[2] R. C. Gonzalez and P. Wintz, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1987.
[3] N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Image Processing*. Berlin, Germany: Springer-Verlag, 1975.
[4] K. G. Beauchamp, *Applications of Walsh and Related Functions*. New York: Academic, 1984.
[5] S. L. Hurst, D. M. Miller, and J. C. Muzio, *Spectral Techniques in Digital Logic*. New York: Academic, 1985.
[6] J. L. Shanks, "Computation of the fast Walsh–Fourier transform," *IEEE Trans. Comput.*, vol. C-18, pp. 457–459, 1969.
[7] J. W. Manz, "A sequency-ordered fast Walsh transform," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 204–205, 1972.
[8] S. Boussakta and A. G. J. Holt, "Fast algorithm for calculation of both the Walsh–Hadamard and Fourier transforms (FWFTS)," *Electron. Lett.*, vol. 25, pp. 1352–1353, Sept. 28, 1989.
[9] S. C. Noble, "A comparison of hardware implementation of the Hadamard transform for real time image coding," in *Proc. Soc. Photo-Optical Instrumentation Engineers*, 1975, pp. 207–211.
[10] P. C. Ching and C. C. Goodyear, "Walsh-transform coding of the speech residual in RELP coders," *Proc. Inst. Elect. Eng. G*, vol. 131, no. 1, pp. 29–34, Feb. 1984. .
[11] Y. Tadokoro and T. Higuchi, "Conversion factors from Walsh coefficients to Fourier coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 231–232, Feb. 1983.

Estimation of Depth Fields Suitable for Video Compression Based on 3-D Structure and Motion of Objects

A. Aydın Alatan and Levent Onural

***Abstract*—Intensity prediction along motion trajectories removes temporal redundancy considerably in video compression algorithms. In three-dimensional (3-D) object-based video coding, both 3-D motion and depth values are required for temporal prediction. The required 3-D motion parameters for each object are found by the correspondence-based E-matrix method. The estimation of the correspondences—two-dimensional (2-D) motion field—between the frames and segmentation of the scene into objects are achieved simultaneously by minimizing a Gibbs energy. The depth field is estimated by jointly minimizing a defined distortion and bit-rate criterion using the 3-D motion parameters. The resulting depth field is efficient in the rate-distortion sense. Bit-rate values corresponding to the lossless encoding of the resultant depth fields are obtained using predictive coding; prediction errors are encoded by a Lempel–Ziv algorithm. The results are satisfactory for real-life video scenes.**

***Index Terms*—Dense depth estimation, depth encoding, motion analysis, object-based video coding, rate-distortion theory, 3-D motion, 3-D structure.**

I. Introduction

In very low bit-rate coding applications, the current trend is shifting from motion compensated discrete cosine transform (DCT) type algorithms, like MPEG-X, H.26X, to object-based methods [1]. In most of the current object-based algorithms, two-dimensional (2-D) motion models are used, although such motion models have limited performance due to lack of representation of three-dimensional (3-D) world dynamics. Currently, 3-D motion models are rarely used in video compression systems [1]–[5], and these approaches are usually far from representing general solutions. However, in such algorithms compression is still possible after removal of the temporal redundancy by predicting intensities along motion trajectories. Both 3-D motion and depth information are necessary to achieve this goal.

A 3-D motion model is the "simplest" way to describe any physical motion, especially when the moving object is rigid, because any rigid 3-D motion is represented by only six degrees of freedom, i.e., six parameters. Estimation of the 3-D motion parameters for a rigid body observed through two consecutive 2-D frames has well-developed solutions [6], [7] and, hence, this estimation problem can be easily overcome. Although depth estimation using these methods can be achieved, the obtained depth fields are usually sparse, whereas for coding purposes it is preferable to have a dense depth field in order to predict the intensities by motion compensation at each pixel. Given two 2-D consecutive video frames, one or more 3-D structures may give perfect intensity match by 3-D motion compensation. A structure that results in perfect intensity match, if it exists, may not be suitable for efficient encoding. Furthermore, one can find structures that are easier to code by allowing some intensity mismatch during the motion compensation. Estimating a dense depth field (structure) suitable for very low bit-rate video compression is the primary issue in this paper.

