Perceptual Feedback in Multigrid Motion Estimation Using an Improved DCT Quantization

Jesús Malo, Juan Gutiérrez, I. Epifanio, Francesc J. Ferri, and José M. Artigas

***Abstract*—In this paper, a multigrid motion compensation video coder based on the current human visual system (HVS) contrast discrimination models is proposed. A novel procedure for the encoding of the prediction errors has been used. This procedure restricts the maximum perceptual distortion in each transform coefficient. This subjective redundancy removal procedure includes the amplitude nonlinearities and some temporal features of human perception. A perceptually weighted control of the adaptive motion estimation algorithm has also been derived from this model. Perceptual feedback in motion estimation ensures a perceptual balance between the motion estimation effort and the redundancy removal process. The results show that this feedback induces a scale-dependent refinement strategy that gives rise to more robust and meaningful motion estimation, which may facilitate higher level sequence interpretation. Perceptually meaningful distortion measures and the reconstructed frames show the subjective improvements of the proposed scheme versus an H.263 scheme with unweighted motion estimation and MPEG-like quantization.**

***Index Terms*—Entropy constrained motion estimation, nonlinear human vision model, perceptual quantization, video coding.**

I. Introduction

In natural video sequences to be judged by human observers, two kinds of redundancies can be identified: 1) *objective redundancies*, related to the spatio-temporal correlations among the video samples and 2) *subjective redundancies*, which refer to the data that can be safely discarded without perceptual loss. The aim of any video coding scheme is to remove both kinds of redundancy. To achieve this aim, current video coders are based on motion compensation and two-dimensional (2-D) transform coding of the residual error [1]–[4]. The original video signal is split into motion information and prediction errors. These two lower complexity sub-sources of information are usually referred to as displacement vector field (DVF) and displaced frame difference (DFD), respectively.

In the most recent standards, H.263 and MPEG-4 [4], [5], the fixed-resolution motion estimation algorithm used in H.261 and MPEG-1 has been replaced by an adaptive, variable-size block matching algorithm (BMA) to obtain improved motion estimates [6]. Spatial subjective redundancy is commonly reduced through a perceptually weighted quantization of a transform of the DFD. The bit allocation among the transform coefficients is based on the spatial frequency response of simple (linear and threshold) perception models [1]–[3].

In this context, there is a clear tradeoff between the effort devoted to motion compensation and transform redundancy removal. On the one hand, better motion estimation may lead to better predictions and should alleviate the task of the quantizer. On the other hand, better quantization techniques may be able to remove more redundancy, thereby reducing the predictive power needed in the motion estimate. Most of the recent work on motion estimation for video coding has been focused on the adaptation of the motion estimate to a *given quantizer* to obtain an good balance between these elements. Since the introduction of the intuitive (suboptimal) entropy-constrained motion estimation of Dufaux *et al.* [7], [8] several optimal, variable-size BMAs have been proposed [9]–[12]. These approaches put forward their intrinsic optimality, but the corresponding visual effect and the relative importance of the motion improvements versus the quantizer improvements have not been deeply explored, mainly because of their subjective nature.

This paper adresses the problem of the tradeoff between multigrid motion estimation and error quantization in a different way. An improved (nonlinear) perception model inspires the whole design to obtain a coder that preserves no more than the subjectively significant information. The role of the perceptual model in the proposed video coder scheme is twofold. First, it is used to simulate the redundancy removal in the human visual system (HVS) through an appropriate perceptually matched quantizer. Second, this perceptual quantizer is used to control the adaptive motion estimation. This control introduces a perceptual feedback in the motion estimation stage. This perceptual feedback limits the motion estimation effort, avoiding superfluous prediction of details that are perceptually negligible and will be discarded by the quantizer. The bandpass shape of the perceptual constraint to the motion estimation gives a scale-dependent control criterion that may be useful for discriminating between significant and noisy motions. Therefore, the benefits of including the properties of the biological filters in the design may go beyond a better rate-distortion performance but also improve the meaningfulness of the motion estimates. This fact may be important for next generation coders that build models of the scene from the low-level information used in the current standards.

In this paper, a novel subjective redundancy removal procedure [13], [14] and a novel perceptually weighted motion estimation algorithm [12] are jointly considered to present a fully perceptual motion compensated video coder. The aim of the

Manuscript received May 6, 1999; revised June 18, 2001. This work was supported in part by CICYT Projects TIC 1FD97-0279 and TIC 1FD97-1910. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rashid Ansari.

J. Malo and J. M. Artigas are with the Departament d'Òptica, Universitat de València, 46100 Burjassot, València, Spain (e-mail: Jesus.Malo@uv.es; http://taz.uv.es/~jmalo).

J. Gutiérrez, I. Epifanio, and F. J. Ferri are with the Departament d'Informàtica Universitat de València, 46100 Burjassot, València, Spain.

Publisher Item Identifier S 1057-7149(01)08211-2.

paper is to assess the relative relevance of optimal variable-size BMAs and quantizer improvements. To this end, the decoded frames are explicitly compared and analyzed in terms of perceptually meaningful distortion measures [15], [16]. The meaningfulness of the motion information is tested by using it as input for a well established motion-based segmentation algorithm used in model-based video coding [17], [18].

The paper is organized as follows. In Section II, the current methods for quantizer design and variable-size BMA for motion compensation are briefly reviewed. The proposed improvements in the quantizer design, along with their perceptual foundations, are detailed in Section III. In Section IV, the proposed motion refinement criterion is obtained from the requirement of a monotonic reduction of the significant (perceptual) entropy of DFD and DVF. The comparison experiments are presented and discussed in Section V. Some final remarks are given in Section VI.

II. Conventional Techniques for Transform Quantizer Design and Multigrid Motion Estimation

The basic elements of a motion compensated coder are the optical flow estimation and the prediction error quantization. The optical flow information is used to reduce the objective temporal redundancy, while the quantization of the transformed error signal [usually a 2-D discrete cosine transform (DCT)] reduces the remaining (objective and subjective) redundancy to certain extent [1]–[4].

Signal independent JPEG-like uniform quantizers are employed in the commonly used standards [1]–[4]. In this case, bit allocation in the 2-D DCT domain is heuristically based on the threshold detection properties of the HVS [2], [3], but neither amplitude nonlinearities [19] nor temporal properties of the HVS [20]–[22] are taken into account. The effect of these properties is not negligible [23], [24]. In particular, the nonlinearities of the HVS may have significant effects on bit allocation and improve the subjective results of the JPEG-like quantizers [13], [14], [25], [26].

