Perceptually weighted optical flow for motion-based segmentation in MPEG-4 paradigm

J. Malo*, J. Gutierrez†, I. Epifanio† and F.J. Ferri†

(*) Dept. d'Òptica, (†) Dept. d'Informàtica, Universitat de València
Dr. Moliner, 50. 46100 Burjassot, València. Spain
jesus.malo@uv.es

In the MPEG-4 paradigm, the sequence must be described in terms of meaningful objects. This meaningful, high-level, representation should emerge from low-level primitives such as optical flow and prediction error which are the basic elements of previous-generation video coders. The accuracy of the high-level models strongly depends on the robustness of the primitives used. In this paper, we show how perceptual weighting in optical flow computation gives rise to better motion estimates which consistently improve motion-based segmentation compared to equivalent unweighted motion estimates.

The authors wish to thank E.H. Adelson for his fruitful comments. This work has been supported by CICYT (TIC) projects 1FD97-0279 and 1FD97-1910

Introduction.

MPEG-4 recommendations stress the relevance of high-level sequence representation [1]. Motion-based segmentation algorithms are then crucial to identify meaningful regions and to obtain models to predict their evolution. In this context, it is interesting to analyze whether the low-level motion information used in the previous generation video coders may be useful as a starting point for the segmentation. This is not a trivial question because the Block Matching Algorithms (BMA) used in MPEG-1,2 or H.263 for motion estimation, were not intended to give a qualitatively meaningful flow, but a good prediction tool in a rate-distortion sense [2].

A great effort has been devoted to optimize multigrid locally adaptive BMAs for its use in H.263-like (low-level) coders [2–5]. The *entropy constraint* for multigrid motion estimation was originally introduced in [2] to obtain a balanced bit allocation between motion information and prediction errors. Most of the posterior works in adaptive BMA including the perceptual DCT quantisation have been motivated by purely quantitative rate-distortion reasonings [3, 4]. Beyond the optimal rate-distortion result, which has been shown to be visually negligible [6], the band-pass nature of the perceptual quantiser introduces desirable properties on the flow [5]. The coherence and robustness of this flow (which could make it suitable for higher level purposes) appears to be of capital importance compared to its optimality.

The aim of this work is to point out that perceptual weighting in the flow computation makes the (low-level) sparse H.263 motion information robust enough to be used for MPEG4-like coding. To test this hypotesis, a well-known motion-based segmentation algorithm for high-level video coding [7, 8] has been used with perceptually weighted and perceptually unweighted BMAs used in low-level H.263-like coders.

Perceptually weighted multigrid BMA.

The general idea behind BMAs is searching the block displacement which minimizes the prediction error. Multigrid BMA is a coarse-to-fine procedure in which the resolution is locally increased by splitting the blocks used in the displacement search. The splitting criterion determines where to focus the motion estimation effort. According to entropy-constrained motion estimation, the resolution of the optical flow should be increased in a particular position if the volume of the final signal is reduced [2, 5]. In particular, a block should be split if the sum of the entropies of the motion flow and the prediction error or *Displacement Frame Difference*, $H(\vec{D}) + H(DFD)$, is reduced. The entropy $H(DFD)$ can be computed either in the spatial domain [2] or in the frequency domain after the application of the perceptual band-pass DCT quantizer [5]. The later option is optimal because it takes into account the actual volume of the encoded signal. Beyond this quantitative difference, while the first *unweighted* criterion operates in the same way for every resolution, the second *perceptually weighted* criterion behaves in a scale-dependent way. This scale-dependent behaviour can be explained in terms of the uncertainty relation (see figure **1**): as the spatial support of a signal, Δx, and its bandwidth, Δf, are related by, $\Delta x \cdot \Delta f = k$, the spectrum of the prediction errors depends on the resolution. Large spatial support in low resolution implies low frequency content errors. Small spatial support in high resolution spreads the error spectrum. In this context, the low band-pass nature of the perceptual DCT quantizer gives high entropy values for low resolution DFDs and low entropy values for high resolution DFDs. As a result, while a permisive splitting criterion is obtained at low resolution, a restrictive criterion is obtained at high resolution. This scale-dependent splitting strategy gives rise to a more robust flow focused in perceptually relevant motions [5]. This qualitative difference of the perceptually weighted flow should improve the results of the motion-based segmentation algorithms.