None of the current video coding methods with 3-D motion models propose a method for estimating a depth field that is suitable for encoding. Some depth encoding algorithms exist for stereo video coding applications [8] in which the depth field is simply obtained by using the disparity information between stereo frames. In these methods the obtained depth map is either DPCM-coded after quantization or fitted onto a wireframe [8]. However, such methods do not take distortion and bit rate into account simultaneously while estimating the depth field.

It should be noted that if the number of bits to encode the depth field is reduced to reach a target rate, some distortion in the depth field, compared to the one which yields perfect intensity matches, may be inevitable. Rate-distortion theory [9] gives a relationship between the minimum number of bits to encode a distorted symbol sequence from a source and the distortion between the true and encoded versions of that sequence. Using similar ideas, a lossy version of the depth field can be found by jointly minimizing the required number of bits and a distortion measure. Such approaches are also used to estimate 2-D motion vectors between video frames [10].

The main focus of this paper is to formulate a novel method for estimating (and thus generating) a depth field that is convenient for encoding. In order to estimate the desired depth field, the frames should be segmented into a number of moving objects and the 3-D motion parameters of the objects should be found. Dense 2-D motion vectors are needed for both object segmentation and correspondence-based 3-D motion estimation. In order to carry out simulations, a simultaneous 2-D motion estimation and segmentation algorithm, and a 3-D motion estimation algorithm, are proposed in Sections II-A and II-B, respectively. Moreover, in order to give an idea about the actual bit requirements associated with the coding of the estimated depth fields, a lossless encoder is utilized in Section IV. Algorithms in Sections II and IV are not the main concern of the paper; they cannot be claimed to have the best performance. However, they do give satisfactory results.

Manuscript received January 14, 1996; revised March 4, 1997. This work was supported by TÜBİTAK of Turkey under the COST 211 Project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Janusz Konrad.

The authors are with the Department of Electrical and Electronics Engineering, Bilkent University, TR-06533 Ankara, Turkey (e-mail: alatan@ee.bilkent.edu.tr; onural@ee.bilkent.edu.tr).

Publisher Item Identifier S 1057-7149(98)03997-9.

1057–7149/98$10.00 © 1998 IEEE

II. Motion Estimation

In this application, the E-matrix method [6], which requires robust 2-D motion estimates (correspondences) between consecutive frames as inputs, is chosen to estimate the rigid 3-D motions of objects and their depth variations. Using the E-matrix method, the depth values can only be estimated at the locations corresponding to those for the robust (usually sparse) 2-D motion vectors. These vectors are not only required for 3-D motion and depth estimation within the E-matrix method, but they are also utilized to segment the scene into a number of objects. Since moving object segmentation and motion estimation are coupled with each other [12], segmentation and finding correspondences are achieved simultaneously before 3-D motion and depth estimation.

A. 2-D Motion Estimation and Object Segmentation

Two-dimensional motion analysis using Gibbs formulation has been shown to be successful for both estimation [11] and segmentation [12]. The Gibbs energy function $\mathcal{U}$, which is the negative exponent of the exponential joint probability density function (pdf), can be formulated in terms of 2-D motion field $\mathcal{D}$, *segmentation* field $\mathcal{R}$ and *temporally unpredictable* (TU) regions $\mathcal{S}$, as follows:

$$\mathcal{U}(\mathcal{D},\mathcal{R},\mathcal{S} \mid \mathcal{I}_t,\mathcal{I}_{t-1}) = \mathcal{U}_n + \lambda_D\mathcal{U}_D + \lambda_R\mathcal{U}_R + \lambda_S\mathcal{U}_S. \quad (1)$$

