OBJECT ORIENTED CODING USING 3D MOTION ESTIMATION

G. Calvagno, V. Orsatti, R. Rinaldo, L. Sbaiz
Dipartimento di Elettronica e Informatica
Via Gradenigo 6/a, 35131 Padova, Italy
Tel: +39 49 828 7731, Fax: +39 49 828 7699
e-mail: rinaldo@dei.unipd.it

ABSTRACT

In this work, we apply 3D motion estimation to the problem of motion compensation for video coding. We model the video sequence as the perspective projection of a collection of rigid bodies which undergo a roto-translational motion. Motion compensation of the sequence frames can be performed once the shape of the objects and the motion parameters are determined. 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 then used to estimate both the motion and the object shape parameters simultaneously. We found that the inclusion of the shape parameters in the estimation procedure is essential for reliable motion estimation. Our experiments show that the proposed approach gives the following advantages: the filter gives more reliable estimates in the presence of measurement noise in comparison with other motion estimators that separately compute motion and structure; the filter can effectively track abrupt motion changes; the structure imposed by the model implies that the reconstructed motion is very natural as opposed to more common block-based schemes; the parametrization of the model allows for a very efficient coding of motion information.

1 INTRODUCTON

Object oriented coding is a promising technique to code video sequences at very low bit rates [1, 2]. According to this concept, each object of an image is described by three sets of parameters defining its motion, shape and surface color. In a typical framework, the input sequence is first analyzed and each region is classified according to several classes.

In this work we propose an object oriented coder for head and shoulder video sequences. In our scheme, the image sequence is subdivided into three regions corresponding to the head, the neck and shoulders, and the background. These regions are modeled as projections of 3D rigid objects onto the camera image plane. Model failure regions are associated with the eyes and the mouth, that are coded and treated differently. The main contribution of this work is the adaptation of the recursive 3D motion estimator proposed in [4] to the framework of a complete video coding scheme. The proposed approach results to be efficient both because the reconstructed motion appears to be very natural and because it allows for effective coding of motion information. The visual quality of the reconstructed sequences is indeed superior to that obtained with standard coders like H.263 at comparable bit rates.

2 MOTION ESTIMATION PROCEDURE

In this work, the video sequence is modeled as the perspective projection of rigid bodies which move in a 3D space. In particular, we focus our attention on video conference sequences which typically show the head and the shoulders of the speaker and a fixed background.

In the present implementation of the coder, we actually consider only three regions, corresponding to the head and the neck and shoulders, while the background is replicated from one image to the other. Those parts of the scene, like the mouth and the eyes, which do not comply with the rigid body model, are considered as model failure objects and coded apart.

The system is initialized by coding the first frame of the sequence in intra frame mode, using the scheme of [5]. The video coding algorithm estimates the three-dimensional motion and structure [4] of the rigid objects, and the scene is reconstructed at the decoder by mapping the texture of the first image onto the successive frames. This is done by taking into account object shape descriptions and motion information, as explained in the next sections.

2.1 Recursive motion estimation

The first step of the procedure requires that the scene is segmented into objects, namely, the static background, the head of the speaker, and the neck and shoulders. The segmentation is carried out using a standard region growing technique, together with the rough motion information provided by standard block matching to separate the static background. The eyes and the mouth are detected using the algorithm described in [6]. These parts have to be treated as model failure regions, since their 3D motion is not rigid, even at a first approximation. Fig. 1 shows a typical segmentation result.

In order to perform motion estimation, we follow an approach which is usually considered in the computer vision literature. We extract a set of characteristic points on each object, called *features*, establish a correspondence between points in adjacent frames by matching the features, and finally compute the structure and motion parameters based on the feature matches. The feature points are selected within blocks with high luminance activity and their position is tracked using a correlation algorithm. The correlation algorithm adopts a 9×9 pixel window and half pixel accuracy.