The conventional design of a generic transform quantizer is based on the minimization of the *average* quantization error over a training set [27]. However, the techniques based on average error minimization have some subjective drawbacks in image coding applications. The optimal quantizers (in an average error sense) may underperform on individual blocks or frames [9] even if the error measure is perceptually weighted [28]: the accumulation of quantization levels in certain regions in order to minimize the average perceptual error does not ensure good behavior on a particular block of the DFD. This suggests that the subjective problems of the conventional approach are not only due to the use of perceptually unsuitable metrics, as usually claimed, but are also due to the use of an inappropriate *average error* criterion. In addition to this, quantizer designs that depend on the statistics of the input have to be re-computed as the input signal changes. These factors favor the use of quantizers based on the threshold frequency response of the HVS instead of the conventional, average error-based quantizers.

Multigrid motion estimation techniques are based on matching between variable-size blocks of consecutive frames of the sequence [6]. The motion estimation starts at a coarse resolution (large blocks). At a given resolution, the best displacement for each block is computed. The resolution of the motion estimate is locally increased (a block of the quadtree is split) according to some refinement criterion. The process ends when no block of the quadtree can be split further.

The splitting criterion is the most important part of the algorithm because it controls the local refinement of the motion estimate. The splitting criterion has effects on the relative volumes of DVF and DFD [7]–[12], and may give rise to unstable motion estimates due to an excesive refinement of the quadtree structure [12], [29]. The usefulness of the motion information for higher-level purposes (as in model-based video coding [5], [30], [31]) highly depends on its robustness (absence of false alarms) and hence on the splitting criterion. Motion-based segmentation algorithms [17], [18] require reliable initial motion information, especially when using sparse (nondense) flows such as those given by variable-size BMA. Two kinds of splitting criteria have already been used: 1) the magnitude of the prediction error, e.g., energy, mean-square error or mean-absolute error [6], [29], [32], [33], and 2) the complexity of the prediction error. In this case, the zeroth-order spatial entropy [7], [8] and the entropy of the encoded DFD [9]–[12] have been reported. While the magnitude-based criteria were proposed without a specific relation to the encoding of the DFD, the entropy-based criteria make explicit use of the trade off between DVF and DFD.

Since the first entropy-constrained approach was introduced [7], [8], great effort has been devoted to obtaining analytical [9]–[11] or numerical [12] optimal entropy-constrained quadtree DVF decompositions. These approaches criticize the (faster) entropy measure of the DFD in the spatial domain of Dufaux *et al.* because it does not take into account the effect of the selective DCT quantizer. This necessarily implies a suboptimal bit allocation between DVF and DFD. The literature [9]–[12] reports the optimality of the proposed methods, but the practical (subjective) effect of this gain on the reconstructed sequence is not analyzed. In particular, only perceptually unweighted SNR or MSE distortion measures are given and no explicit comparison of the decoded sequences is shown.

III. Perceptually Uniform DCT Quantization

Splitting the original signal into two lower complexity signals (DVF and DFD) does reduce their redundancy to a certain extent. However, the enabling fact behind very-low-bit-rate coding is that not all the remaining data are significant to the human observer. This is why more than just the strictly predictable data can be safely discarded in the DFD quantization.

According to the current models of human contrast processing and discrimination [34], [35], the input spatial patterns are first mapped onto a local frequency domain through a set of bandpass filters with different relative gains. After that, a log-like nonlinearity is applied to each transform coefficient to obtain the response representation. Let us describe this two-step process as

$$\mathbf{A} \xrightarrow{T} \mathbf{a} \xrightarrow{R} \mathbf{r} \tag{1}$$

where

vector $\mathbf{A}$	input image;
matrix T	filter bank;
vector $\mathbf{a}$	local frequency transform;
function R	nonlinearity;
vector $\mathbf{r}$	response to the input.

The n components of the image vector, $\mathbf{A} = \{A_x\}_{x=1}^{n}$, represent the samples of the input luminance at the discrete positions $x = 1, \ldots, n$. T is a $m \times n$ matrix constituted by the impulse responses of the m bandpass filters. The local frequency transform is $\mathbf{a} = T \cdot \mathbf{A}$. Each coefficient of the transform $\mathbf{a} = \{a_f\}_{f=1}^{m}$, represent the output of the filter f with $f = 1, \ldots, m$. Each local filter f is tuned to a certain frequency. In general [34], each coefficient r_f of the response $\mathbf{r} = R(\mathbf{a}) = \{r_f\}_{f=1}^{m}$ will depend on several transform coefficients $a_{f'}$. However, at a first approximation [19], the contributions of $a_{f'}$ with $f' \neq f$ can be neglected.

The effect of the response R in the transform $\mathbf{a}$ can be conveniently modeled by a nonuniform perceptual quantizer Q_p. This interpretation as a quantizer is based on the limited resolution of the HVS. If the amplitude of a basis function of the transform T is modified, the induced perception will remain constant until the just noticeable difference (JND) is reached. In this case, as in quantization, a continuous range of amplitudes gives rise to a single perception [36], [37]. This perceptual quantizer has to be nonuniform because the empirical JNDs are nonuniform [19], [21], [22], [34]. The similarity between the impulse responses of the perceptual filters of the transform T and the basis functions of the local frequency transforms used in image and video coding has been used to apply the experimental properties of the perceptual transform domain to the block DCT transform as a reasonable approximation [13], [14], [25], [26], [38], [39]. In this paper, Q_p is formulated in the DCT domain through an explicit design criterion based on a distortion metric that includes the HVS nonlinearities [15], [16] and some temporal perceptual features [20]–[22].

A. Maximum Perceptual Error (MPE) Criterion for Quantizer Design

The natural way of assessing the quality of an encoded picture (or sequence) involves a one-to-one comparison between the original and the encoded version. The result of this comparison is related to the ability of the observer to notice the particular quantization noise in the presence of the original (masking) pattern. This one-to-one noise detection or assessment is clearly related to the tasks behind the standard pattern discrimination models [34], [35], in which an observer has to evaluate the distortion from a masking stimulus. In contrast, a hypothetical request of assessing the global performance of a quantizer over a set of images or sequences would involve a sort of averaging of each one-to-one comparison. It is unclear how a human observer does this kind of averaging to obtain a global feeling of performance and the task itself is far from the natural one-to-one comparison that arises when one looks at a particular picture.

The conventional techniques of transform quantizer design use average design criteria in such a way that the final quantizer achieves the minimum average error over the training set (sum of the one-to-one distortions weighted by their probability) [27]. However, the minimization of an average error measure does not guarantee a satisfactory subjective performance on individual comparisons [9]. Even if a perceptual weighting is used, the average criteria may bias the results. For instance, Macq [28] used uniform quantizers instead of the optimal Lloyd-Max quantizers [27], [40], due to the perceptual artifacts caused by the outliers on individual images.

