# STATISTICAL BASED MOTION ESTIMATION FOR VIDEO CODING

*G. Calvagno, L. Celeghin, R. Rinaldo, L. Sbaiz*

Dipartimento di Elettronica e Informatica
Via Gradenigo 6/a, 35131 Padova, Italy
Tel: +39-49-827 7731, Fax: +39-49-827 7699, E-mail: calvagno@dei.unipd.it

## ABSTRACT

In this work, statistical based motion estimation is applied to the problem of motion estimation for video coding. We show that the motion equations of a rigid body can be formulated as a non linear dynamic system whose state is represented by the motion parameters and by the scaled depths of the object feature points. An extended Kalman filter is used to estimate the global motion, from which successive frames can be predicted in a motion compensated video coding system. The structure imposed by the model implies that the reconstructed motion is very natural in comparison to more common block-based schemes. Moreover, the parametrization of the model allows for a very efficient coding of motion information.

## 1. INTRODUCTION

Typical video sequences consist of few moving rigid objects and a static background. In particular, video-conference scenes have an almost fixed scene content, consisting of the head-and-shoulder of the speaker and of the background. The movement of the speaker mainly consists of the global movement of the shoulder and head, which can be approximated as rigid objects, and of the local motion due to facial expression changes and speech.

Statistical based motion estimation has been widely used in computer vision [1] and more recently in video coding [2]. In this work, a modification of the scheme of [1] is applied to the problem of motion estimation for video coding. The estimated motion parameters for each object in the scene, modeled as the projections of a 3D rigid body, can be used to reconstruct the image sequence at the decoder. The constraints imposed by the model guarantee that the reconstructed motion is very natural compared to simpler and more common block-based schemes. Moreover, the simple parametrization of the model allows for a very efficient coding of motion information.

## 2. PROBLEM FORMULATION

We suppose that the cartesian reference system is centered at the pupil of the observer, the $Z$ axis points forward and coincides with the optical axis, while $X$ and $Y$ are parallel to the image plane and form with $Z$ a right handed reference. Let $\mathbf{X}_i(t) = [X_i(t), Y_i(t), Z_i(t)]^T$ denote the coordinates of the generic point $i$ of a rigid body at time $t$.

The velocity of any point $i$ of the rigid body can be represented by the sum of a translation velocity $\dot{\mathbf{X}}_O(t)$ and of a rotation velocity, namely

$$\dot{\mathbf{X}}_i(t) = \mathbf{\Omega}(t) \wedge \mathbf{X}_i(t) + \dot{\mathbf{X}}_O(t) \tag{1}$$

where $\mathbf{\Omega}(t) = [\Omega_X(t), \Omega_Y(t), \Omega_Z(t)]^T$ is the vector of the angular velocities. Thus, 6 parameters are sufficient to characterize the motion. We can rewrite equation (1) in matrix form as

$$\dot{\mathbf{X}}_i(t) = \tilde{\mathbf{\Omega}}(t)\mathbf{X}_i(t) + \dot{\mathbf{X}}_O(t), \tag{2}$$

where

$$\tilde{\mathbf{\Omega}}(t) = \begin{bmatrix} 0 & -\Omega_Z(t) & \Omega_Y(t) \\ \Omega_Z(t) & 0 & -\Omega_X(t) \\ -\Omega_Y(t) & \Omega_X(t) & 0 \end{bmatrix} \tag{3}$$

is a skew symmetric matrix.

The continuous time equation (2) can be solved to derive a discrete time equation for $\mathbf{X}_i(t)$, namely

$$\mathbf{X}_i(t+1) = \mathbf{R}(t)\mathbf{X}_i(t) + \mathbf{T}(t). \tag{4}$$

If $\mathbf{\Omega}(t)$ is constant between $t$ and $t+1$, as we will assume in the following, we have in particular

$$\mathbf{R}(t) = e^{\tilde{\Omega}(t)},$$

$$\mathbf{T}(t) = [T_X(t), T_Y(t), T_Z(t)]^T = \int_t^{t+1} e^{\tilde{\Omega}(t)(t+1-\tau)} \dot{\mathbf{X}}_O(\tau)\, d\tau. \tag{5}$$

Suppose we are given the perspective projections of $N$ points, or features, of the rigid body in a set of consecutive frames. We will show that we can use these projections to estimate the state of a discrete time non linear system which describes the 3-D motion of the object and its shape. Moreover we can use such estimate to predict the feature positions at time $t+1$ from the positions at time $t$.

