Adaptive Motion Estimation and Video Vector Quantization Based on Spatiotemporal Non-linearities of Human Perception

J. Malo, F. Ferri*, J. Albert* & J.M. Artigas
*Departament d' Òptica, *Institut de Robòtica,*
Facultat de Física, Universitat de València,
C/ Dr. Moliner 50, 46100, Burjassot, València (Spain)
e-mail: Jesus.Malo@uv.es

Abstract

The two main tasks of a video coding system are motion estimation and vector quantization of the signal. In this work a new splitting criterion to control the adaptive decomposition for the non-uniform optical flow estimation is exposed. Also, a novel bit allocation procedure is proposed for the quantization of the DCT transform of the video signal. These new approaches are founded on a perception model that reproduce the relative importance given by the human visual system to any location in the *spatial frequency, temporal frequency and amplitude* domain of the DCT transform. The experiments show that the proposed procedures behave better than their equivalent (fixed-block-size motion estimation and fixed-step-size quantization of the spatial DCT) used by MPEG-2.

Key Words: Video Coding, Motion Estimation, Perceptual oriented Quantization.

1 Introduction

The objective of video coding is to find a lossy non-redundant expression of the signal to minimize the transmission or storage requirements in such a way that the retrieved sequence and the original one satisfy some criterion of similarity. In natural video sequences two kinds of redundacies can be indentified: objective (related to spatio-temporal correlations) and subjective (related to the human visual system characteristics). These strong redundancies will give rise to strong compression ratios if they are removed. In practical applications judged by a human observer the compression procedure has to be designed to minimize the *subjective* distortions of the reconstructed sequence. In those cases, perceptual properties of the human viewer must be included in the coder design. The redundancy removal in today widely accepted video compressors (H.261, MPEG-1, MPEG-2 [1,2]) is based in the temporal predictability of the signal [3]. In this encoding scheme a motion estimation block extracts the motion information, and a predictor block uses this to make a prediction of the next frame. In general, the motion estimation/prediction process cannot exactly reproduce the next frame, so an error signal has to be transmitted to the decoder. In this error sequence there still exists a certain degree of redundancy which must be reduced through a quantization module. Research effort in the field has been focused in the motion estimation and error quantization modules considered in an *isolated way*. In this sense, better motion estimation techniques [4,5] try to give better predictions and intrinsically lower complexity error signals. On the other hand, better quantization

Acknowledgements: This work has been supported by CICYT projet TIC95-676-C02-01 and IVEI (Generalitat Valenciana) project N°96/003-035.

techniques [6] are intended to distribute the quantization noise in such a way that the required quality is preserved. In order to achieve a lower complexity error signal, Dufaux et al. [4] have proposed a multiresolution motion estimation algorithm based on minimizing the spatial entropy of the corrections and displacement field, but this approach does not take into account the frequency-dependent transform coding that follows motion estimation. On the other hand, most common implementations of perceptual quantizers are based on the human visual system still-threshold spatial frequency response [1,2,7]. 3D transform coders have been recently proposed [8,9], but in all these approaches, the spatio-temporal suprathreshold characteristics of human perception have not been deeply exploited [6].

Following this, a new video coding scheme is introduced in this work to improve in several ways the temporal and spatial redundancy removal used in MPEG-2. First, a new adaptive-block-size matching algorithm for motion estimation is presented. In the proposed scheme, the splitting criterion is based on minimizing the entropy *in the frequency domain*, taking into account the frequency selective quantization made after the motion estimation stage. In this way, the motion estimation is modified according to the frequency selective encoding requirements. And second, a novel non-linear bit allocation procedure is used for the quantization of the DCT transform of the video signal. This new approach is founded on a perception model that reproduce the relative importance given by the human visual system to any location in the 4-dimensional *spatial frequency, temporal frequency and amplitude* domain of the DCT [10,11].