To prevent large perceptual distortions on individual images arising from outlier coefficients, the coder should restrict the *maximum perceptual error (MPE)* in each coefficient and amplitude [13], [14]. This requirement is satisfied by a perceptually uniform distribution of the available quantization levels in the transform domain. If the perceptual distance between levels is constant, the MPE in each component is bounded regardless of the amplitude of the input.

In this paper, the restriction of the MPE will be used as a design criterion. This criterion can be seen as a perceptual version of the minimum maximum error criterion [9]. This idea has been implicitly used in still image compression [25], [26] to achieve a constant error contribution from each frequency component on an individual image. It has been shown that bounding the perceptual distortion in each DCT coefficient may be subjectively more effective than minimizing the average perceptual error [13], [14]. Moreover, the MPE quantizers reduce to the JPEG and MPEG quantizers if a simple (linear) perception model is considered.

B. Optimal Spatial Quantizers Under the MPE Criterion

The design of a transform quantizer for a given block transform involves finding the optimal number of quantization levels for each coefficient (bit allocation) and the optimal distribution of these quantization levels in each case [27].

Let us assume that the squared perceptual distance between two similar patterns in the transform domain $\mathbf{a}$ and $\mathbf{a} + \Delta\mathbf{a}$ is given by a weigthed sum of the distortion in each coefficient

$$D^2(\mathbf{a}, \mathbf{a} + \Delta\mathbf{a}) = \sum_{f=1}^{m} D_f^2 = \sum_{f=1}^{m} W_f(a_f)\, \Delta a_f^2 \qquad (2)$$

where $W_f(a_f)$ is a frequency and amplitude-dependent perceptual metric.

In order to prevent large perceptual errors on individual images coming from outlier coefficient values, the coder should be designed to bound the MPE for every frequency f and amplitude a_f.

If a given coefficient (at frequency f) is represented by N_f quantization levels distributed according to a density $\lambda_f(a_f)$ the maximum Euclidean quantization error at an amplitude a_f will be bounded by half the Euclidean distance between two levels

$$\Delta a_f(a_f) \leq \frac{1}{2N_f\, \lambda_f(a_f)}. \qquad (3)$$

The MPE for that frequency and amplitude will be related to the metric and the density of levels:

$$\mathrm{MPE}_f(a_f) = W_f(a_f) \cdot \max\left(\Delta a_f(a_f)\right)^2 = \frac{W_f(a_f)}{4N_f^2\, \lambda_f^2(a_f)}. \qquad (4)$$

The only density of quantization levels that gives a constant MPE bound over the amplitude range is the one that varies as the square root of the metric

$$\lambda_{f\,\mathrm{opt}}(a_f) = \frac{W_f(a_f)^{1/2}}{\int W_f(a_f)^{1/2} da_f}. \tag{5}$$

With these optimal densities, the MPE in each coefficient f will depend on the number of allocated levels and on the integrated value of the metric

$$\mathrm{MPE}_{f\mathrm{opt}} = \frac{1}{4N_f^2}\left(\int W_f(a_f)^{1/2} da_f\right)^2. \tag{6}$$

Fixing the same maximum distortion for each coefficient $\mathrm{MPE}_{f\mathrm{opt}} = k^2$ and solving for N_f, the optimal number of quantization levels is obtained

$$N_{f\,\mathrm{opt}} = \frac{1}{2k}\int W_f(a_f)^{1/2}\, da_f. \tag{7}$$

The general form of the optimal MPE quantizer is given by (5) and (7) as a function of the perceptual metric. Thus, the behavior of the MPE quantizer will depend on the accuracy of the selected W. Here, a perceptual metric related to the gradient of the nonlinear response R and to the amplitude JNDs has been considered [15], [16]

$$\begin{aligned} W_f(a_f) &= \left(\frac{\partial R(\mathbf{a})}{\partial a_f}\right)^2 \propto \mathrm{JND}(a_f)^{-2} \\ &= \left(\mathrm{CSF}_f^{-1} + \frac{a_f}{L}\, G_f(a_f)\right)^{-2} \end{aligned} \tag{8}$$

where L is the local mean luminance, CSF_f is the *contrast sensitivity function* (the bandpass linear filter which characterizes the HVS performance for low amplitudes [20], [24], [28], [41]), and $G_f(a_f)$ are empirical monotonically increasing functions of amplitude for each spatial frequency to fit the amplitude JND data [15]. In particular, we have used the CSF of Nygan *et al.* [41]

$$\begin{aligned} \mathrm{CSF}_f = &\left(\frac{1}{4} + \frac{1}{\pi^2}\log^2\left(\frac{2\pi f}{11.6} + \sqrt{\frac{4\pi^2 f^2}{11.6^2} + 1}\right)\right)^{\frac{1}{2}} \\ &\times (115.9 + 258.1f)\cdot e^{-0.29f} \end{aligned} \tag{9}$$

and the following nonlinear functions $G_f(a_f)$ [15] (frequency f in cycles/degrees)

$$G_f(a_f) = \frac{(-0.03\log f + 0.3)\left(\frac{a_f}{L}\right)^{\frac{0.8f^{1.7}}{0.5+f^{1.7}}} - \mathrm{CSF}_f^{-1}}{((-0.03\log f + 0.3)\mathrm{CSF}_f)^{-\frac{0.5+f^{1.7}}{0.8f^{1.7}}} + \frac{a_f}{L}}. \tag{10}$$

It is important to note that the metric weight for each coefficient in (8) has two contributions: one constant term (the CSF) and one amplitude-dependent term that vanishes for low amplitudes. This second term comes from a nonlinear correction to the linear threshold response described by the CSF. These two terms in the metric give two interesting particular cases of the MPE formulation. First, if a simple linear perception model is assumed, a CSF-based MPEG-like quantizer is obtained. If the nonlinear correction in (8) is neglected, uniform quantizers are obtained for each coefficient and N_f becomes proportional to the CSF, which is one of the recommended options in the JPEG and MPEG standards [1]–[3]. Second, if both factors of the metric are taken into account, the algorithm of [13], [14], [26] is obtained: the quantization step size is input-dependent and proportional to the JNDs and bit allocation is proportional to the integral of the inverse of the JNDs. From now on, these two cases will referred to as linear and nonlinear MPE, respectively.

