JOINT ESTIMATION AND OPTIMUM ENCODING OF DEPTH FIELD FOR 3-D OBJECT-BASED VIDEO CODING

A. Aydın Alatan and Levent Onural

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

ABSTRACT

3-D motion models can be used to remove temporal redundancy between image frames. For efficient encoding using 3-D motion information, apart from the 3-D motion parameters, a dense depth field must also be encoded to achieve 2-D motion compensation on the image plane. Inspiring from Rate-Distortion Theory, a novel method is proposed to optimally encode the dense depth fields of the moving objects in the scene. Using two intensity frames and 3-D motion parameters as inputs, an encoded depth field can be obtained by jointly minimizing a distortion criteria and a bit-rate measure. Since the method gives directly an encoded field as an output, it does not require an estimate of the field to be encoded. By efficiently encoding the depth field during the experiments, it is shown that the 3-D motion models can be used in object-based video compression algorithms.

1. INTRODUCTION

Even though many video compression standards exist, very low bit-rate coding is still a very challenging problem. Since coding of still images has almost reached to its limits, more compression might be possible for video in the temporal domain. Currently, most of the video compression algorithms reduces the temporal redundancy by using 2-D motion models. Since the performance of these algorithms has been saturated, the motion models should be re-examined to obtain better description, prediction and compression.

Recently, 3-D motion models are being utilized in some video coding algorithms [1, 2, 3, 4]. Although these methods obtain acceptable 3-D motion estimates, they do not propose any scheme on how to encode a dense depth field which is necessary to motion compensate the intensities on 2-D image frames. There are also some suboptimal approaches for encoding the dense depth fields in stereo coding applications [5].

In the following sections, after some necessary initial steps (2,3-D motion estimation and segmentation), a novel object-based depth encoding method will be examined.

2. MOTION ESTIMATION AND SEGMENTATION

Feature-based 3-D motion estimation methods [6] need 2-D correspondences between frames. These matches are usually found between features which are invariant to the relative motion between the surface and light sources [7]. However for object-based video coding purposes segmentation should also be achieved. A possible approach is to apply motion-based segmentation to obtain 2-D motion vectors for each object and choose "trustable" ones among this dense set to be used for 3-D motion parameter estimation. Hence the first step is jointly estimating 2-D motion and segmentation fields.

2.1. Finding 2-D Motion of Objects

Gibbs modeled motion estimation and segmentation has been proven to be successful [8]. Given two intensity frames, $\mathcal{I}_{t,t-1}$, to obtain the unknown 2-D motion, $\mathcal{D}$, segmentation, $\mathcal{R}$, and temporally unpredictable (TU), $\mathcal{S}$, fields, a cost function (also the energy function of a Gibbs distribution) can be minimized with respect to these unknowns. This function can be written as

$$\mathcal{U}(\mathcal{D},\mathcal{R},\mathcal{S} \mid \mathcal{I}_t,\mathcal{I}_{t-1}) = \mathcal{U}_n + \lambda_m\, \mathcal{U}_m + \lambda_R\, \mathcal{U}_R + \lambda_s\, \mathcal{U}_s \quad (1)$$

$$\begin{aligned}
\mathcal{U}_n &= \sum_{\mathbf{x}\in\Lambda} \left(I_t(\mathbf{x}) - I_{t-1}\left(\mathbf{x}-\mathbf{D}(\mathbf{x})\right)\right)^2 \left(1-S(\mathbf{x})\right) + S(\mathbf{x})T_s \\
\mathcal{U}_m &= \sum_{\mathbf{x}\in\Lambda} \sum_{\mathbf{x}_c\in\eta_\mathbf{x}} ||\mathbf{D}(\mathbf{x})-\mathbf{D}(\mathbf{x}_c)||^2 \;\; \delta\left(R(\mathbf{x})-R(\mathbf{x}_c)\right) \\
\mathcal{U}_R &= \sum_{\mathbf{x}\in\Lambda} \sum_{\mathbf{x}_c\in\eta_\mathbf{x}} \left[1-\delta\left(R(\mathbf{x})-R(\mathbf{x}_c)\right)\right] + \\
& \qquad \lambda_t\, \frac{\left[1-\delta(R(\mathbf{x})-R(\mathbf{x}_c))\right]}{1+\left(I_t(\mathbf{x})-I_t(\mathbf{x}_c)\right)^2} + \theta\left(R(\mathbf{x})\right) \\
\mathcal{U}_s &= \sum_{\mathbf{x}\in\Lambda} \sum_{\mathbf{x}_c\in\eta_\mathbf{x}} \left[1-\delta\left(S(\mathbf{x})-S(\mathbf{x}_c)\right)\right]
\end{aligned}$$

The reason for choosing such a cost function and some other details can be found in [4].

