JOINT MOTION ESTIMATION/SEGMENTATION FOR OBJECT-BASED VIDEO CODING

Soo-Chul Han, Lilla Böröczky*, and John W. Woods
Center for Image Processing Research & ECSE Department
Rensselaer Polytechnic Institute
Troy NY 12180-3590, USA
e-mail: sooch@ipl.rpi.edu, lboroczky@vnet.ibm.com, woods@ecse.rpi.edu

ABSTRACT

A video coding scheme is presented in which the coding is performed on individual moving objects. A Markov Random Field model is employed in finding the motion and boundaries of the objects. By guiding the object segmentation process with the spatial color information, meaningful objects representative of the real video scene are extracted. Furthermore, this enables a systematic treatment in handling the covered/uncovered regions, as well as the appearance/disappearance of moving objects. The rate for transmitting object motion and boundary is greatly reduced by use of temporal updating. The interior coding is performed by object-based subband decomposition. Simulations indicate promising results for low bitrate applications.

1 INTRODUCTION

With very low bitrate coding emerging as an important field of video compression, object-based video coding has received considerable attention recently [1, 2, 3, 4]. This is evidenced further by the on-going MPEG-4 standardization process. In object-based analysis/synthesis coding the scene content is analyzed at the transmitter. The reconstruction is carried out at the receiver by synthesizing the image sequence using the transmitted information. Consequently, efficient and reliable extraction of moving objects and accurate estimation of their motion are essential. Although numerous publications have appeared on the subject, few of them carry out the entire analysis-coding process from start to finish. Thus, the widespread belief that object-based methods could outperform standard techniques at low rates has yet to be firmly established. We attempt to take the step in that direction with new ideas in both the motion analysis and the source encoding procedures.

In this paper an enhanced motion estimation and segmentation algorithm is proposed. It is based on a coupled MRF model with the usual constraints such as spatiotemporal smoothness and consistency of the motion vectors and segmentation labels [2, 5, 6]. In addition, we constrain the motion boundaries to coincide with spatial intensity boundaries. We link the segmentations at consecutive frames to construct objects in the space-time domain. Furthermore, extraction of the covered/uncovered regions is more robust and meaningful because the segmentation represents objects in the real video scene. A relatively smooth motion field is obtained within each object. The contour information can also be encoded efficiently by means of temporal updating using the linked object labels. This alleviates the bottleneck created by the contour information in region-based coding schemes [7] where contours are coded frame by frame.

The rest of this paper is organized as follows. The joint motion estimation/segmentation algorithm based on the MRF model is presented in detail in section 2. We compress the motion fields by fitting an affine parametric model to each object. The parametric representation leads to an efficient temporal updating technique to encode the contour information. These two methods are described in section 3. In section 4, we discuss the interior coding (color/texture) of objects, an important ingredient of object-based coding systems which has been overlooked for the most part. We introduce object-based motion compensated subband/wavelet coding, and present simulation results of our overall coding system at low bitrates. Section 5 contains concluding remarks and discusses an alternative object-based encoding method that is currently under investigation.

2 MOTION ANALYSIS

2.1 Problem formulation

At time t, let $\mathbf{I}^t$ represent the discretized sequence of images, $\mathbf{d}^t$ the motion field, and $\mathbf{z}^t$ the segmentation field consisting of numerical labels at every pixel, each label representing one moving object. Then, the goal of motion estimation/segmentation is to find $\{\mathbf{d}^t, \mathbf{z}^t\}$ given $\mathbf{I}^t$ and $\mathbf{I}^{t-1}$. We adopt the maximum *a posteriori* (MAP) formulation:

$$\{\hat{\mathbf{d}}^t, \hat{\mathbf{z}}^t\} = \arg \max_{\{\mathbf{d}^t, \mathbf{z}^t\}} p(\mathbf{d}^t, \mathbf{z}^t | \mathbf{I}^t, \mathbf{I}^{t-1}) \qquad (1)$$

[*] Now with IBM Corporation, Endicott, NY

which can be rewritten via Bayes rule as

$$\{\hat{\mathbf{d}}^t, \hat{\mathbf{z}}^t\} = \arg \max_{\{\mathbf{d}^t, \mathbf{z}^t\}} p(\mathbf{I}^{t-1}|\mathbf{d}^t, \mathbf{z}^t, \mathbf{I}^t) p(\mathbf{d}^t|\mathbf{z}^t, \mathbf{I}^t) p(\mathbf{z}^t|\mathbf{I}^t) \tag{2}$$

Given the formulation of (2), the rest of the work amounts to specifying the probability densities (or the corresponding energy functions) involved and solving for the solution.