2 Improvements and Alternatives to the Standard Coding Schemes

In the present standards (up to MPEG-2) the motion estimation module is just an optical flow estimator [3,4,5]. There are several methods to compute the image flow (eg. *differential methods* [4,5], *frequency domain* methods [5]), but standard video coders estimate the displacement field through a *matching technique* [3,4,5]. The usual MPEG implementation of this general matching idea consists of: using rectangular neighbourhoods (blocks), searching for displacement candidates in a restricted set, and using standard correlation as a measure of similarity [1,2,4]. This is called fixed-block-size Block Matching Algorithm (BMA). Regarding to the quality of the predicted frame, the most important parameter in the BMA is the block size. Decreasing the block size leads to a more accurate prediction and to a lower complexity error signal, but it implies a more exhaustive optical flow computation increasing superfluous motion information and reducing the robustness of the displacement estimate.

Standard still image compressors [7] and video coders [1,2] introduce human visual system characteristics at the quantization stage. A transform of the error signal to a frequency domain (DCT) is done because human visual sensitivity is highly uneven in these domains. The bit allocation procedure is founded in human visual system models that consider the perception as a linear filtering process in the Fourier domain. The quantization matrix is inspired in the threshold frequency response of the human viewer [7] and then a variable number of quantization levels per coefficient is allowed. These levels are uniformly distributed in the amplitude range [7,10]. The threshold linear filter model reproduces the relative importance given by the human visual system to the different frequencies, but its validity is restricted to near-threshold contrasts. Neither amplitude sensitivity non-linearities, nor temporal frequency selectivity are taken into account in such a quantization scheme.

2.1 Adaptive-Block-Size BMA with Frequency Domain Splitting Criterion

It has been shown [4,12,13] that an adaptive block size algorithm reduces the number of necessary vector displacement computations concentrating the effort in highly moving areas and improves the quality of moving objects contours in the predicted frames. The adaptive-block-size BMA starts with a motion estimation at a coarse resolution level (big block size). Then, subsequent local refinements of this estimation are made, increasing the resolution in certain areas (splitting selected blocks). The key parameter in this adaptive-block-size BMA is the *splitting criterion*, which is usually based in some *objective* difference measurement related to the energy of the error [12,13]. Taking into account that compression is the main objective of a video coder, Dufaux et al. [4] proposed a splitting criterion for adaptive-block-size motion estimation based on minimizing the entropy of the corrections and displacement field. A coarse block was divided if the entropy of the divided block error signal and its associated flow field was lower than the entropy of the full block error signal and its displacement field:

$$\text{if } H_{\text{error (split)}} + H_{\text{displacements (split)}} < H_{\text{error (nosplit)}} + H_{\text{displacements (nosplit)}} \Rightarrow \text{split} \quad (1)$$

In this scheme the entropy of the error signal is calculated in the spatial domain before the quantization stage. It means that this *spatially measured entropy* splitting criterion is independent of the quantization process. Splittings that reduce the error signal complexity can be superfluous depending on the type of quantizer used. Following this, and assuming a certain quantization scheme Q in the frequency DCT domain, a novel splitting criterion is proposed:

$$\text{if } H_{\text{Q[error (split)]}} + H_{\text{displac.(split)}} < H_{\text{Q[error (nosplit)]}} + H_{\text{displac.(nosplit)}} \Rightarrow \text{split} \quad (2)$$