In order to find robust correspondences between consecutive frames, a selection process should be applied to dense 2-D motion field. By simply thresholding spatial gradients and local Gibbs energies, outliers of the 2-D motion field can be removed and a sparse subset of dense 2-D motion vector field is obtained. This sparse and robust set can be used in 3-D motion estimation algorithm which is explained in the next section.

2.2. Estimation of 3-D Motion

E-matrix method [9] is one of the most popular 3-D motion parameter estimation algorithm. This linear algorithm is susceptible to noise, but a nonlinear version, which takes into account noise and errors is proposed, too [10]. For each segmented object, using the robust 2-D correspondences and assuming that the object is rigid, 3-D motion parameters (rotation matrix $\mathbf{R}$ and translation vector $\mathbf{T}$) is estimated using improved E-matrix approach [10, 4].

Since the robust correspondences are sparse, dense depth estimates can not be obtained using this algorithm. If all the dense motion estimates which are obtained after minimizing Equation 1 are used, then the depth estimates at "untrustable" points will be quite sensitive to errors. Hence robust *dense* depth estimation is not possible using improved E-matrix method, although 3-D motion parameters are available for each object. However, for motion compensating intensities for every object, a depth value must be estimated for each image point. Moreover these depth values should be encoded efficiently. These two goals can be achieved at the same time by using the proposed method, explained in the next section.

3. JOINT ESTIMATION AND ENCODING OF DEPTH

Since 3-D motion parameter encoding is ultimately efficient for a rigid object (6 parameters/object), the compression performance of a 3-D object-based scheme depends on encoding of the depth field. For very low bit-rate video coding applications, the depth field should be encoded with some loss, since it is very expensive to transmit the "true" depth field.

Rate-distortion theory gives the minimum required bits to encode a source symbol at a given distortion (or vice versa) with some probability distribution and a given distortion measure [11]. Hence the encoded symbol is optimal for the corresponding distribution and distortion measure. By properly selecting an encoding criteria, $\mathcal{J}(\Delta, \mathcal{B})$ and minimizing this criteria with respect to depth, the optimal depth field to be encoded can be obtained. Δ is the distortion measure between the true, $\mathcal{Z}$, and lossy encoded depth, $\hat{\mathcal{Z}}$, fields and $\mathcal{B}$ is the number of bits to encode $\mathcal{Z}$ to obtain $\hat{\mathcal{Z}}$.

Since Δ and $\mathcal{B}$ are two different quantities to be jointly minimized, method of *objective weighting* [12] is an approach to solve this vector optimization problem, which is written as

$$\mathcal{J}(\Delta, \mathcal{B}) = \Delta + \lambda_0 \cdot \mathcal{B} \tag{2}$$

where λ_0 is a constant which reflects the weighting between two quantities Δ and $\mathcal{B}$. Before achieving joint optimization of bit-rate and distortion, a distortion criteria and a measure of bit-rate should be defined.

3.1. Distortion Criteria

Although the true dense depth field is not explicitly known, it is implicitly available in the intensities of consecutive frames. The true depth field should make intensity matches between consecutive frames by using 3-D motion parameters. For each object R_i with N object points, the distortion criteria can be defined as

$$\Delta = \frac{1}{N} \sum_{\mathbf{x} \in R_i} \left(I_t(\mathbf{x}) - \hat{I}_t(\mathbf{x}) \right)^2 \tag{3}$$

where the reconstructed frame, $\hat{I}_t$ is also equal to

$$\hat{I}_t(\mathbf{x}) = I_{t-1} \left(\mathbf{x} - \mathbf{D}_{2D} \left(\hat{Z}(\mathbf{x}, t) \right) \right) \tag{4}$$

As shown in Figure 1, $\mathbf{D}_{2D}$ is the perspectively projected 3-D object point motion, which also depends on $\hat{Z}(\mathbf{x})$. Since the true depth field information is available in $I_t(\mathbf{x})$ with a similar formulation to Equation 4, a nonlinear distortion function is obtained between the true and encoded fields.