Motion-based segmentation and its evaluation.

The algorithms to obtain a layered representation of a video sequence [7, 8] use a displacement field as input low-level motion information. From this flow, an arbitrary initial segmentation is iteratively refined estimating an affine motion model in each region and merging the regions with similar affine parameters. Given a pair of frames from a sequence with n moving objects, the segmentation algorithm gives m masks, representing the identified objects and the corresponding affine models.

In order to assess the relative performance of the different considered flows, the following dissimilarity measure between the (manually segmented) object O_i and the mask M (which maximises $M_j \cap O_i$ for all the obtained masks M_j) has been used,

$$\epsilon_i = \int_{M \cap O_i^C} D_i(x)\, dx + \int_{O_i \cap M^C} D_i(x)\, dx \tag{1}$$

where $M \cap O_i^C$ represents the pixels which belong to the mask but are not included in the object, $O_i \cap M^C$ represents the pixels which belong to the object but are not included in the mask, and the chamfer distance, $D_i(x)$, to the object O_i [9] progressively penalises unmatched pixels far from the object boundaries. The segmentation error for a frame, ϵ, is simply the sum of all ϵ_i. It has been empirically found that this error measure adequately describes the quality of the segmentation.

Results and discussion.

Several experiments have been carried out using the referred layer identification algorithm [7, 8] initialised with three different sparse flows (one motion vector per block): fixed-size BMA, unweighted variable-size BMA [2], and perceptual variable-size BMA [5]. Results regarding two standard sequences, TAXI and RUBIK, have been considered here because they represent extreme cases for the segmentation algorithm. Little moving objects with simple translations, and one big object rotating around an axis, respectively. While in the first sequence even simple velocity clustering could solve the problem, in the second one, regions with very different velocities have to be merged in a single complex-motion object.

Table **1** shows the average segmentation error (normalised by the maximum in each sequence), the number of iterations and the number of identified regions in each case. Figures **2** and **3** show examples of the segmentations achieved. The block appearance of the layers is obviously due to the sparseness of the input flows .

The following trends can be identified from the obtained results: a) the segmentation is very sensitive to the robustness and coherence of the sparse flow, which implies that the block size cannot be arbitrarily reduced to improve the resolution, and b) the segmentation is better when the blocks are fairly small compared to the size of the moving regions (see the variable-size BMA error results for RUBIK –big object– and TAXI –small objects–).

This general behaviour may be explained in terms of the data needed to estimate the six parameters of the affine models. Ideally, a minimum of three independent vectors are needed to segment a region. In real situations more vectors per region will be needed. Consequently, the segmentations of big objects are comparatively better. On the other hand, as the number of motion measurements is reduced when using sparse flows (with regard to dense flows), their robustness becomes critical. This is why the results based on sparse fixed-size BMA are extremely poor. However, if the variable-size BMA used is robust enough, reliable rough segmentations are still possible. In particular, the increased robustness of the perceptually weighted flow speeds up the convergence of the segmentation algorithm and minimizes the segmentation error (see table **1**). Moreover, a closer estimation of the actual number of moving objects is obtained due to the smoothness of the flow within moving regions, which avoids false splittings.

In conclusion, despite the optimal (perceptual) variable-size BMA [3–5] does not improve the quality of the reconstructed sequence in the context of H.263 [6], the results presented here show that the perceptual weighting makes this flow suitable to obtain rough segmentations required in MPEG-4 from sparse motion information.

The segmentation algorithm is usually initialised with one motion vector per pixel [7, 8]. Note how the resolution of the segmentation is increased if a dense flow is used (in the example of figure **2** Horn and Schunck [10] was used)

Table 1: *Segmentation errors, number of iterations and identified regions.*

	RUBIK (2 objects)			TAXI (4 objects)		
	ϵ	Iterat.	Regions	ϵ	Iterat.	Regions
FSBMA	1.00 ± 0.05	25	21.0 ± 4.8	1.00 ± 0.14	25	15.8 ± 3.8
Unweight. VSBMA	0.46 ± 0.06	21 ± 5	4.4 ± 2.2	0.52 ± 0.15	11 ± 9	5.8 ± 2.0
Percept. VSBMA	0.27 ± 0.03	7 ± 6	2.1 ± 0.3	0.49 ± 0.14	4 ± 1	4.3 ± 0.5

References

[1] ISO/IEC JTC1/SC29/WG11 N1909. Overview of the MPEG-4 version 1 standard, 1997.