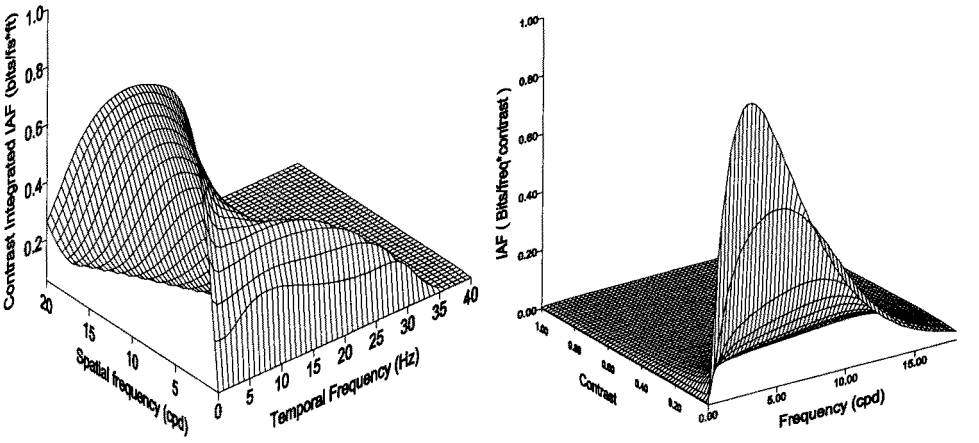
With this *spectral entropy* criterion, motion estimation takes into account the particular quantizer used and some theoretical adventages of this new criterion can be outlined:

1) The *spectral entropy* criterion should be more restrictive than the *spatial entropy* criterion: if the quantization is severe, the action of the quantizer over the non-split block significantly reduce its complexity, so no benefit will be obtained from the splitting. The practical consequence of this is that a simpler quadtree decomposition (fewer blocks) is obtained and usually, a more robust motion estimate is associated with this decomposition.
2) The splitting criterion is matched with the quantizer requirements. The motion estimation is refined only if it improves the error signal in the areas which are significant to the quantizer. If the quantizer is designed with a certain objective, the *spectral entropy* criterion controls the motion estimation to optimize the performance in the same direction. In this case, the *spatial entropy* criterion will minimize the entropy measured in the spatial domain, but the final entropy (the entropy of the transform coefficients) should be minimized by the *spectral entropy* criterion.

2.2 The Information Allocation Function Perception Model

Non-linear response of the human visual system to the amplitude of sinusoidal patterns has lead to question the threshold linear filter based models. The existence of perceptual tolerances to changes in the amplitude and frequency of still gratings implies that the visual system maps the continuous spatial frequency/amplitude range into a finite set of non-uniformly distributed discrete perceptions. This fact has been used to propose a new suprathreshold and non-linear perception model [10,11]. In this model, low level perception is considered as an information removal process characterized by an Information Allocation Function (*IAF*) giving the amount of information used by the

system to encode each area of the spatial frequency/amplitude domain. This perceptual bit allocation function has been succesfully used to design a non-uniform step size quantizer for image compression improving the performance of the JPEG-like uniform quantization [10]. The *IAF* model has been used to define a subjective quality measure which accurately reproduce the opinion of the observers under a variety of noise conditions [11]. In this work we generalize this non-linear vector quantizer model to non-zero temporal frequencies, modifiying the threshold spatio-temporal filter [14] with suprathreshold contributions as in the 2-dimensional case, to postulate a spatio-temporal *IAF* of the human observer. Figure 1 show different views of this function.



(a) *(b)*

Fig. 1. (a) Contrast integrated *IAF* (relative amount of information used by the human visual system to encode each spatio-temporal component of an input sequence). (b) Spatial frequency and amplitude dependency of the *IAF* at a particular temporal frequency (non-uniform amplitude bit allocation for different spatial coefficients).

2.3 New Bit Allocation in the DCT Transform Domain

Assuming the *IAF* model, the non-uniformities of perceptual sensitivity in the spatial frequency, temporal frequency and amplitude domain allows us to concentrate the encoding effort in the middle spatial frequencies, low amplitudes and low temporal frequencies. In this work the spatio-temporal *IAF* has been used in two ways. First in a frame-by-frame way, without considering temporal characteristics of the human visual system. We have applied the *IAF* at zero temporal frequency to quantize each frame of the error sequence as does MPEG with its Q matrix. And second, considering both spatial and temporal characteristics of the human viewer. In this later case a quantization of the coefficients of the 3D DCT of the error sequence is performed. In both cases, the codebook design is as follows [10,11]: in the 2D case, one starts deciding the number of quantization levels per coefficient. This is done by means of contrast integrated IAF at zero temporal frequency. Then, the quantization steps and the decision boundaries are non-uniformly distributed in amplitude for each coefficient by a process similar to non-uniform random number generation.