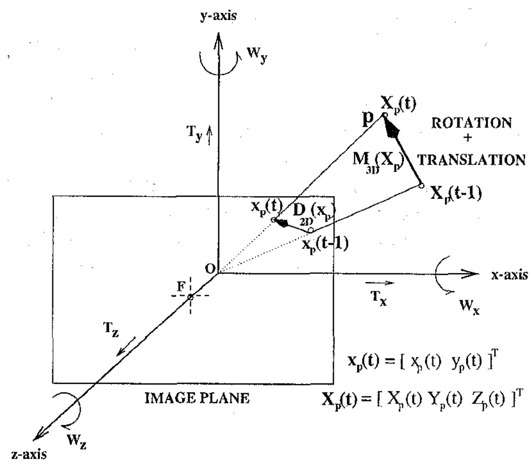

Figure 1: 3-D coordinate system

3.2. Bit-rate Measure

Since any scene is assumed to be the output of a random source, the depth field of a scene is a random field with some associated probability distribution. Using this probability measure, the number of bits to encode this depth field can be determined according to the basic principles of information theory [11]. Assuming that indoor scenes are observed through frames, it is expected to have smooth surfaces frequently. A Gibbs distribution can be written taking into account this a priori information with an associated energy function as below.

$$\mathcal{U}_Z(\mathcal{Z}) = \sum_{\mathbf{x} \in R_i} \sum_{\mathbf{x}_c \in \eta_{\mathbf{x}}} \left(\hat{Z}(\mathbf{x}) - \hat{Z}(\mathbf{x}_c) \right)^2 \tag{5}$$

The sum is over all points $\mathbf{x}$ of the ith object, segmented by the region R_i and $\eta_{\mathbf{x}}$ is the neighborhood of $\mathbf{x}$. By taking the logarithm of base 2 of the corresponding probability, the number of bits to encode the depth field is obtained as

$$\mathcal{B} = k \cdot log_2 e \cdot \sum_{\mathbf{X} \in R_i} \sum_{\mathbf{X}_c \in \eta_{\mathbf{X}}} \left(\hat{Z}(\mathbf{x}) - \hat{Z}(\mathbf{x}_c)\right)^2 + c(k) \quad (6)$$

where $c(k)$ parameter is simply equal to $\frac{N}{2} log_2(\frac{\pi}{4k})$.

3.3. Minimization of Encoding Criteria

Distortion and bit-rate is jointly optimized using Equations 3,6 which give

$$\min_{\mathcal{Z}} \quad \left\{ \left(\frac{1}{N} \sum_{\mathbf{X} \in R_i} \left(I_t(\mathbf{x}) - I_{t-1}\left(\mathbf{x} - \mathbf{D}_{2D}\left(\hat{Z}(\mathbf{x})\right)\right)\right)^2 \right) + \lambda \left(\sum_{\mathbf{X} \in R_i} \sum_{\mathbf{X}_c \in \eta_{\mathbf{X}}} \left(\hat{Z}(\mathbf{x}) - \hat{Z}(\mathbf{x}_c)\right)^2 \right) \right\} \quad (7)$$

By minimizing Equation 7 with respect to depth, an optimal lossy depth field with respect to the defined distortion and bit-rate measure is obtained. $c(k)$ parameter is removed from Equation 7, since it does not effect minimization. k and $log_2(e)$ constants can be multiplied with λ_0 constant and hence this product is defined to be λ . The minimization can be achieved by using a Multiscale Constrained Relaxation (MCR) method [13]. For different values of λ, different optimal rate-distortion pairs are obtained and λ can not be determined without extra constraints on rate and/or distortion. Such constraints might be available for video coding applications.

Since it is impossible to give a codeword to all existing depth fields according to their probabilities, in practice another coding strategy must be followed. Simple predictive coding can be used to remove redundancy from the obtained depth field. After linearly predicting a depth value by its causal neighbors, the prediction error can be encoded using a "lossless" compression algorithm (e.g. Lempel-Ziv). In this way, a codeword for the optimal dense depth field can be obtained.

3.4. Proposed Depth Encoder

The proposed depth encoder can be summarized as below :

1. Find 3-D motion parameters for each segmented object.
2. For a given λ, minimize Equation 7 to obtain the distorted depth field to encode.
3. Encode the prediction error of depth values using lossless Lempel-Ziv coding.

If λ is not given externally, for various values of λ repeat part 2 of the above algorithm to choose the best λ for a "target" distortion.

4. EXPERIMENTAL RESULTS

Simulations are conducted in two phases. In the first phase, an artificial sequence is used whose 3-D motion parameters and segmentation are known. Two frames from the artificial "Cube" sequence are presented in Figure 2. Minimizing Equation 7 for the value $\lambda = 1000$, an encoded depth field is obtained for the current frame. In Figure 3, the true and encoded depth fields ($\lambda = 1000$) are shown. Note that the encoded depth is a smoother version of the "true" one.