[2] F. Dufaux and F. Moscheni. Motion estimation techniques for digital TV: A review and new contribution. *Proc. IEEE*, 83(6):858–876, 1995.

[3] G.M. Schuster and A.K. Katsaggelos. A video compression scheme with optimal bit allocation among segmentation, motion and residual error. *IEEE Trans. Im. Proc.*, 6(11):1487–1502, 1997.

[4] J. Lee. Joint optimization of block size and quantization for quadtree-based motion estimation. *IEEE Trans. Im. Proc.*, 7(6):909–912, 1998.

[5] J. Malo, F. Ferri, J. Albert, and J.M. Artigas. Splitting criterion for hierarchical motion estimation based on perceptual coding. *Electr. Lett.*, 34(6):541–543, 1998.

[6] J. Malo, F. Ferri, J. Gutierrez, and I. Epifanio. Importance of quantiser design compared to optimal multigrid motion estimation in video coding. *Elect. Lett.*, 36(9):807–809, 2000.

[7] J.Y.A. Wang and E.H. Adelson. Representing moving images with layers. *IEEE Trans. Im. Proc.*, 3(5):625–638, 1994.

[8] L. Torres, D. García, and A. Mates. A robust motion estimation and segmentation approach to represent moving images with layers. In *IEEE Intl. Conf. Acoust. Speech Sig. Proc.*, pages 2981–2984, Munich, Germany, 1997.

[9] G. Borgefors. Distance transformations in digital images. *CVGIP*, 34:344–371, 1986.

[10] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artif. Intell.*, 16(1):185–203, 1981.

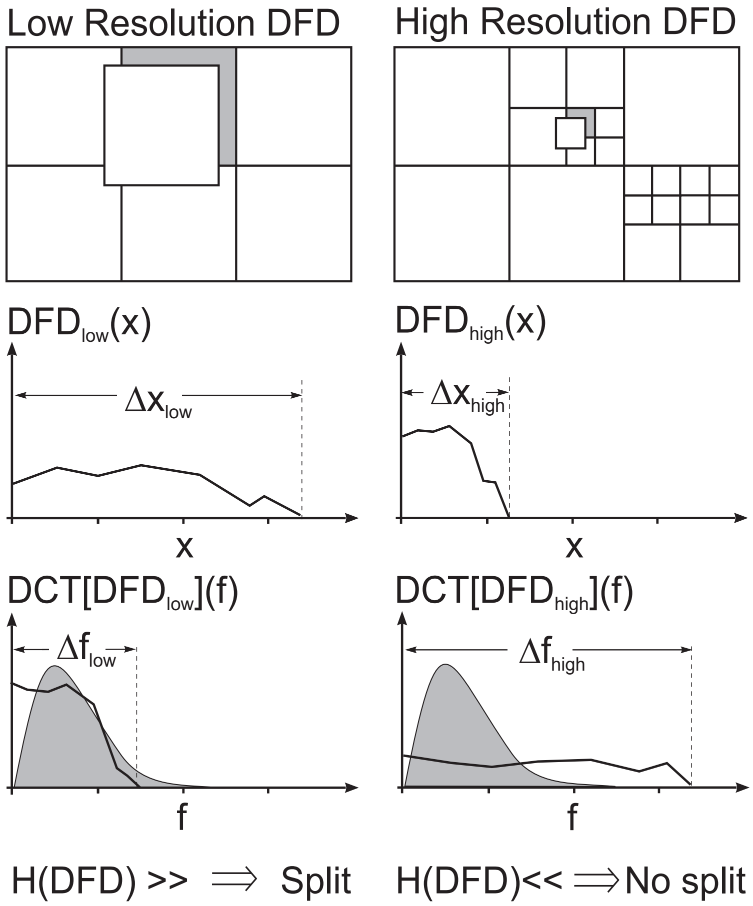

Fig. 1: *The band-pass DCT bit allocation represented by the dark area in the frequency domain determines a different splitting strategy at different resolutions.*