2.2 Probability models

The first term on the right-hand side of (2) is the likelihood functional that describes how well the observed images match the motion field data. We adopt the additive noise model used in [8]:

$$U_l(\mathbf{I}^{t-1}|\mathbf{d}^t, \mathbf{I}^t) = \sum_{\mathbf{x}} (\mathbf{I}^t(\mathbf{x}) - \mathbf{I}^{t-1}(\mathbf{x} - \mathbf{d}^t(\mathbf{x})))/2\sigma^2 \tag{3}$$

The *a priori* density of the motion, $p(\mathbf{d}^t|\mathbf{z}^t, \mathbf{I}^t)$, enforces prior constraints on the motion field. We adopt a coupled MRF model to govern the interaction between the motion field and segmentation field both spatially and temporally. The corresponding energy function is given as

$$\begin{aligned} U_d(\mathbf{d}^t|\mathbf{z}^t) = \\ \lambda_1 \sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}} \| \mathbf{d}^t(\mathbf{x}) - \mathbf{d}^t(\mathbf{y}) \|^2 \delta(z^t(\mathbf{x}) - z^t(\mathbf{y})) \\ + \lambda_2 \sum_{\mathbf{x}} \| \mathbf{d}^t(\mathbf{x}) - \mathbf{d}^{t-1}(\mathbf{x} - \mathbf{d}^t(\mathbf{x})) \|^2 \\ - \lambda_3 \sum_{\mathbf{x}} \delta(z^t(\mathbf{x}) - z^{t-1}(\mathbf{x} - \mathbf{d}^t(\mathbf{x}))). \end{aligned} \tag{4}$$

where $\delta(\cdot)$ refers to the usual Kronecker delta function, $\| \cdot \|$ is Euclidean norm in $\mathbf{R}^2$, and $\mathcal{N}_{\mathbf{x}}$ indicates a spatial neighborhood system with respect to $\mathbf{x}$. The first two terms of (4) are similar to those in [2], while the third term encourages consistency of the object labels along motion trajectories. This constraint allows a framework for the object labels to be linked in time.

The object label field is also modeled in a novel manner so that the object discontinuities coincide with spatial intensity boundaries. The energy function is modeled as

$$U_z(\mathbf{z}^t|\mathbf{I}^t) = \sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}} V_c(z(\mathbf{x}), z(\mathbf{y})|\mathbf{I}^t), \tag{5}$$

where the clique potential is given by

$$V_c(z(\mathbf{x}), z(\mathbf{y})|\mathbf{I}^t) = \begin{cases} -\gamma & \text{if } z(\mathbf{x}) = z(\mathbf{y}),\ s(\mathbf{x}) = s(\mathbf{y}) \\ 0 & \text{if } z(\mathbf{x}) = z(\mathbf{y}),\ s(\mathbf{x}) \neq s(\mathbf{y}) \\ +\gamma & \text{if } z(\mathbf{x}) \neq z(\mathbf{y}),\ s(\mathbf{x}) = s(\mathbf{y}) \\ 0 & \text{if } z(\mathbf{x}) \neq z(\mathbf{y}),\ s(\mathbf{x}) \neq s(\mathbf{y}) \end{cases} \tag{6}$$

Here, $\mathbf{s}$ refers to the spatial segmentation field that is pre-determined from $\mathbf{I}$. A simple region-growing method [9] was used in our experiments. This slightly more complex model ensures that the moving object segments have some sort of spatial cohesiveness as well. This can be a very important property in certain coding situations.