Let $\mathbf{x}_i(t)$ denote the vector of the $i$-th feature coordinates on the image plane at time $t$. The coordinates on the image plane are related to the 3-D coordinates by perspective projection, i.e., assuming a focal length equal to 1,

$$\mathbf{x}_i(t) = \frac{\mathbf{X}_i(t)}{Z_i(t)} = \begin{bmatrix} X_i(t)/Z_i(t) \\ Y_i(t)/Z_i(t) \\ 1 \end{bmatrix}. \tag{6}$$

We define by $\bar{Z}(t) = \sum_{i=1}^{N} Z_i(t)/N$ the average depth and by $s_i(t) = Z_i(t)/\bar{Z}(t)$ the scaled depth. From equation (4),

0-7803-3258-X/96/$5.00 © 1996 IEEE

we derive the following equation for the $Z$ component

$$\begin{aligned} Z_i(t+1) &= \mathbf{R}_3(t)\mathbf{X}_i(t) + T_Z(t) \\ &= (\mathbf{R}_3(t)s_i(t)\mathbf{x}_i(t) + \tilde{T}_Z(t))\bar{Z}(t) \end{aligned} \quad (7)$$

where $\tilde{\mathbf{T}}(t) = \mathbf{T}(t)/\bar{Z}(t)$ is the scaled translation and $\mathbf{R}_3(t)$ denotes the third row of $\mathbf{R}(t)$. From equations (4) and (7) we obtain

$$\mathbf{x}_i(t+1) = \frac{\mathbf{R}(t)s_i(t)\mathbf{x}_i(t) + \tilde{\mathbf{T}}(t)}{\mathbf{R}_3(t)s_i(t)\mathbf{x}_i(t) + \tilde{T}_Z(t)} \quad (8)$$

which gives the feature frame coordinates at time $t+1$ as a function of the coordinates at time $t$, the motion parameters $\mathbf{R}(t)$, $\mathbf{T}(t)$ and the scaled depth $s_i(t)$. Note that the set of scaled depths $s_i(t)$ gives information about the 3-D shape of the object. We found that the inclusion of $s_i(t)$ in the motion model is essential for the robustness of the estimates.

We can interpret equation (8) as an implicit relation between the coordinates $\mathbf{x}_i(t)$ and the state $\mathbf{R}(t)$, $\mathbf{T}(t)$, $s_i(t)$, $i = 1, ..., N$ of a non-linear system governing the motion of the rigid body. Our objective is to estimate the system state, i.e., the object motion parameters and the scaled depths of the features, from the feature projections $\mathbf{x}_i(t)$. In the following, we derive the state update equations for the system.

Using (4) and (5), one can derive an expression for $\mathbf{R}(t)$ as a function of $\boldsymbol{\Omega}(t)$ (Rodrigues' formula [3]). For this reason, we will use $\boldsymbol{\Omega}(t)$ instead of $\mathbf{R}(t)$ in the state equations and assume for its dynamics a random walk model

$$\boldsymbol{\Omega}(t+1) = \boldsymbol{\Omega}(t) + \mathbf{n}_\Omega(t), \quad (9)$$

where $\mathbf{n}_\Omega(t)$ is a zero mean white noise. The update equation for the scaled translation $\tilde{\mathbf{T}}(t)$ can be derived by assuming a random walk model also for $\mathbf{T}(t)$ and by averaging equation (7) for $i = 1, ..., N$. We obtain

$$\tilde{\mathbf{T}}(t+1) = \frac{\mathbf{T}(t+1)}{\bar{Z}(t+1)} = \frac{\tilde{\mathbf{T}}(t)}{\mathbf{R}_3(t)\bar{\mathbf{x}}(t) + \tilde{T}_Z(t)} + \mathbf{n}_{\tilde{\mathbf{T}}}(t), \quad (10)$$

where $\bar{\mathbf{x}}(t) = \frac{1}{N}\sum_{i=1}^{N} s_i(t)\mathbf{x}_i(t)$.