The CSF-based (linear MPE) quantizer used in MPEG [1]–[3] and the proposed nonlinear MPE quantizer [13], [14], [26], represent different degrees of approximation to the actual quantization process, Q_p, eventually carried out by the HVS. The scheme that takes into account the perceptual amplitude nonlinearities will presumably be more efficient in removing the subjective redundancy from the DFD.

Fig. 1 shows the product $N_f \cdot \lambda_f(a_f)$ for the linear (MPEG-like) and the nonlinear MPE quantizers. This product represents the number of quantization levels per unit of area in the frequency and amplitude plane. This surface is a useful description of where a quantizer concentrates the encoding effort [14]. Fig. 2 shows the bit allocation solutions (number of quantization levels per coefficient N_f) in the linear and the nonlinear MPE cases. Note how the amplitude nonlinearities enlarge the bandwidth of the quantizer in comparison to the CSF-based case. This enlargement will make a difference when dealing with wide spectrum signals like the DFD.

C. Introducing HVS Temporal Properties in the Prediction Loop

The previous considerations about optimal MPE 2-D transform quantizers can be extended to three-dimensional (3-D) spatio-temporal transforms. The HVS motion perception models extend the 2-D spatial filter bank to nonzero temporal frequencies [42], [43]. The CSF filter is also defined for moving gratings [20] and the contrast discrimination curves for spatio-temporal gratings show roughly the same shape as the curves for still stimuli [21], [22]. By using the 3-D CSF and similar nonlinear corrections for high amplitudes, the expression of (8) could be employed to measure differences between local moving patterns. In this way, optimal MPE quantizers could be defined in a spatio-temporal frequency transform domain. However, the frame-by-frame nature of any motion compensated scheme makes the implementation of a 3-D transform quantizer in the prediction loop more difficult.

In order to exploit the subjective temporal redundancy removal to some extent, the proposed 2-D MPE quantizer can be complemented with one-dimensional (1-D) temporal filtering based on the perceptual bit allocation in the temporal dimension. This temporal filter can be implemented by a simple finite impulse response weighting of the incoming error frames. The temporal frequency response of the proposed 1-D filter is set proportional to the number of quantization levels that should be allocated in each temporal frequency frame of a 3-D MPE optimal quantizer. For each spatio-temporal coefficient $\mathbf{f} = (f_x, f_t)$, the optimal number of quantization levels is given

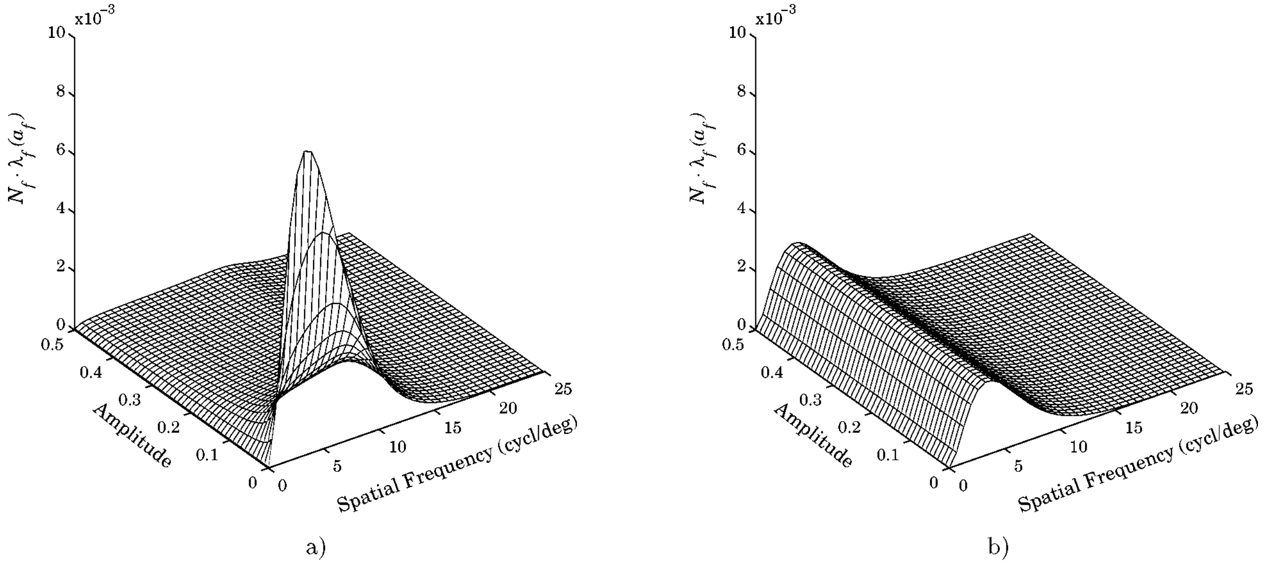

Fig. 1. Relative number of quantization levels allocated in the frequency and amplitude plane for a) nonlinear MPE and b) linear MPE quantizers. The surfaces are scaled to have unit integral (the same total number of quantization levels). The distribution of the quantization levels in amplitude for a certain coefficient is just the corresponding slice of the surface at the desired frequency. The MPE design [(5) and (7)] implies that this surface is proportional to the metric $N_f \cdot \lambda_f(a_f) \propto W(a_f)^{1/2}$ so different perception models (different metrics) give rise to a different distribution of quantization levels. Note that the distribution is uniform for every frequency in the linear MPE (MPEG-like) case and nonuniform (peaked at low amplitudes) in the nonlinear MPE case.

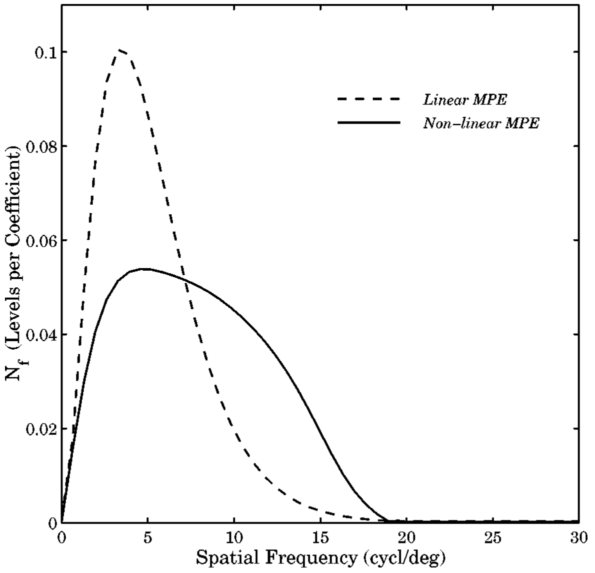

Fig. 2. Bit allocation results (relative number of quantization levels per coefficient) for the linear MPE (MPEG-like case) and for the nonlinear MPE case. The curves are scaled to have unit integral (the same total number of quantization levels). In the linear case, the metric is just the square of the CSF and then $N_f \propto \mathrm{CSF}_f$ as recommended by JPEG and MPEG. A more complex (nonlinear) model gives rise to a wider quantizer bandpass.