In (1), the $\mathcal{U}_n$ term supports intensity matching between consecutive frames with correct 2-D motion vectors according to optical flow. The error measures of intensity matches can be higher than a predetermined threshold only in occlusion, i.e., TU regions. The $\mathcal{U}_D$ term favors smooth variations between neighboring 2-D motion vectors, except at object boundaries. The projections of the 3-D motions of rigid and even deformable bodies are expected to obey such a constraint. The $\mathcal{U}_R$ term supports objects that have projected broad regions on the 2-D image plane rather than some individual points. Similar to the $\mathcal{U}_R$ term, the $\mathcal{U}_S$ term supports $\mathcal{S}$ field to consist of regions. All λ's in (1) are constants that determine the weighting between these different terms. Further details of the energy terms in (1) can be found in [5]. A maximum *a posteriori* (MAP) estimate of the unknown 2-D motion field, segmentation field and TU regions can be obtained simultaneously by minimizing the energy function, $\mathcal{U}$. The $\mathcal{R}$ field segments the scene into the objects and then 3-D motion analysis is performed on these objects separately. However, it should be noted that this minimization is a nonconvex problem.

B. 3-D Motion Estimation

As shown in [7], for any rigid motion from time $t-1$ to t, the 3-D coordinates of object point $\mathbf{p}$ at time $t-1$ can be written in terms of $\mathbf{X_p}(t)$ as $\mathbf{X_p}(t-1) = \mathbf{R}\mathbf{X_p}(t) + \mathbf{T}$, where $\mathbf{R}$ is a 3×3 rotation matrix and $\mathbf{T}$ is a 3×1 translation vector. It should be noted that $\mathbf{R}$ and $\mathbf{T}$ do not reflect the "real" motion from time $t-1$ to t, but rather an "inverse" motion from time t to $t-1$. After *perspective projection* of the 3-D object points onto the 2-D image plane, the following equations are obtained [6]:

$$x_{\mathbf{p}}(t-1) = f \cdot \frac{r_{11}\cdot x_{\mathbf{p}}(t) + r_{12}\cdot y_{\mathbf{p}}(t) + r_{13}\cdot f + \frac{T_x\cdot f}{Z_{\mathbf{p}(\mathbf{x_p},t)}}}{r_{31}\cdot x_{\mathbf{p}}(t) + r_{32}\cdot y_{\mathbf{p}}(t) + r_{33}\cdot f + \frac{T_z\cdot f}{Z_{\mathbf{p}(\mathbf{x_p},t)}}}$$

$$y_{\mathbf{p}}(t-1) = f \cdot \frac{r_{21}\cdot x_{\mathbf{p}}(t) + r_{22}\cdot y_{\mathbf{p}}(t) + r_{23}\cdot f + \frac{T_y\cdot f}{Z_{\mathbf{p}(\mathbf{x_p},t)}}}{r_{31}\cdot x_{\mathbf{p}}(t) + r_{32}\cdot y_{\mathbf{p}}(t) + r_{33}\cdot f + \frac{T_z\cdot f}{Z_{\mathbf{p}(\mathbf{x_p},t)}}} \quad (2)$$

where f is the focal length of the camera, r_{ij} is an element of the rotation matrix, and (T_x, T_y, T_z) are the elements of the translation vector. $\mathbf{x_p}(t-1) = [x_{\mathbf{p}}(t-1)\ \ y_{\mathbf{p}}(t-1)]^T$ are the projected 2-D coordinates of the object point $\mathbf{p}$ at time $t-1$ (Fig. 1). Notice that $Z_{\mathbf{p}}(\mathbf{x_p}, t)$ is the third component of the vector $\mathbf{X_p}(t)$ whose perspective projection gives $\mathbf{x_p}(t)$ and is simply called the *depth value*. Equation (2) shows that the displacements of pixels on the 2-D image plane depend on both the 3-D motion parameters (r_{ij} and $T_{x,y,z}$) and the depth values.