In the following, we consider a cartesian reference system centered at the pupil of the observer, with the Z axis pointing forward and aligned with the optical axis. The X and Y axes 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, six parameters are sufficient to characterize the motion. The continuous time equation (1) can be solved to derive a discrete time equation for $\mathbf{X}_i(t)$, namely

$$\mathbf{X}_i(t+1) = \mathbf{R}(t)\mathbf{X}_i(t) + \mathbf{T}(t). \tag{2}$$

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

$$\mathbf{x}_i(t) = \frac{\mathbf{X}_i(t)}{Z_i(t)} = \begin{bmatrix} X_i(t)/Z_i(t) \\ Y_i(t)/Z_i(t) \\ 1 \end{bmatrix}. \tag{3}$$

Using equation (2) one can easily derive:

$$\mathbf{x}_i(t+1) = \frac{\mathbf{R}(t)Z_i(t)\mathbf{x}_i(t) + \mathbf{T}(t)}{\mathbf{r}_3(t)Z_i(t)\mathbf{x}_i(t) + T_Z(t)}, \tag{4}$$

where $\mathbf{r}_3(t)$ denotes the third row of matrix $\mathbf{R}(t)$ and $T_Z(t)$ is the third component of vector $\mathbf{T}(t)$. Equation (4) gives the position on the image plane of the projected point i at time $t+1$ when one knows its position at time t, the rotation matrix $\mathbf{R}(t)$, the translation vector $\mathbf{T}(t)$ and the depth $Z_i(t)$.

As described above, a correlation procedure permits to determine the projections onto the image plane of a set points of the three-dimensional object. From the measured feature coordinates $\mathbf{x}_i(t)$ and $\mathbf{x}_i(t+1)$ in successive frames, it is indeed possible to compute the motion parameters $\mathbf{R}(t)$, $\mathbf{T}(t)$ and the three-dimensional structure parameters $Z_i(t)$, up to a scale factor [4]. In a practical environment, projections $\mathbf{x}_i(t)$ and $\mathbf{x}_i(t+1)$ will be affected by observation noise and therefore the measurement of parameters becomes a typical estimation problem.

We define by $\bar{Z}(t) = \sum_{i=1}^{N} Z_i(t)/N$ the average depth, by $s_i(t) = Z_i(t)/\bar{Z}(t)$ the scaled depth and by $\tilde{\mathbf{T}}(t) = \mathbf{T}(t)/\bar{Z}(t)$ the scaled translation. Using these positions and assuming a random walk model for $\mathbf{\Omega}(t)$ and $\tilde{\mathbf{T}}(t)$, we obtain the system equations

$$\left\{ \begin{array}{lcl} \mathbf{\Omega}(t+1) & = & \mathbf{\Omega}(t) + \mathbf{n}_{\Omega}(t) \\ \tilde{\mathbf{T}}(t+1) & = & \frac{\tilde{\mathbf{T}}(t)}{\mathbf{r}_{3\cdot}(t)\bar{\mathbf{X}}(t)+\tilde{T}_3(t)} + \mathbf{n}_{\tilde{\mathbf{T}}}(t) \\ s_i(t+1) & = & \frac{\mathbf{r}_{3\cdot}(t)s_i(t)\mathbf{X}_i(t)+\tilde{T}_3(t)}{\mathbf{r}_{3\cdot}(t)\bar{\mathbf{X}}(t)+\tilde{T}_3(t)} + n_{s_i}(t) \\ \sum_{i=1}^{N} s_i(t) & = & N \\ \mathbf{x}_i(t+1) & = & \frac{\mathbf{R}(t)s_i(t)\mathbf{X}_i(t)+\tilde{\mathbf{T}}(t)}{\mathbf{r}_{3\cdot}(t)s_i(t)\mathbf{X}_i(t)+\tilde{T}_3(t)} + \mathbf{n}_x(t) \end{array} \right. \tag{5}$$

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 $\xi(t) = [\mathbf{\Omega}(t)^T, \tilde{\mathbf{T}}(t)^T, s_1(t), \ldots, s_N(t)]^T$ 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 (5) as

$$\left\{ \begin{array}{l} \xi(t+1) = f(\xi(t), \mathbf{y}(t)) + \tilde{\mathbf{n}}(t) \\ h(\xi(t), \mathbf{y}(t) - \mathbf{w}(t)) = 0 \end{array} \right. \tag{6}$$

where $\tilde{\mathbf{n}}(t)$ is a function of the noises in (5) and of $\mathbf{w}(t)$.