In the 3D case, the first step is deciding how much information is used to encode each temporal frequency frame of the 3D DCT of the error signal. This is done through the contrast and spatial frequency integrated *IAF*. After the temporal frequency dependent

bit allocation, there is a spatial frequency dependent and amplitude dependent bit allocation process, analogous to the 2D case for every temporal frequency frame.

3 Results

Both classical and proposed video coding schemes have been applied to different natural sequences. The flow fields and the decoded frames are compared using the exposed techniques at similar bit rates (below 0.5 bpp) on the *Taxi* test sequence [5]. The size of the estimation blocks is 8*8 in the fixed-block-size BMA while this size ranges from 64*64 to 4*4 in the adaptive cases (resolution level of the quadtree decomposition ranging from 2 to 6). The size of the quantization blocks is 16*16 in every case. The splitting criterion is applied in such a way that no interaction between blocks or resolutions is considered [4]. Only forward prediction is used for motion compensation.

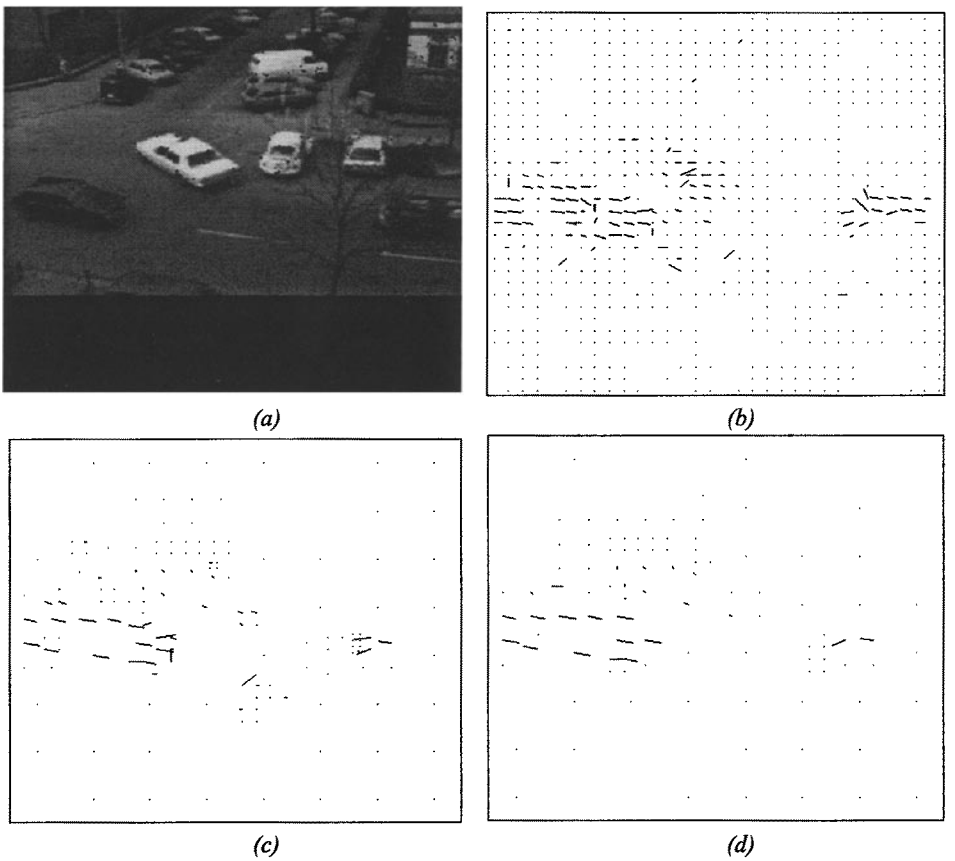


Fig. 2. (a) Original frame 2 of the test sequence. (b) Fixed-block-size (8*8) BMA flow field between frames 2 and 3. (c) Adaptive-block-size (64*64 to 4*4) BMA flow field with spatially measured entropy criterion. (d) The same with spectral entropy criterion.