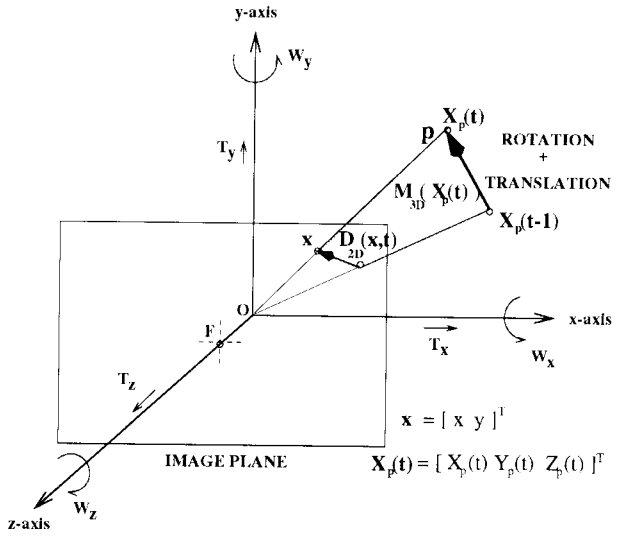


Fig. 1. Three-dimensional coordinate system.

There are different approaches to the 3-D motion and structure estimation problem, and it is shown that the linear E-matrix approach [6] has given good results for estimating global motion of a camera and depth of the stationary environment using some 2-D point correspondences between frames. In the E-matrix approach, the depth term is simply dropped from (2), and the resulting single equation without depth information is solved linearly with the help of at least eight robust correspondences for 3-D motion parameters [6]. In object-based coding applications, the E-matrix method can be applied to individual objects rather than to the whole image by using the segmented 2-D motion vectors obtained as in Section II-A. These 2-D motion vectors give more correspondences than the minimum required of eight. However, in order to improve the performance of this error-prone algorithm, instead of using all the correspondences ($\mathcal{D}$ field), "reliable" estimates are chosen by simply thresholding their low intensity matching error and high spatial image gradient. Such an approach is almost equivalent to finding good matches between edges and corners. Since 2-D motion vectors have already been found in segmentation step, this selection mechanism is more efficient rather than applying an extra feature-matching step. Finally, a rotation matrix and a translation vector are obtained for each segmented object. Using the estimated 3-D motion parameters and available 2-D correspondences, depth values can be obtained at the corresponding locations using (2).

III. Depth Estimation in Rate-Distortion Sense

Since any 3-D scene can be assumed to be an output of a random source, the depth field of the scene will be a random field with a corresponding probability. The assignment of probability to a depth field is meaningful if it matches the frequency of occurrence of that field in the real world; it is assumed that such an assignment is made. Using this probability measure, the number of bits required to encode any depth field can be determined according to the basic principles of information theory [9]. Rate-distortion theory seeks

the minimum achievable rate for a source to be encoded under a distortion constraint. Based on this theory, an algorithm to find the dense depth field to be encoded can be found. A possible approach is to minimize a function $\mathcal{J}(\Delta, \mathcal{B})$ that takes both distortion Δ and bit-rate $\mathcal{B}$ into account, with respect to the depth field to be encoded. There are many different ways to approach this *vector optimization* problem; the method of *objective weighting* [13] is one possible choice, where $\mathcal{J}(\Delta, \mathcal{B}) = \Delta + \lambda_0 \cdot \mathcal{B}$, with λ_0 being a constant which reflects weighting between two different quantities Δ and $\mathcal{B}$. Before achieving joint optimization of bit rate and depth, a distortion criterion and a measure of bit rate should be defined.