From equation (7) we derive the update equation for $s_i(t)$

$$s_i(t+1) = \frac{Z_i(t+1)}{\bar{Z}(t+1)} = \frac{\mathbf{R}_3(t)s_i(t)\mathbf{x}_i(t) + \tilde{T}_Z(t)}{\mathbf{R}_3(t)\bar{\mathbf{x}}(t) + \tilde{T}_Z(t)}. \quad (11)$$

Moreover we have the constraint

$$\frac{1}{N}\sum_{i=1}^{N} s_i(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{Z_i(t)}{\bar{Z}(t)} = 1. \quad (12)$$

In summary, the system equations are

$$\begin{cases} \boldsymbol{\Omega}(t+1) &= \boldsymbol{\Omega}(t) + \mathbf{n}_\Omega(t) \\ \tilde{\mathbf{T}}(t+1) &= \frac{\tilde{\mathbf{T}}(t)}{\mathbf{R}_3(t)\bar{\mathbf{x}}(t)+\tilde{T}_3(t)} + \mathbf{n}_{\tilde{\mathbf{T}}}(t) \\ s_i(t+1) &= \frac{\mathbf{R}_3(t)s_i(t)\mathbf{x}_i(t)+\tilde{T}_3(t)}{\mathbf{R}_3(t)\bar{\mathbf{x}}(t)+\tilde{T}_3(t)} + n_{s_i}(t) \\ \sum_{i=1}^{N} s_i(t) &= N \\ \mathbf{x}_i(t+1) &= \frac{\mathbf{R}(t)s_i(t)\mathbf{x}_i(t)+\tilde{\mathbf{T}}(t)}{\mathbf{R}_3(t)s_i(t)\mathbf{x}_i(t)+\tilde{T}_3(t)} + \mathbf{n}_x(t) \end{cases} \quad (13)$$

where $n_{s_i}(t)$ and $\mathbf{n}_x(t)$ are model noises that may take into account slow deformations of the object.

Defining the system state by $\xi(t) = [\boldsymbol{\Omega}(t)^T, \tilde{\mathbf{T}}(t)^T, s_1(t), \ldots, s_N(t)]^T$ and observations by $\mathbf{y}(t) = [\mathbf{x}_1(t)^T, \ldots \mathbf{x}_N(t)^T, \mathbf{x}_1(t+1)^T, \ldots, \mathbf{x}_N(t+1)^T]^T + \mathbf{w}(t)$, where $\mathbf{w}(t)$ is the observation noise, we may rewrite (13) as

$$\begin{cases} \xi(t+1) &= f(\xi(t), \mathbf{y}(t)) + \tilde{\mathbf{n}}(t) \\ h(\xi(t), \mathbf{y}(t) - \mathbf{w}(t)) &= 0 \end{cases} \quad (14)$$

where $\tilde{\mathbf{n}}(t)$ is a function of the noises in (13) and of $\mathbf{w}(t)$.

## 3. MOTION ESTIMATION

System (14) is non linear and implicit, therefore we can estimate and predict its state by means of the Implicit Extended Kalman Filter (IEKF) [4].

We will denote with $\hat{\xi}(t|t)$ the estimate of the state at time $t$ from the measurements $\{\mathbf{y}(\tau) : \tau \leq t\}$ and with $\hat{\xi}(t+1|t)$ the prediction of the state at time $t+1$ from the measurements $\{\mathbf{y}(\tau) : \tau \leq t\}$. Then, if one defines

$$\mathbf{F} \triangleq \left.\frac{\partial f}{\partial \xi}\right|_{\hat{\xi}(t|t),\mathbf{y}(t)} \quad \mathbf{C} \triangleq \left.\frac{\partial h}{\partial \xi}\right|_{\hat{\xi}(t+1|t),\mathbf{y}(t)} \quad \mathbf{D} \triangleq \left.\frac{\partial h}{\partial \mathbf{y}}\right|_{\hat{\xi}(t+1|t),\mathbf{y}(t)}, \quad (15)$$

the equations of the IEKF become

- **Prediction Step**

$$\hat{\xi}(t+1|t) = f(\hat{\xi}(t|t)) \quad (16)$$

$$\mathbf{P}(t+1|t) = \mathbf{F}\mathbf{P}(t|t)\mathbf{F}^T + \tilde{\mathbf{Q}}_{\tilde{n}} \quad (17)$$