3.1 Motion Estimation

A number of interesting remarks arise from the experimental results that confirm the statements made in previous sections.

Fixed versus adaptive schemes: Figure 2 shows how the adaptive algorithm reduces the motion estimation effort and improves it only in the complex motion areas (moving cars). The fixed-block-size algorithm is more sensitive to noise (see the false alarms at the bottom, a static area). Little moving structures are not resolved in any case.

	Blocks		*Displacement estimations*		*Entropy computations*	
Initial resol. Level	2	3	2	3	2	3
Spatial criterion	97.3	161.2	484.8	795.0	234.5	365.7
Spectral criterion	68.5	137.5	352.0	700.1	168.0	318.4

Table 1. Average number (per frame) of blocks -size of the vector field-, displacement estimations and entropy computations as a function of the splitting criterion and the initial resolution level.

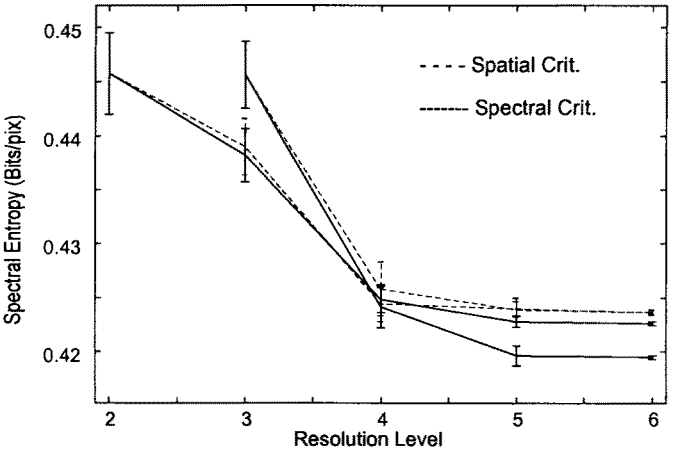


Fig. 3. Average spectral entropy (and standard deviations) of the error signal as a function of the resolution level, starting from level 2 and level 3.

Spatial entropy criterion versus spectral entropy criterion: The spectral entropy criterion is more restrictive than the spatial one. Figure 2 shows how the spatial criterion give rise to further block splitting and a more complex motion information. Note how the spatial criterion splits some blocks in static areas. The data of table 1 show that spectral criterion reduces the number of split blocks and hence the final vector field size and the number of displacement estimations and entropy computations. As the motion estimation blocks are kept bigger, the spectral criterion gives a more robust motion estimation. Note (figure 2) that in the spatial case when splitting occurs in a static area the block matching search gives rise to the same false alarms as in the exhaustive computation. The relation between motion estimation and vector quantization in the spectral criterion case implies a better reduction of the entropy of the encoded error. In figure 3, the decrease of spectral entropy of the error signal with different inial resolution levels and different splitting criteria is compared. This figure

shows that spectral criterion always achieves a final lower entropy. It means that minimizing the entropy in the spatial domain does not ensure that it will be minimized after the transform coding.

Effect of the initial resolution level: It is clear that increasing the initial resolution level increases the complexity: if a lower resolution is taken, both criterion keep some blocks unsplit, so the complexity is reduced. (See the values of the table 1). Figure 3 shows that a more exhaustive initial level achieve better entropy values of the error signal (at a cost of bigger motion information complexity).