System (6) is non linear and implicit, therefore we can estimate and predict its state by means of an Implicit Extended Kalman Filter (IEKF) [7].

The state estimate $\hat{\xi}(t|t)$ can be used to predict the feature positions at time $t+1$ from their positions at time t by means of equations (5). We remark that the inclusion of the scaled depths $s_i(t)$ in the filter state is *essential* to obtain reliable motion estimates in the case of very noisy observations like those relative to real video sequences [4].

2.2 Motion compensation

In the present implementation of the coder, the luminance of the pixel at position $\mathbf{x}_i(t+1)$ in the current frame is taken to be equal to the luminance of the pixel $\mathbf{x}_i(t)$ in the frame at time t, except for the model failure regions that are coded apart. This luminance prediction does not take into account illumination or object reflection properties, but is adequate for the application at hand.

As a matter of fact, the scaled depths are estimated by the Kalman filter only for the feature points, while the depth information is necessary to update the position $\mathbf{x}_i(t)$ of every pixel using (5). Our approach is to suppose that the objects of the scene have smooth surfaces. As a consequence, the depth of a generic point is approximated by means of a weighted sum of the depths of neighbor feature points. In particular, to all the image pixels that are not features, we assign a scaled depth obtained as a weighted average of the estimates of the scaled depths $\hat{s}_i(t|t)$ of the feature points, namely

$$s(t) = \frac{\sum_{i=1}^{N} w_i \hat{s}_i(t|t)}{\sum_{i=1}^{N} w_i}. \tag{7}$$

The weights were empirically set to $w_i = (|x - x_i| + |y - y_i|)^{-3}$ to take into account the distance between the generic pixel coordinates (x, y) and the coordinates (x_i, y_i) of feature i. In summary, to each pixel $\mathbf{x}(t)$ of the segmented regions we apply equations (5) using the corresponding motion parameters and estimated scaled depth to calculate the pixel coordinates $\mathbf{x}(t+1)$. The luminance value of pixel $\mathbf{x}(t+1)$ in model compliance regions is set to the same luminance value of $\mathbf{x}(t)$. We assume no motion in the background and the corresponding pixels are simply replicated from time t to $t+1$.

3 RESULTS

The system is initialized by coding the first frame of the sequence in intra frame mode, using the scheme of [5]. The video coding algorithm estimates the three-dimensional motion and structure of the rigid objects, and the scene is reconstructed at the decoder by mapping the texture of the first image onto the successive frames. This is done by taking into account object shape descriptions and motion information. To avoid error propagation in the motion estimation procedure, we introduce a feedback scheme and use the feature positions relative to the *decoded* previous image. Of course, when the error between the original and reconstructed sequences becomes unacceptable, due to illumination or abrupt scene changes, the process has to be reinitialized by coding an intra image.

The estimated state is an efficient way to code model compliance objects. We use arithmetic coding to code the innovation of the Kalman filter, which gives sufficient information to reconstruct the motion parameters at the decoder.

Model failure regions, corresponding to the mouth and the eyes, are coded using an adaptation of the algorithm of [5] capable of handling regions of arbitrary shape.

We tested the proposed algorithm with the sequence Miss America at 15 frames/s. Fig. 1 shows the result of the segmentation and of the eye and mouth location procedure on frame #110 of Miss America. The region borders are approximated by a polygonal line, and coded with 300 bits on average.