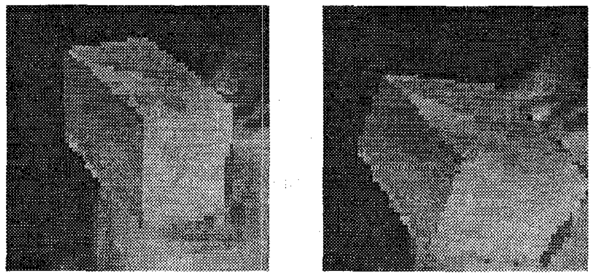

Figure 2: Original previous and current frame of the "Cube" sequence.

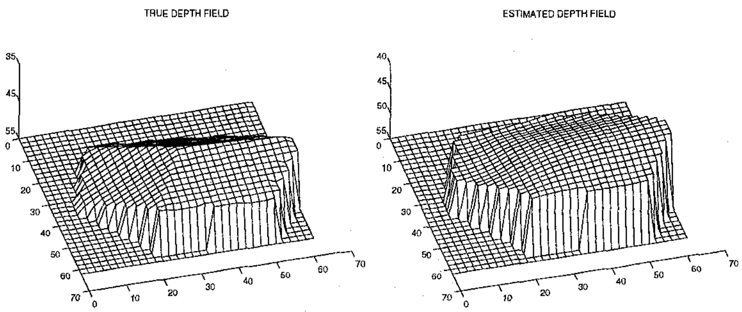

Figure 3: The mesh representations of the true and encoded depth field of the current frame of the "Cube" sequence.

In the second phase of the experiments, two frames (100 and 103) from *Foreman* sequence (176×144) are used (Figure 4) to find the 3-D motion parameters and the depth field to encode. The results of 2-D motion estimation and segmentation is shown in Figure 5. The 3-D motion parameters of the segmented head are found as

$$R = \begin{bmatrix} 0.9993 & 0.0242 & 0.0251 \\ -0.0242 & 0.9997 & 0.0003 \\ -0.0251 & -0.0003 & 0.9996 \end{bmatrix}, T = \begin{bmatrix} -0.0117 \\ 0.5585 \\ 0.8293 \end{bmatrix}$$

Minimizing Equation 2 for different values of λ (Table 1), the rate-distortion plot is obtained, shown in Figure 6. For $\lambda = 5$, the encoded depth field and reconstructed current frame (inside head region SNR_{peak} is over $38dB$) are also shown in Figure 6.

5. CONCLUSIONS

Since 3-D motion description is efficient for rigid bodies, a powerful depth encoding strategy is necessary for compression using 3-D motion models. Joint minimization of distortion and bit-rate measures gives optimal encoded depth, which has minimum distortion for a given bit-rate (or vice versa). By properly selecting a distortion criteria, the encoding of depth field is achieved without explicitly having the true depth, since this information is implicitly available in the intensities of consecutive frames. The encoded depth, which is a distorted and usually a smoother version of the true field, is definitely encoded with less number of bits with respect to the undistorted true depth. This is a desired situation in very low bit-rate coding, since the main purpose is efficient coding rather than finding the true values, while sacrificing from intensity distortion. In this study, the optimal depth fields are found with this aim. Although the number of bits to encode a dense depth field is still high, it should be noted that the structure of a rigid body has considerable amount of redundancy in time and hence very small number of bits should be required once the initial depth field is transmitted. Hence, as the experimental results indicate, 3-D motion models can be used for object-based video coding applications.

6. REFERENCES

[1] A. Zakhor and F. Lari "Edge-Based 3-D Camera Motion Estimation with Applications to Video Coding," *IEEE Trans. on Image Processing*, vol. 2, pp. 481–498, October 1993.

[2] H. Morikawa and H. Harashima "3D Structure Extraction Coding of Image Sequences," *Journal of Visual Communication and Image Representation*, vol. 2, pp. 332–344, December 1991.

[3] N. Diehl "Object-Oriented Motion Estimation and Segmentation in Image Sequences," *Signal Processing : Image Communication*, vol. 3, pp. 23–56, 1991.

[4] A.A. Alatan and Levent Onural "Object-based 3-D motion and structure estimation," in *Proceedings of IEEE Int. Conf. on Image Processing '95, Washington D.C., October*, pp. I 390–393, 1995.