3.2 Vector quantization

Figure 4 shows the reconstructed frame 5 of the sequence by the different schemes. Distorted contours of moving objects are due to errors of the motion estimation algorithms. The high frequency textured noise is due to the frequency shaped quantization. The adaptive motion estimation (fig. 4b,c) gives clearly better motion compensated images. It is apparent that the quantization errors are *less visible* in the perceptually improved cases.

(a)

(b)

(c)

Fig. 4. Detail of reconstructed frame 5 of the TAXI sequence. (a) MPEG-like scheme. Fixed-block-size BMA, amplitude uniform bit allocation (Linear filter model). (b) Adaptive-size BMA, Frame-by-frame non-linear quantization. 2-d *IAF* model. (c) Adaptive-size BMA, spatio-temporal non-linear quantization. 3-d *IAF* model.

4 Concluding Remarks

Some specific improvements of a tipical video scheme have been presented. First, a perceptually based vector quantization has been designed using a recently developed perception model. Furthermore, an adaptive BMA with an alternative splitting criterion specially related to the quantization block has been proposed. This criterion gives rise to simpler and more robust description of the scene motion. The joint application of the proposed methods substantially improves the subjective quality of the reconstructed signal at a given compression ratio. Future research, and more exhaustive experimentation, is still needed to quantify which are the relative importance of each proposed alternative with regard to the final performance of the system.

References

[1] *D. LeGall.* MPEG: A video compression standard for multimedia applications. Comm. ACM Vol. 34, N° 4, 47-58. (1991)

[2] *ISO/IEC* 13818 Draft International Standard: Generic coding of moving pictures and associated audio, Part 2: Video. (1993)

[3] *H.G. Musmann, P. Pirsh & H.J. Grallert.* Advances in video coding. Proc. IEEE Vol.73, N°4, 523-548. (1985).

[4] *F. Dufaux & F. Moscheni.* Motion estimation techniques for digital TV: A review and new contribution. Proc IEEE Vol.83, N°6, 858-876. (1995)

[5] *J.L. Barron, D.J. Fleet & S.S. Bauchemin.* Performance of optical flow techniques. IJCV, Vol.12, N°1, 43-77. (1994)

[6] *N. Jayant, J. Johnston & R. Safranek.* Signal compression based on models of human perception. Proc. IEEE, Vol.81, N°10, 1385-1422. (1993)

[7] *G.K. Wallace.* The JPEG still picture compression standard. Comm. ACM. Vol.34, N°4, 31-43. (1991)

[8] *F. Bosveld, R.L. Lagendijk & J. Biemond.* Compatible spatio-temporal subband encoding of HDTV. Signal Proc. Vol.28, 271-290 (1992)

[9] *J. Luo, C.W. Chen, K.J. Parker & T.S. Huang.* Three dimensional subband video analysis and synthesis with adaptive clustering in high-frequency subbands. Proc IEEE Int. Conf. Im. Proc. Austin TX (1994)

[10] *J. Malo, A.M. Pons & J.M. Artigas.* Bit allocation algorithm for codebook design in vector quantization fully based on human visual system non-linearities for suprathreshold contrasts. Electr. Lett. Vol.24, 1229-1231. (1995)

[11] *J. Malo, A.M. Pons & J.M. Artigas.* Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain. (Accepted in Image Vis. Comp.)

[12] *M.H. Chan, Y.B. Yu, & A.G. Constantinides.* Variable size block matching motion compensation with applications to video coding. Proc. IEE, Vol.137, N°4, 205-212 (1990)

[13] *F. Duffaux & M. Kunt.* Multigrid block matching motion estimation with an adaptive local mesh refinement. SPIE Proc. Visual Commun. and Image Process 92', Vol. 1818. (1992)

[14] *D.H. Kelly.* Motion and vision II: Stabilized spatio-temporal threshold surface. JOSA, Vol.69, N°10, 1340-1349. (1979)