- **Update Step**

$$\hat{\xi}(t+1|t+1) = \hat{\xi}(t+1|t) + \mathbf{L}(t+1)h(\hat{\xi}(t+1|t), y(t+1)) \quad (18)$$

$$\mathbf{L}(t+1) = -\mathbf{P}(t+1|t)\mathbf{C}^T(\mathbf{C}\mathbf{P}(t+1|t)C^T + \mathbf{R})^{-1} \quad (19)$$

$$\mathbf{R} = \mathbf{D}\mathbf{R}_w\mathbf{D}^T \quad (20)$$

$$\begin{aligned} \mathbf{P}(t+1|t+1) &= (\mathbf{I}_N - \mathbf{L}(t+1)\mathbf{C})\mathbf{P}(t+1|t) \\ &(\mathbf{I}_N - \mathbf{L}(t+1)\mathbf{C})^T + \mathbf{L}(t+1)\mathbf{R}\mathbf{L}(t+1)^T \end{aligned} \quad (21)$$

where $\mathbf{P}(t|t) = \mathrm{E}[(\xi(t) - \hat{\xi}(t|t))(\xi(t) - \hat{\xi}(t|t))^T]$ is the estimation error variance, $\mathbf{P}(t+1|t) = \mathrm{E}[(\xi(t+1) - \hat{\xi}(t+1|t))(\xi(t+1) - \hat{\xi}(t+1|t))^T]$ is the prediction error variance, $\mathbf{R}_w$ is the covariance matrix of $\mathbf{w}(t)$ and $\tilde{\mathbf{Q}}_{\tilde{n}}$ is the covariance matrix of $\tilde{\mathbf{n}}(t)$.

The estimation $\hat{\xi}(t|t)$ can be used to predict the feature positions at time $t+1$ from their positions at time $t$ by means of equation (8). Moreover, from equation (8), we can predict the position of points that are not features by assigning to the generic point $\mathbf{x}(t)$ a scaled depth $s(t)$ obtained by averaging the estimates of the scaled depths $\hat{s}(t|t)$. For example, we can use the weighted sum

$$s(t) = \frac{\sum_{i=1}^{N} w_i \hat{s}_i(t|t)}{\sum_{i=1}^{N} w_i} \quad (22)$$

where the weight $w_i$ takes into account the distance between the point $\mathbf{x}(t)$ and the feature coordinate $\mathbf{x}_i$.

## 4. EXPERIMENTAL RESULTS

In this section we will present some simulation results, relative to both synthetic and real image sequences.

In the first experiment, a synthetic sequence of features was used. We considered a set of 33 points (obtained using a uniform random generator) placed inside a cube of side 1m with centroid positioned 1.5m ahead of the viewer. We suppose to observe the scene using a camera with a visual field of 52° and CIF resolution ($288 \times 352$ pixels). The cloud of points has been projected on the image plane and corrupted by Gaussian noise with a standard deviation $\sigma$ satisfying $3\sigma = 1$ pixel. The cloud undergoes a rotational motion around its center of mass. In the camera coordinate system, this corresponds to a rotation around the horizontal axis with angular velocity $\Omega_X = 3°$/frame, and a translation with velocity $1.5(1-\cos\Omega_X)$m/frame along the $Z$ axis and $1.5\sin\Omega_X$m/frame along the $Y$ axis. After 50 frames the cloud inverts its direction of rotation. The Kalman filter described in the previous section was used to estimate the motion. We used an initial null estimate for $\tilde{\mathbf{T}}$ and $\boldsymbol{\Omega}$ and the initial estimates of the scaled depths was set to 1. Fig. 1 shows the estimates of the velocities as a function of the frame number. The ground truth is plotted in dotted line. We can see that the filter takes about 20 frames to converge. After that, it follows the cloud even after the abrupt inversion of motion at frame 50. We found that the use of $s_i(t)$ as state variables is essential to make the tracking procedure effective and robust.