by (7). Integrating over the spatial frequency, the number of quantization levels for that temporal frequency is

$$N_{f_t} = \sum_{f_x} N_{\mathbf{f}} = \frac{1}{2k} \sum_{f_x} \int W_{\mathbf{f}}(a_{\mathbf{f}})^{1/2} \, da_{\mathbf{f}}. \qquad (11)$$

Fig. 3(a) shows the number of quantization levels for each spatio-temporal frequency of a 3-D nonlinear MPE quantizer. This is the 3-D version of the 2-D nonlinear bit allocation of Fig. 2. Note that (except for a scale factor) the spatial frequency curve for the zero temporal frequency is just the solid curve of Fig. 2. Fig. 3(b) shows the temporal frequency response that is obtained by integrating over the spatial frequencies.

IV. Perceptual Feedback in the Motion Estimation

Any approximation to the actual perceptual quantization process Q_p has an obvious application in the DFD quantizer design, but it may also have interesting effects on the computation of the DVF if the proper feedback from the DFD quantization is established in the prediction loop.

If all the details of the DFD are considered to be of equal importance, we would have an *unweighted* splitting criterion as in the difference-based criteria [6], [29], [32], [33] or as in the spatial entropy-based criterion of Dufaux *et al.* [7], [8]. However, as the DFD is going to be simplified by some nontrivial quantizer Q_p, which represents the selective bottleneck of early perception, not every additional detail predicted by a better motion compensation will be significant to the quantizer. In this way, the motion estimation effort has to be focused on the moving regions that contain *perceptually significant motion information*. In order to formalize the concept of perceptually significant motion information, the work of Watson [36] and Daugman [37] on entropy reduction in the HVS should be taken into account. They assume a model of early contrast processing based on a pair (T, Q_p) and suggest that the entropy of the cortical scene representation (a measure of the perceptual entropy of the signal) is just the entropy of the quantized version of the transformed image. Therefore, a measure of the perceptual entropy H_p of a signal $\mathbf{A}$ is

$$H_p(\mathbf{A}) = H\left(Q_p[\mathrm{T} \cdot \mathbf{A}] \right). \qquad (12)$$

Using this perceptual entropy measure (which is simply the entropy of the output of a MPE quantizer), we can propose an explicit definition of what perceptually significant motion information is. Let us motivate the definition as follows. Given a cer-

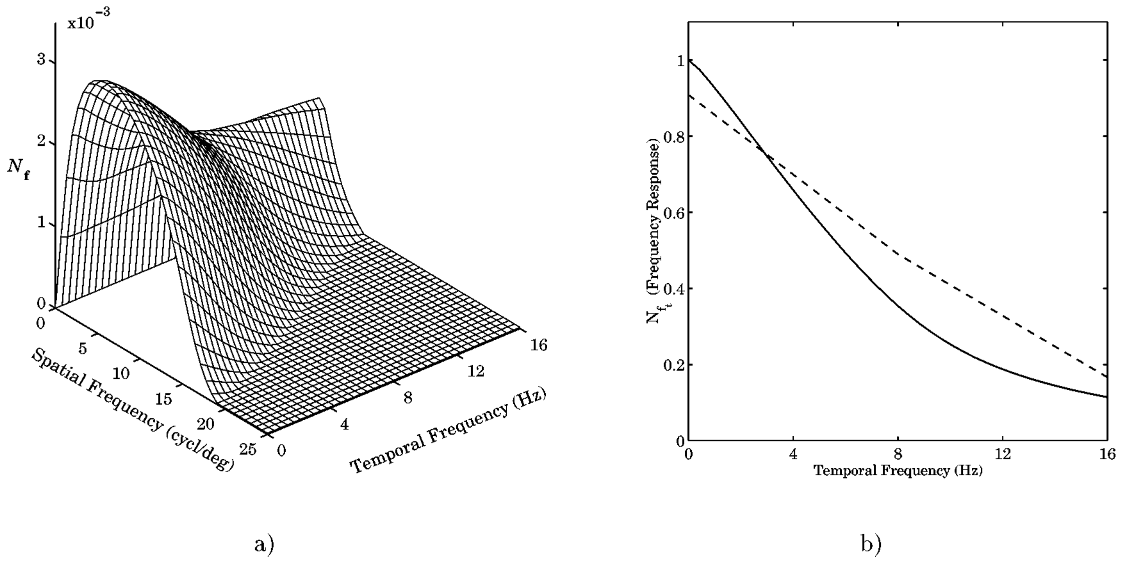

Fig. 3. (a) Nonlinear MPE bit allocation results in the 3-D spatio-temporal frequency domain (relative number of quantization levels per 3-D coefficient). The surface is scaled to have unit integral. (b) Frequency response of the perceptual temporal filter, proportional to N_{f_t}. The solid line corresponds to the theoretical curve and the dashed line stands for the actual frequency response obtained with the fourth-order FIR filter used in the experiments (see Section V). The coefficients of the filter in the temporal domain are: 0.0438, 0.1885, 0.4443, 0.1885, and 0.0438.

tain motion description, DVF, with entropy $H(\mathrm{DVF})$, a prediction of the next frame can be done. Some particular error, DFD, will be obtained, with a perceptual entropy $H_p(DFD)$. If additional motion information is available, $\Delta H(\mathrm{DVF}) > 0$ (more complex quadtree segmentation and more motion vectors), one would expect a reduction of the perceptual information of the remaining DFD, i.e., $\Delta H_p(DFD) < 0$. Let us define this additional motion information as perceptually significant only if it implies a greater reduction in the perceptual entropy of the prediction errors: *$\Delta(DVF)$ is perceptually significant if*

$$\Delta H(DVF) < -\Delta H_p(DFD). \tag{13}$$

Broadly speaking, some additional motion information is perceptually significant if it increments the perceptual information of the prediction more than its own volume.

If each motion refinement in a variable-size BMA, $\Delta H(\mathrm{DVF}) = H(\mathrm{DVF}_{\mathrm{split}}) - H(\mathrm{DVF}_{\mathrm{nosplit}})$, is required to be perceptually significant, the following *perceptually weighted* splitting criterion arises: *a block of the quadtree structure should be split if*

$$\begin{aligned} H(\mathrm{DVF}_{\mathrm{split}}) &+ H_p(\mathrm{DFD}_{\mathrm{split}}) \\ &< H(\mathrm{DVF}_{\mathrm{nosplit}}) + H_p(\mathrm{DFD}_{\mathrm{nosplit}}) \end{aligned} \tag{14}$$

where $H(\mathrm{DVF})$ is the entropy of the DPCM coded DVF plus the information needed to encode the quadtree structure and $H_p(\mathrm{DFD})$ is the perceptual entropy of the residual error signal.