2.3 Solution

Due to the equivalence of MRFs and Gibbs densities, the MAP solution amounts to a minimization of the sum of potentials given by (3), (4), and (5). To ease the computation, a two-step iterative procedure [6] is implemented, where the motion and segmentation fields are found in an alternating fashion assuming the other is given. Mean field annealing [8] is used for the motion field estimation, while the object label field is found by a deterministic iterated conditional modes (ICM) algorithm.

3 MOTION/CONTOUR CODING

Because of the spatial smoothness of the resulting motion field within each object, the object motion could be represented efficiently by parametrization. We chose the six parameter affine model $\mathbf{p}_i^T = [a_{xx} a_{xy} a_{yx} a_{yy} t_x t_y]$ for each object i given by

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{7}$$

where (v_x, v_y) refers to the motion field. The $\mathbf{p}_i$'s can be found by least-squares estimation using the dense motion field for each object. This amounts to fitting a plane in the velocity space. Furthermore, the affine parameters for each object are predicted from the stored parameters of the previous frame. Other waveform coding techniques could be used to take advantage of the spatial and temporal correlation.

The contour information describing the object boundaries is encoded with a novel efficient algorithm. First, for any "new" object, an intra-frame lossless chain coding scheme is implemented. For an object n that is "continuing" at frame t, the location and shape is first predicted from $t-1$ by finding and mapping the affine *displacement* parameters from $t-1$ to t. These parameters can be derived from (7) without explicit computation by assuming that the velocity fields and displacement fields are equivalent, Then at every site in t, we search for the precedent in $t-1$ by trying all the $\mathbf{p}_i$'s using (7). If we point back to object n at $t-1$, and this matches our predicted label for that site, label n is selected. For the remaining pixels, mostly at the boundaries, the ambiguity is resolved by transmitting 1-bit flags (assuming only 2 objects share the boundary). This method proved efficient because the number of such pixels to update usually turned out to be small.

4 VIDEO CODING RESULTS

In object-based video coding, each frame is segmented into the moving objects, and the coding and decoding

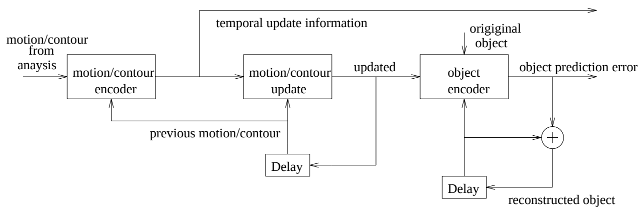


Figure 1: Object encoder

is done object-by-object. The interior coding of objects following motion compensation is performed using a region-based subband scheme, first introduced in [10] for still image compression. A spatial subband analysis/synthesis is performed on each object, with entropy coding using uniform quantizers on the subband coefficients. The overall encoding scheme for one object is shown in Fig 1.

Simulations were performed on the QCIF resolution sequences *Miss America* and *Carphone* at a frame rate of 7.5 Hz. The motion estimation/segmentation produced 5-10 objects for *Miss America* and 17-23 objects for *Carphone*. Observing Fig 2, the object segments indeed were temporally tracked, and the appearance of new moving objects is properly handled. The hair is considered a part of the stationary background in the beginning, and gets labeled as a moving object when it starts to move to the right. The new label is assigned within the context of the MRF formulation. The hair is again merged into the background when it stopped moving. In Fig 3, we can see that our motion estimates provided a smooth motion field that adhered to the true motion. The bits among the objects were allocated according to the residual variances and spatial sizes. Our results were compared with standard 16x16 block matching and DCT encoding, using the H.263 variable length code table [11]. Results in terms of decoded PSNR at 12 kbps for *Miss America* and 24 kbps for *Carphone* are summarized below. The results of our object-based encoder were visually more pleasing, with reduced blurriness and no blocking effects, as demonstrated in Fig 3.

	proposed	block-based
Miss America	35.40 dB	35.31 dB
Carphone	30.12 dB	29.30 dB

5 CONCLUSIONS

We have introduced new results on object-based video coding. An improved motion estimation/segmentation algorithm enables the extraction of moving objects that correspond to the true scene. By following the objects in time, the object motion and contour can be encoded efficiently with temporal updating. The interior of the objects are encoded by 2-D subband analysis/synthesis. No *a priori* assumptions about the image content or motion is needed.