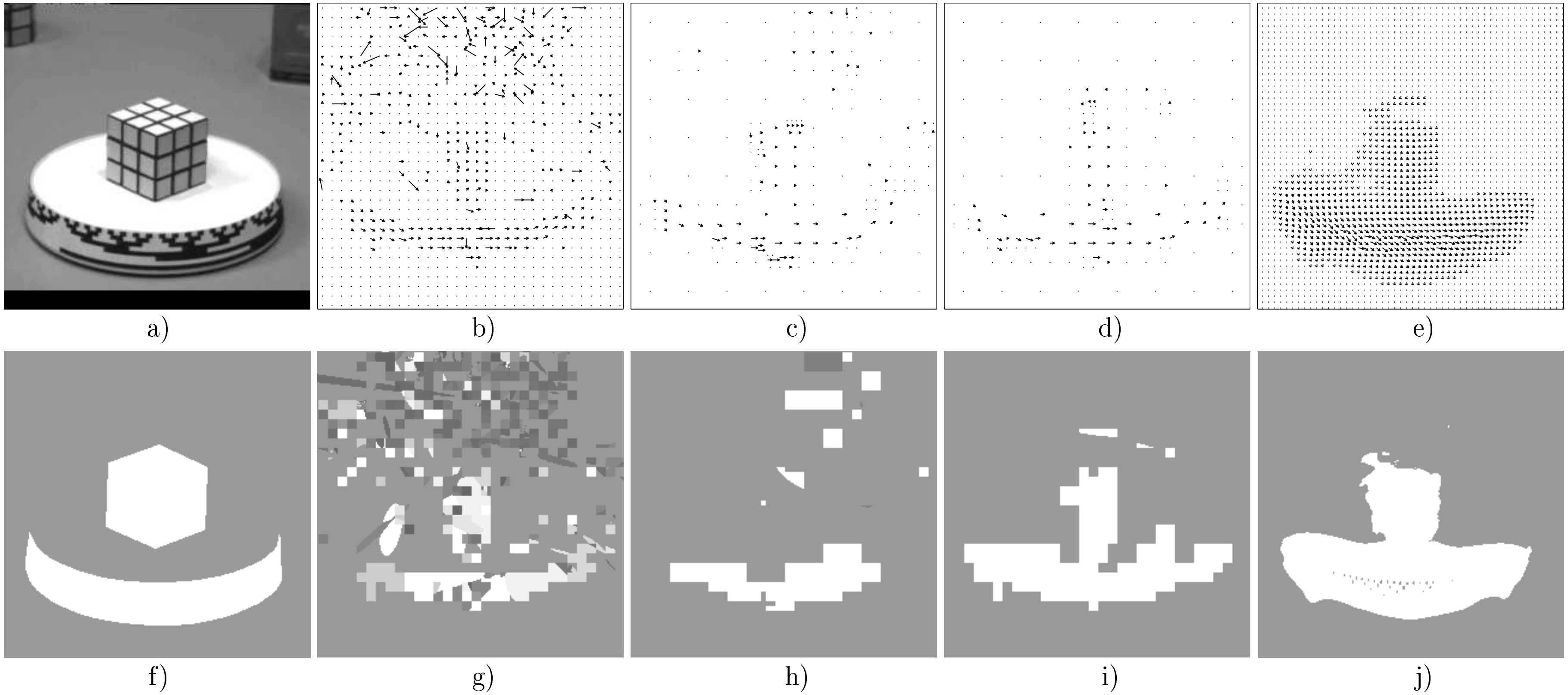

Fig. 2: *Segmentation results (RUBIK, frame 3). a) Original. b) Sparse FSBMA flow. c) Sparse Unweighted VSBMA flow. d) Sparse Perceptual VSBMA flow. e) Dense Horn and Shunck flow (it subsampled in this visualization). f) Ideal segmentation. g) Segmentation with FSBMA, $\epsilon = 0.98$. h) Segmentation with Unweighted VSBMA, $\epsilon = 0.51$. i) Segmentation with Perceptual VSBMA, $\epsilon = 0.25$. j) Segmentation with Horn flow, $\epsilon = 0.22$. Note that dense flows do improve the resolution of the segmentation but they are not used in practical video coding algorithms.*