Equation (14) has the same form as the criterion proposed by Dufaux and Moscheni [7], [8], except for the way in which the entropy of the DFD is computed. In this case, the unweighted entropy of the DFD in the spatial domain is replaced by the perceptually weighted entropy measure in the appropriate encoding domain. The consideration of the entropy of the quantized DFD (or perceptual entropy) is the main difference between the suboptimal approach of Dufaux and Moscheni [7], [8], and the optimal approaches [9]–[12]. The optimal approaches only differ in the way the problem is stated: While in [9]–[11] an explicit rate-distortion sum is minimized, in [12] a fixed distortion is assumed (constant MPE quantizer) and a monotonically decreasing behavior for the rate is imposed. These optimal algorithms do not necessarily find the absolute minimum in the rate-distortion sense but only local minima; however, as the actual DFD entropy is used, the performance of the suboptimal result is always improved. It is interesting to note that the reasoning about the perceptual relevance of the motion information presented here leads to (14), which takes into account the quantized DFD entropy, in a natural way. As long as [9]–[12] use CSF-based quantizers, all these optimal approaches can also be referred to as perceptually weighted variable-size BMAs. In what follows, the algorithm of Dufaux *et al.* [7], [8] and the algorithm from (14) will be compared and referred to as unweighted and perceptually weighted variable-size BMA, respectively.

The perceptual quantizer constraint on the motion estimate comes from the particular video coding application in which the actual bit-rate has to be minimized [9]–[12]. However, the benefits of including the properties of the biological filters in the motion estimation may go beyond the rate-distortion optimization. The bandpass shape of human sensitivity (Fig. 2) gives a scale-dependent measure of the perceptual entropy of the DFD. As some frequency bands (some scales) have more perceptual importance than others, the application of the perceptual criterion results in a different splitting behavior in the different levels of the multigrid structure. Fig. 4 qualitatively shows how a bandpass criterion may give a scale-dependent splitting result. In coarse levels of the multigrid (left side figures), the spatial support of the DFD is large due to the displacement of large blocks. The uncertainty relationship $\Delta x \cdot \Delta f = k$ leads to a DFD with narrow bandwidth in the case of a large DFD support. Conversely, in the fine levels of the multigrid (right-side

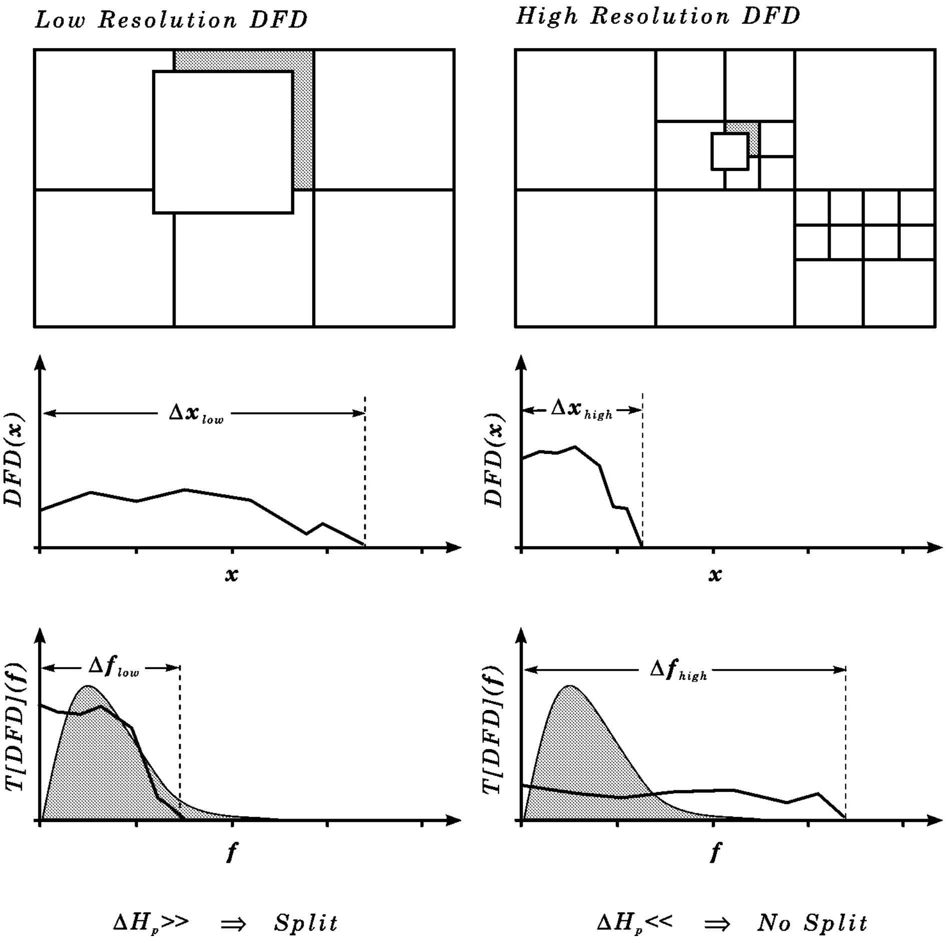

Fig. 4. Scale-dependent splitting strategy due to perceptual feedback. The dashed regions in the frequency domain represent the bit allocation of the MPE quantizer. They determine the frequency band which is considered to compute the perceptual entropy of the signal. For a given energy and resolution level, the spatial extent and the frequency bandwidth of the DFD (thick solid lines) are related by the uncertainty relation $\Delta x \cdot \Delta f = k$. The bandwidth of the DFD will depend on the resolution, giving rise to a different splitting behavior when using a bandpass splitting criterion such as the perceptual entropy.

figures), the DFD is spatially localized, giving rise to a broadband error signal. If the complexity measure is more sensitive to the complexity of the signal in low- and middle-frequency bands, the splitting criterion will be tolerant in the coarse levels of the multigrid and will be strict in the high-resolution levels. The next section will show that this scale-dependent behavior is useful for discriminating between significant and noisy motions.

V. Experiments and Discussion

Four experiments on several standard sequences [44] were carried out at a fixed bit-rate:

- Different quantizers with the same motion estimation.
- Different motion estimations with the same quantizer.
- Relative relevance of the improvements in the motion estimation and the quantization.
- Proposed scheme versus previous comparable schemes (H.263, MPEG-1).