A natural extension to our proposed encoder is 3-D subband coding of each moving object, in light of the fact that our motion estimation and segmentation provides us with a time-space segmentation of each object. Thus, each object can be decomposed into temporal frequencey bands by filtering along the motion trajectories within the boundaries. The filtering in theory can begin when the object first appears and continue until it disappears or stops moving. The resulting temporal frequency bands are arbitrarily shaped regions, for which spatial subband schemes described earlier can be used. This leads to a true 3-D subband decomposition of the objects in the space-time domain.

References

[1] H Mussman, M Hotter, and J Ostermann, "Object-oriented analysis-synthesis coding of moving images", *Signal Processing: Image Communications*, vol. 1, no. 2, pp. 117–138, Oct. 1989.

[2] C Stiller, "Object-oriented video coding employing dense motion fields", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing.* 1994, vol. V, pp. 273–276, Adelaide, Australia.

[3] J Wang and E Adelson, "Representing moving images with layers", *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 625–638, Sept. 1994.

[4] Y Yokoyama, Y Miyamoto, and M Ohta, "Very low bit rate video coding using arbitrarily shaped region-based motion compensation", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 500–507, Dec. 1995.

[5] P Bouthemy and E Francois, "Motion segmentation and qualitative dynamic scene analysis from an image sequence", *International Journal of Computer Vision*, vol. 10, no. 2, pp. 157–182, 1993.

[6] M Chang, I Sezan, and A Tekalp, "An algorithm for simultaneous motion estimation and scene segmentation", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing.* 1994, vol. V, pp. 221–224, Adelaide, Australia.

[7] V Dang, A Mansouri, and J Konrad, "Motion estimation for region-based video coding", in *Proc. IEEE Int. Conf. Image Processing.* 1995, pp. 189–192, Washington, DC.

[8] J Zhang and G G Hanauer, "The application of mean field theory to image motion estimation", *IEEE Trans. Image Process.*, vol. 4, pp. 19–33, 1995.

[9] Robert Haralick and Linda Shapiro, *Computer and Robot Vision*, Addison-Wesley Pub. Co., Reading, MA, 1992.

[10] H J Barnard, *Image and Video Coding Using a Wavelet Decomposition*, PhD thesis, Delft University of Technology, The Netherlands, 1994.

[11] ITU-T Recommendation H.263, *Video coding for low bitrate communication*, Nov. 1995.