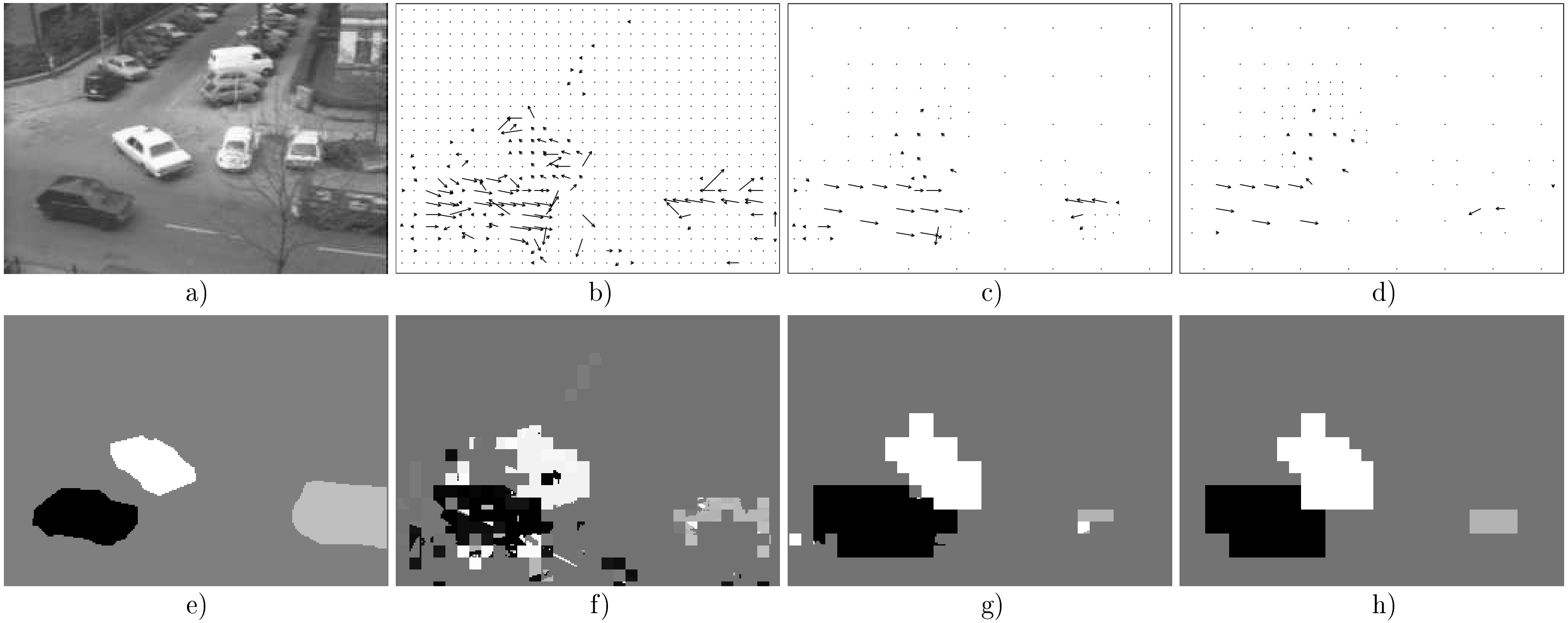

Fig. 3: *Segmentation results (TAXI, frame 9). a) Original. b) FSBMA flow. c) Unweighted VSBMA flow. d) Perceptual VSBMA flow. e) Ideal segmentation f) Segmentation with FSBMA,* $\epsilon = 1.02$. *g) Segmentation with Unweighted VSBMA,* $\epsilon = 0.49$. *h) Segmentation with Perceptual VSBMA,* $\epsilon = 0.35$.