[5] D. Tzovoras, N. Grammailidis and M. G. Strintzis "Depth Map Coding for Stereo and Multiview Image Sequence Transmission," in *Proceeedings of the Inter. Workshop on Stereoscopic and 3-D Imaging, Santorini, Greece, Sept 6-8*, pp. 75–80, 1995.

[6] J.K. Aggarwal and N. Nandhakumar "On the Computation of Motion from Image Sequences-A Review," *IEEE Proceedings*, vol. 76, pp. 917–935, August 1988.

[7] J. Weng, N. Ahuja and T. S. Huang "Matching Two Perspective Views," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 806–825, August 1992.

[8] M. Chang, M.I. Sezan and A.M. Tekalp "A Bayesian Framework for Combined Motion Estimation and Scene Segmentation in Image Sequences," in *Proceedings of IEEE ICASSP 94*, pp. 221–224, 1994.

[9] R.Y. Tsai and T.S. Huang "Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 13–27, January 1984.

[10] J. Weng, N. Ahuja and T.S. Huang "Optimal Motion and Structure Estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 864–884, September 1993.

[11] T. Cover. *Elements of Information Theory.* Wiley, 1991.

[12] W. Stadler. *Multicriteria Optimization in Engineering and in the Sciences.* Plenum Press, 1988.

[13] F. Heitz, P. Perez and P. Bouthemy "Multiscale Minimization of Global Energy Functions in Some Visual Recovery Problems," *CVGIP-Image Understanding*, vol. 59, pp. 125–134, January 1994.

Table 1: For different values of λ, Equation 2 is minimized to obtain Δ and $\mathcal{B}$ (with arbitrary $k = 0.5$) values. Bit-rate is obtained after encoding of the prediction error.

λ	Δ	$\mathcal{B}$	Bit-rate(bits/object)
1	33	9200	14928
5	60	4586	10312
10	65	4147	9752
50	93	2455	6288
100	118	2288	5656

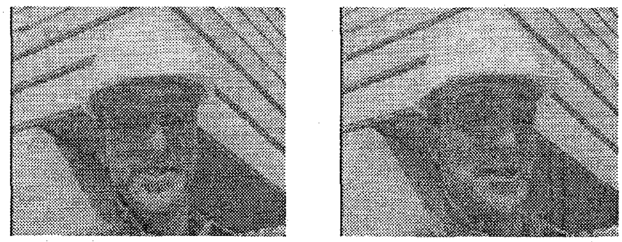

Figure 4: *100th* and *103th* frames of *Foreman*

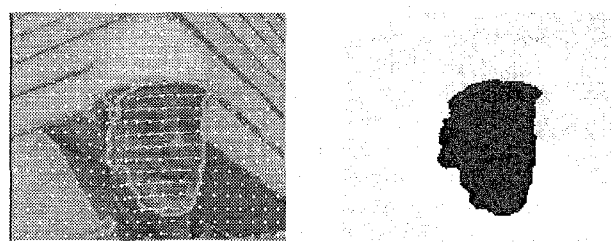

Figure 5: (a) 2-D motion estimation and (b) segmentation

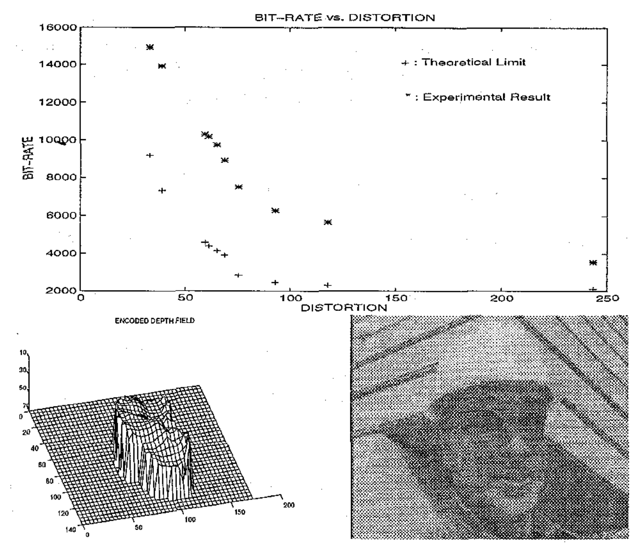

Figure 6: For the segmented head, (a) For different values of λ, corresponding rate-distortion pairs; (b) Encoded depth field and (c) reconstructed frame using the encoded depth field and motion parameters, for $\lambda = 5$