We are interested in two different comparisons: 1) quality of the reconstructed signal and 2) usefulness of the motion flows for higher-level purposes. To this end, examples of the reconstructed frames and subjective distortion measures using a perceptually meaningful metric [15], [16] are given in each case. Also, a well-known motion-based segmentation algorithm for high-level video coding [17], [18] has been used with both perceptually weighted and perceptually unweighted BMAs to test their respective usefulness.

In every experiment, the quantizer was adapted for each group of pictures to achieve the desired bit-rate (200 kb/s with QCIF format). In order to highlight the relative differences among the different approaches considered, only the first frame was intracoded and only forward prediction was used in the remaining frames (i.e., no bidirectional interpolated frames nor additional intracoded frames were introduced). As a consequence, the results may seem abnormally distorted at this rate. The DC coefficient of error signals is basically zero [45], so the luminance information of the original blocks was used to normalize the amplitudes of the error blocks. These luminance values were not quantized but DPCM coded from block to block as in the JPEG standard. The 1-D temporal filter was implemented by a linear-phase FIR filter using least-squares error minimization in the frequency response. A simple fourth-order filter was used to restrict the buffer requirements (see Fig. 3 for the achieved frequency response and the filter coefficients).

A maximum of five resolution levels (blocks from $64 \times 64$to 4×4) were used in the variable-size BMA quadtrees. Blocks of size 8×8 were used in the fixed-size BMA. The *n*-step displace-

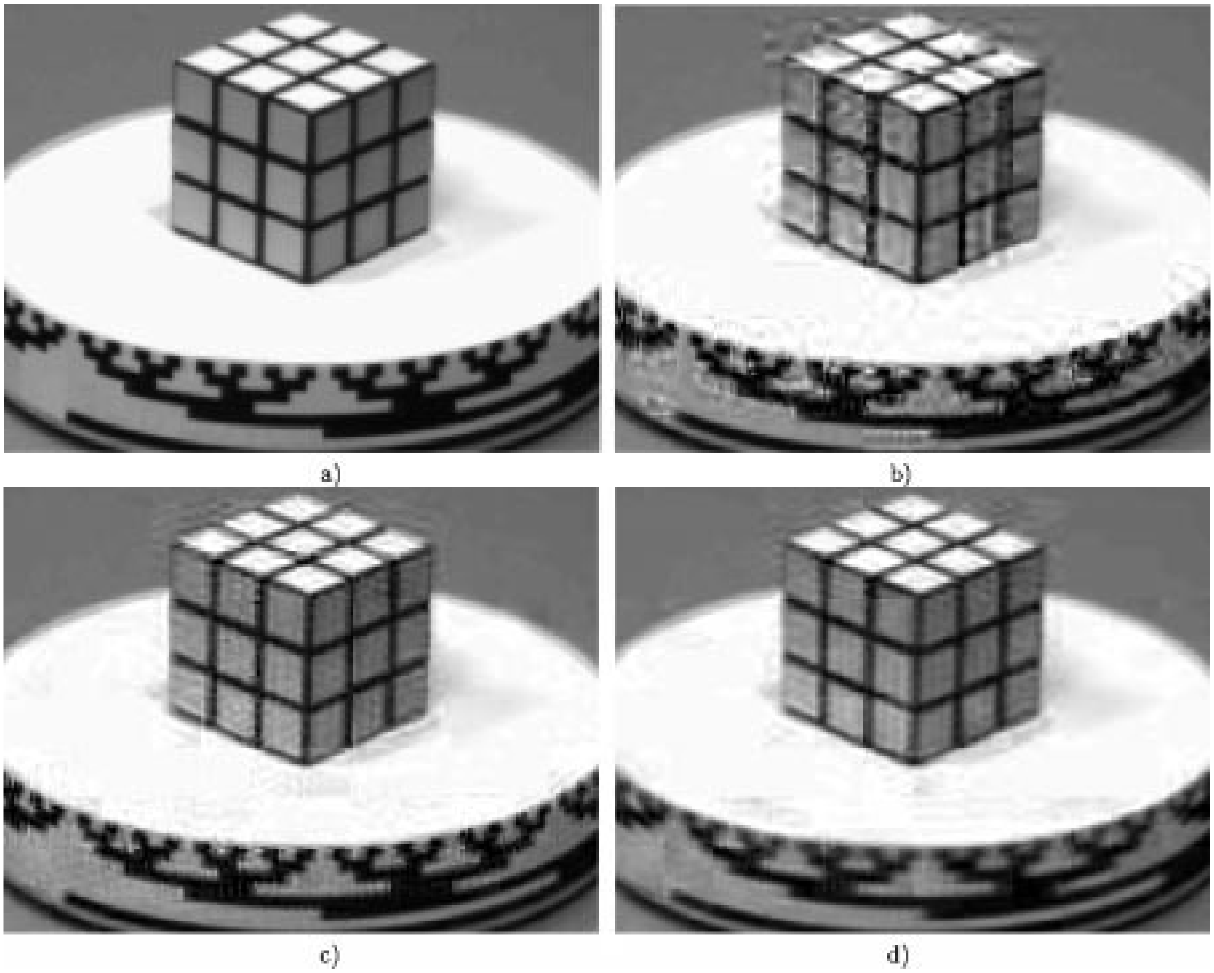

Fig. 5. Quantization results with a fixed motion estimation algorithm (unweighted variable-size BMA). (a) Original (detail of frame 7 of the *Rubik* sequence). (b) Two-dimensional linear MPE, uniform MPEG-like quantization. (c) Two-dimensional nonlinear MPE. (d) Two-dimensional nonlinear MPE and temporal filtering.

ment search [8] and integer-pixel accuracy was used in every resolution level of the BMAs. The usual correlation was used as a similarity measure.

The definitive proof of the subjective benefits of an algorithm is a set of psychophysical experiments on the quality of the reconstructed sequences, but it requires time-consuming experiences involving several observers. An approximate, but more practical, approach is to use perceptually meaningful distortion measures. This is not a simple issue [46], [47]. However, basic facts as the spatial frequency sensitivity make a fundamental difference between plain Euclidean distortion metrics (e.g MSE or SNR) and any (even the simplest) perceptually weighted metric. In the experiments, the perceptual distortions were computed through (2) using the nonlinear perceptual metric of (8) [15], [16] on a frame-by-frame basis. This metric incorporates the basic elements of early achromatic visual processing [46], [47]: luminance adaptation, spatial frequency channels, frequency dependent filtering, contrast masking, and (quadratic) probability summation. The squared distortion was computed for each DCT block in a frame and averaged across the frame. This straighforward frame-by-frame implementation may neglect some temporal factors, but it still gives a rough approximation to the observers opinion and is thus useful for confirming the results.