A. Distortion Criterion

It is possible to define the distortion between the true and reconstructed depth values using input frame intensities. The distortion criterion Δ can be defined as the average error between the original and reconstructed frames computed region-by-region, as follows:

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

where N is the total number of object pixels in region R_i. I_t is the original frame, which can also be written as

$$I_t(\mathbf{x}) = I_{t-1}(\mathbf{x} - \mathbf{D}_{2D}(\mathbf{x}, t)) \tag{4}$$

with the assumptions that the corresponding point is in a noise-free nonoccluding region with no illumination change, and the object is opaque. As can be seen in Fig. 1, for an object point $\mathbf{p}$, $\mathbf{D}_{2D}(\mathbf{x}, t)$ is equal to

$$\mathbf{D}_{2D}(\mathbf{x}, t) = \mathcal{P}[\mathbf{M}_{3D}(\mathbf{X}_{\mathbf{p}}(t))]\big|_{\mathcal{P}[\mathbf{X}_{\mathbf{p}}(t)]=\mathbf{x}} \tag{5}$$

where $\mathcal{P}$ denotes the perspective projection. Consequently, $\mathbf{D}_{2D}(\mathbf{x}, t)$ is a function of $Z(\mathbf{x}) = Z_{\mathbf{p}}(t)$, which is the depth value for perfect intensity match corresponding to location $\mathbf{x}$. The reconstructed frame $\hat{I}_t$ can be expressed similarly to (4) by using the resultant depth value $\hat{Z}(\mathbf{x})$ that would yield $\hat{\mathbf{D}}_{2D}(\mathbf{x}, t)$. Hence, (3) defines the distortion in a nonlinear way between the resulting depth field and the depth field which would give a perfect match.

B. Bit Rate of Encoded Depth

In many indoor scenes, objects normally have smooth depth variations, except at their boundaries. Although other smoothness definitions are possible, a Gibbs energy taking this observation into account can be written as

$$\mathcal{U}_Z(\mathcal{Z}) = \sum_{\mathbf{x} \in R_i} \sum_{\mathbf{x}_c \in \eta_{\mathbf{x}}} (\hat{Z}(\mathbf{x}) - \hat{Z}(\mathbf{x}_c))^2 \tag{6}$$

where the sum is over all points $\mathbf{x}$ of the ith object, segmented by the region R_i; $\eta_{\mathbf{x}}$ is the neighborhood of $\mathbf{x}$. The required number of bits, $\mathcal{B}$, to encode the depth field is simply equal to $-\log_2(\mathbf{P}(\mathcal{Z}))$, where $\mathbf{P}(\mathcal{Z})$ is the probability distribution of the depth field. Hence, using (6)

$$\mathcal{B} = k \cdot (\log_2 e) \cdot \left(\sum_{\mathbf{x} \in R_i} \sum_{\mathbf{x}_c \in \eta_{\mathbf{x}}} (\hat{Z}(\mathbf{x}) - \hat{Z}(\mathbf{x}_c))^2 \right) + c(k) \tag{7}$$

where k is the Gibbs energy constant, and $c(k)$ constant does not depend on $\mathcal{Z}$.

C. Minimization of the Encoding Criterion

Distortion and bit-rate are jointly minimized with respect to $\mathcal{Z}$ and this is written as

$$\min_{\mathcal{Z}} \left\{ \left(\frac{1}{N} \sum_{\mathbf{x} \in R_i} (I_t(\mathbf{x}) - I_{t-1}(\mathbf{x} - \hat{\mathbf{D}}_{2D}(\mathbf{x}, t)))^2 \right) + \lambda \left(\sum_{\mathbf{x} \in R_i} \sum_{\mathbf{x}_c \in \eta_{\mathbf{x}}} (\hat{Z}(\mathbf{x}) - \hat{Z}(\mathbf{x}_c))^2 \right) \right\}. \tag{8}$$

Since $c(k)$ does not depend on $\mathcal{Z}$, it is removed from (8). The constants k and $\log_2(e)$ are multiplied with λ_0, and this product is defined as λ. For different choices of λ, different values for rate and distortion can be obtained. For a given bit rate (or distortion), the corresponding distortion (or bit rate) is optimal, if the defined pdf-model for the depth field matches the frequency of occurrence of such a field in the real world. λ may be specified externally or, equivalently, some external constraints on distortion or bit rate may be used to imply a λ.