Within the head and the neck-and-shoulder regions, we selected and tracked 27 and 13 features respectively. In Table 1 the number of bits needed to code each object are reported. We note that model compliance objects, which need only motion information, are coded with a small amount of bits in comparison with model failure objects.

The overall resulting bit rate, including intra frames, was about 30 kbit/s. There was an intra image approximately every 15 coded frames, and each intra image was coded using about 12 kbits. Fig. 2 shows the original frame #96 and the corresponding reconstructed frame, while Fig. 3 shows the PSNR for the frames 60 to 140 of the reconstructed sequence "Miss America."

4 CONCLUSION

In this paper, we described the application of the recursive 3D motion estimator proposed in [4] to the framework of a complete video coding scheme. The proposed video coder adopts a statistical based motion estimation procedure. In particular, the video sequence is segmented into regions that are modeled as the projections of 3D moving objects onto the camera plane. An extended implicit Kalman filter is used to estimate the state of a dynamical system that includes object structure and motion parameters. The eyes and the mouth are automatically detected and treated as model failure regions. Besides the limited amount of information sent to the decoder, the visual quality of the reconstructed sequences is quite good, especially when compared to block based coding systems.

Table 1: Average number of bits required to code each object of a predicted frame.

object	average # of bits
eyes and mouth	754
head	168.3
shoulders	82.4
total	1004.7

References

[1] H. G. Musmann, M. Hötter, J. Ostermann, "Object-oriented analysis-synthesis coding of moving images," *Signal Processing: Image Communication,* Vol. 1, No. 2, pp. 117-138, Oct. 1989.

[2] 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.

[3] S. Soatto, R. Frezza e P. Perona "Motion estimation via dynamic vision," *IEEE Transactions on Automatic Control*, Vol. 41, No. 3, March 1996, pp. 393-413.

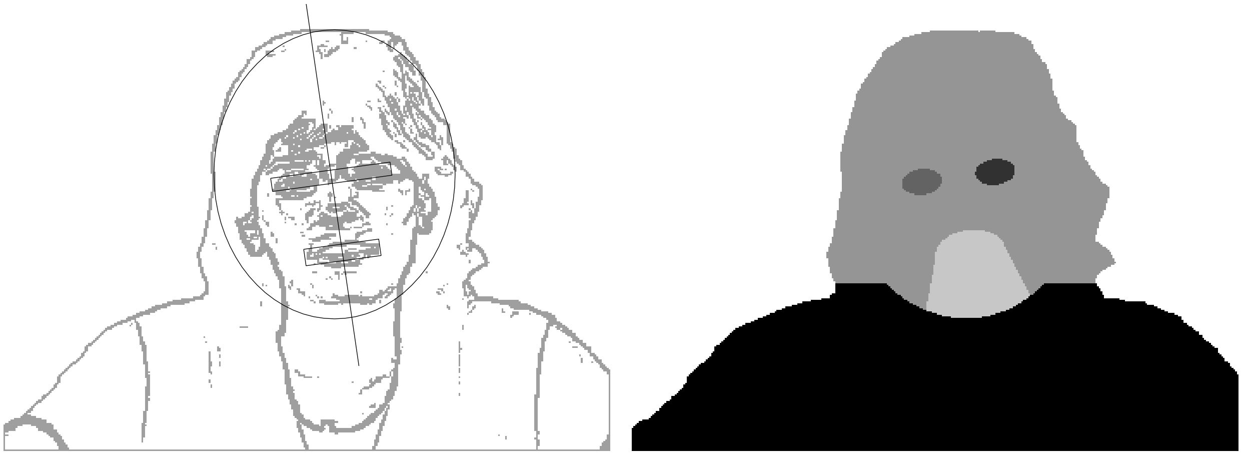

Figure 1: Result of segmentation and eye–mouth location procedure for frame #110 of Miss America.