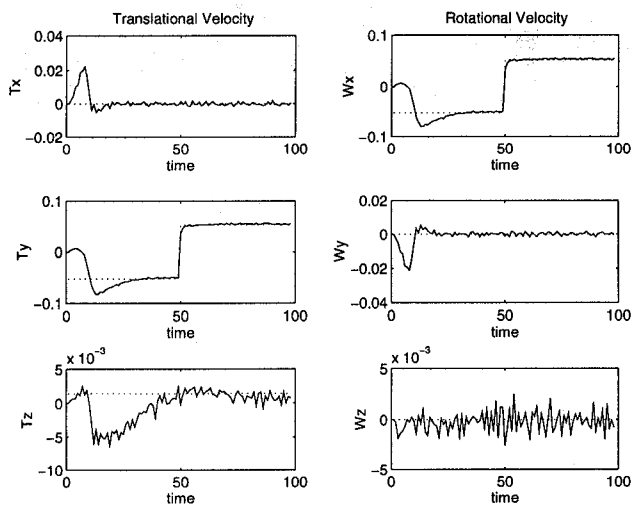


Figure 1: Estimates of the motion parameter for the synthetic sequence.

To test the estimator with a real video sequence, we need to choose the feature points in the first frame and track them in the following frames. To this purpose we used a multiresolution version of Lucas-Kanade's algorithm [5, 6]. This procedure consists in approximating the luminance at time $t$ around the point at position $\mathbf{x}$ with a differentiable function $I(\mathbf{x}, t)$. In addition, one supposes that the luminance variations are due only to translations. Therefore, denoting with $\mathbf{d}$ the displacement of $\mathbf{x}$ from time $t$ to $t+1$, one can write

$$I(\mathbf{x}, t) = I(\mathbf{x} - \mathbf{d}, t+1) \simeq I(\mathbf{x}, t+1) - \mathbf{g}^T\mathbf{d} \qquad (23)$$

where $\mathbf{g} = \text{grad}I$. Solving (23) with respect to $\mathbf{d}$ and minimizing the mean squared error on a window $\mathcal{W}$, one finds

$$\mathbf{G}\mathbf{d} = \mathbf{e} \qquad (24)$$

where

$$\mathbf{G} = \int_{\mathcal{W}} \mathbf{g}^T\mathbf{g}\, d\mathbf{x} \quad \mathbf{e} = \int_{\mathcal{W}} (I(\mathbf{x}, t) - I(\mathbf{x}, t+1))\mathbf{g}\, d\mathbf{x}. \qquad (25)$$

System (24) allows one to find the displacement of a feature and, if the eigenvalues of the matrix $\mathbf{G}$ are nonzero, it has a unique solution. In practice, due to the presence of noise, the eigenvalues must be greater than a threshold. This suggests to consider only those feature points that correspond to local maxima of the minimum eigenvalue of $\mathbf{G}$.

This procedure was applied to frames 10 to 58 (with step 4) of the video sequence "Miss America". The image sequence is segmented, as suggested for MPEG 4 [7]. The features obtained using Lucas Kanade's algorithm were classified into three groups corresponding to the regions of the hair, the face and the shoulders (see Fig. 2). For each region the proposed filter was used to estimate the motion and the shape. Frame at time $t+1$ was predicted from frame at time $t$ using the estimated parameters. For this purpose we used equation (8) to predict the pixel positions in frame at time $t+1$ from their positions in frame at time $t$. We assigned a scaled depth to all the other image pixels using equation (22) with weights $w_i = (|x - x_i| + |y - y_i|)^{-3}$ where $(x, y)$ are the generic pixel coordinates and $(x_i, y_i)$ are the coordinates of feature $i$. In particular, to each pixel $\mathbf{x}(t)$ we applied equation (8) using the corresponding motion parameters and estimated scaled depth. This permits to predict the pixel coordinates $\hat{\mathbf{x}}(t+1)$ at time $t+1$. The luminance value of pixel $\mathbf{x}(t+1)$ at time $t+1$ is set to the same luminance value of $\hat{\mathbf{x}}(t)$ at time $t$. We assumed no motion in the background and the corresponding pixels are simply replicated from time $t$ to $t+1$.

We report the results relative to the prediction of frame 58 from frame 54 using the proposed algorithm and, for comparison purposes, a block matching procedure. Block matching was performed using $16 \times 16$ blocks, motion vectors in the range $-15$ $+15$ and half pixel refinement. We also consider the case of using frame 54 as an estimate of frame 58, with no motion compensation.