After minimizing (8), a depth field is obtained. Compared to the depth fields that are estimated using different algorithms, this field is more suitable for encoding since bit rate and distortion are minimized simultaneously. In other words, the best bit-rate savings are obtained for a given distortion. This is a significant result with useful applications in low bit-rate video coding.

IV. Entropy Coding of Depth

Lossless (entropy) coding of the resultant depth field is essential. Since the depth field found in Section III-C is optimal in the sense of minimizing (8), any alteration in bit rate (or distortion) should be achieved during the minimization of (8) instead of a subsequent lossy encoder. Note that higher values of λ would yield lower bit rate.

Although finding a depth field for efficient encoding is explained, the method by which this depth field can be encoded to approach the theoretical bit-rate (entropy) limit is still not specified. Since it is impossible to give a codeword to all existing depth fields according to their probabilities, another coding strategy must be followed. In order to get an idea about the actual bit requirements associated with the coding of the estimated depth fields, a heuristic lossless encoder is proposed as follows. Predictive coding is applied to remove the redundancy existing in the depth field. Each depth value is predicted from its casual horizontal and vertical neighbors ($\mathbf{x_{hor}}$ and $\mathbf{x_{ver}}$, respectively) as $\hat{Z}_e(\mathbf{x}) = 0.5(\hat{Z}(\mathbf{x_{ver}}) + \hat{Z}(\mathbf{x_{hor}}))$. The prediction error is coded in a lossless fashion using a Lempel–Ziv algorithm [9]. This predictor can be justified by the fact that our quadratic energy function leads to a linear predictor, and that the symmetry between horizontal and vertical dependencies favors equal weighting of the neighbors.

V. Experimental Results

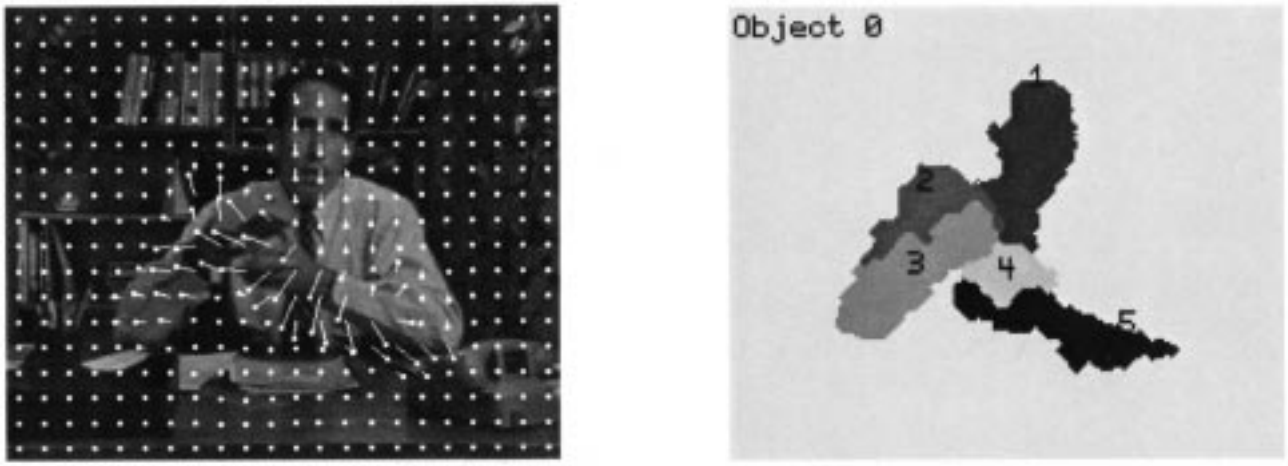
Two frames (10 and 16) from the salesman sequence are used to test the proposed algorithm (Fig. 2). In these frames, the man moves both of his arms and his head. The size of the frames is 176 $\times$144 (QCIF) and it is assumed that the unknown focal length of the camera is equal to 250 *pixels* (this selection corresponds to approximately 50 mm focal length of a 35 mm camera). Although this assumption is coarse, it gives acceptable results. Similar to Fig. 1, it is assumed that the optical axis passes through the center of these images.

