x4 for this case. [bottom row: left to right]

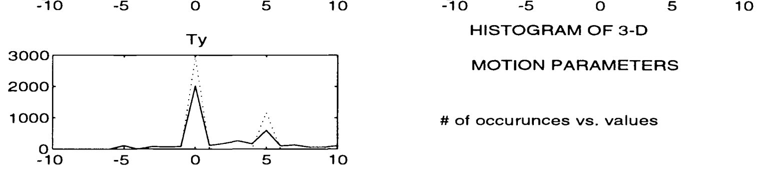

Figure 8. The histogram of the motion parameters in Figure 7. Solid lines are the estimated values, whereas the dotted lines give the correct solution.

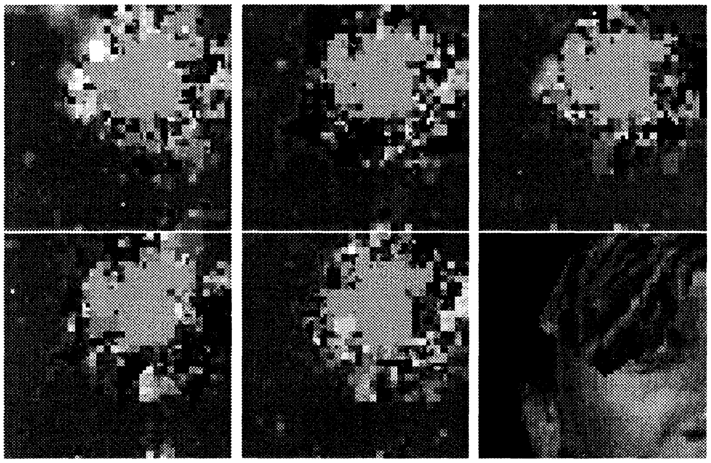

Figure 9. Estimated w_x, w_y and w_z parameter values of the motion. [up row: left to right] Estimated T_x and T_y values. The reconstructed first frame after the estimation of 3-D motion parameters. PSNR is 51dB in the reconstructed image. The block size is 1x1 for this case and the result of 2x2 block size estimation are used as initial estimates. Transition from block size of 4x4 to 2x2 is not shown. [bottom row: left to right]

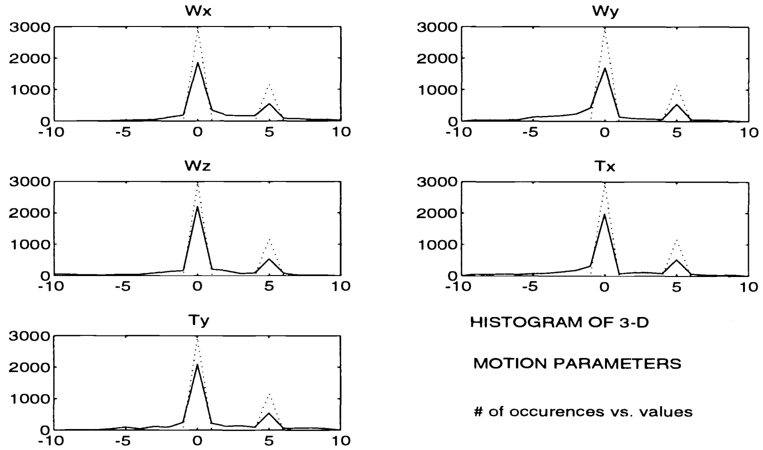

Figure 10. The histogram of the motion parameters in Figure 9. Solid lines are the estimated values, whereas the dotted lines give the correct solution.