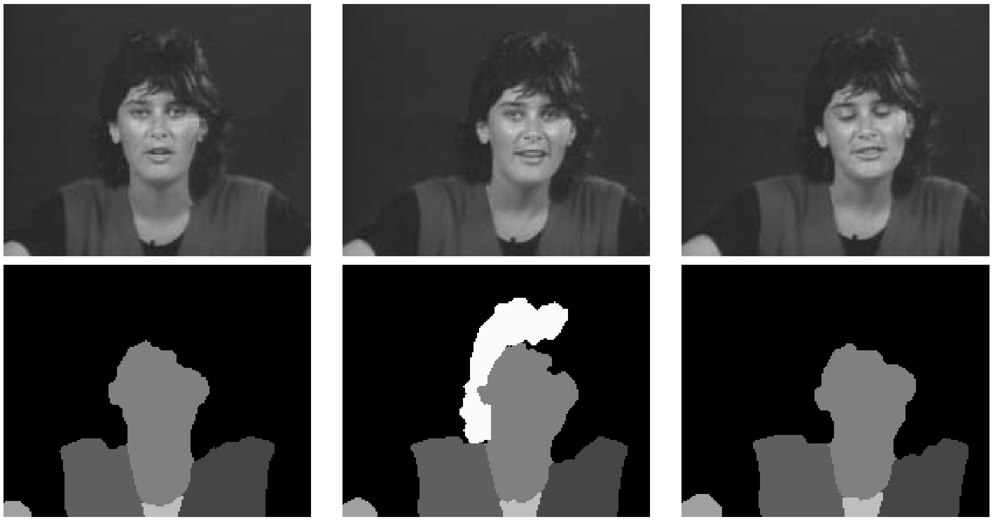
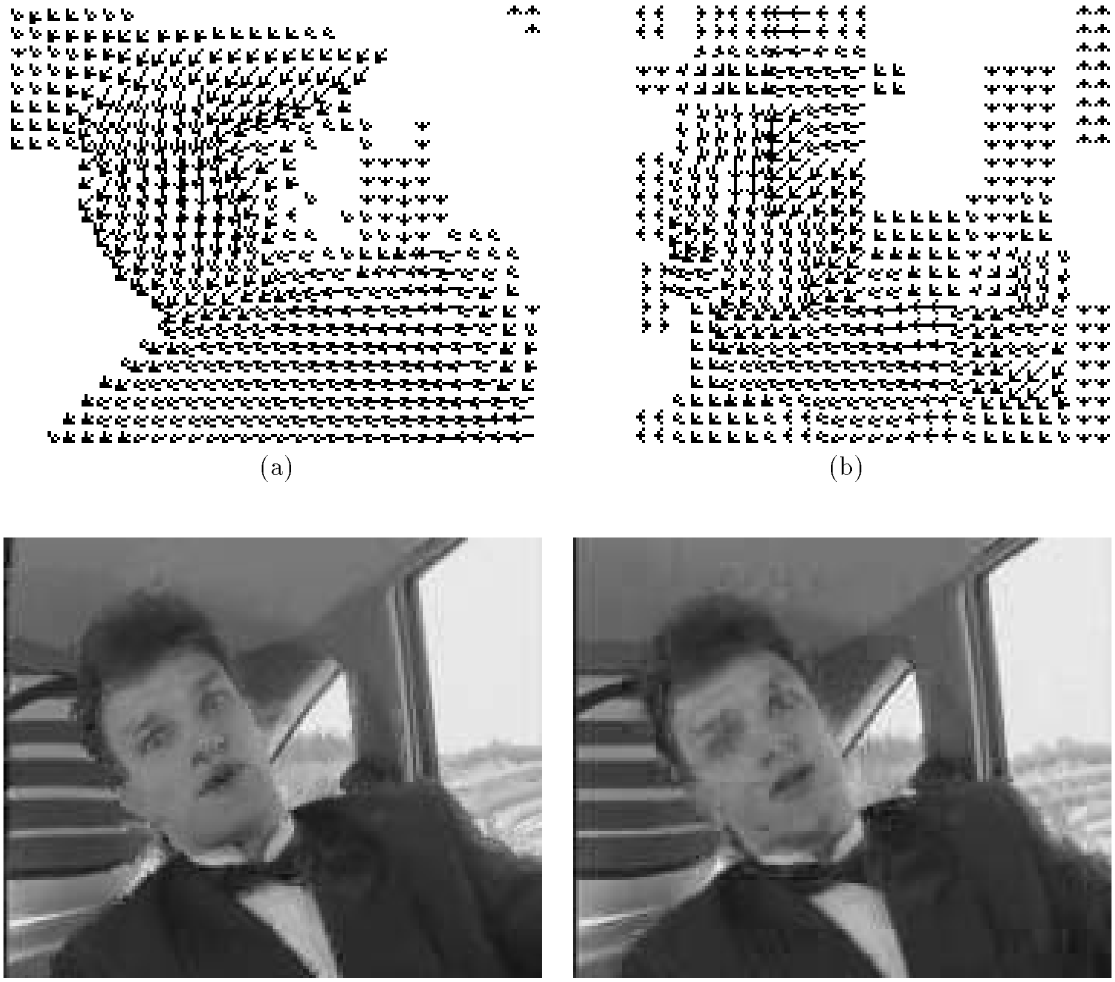


Figure 2: Top row: *Miss America* frames 60, 84, and 92. Second row: corresponding object label field



(a)

(b)

(c)

(d)

motion field by (a) MRF model (b) hierarchical block matching
decoded frames (c) object-based coding (PSNR 30.8 dB) (d) block-based coding (PSNR 30.4 dB)

Figure 3: Comparison of motion estimation and coding: *Carphone* frame 88