The results of 2-D motion estimation are shown in Fig. 3(a). The minimization of (1) is achieved by using the multiscale constrained relaxation (MCR) [5] algorithm with four scales and two iterations of iterated conditional modes (ICM) [11] at each scale. ICM requires good initial estimates for better performance. Hence, a hierarchical



(a) (b)

Fig. 2. Original (a) tenth and (b) sixteenth frames of salesman sequence.



(a) (b)

Fig. 3. Experimental results of 2-D motion analysis and segmentation for salesman sequence. (a) Needlegram of the 2-D motion estimates. (b) Segmentation field areas.

block matching algorithm is used to initialize the 2-D motion field. Similarly, in order to improve segmentation, the result of a region-based segmentation algorithm [14] is used as an initial estimate for the segmentation field before minimization. After the minimization, the resulting segmentation of the moving objects is shown in Fig. 3(b). Objects 2 and 5 represent the occluding regions of right and left arms, respectively. After obtaining a set of reliable 2-D correspondences, which have high intensity gradient and low intensity matching errors, the E-matrix is solved using least squares for this sparse set of 2-D motion vectors. The rotation matrices and translation vectors are found for each segmented object, respectively. A sparse set of depth values is also obtained as a result of the E-matrix method.

The depth values that are obtained from the E-matrix method are used as initial estimates for the proposed depth field estimation method. Minimization of (8) is performed using the MCR method for various values of λ. Table I shows, for each object, the distortion values as well as the bit-rate values after encoding of the depth prediction error using Lempel–Ziv algorithm. As expected, the distortion decreases as the number of bits to encode the depth field increases. The last row of Table I is related to the encoding of the dense depth values that are obtained using the plain E-matrix method. The dense 2-D correspondence set is utilized in the depth estimation step of the E-matrix method to obtain a dense depth map. The proposed entropy coding method explained in Section IV is used to encode this dense depth field resulting from the E-matrix method. The simulation results in Table I show that the proposed depth estimation algorithm performs better than the E-matrix method. Although both algorithms use the same 3-D motion parameters, the depth field of the proposed method yields superior performance, for any λ value, over the E-matrix method in the rate-distortion sense.

In Fig. 4, the reconstructed current frame, which is obtained by using the estimated 3-D motion parameters, previous frame and the encoded depth field, is shown for $\lambda = 100$. The TU areas have been segmented using (1); the visual quality of the reconstructed frame is acceptable. A significant part of object 5 is successfully segmented as TU. As expected, the projections of the 3-D motions are meaningful for the rigid objects 1, 3, and 4. The obtained depth fields for the objects are also represented in the same figure for the same value of λ.

Due to nonlinear minimization, the computational complexity of the encoding procedure is significant. However, compared to the well-known Markov random field (MRF) based 2-D motion estimation algorithms [11], the complexity is lower by a factor of $N \times N$ to N, where N is the number of quantized levels of the search space for each unknown. Therefore, the computational complexity

TABLE I
EXPERIMENTAL RESULTS FOR SALESMAN SEQUENCE. FOR EACH OBJECT AND DIFFERENT VALUES OF λ, (8) IS MINIMIZED TO OBTAIN THE CORRESPONDING Δ AND BIT-RATE VALUES

	Object 1		Object 2		Object 3		Object 4		Object 5	
λ	Δ	Bits	Δ	Bits	Δ	Bits	Δ	Bits	Δ	Bits
1	70.6	5432	118.8	3904	221.8	6592	23.3	2536	2316.2	1240
10	155.4	2656	191.1	3601	227.6	5696	26.8	2320	2317.5	1200
50	176.4	1472	199.2	3248	258.3	4704	34.4	2201	2317.3	1144
100	177.2	1402	203.4	3224	281.6	4344	49.2	2152	2318.5	1136
1000	184.3	1304	836.2	2512	442.9	2608	216.5	1848	2318.2	1168
10000	201.2	1264	1524.9	1272	590.6	1288	981.8	1296	2319.0	1160
E-Matrix	200.4	5392	881.7	3888	623.2	8264	1446.8	2432	2262.5	1280