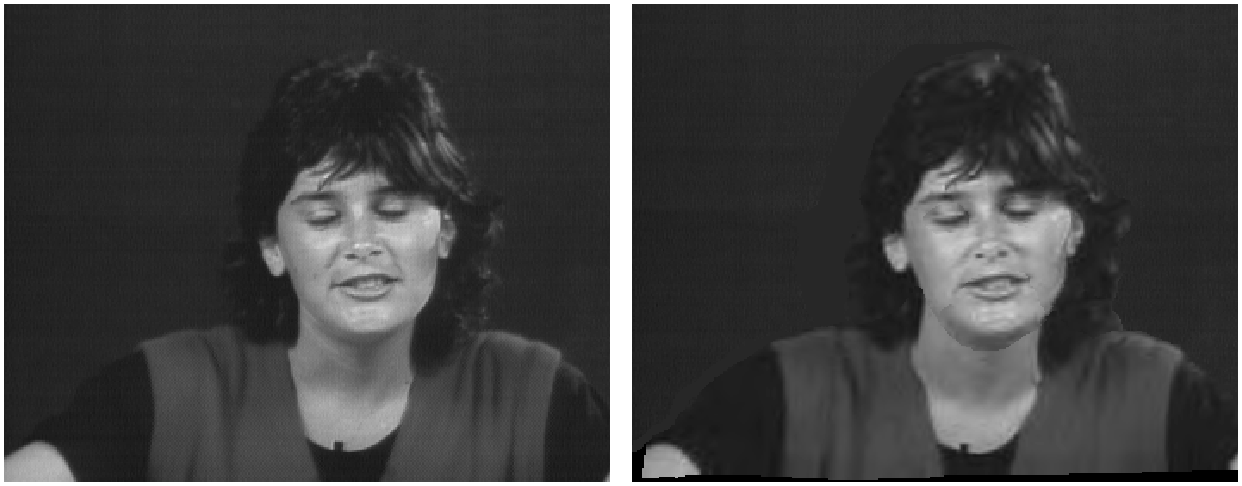

Figure 2: Original and reconstructed frame #96 of the test sequence Miss America.

[4] G. Calvagno, L. Celeghin, R. Rinaldo, L. Sbaiz, "Statistical Based Motion Estimation for Video Coding," *Proceedings of the 1996 IEEE International Conference on Image Processing (ICIP 1996)*, Lausanne, 16–19 Sept. 1996, Vol. I, pp. 105–108.

[5] R. Rinaldo, G. Calvagno, "Hybrid Vector Quantization for Multiresolution Image Coding," IEEE Transactions on Image Processing, Vol. 6, No. 5, May 1997, pp. 753-758.

[6] A. Eleftheriadis, A. Jacquin, "Automatic face location detection for model-assisted rate control in H.261-compatible coding of video," *Signal Processing: Image Communication,* no. 7, pp. 435-455, 1995.

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

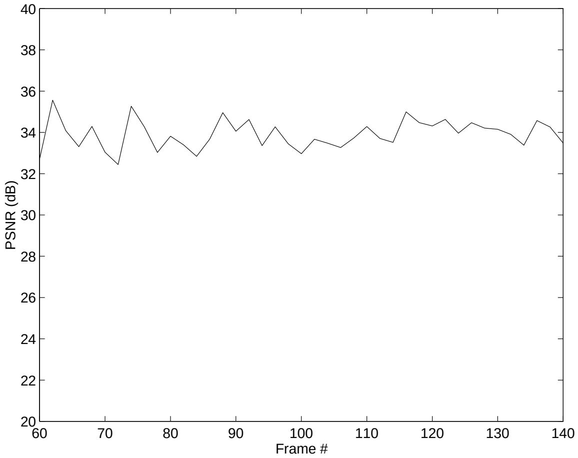

Figure 3: PSNR for the frames 60 to 140 of the reconstructed sequence "Miss America."