In the first row of Table 1 the mean squared error (MSE) between frame 58 and its prediction is given for the three cases. We can see that the MSE of the proposed solution is slightly greater than the MSE obtained with block matching. On the other hand, we can compare the number of bits necessary to code the motion vectors in the block matching algorithm with the number of bits required to code the motion and shape parameters of the estimation filter. In the second row of Table 1 the uncompressed number of bits is reported for both cases. Block matching requires 12 bits per vector while, in the proposed algorithm, we need to code the estimated state for each filter. This requires six floats for the motion parameters and one float for each scaled depth. In Table 1, we assume to use 16 bits per float. Appropriate

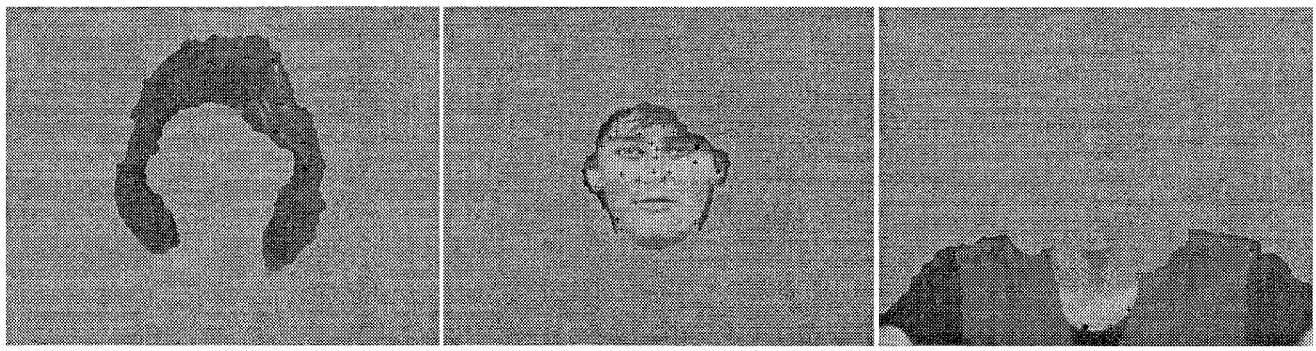

Figure 2: The three regions used for the test on the sequence "Miss America", with the features computed by the Lucas Kanade's algorithm (crosses).

Figure 3: Original frame #58 of "Miss America."

Figure 4: Predicted frame #58 of "Miss America."

| | Proposed method | Block matching | No comp. |
|---|---|---|---|
| MSE | 13.2 | 10.4 | 73.6 |
| bits | 1376 | 4752 | 0 |

Table 1: Results for different prediction methods.

coding can be used to reduce the required number of bits. Fig. 3 shows the original frame #58 of "Miss America," while Fig. 4 shows the predicted frame using our method. In the presence of relevant local motion around the lips and the eyes, the prediction obtained using the proposed method can give less satisfactory results, as expected. For instance, the prediction of frame 50 from frame 46 gives an MSE=32.6 with our method and an MSE=22.1 with block matching.

## 5. ACKNOWLEDGEMENT

The authors would like to thank Stefano Soatto for his useful suggestions and for kindly supplying the feature selection and tracking program.

## 6. REFERENCES

[1] S. Soatto, R. Frezza e P. Perona "Motion estimation via dynamic vision," *IEEE Trans. on Automatic Control*, March 1996, pp. 393-413.

[2] R. J. Crinon and W. J. Kolodziej "Adaptive Model-Based motion Estimation," *IEEE Trans. on Image Processing*, September 1994, pp. 469-481.

[3] Z. Zhang and O.D. Faugeras, "Three-dimensional motion computation and object segmentation in a long sequence of stereo frames," Rapports de Recherche, INRIA, Juillet 1991.

[4] P. S. Maybeck, *Stochastic Models, Estimation and Control. Volume 1 and 2*, Academic Press, 1979-1982.

[5] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. of the 7th Int. Joint Conf. on Art. Intell.*, 1981.

[6] J. Bouquet and P. Perona, "Visual navigation using a single camera," *Proc. of the 5th Inter. Conf. on Computer Vision*, 1995.

[7] Ad hoc group on MPEG-4 video VM editing, *MPEG-4 Video Verification Model Version 1.22.0*, ISO/IEC JTC1/SC29/WG11, n. 1260, Firenze, March 1996.