(a) (b)

Fig. 4. Results of 3-D motion and depth estimation for salesman sequence. (a) Motion-compensated current frame using 3-D motion parameters and encoded depth field (TU areas are segmented). (b) Needlegram of 2-D projection of 3-D motion. Encoded depth field with (c) mesh and (d) intensity representations.

is less prohibitive compared to MRF-based 2-D motion estimation algorithms.

VI. CONCLUSION

A novel depth estimation algorithm that generates dense depth fields that are easy to encode, is proposed. The utilization of such an algorithm within object-based video coders based on 3-D motion and structure, should be more preferable than conventional depth estimation algorithms, since bit rate and distortion are taken into account together. During experiments, it was observed that better compression and quality can be obtained whenever the 3-D motion parameter set represents an acceptable motion between the two frames. Hence, 3-D motion estimation is a critical factor that determines the overall performance. The simulation results show that the required number of bits to encode a depth field is still too high for very low bit-rate applications. However, it should be noted that the encoded depth field belongs to a rigid object and the temporal redundancy in this field is high. Therefore, the real benefits will be achieved when longer sequences with more than two frames are encoded.

REFERENCES

[1] M. Hotter and R. Thoma, "Image segmentation based on object oriented mapping parameter estimation," *Signal Process.*, vol. 15, pp. 315–334, 1988.

[2] A. Zakhor and F. Lari, "Edge-based 3-D camera motion estimation with applications to video coding," *IEEE Trans. Image Processing*, vol. 2, pp. 481–498, Oct. 1993.

[3] H. Morikawa and H. Harashima, "3D structure extraction coding of image sequences," *J. Vis. Commun. Image Represent.*, vol. 2, pp. 332–344, Dec. 1991.

[4] N. Diehl, "Object-oriented motion estimation and segmentation in image sequences," *Signal Process.: Image Commun.*, vol. 3, pp. 23–56, 1991.

[5] A. A. Alatan and L. Onural, "Object-based 3-D motion and structure estimation," in *Proc. IEEE Int. Conf. Image Processing'95*, Washington D.C., pp. I 390–393.

[6] J. Weng, N. Ahuja, and T. S. Huang, "Optimal motion and structure estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 864–884, Sept. 1993.

[7] T. S. Huang and A. N. Netravali, "Motion and structure from feature correspondences: A review," *Proc. IEEE*, vol. 82, pp. 252–268, Feb. 1994.

[8] D. Tzovoras, N. Grammailidis, and M. G. Strintzis, "Depth map coding for stereo and multiview image sequence transmission," in *Proc. Int. Workshop on Stereo and 3-D Imaging*, Santorini, Greece, 1995, pp. 75–80.

[9] T. Cover, *Elements of Information Theory*. New York: Wiley, 1991.

[10] D. Tzovaras and M. G. Strintzis, "Motion estimation using rate distortion theory for very low bit-rate image sequence coding," in *Proc. Int. Conf. Telecommunications*, Istanbul, Turkey, Apr. 1996, vol. 2, pp. 608–611.

[11] J. Konrad and E. Dubois, "Bayesian estimation of motion vector fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 910–927, Sept. 1992.

[12] M. Chang, M. I. Sezan, and A. M. Tekalp, "A Bayesian framework for combined motion estimation and scene segmentation in image sequences," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing'94*, pp. 221–224.

[13] W. Stadler, *Multicriteria Optimization in Engineering and in the Sciences*. New York: Plenum, 1988.

[14] M. J. Biggar, O. J. Morris, and A. G. Constantinides, "Segmented-image coding: Performance comparison with the discrete cosine transform," *Proc. Inst. Elect. Eng.*, vol. 135, pp. 121–132, Apr. 1988.