A. Experiment 1: Different Quantizers with the Same Motion Estimation

The performance of the linear MPE (MPEG-like) quantizer and the proposed nonlinear MPE quantizers was compared using the same (perceptually unweighted H.263-like) motion estimation. Fig. 5 shows a representative example of the kind of errors obtained with the different quantizers at a fixed bit-rate. Fig. 6 shows the increase of the perceptual distortion in the different reconstructions of the *Rubik* sequence.

The consideration of the (2-D or 3-D) nonlinearities introduces a substantial improvement in comparison to the linear MPE quantizer. The temporal filtering smooths the reconstructed sequence and reduces to some extent the remaining blocking effect and busy artifacts of the 2-D nonlinear approach. However, despite the eventual visual advantages of this temporal filtering, the key factor in the improvements of the proposed quantizers is the consideration of the amplitude nonlinearities and the corresponding enlargement of the spatial quantizer bandwidth (Fig. 2). The reconstructed examples and the computed distortion confirm this point. This enlargement implies that the quantized signal keeps some significant details otherwise discarded, avoiding the rapid degradation of the reconstructed signal.

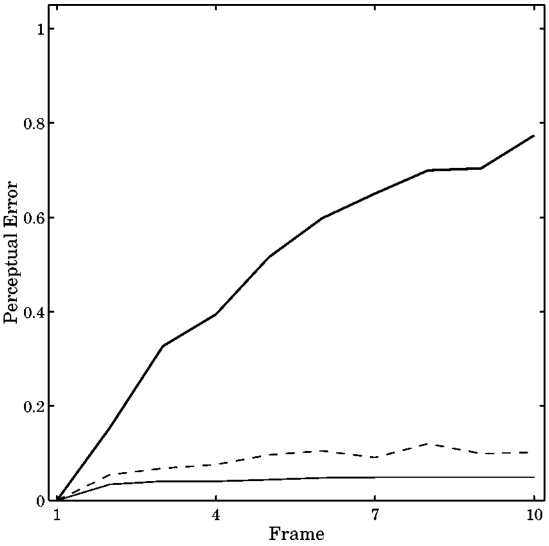

Fig. 6. Perceptual distortion measures for the frames of the *Rubik* sequence using different quantizers with the same (unweighted H.263-like) motion estimation. All the distortion results presented in the paper are normalized by the worst distortion obtained at the last frame of the *Rubik* sequence using an H.261 or MPEG-1 approach. The thick solid line correspond to the linear MPE quantizer. The thin solid line correspond to the 2-D nonlinear MPE quantizer and the thin dashed line correspond to the 2-D nonlinear quantizer with temporal filtering. In this 3-D case, the frame-by-frame implementation of the perceptual distortion measure may slightly overestimate the visual effect of the frame blurring introduced by the temporal filter. Despite this fact, the perceptual distortions reveal that the amplitude nonlinearities substantially improve the reconstruction results.

A very interesting consequence is that the (2-D or 3-D) nonlinear quantizers keep the distortion bounded over a large group of frames (compare the behavior of the linear and nonlinear distortion curves). In this way, the need to introduce bit-consuming intracoded frames is reduced.

B. Experiment 2: Different Motion Estimations with the Same Quantizer

The proposed perceptually weighted variable-size BMA and the unweighted variable-size BMA were compared in this experiment using the same linear MPE (MPEG-like) quantizer. The results with a fixed-size 8×8 BMA are also included as a reference.

The quantitative rationale behind entropy constrained approaches is saving motion information to improve the DFD encoding and have a better reconstruction. From this point of view, the only reason to take into account the perceptual quantization is an improvement of the reconstructed sequence. The qualitative rationale to take into account the perceptual quantization is to include a perceptual criterion to decide when some motion information is significant. Accordingly, the effects of the perceptual weight in the motion estimation were tested in two different ways. First, the effect of the savings in motion information on the signal quality was analyzed. Second, the usefulness of the motion information for higher-level purposes was tested.

The Table I and Fig. 7 analyze the bit allocation performance of the motion estimation algorithms. Table I shows the percentage of the total bit-rate used for the motion flow. Fig. 7 shows the reduction of the bit-rate in the *Taxi* sequence while refining the motion estimate with the weighted and unweighted variable-size BMAs. The consideration of the perceptual quantizer in the motion estimation certainly minimizes the entropy of the motion flow and the total entropy as claimed in [9]–[12]. The problem with the unweighted criterion is that it is too permissive at high-resolution levels. In this way, too many blocks are split increasing the motion information without decreasing the information content of the DFD, i.e., perceptually negligible motion information is being added.

However, do these motion information savings in [9]–[12] have a significant effect on the reconstructed quality? Fig. 8 shows a representative example of the decoded results using the same MPEG-like quantizer and the different considered motion estimations at a fixed bit-rate (frame 7 of the *Taxi* sequence). Fig. 9 shows the increase of the perceptual distortion in the different reconstructions of the *Taxi* sequence. The distortions here are relatively lower than in the *Rubik* sequence because the area of the moving regions (and hence the distorted area) is also lower. These results show that both variable-size BMAs make some difference in quality with regard to the fixed-size H.261 BMA due to the savings in DVF information. However, the practical advantages of the better bit allocation between DVF and DFD over the suboptimal variable-size BMA are not evident. The benefits of the better bit allocation in comparison to the suboptimal variable-size BMA are so small (Table I and Fig. 7) that in practice (Figs. 8 and 9), they cannot be exploited by the quantizer to give a better encoded DFD. That is, any adaptive DVF consumes such a small portion of the total bit-rate that there is no significant difference in the reconstruction between the different variable-size BMA algorithms.

The scale-independent behavior of the unweighted algorithm leads to too many splittings at high resolution. This suboptimal result has no effect in the reconstructed sequence, but it may give rise to noisy, less meaningful, motion information. In order to test the meaningfulness of the motion information obtained with the different algorithms, a well-known motion-based segmentation algorithm [17], [18] was initialized with the different flows.

The layer identification algorithm [17], [18] starts from an arbitrary initial segmentation, which is refined by estimating the affine motion model of 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 flows under consideration, the following segmentation error measure has been used. Let ϵ_i be the dissimilarity between each (manually segmented) object O_i and its corresponding mask M (M is the mask that maximizes $M_j \cap O_i$ for all the obtained masks M_j)

$$\epsilon_i = \int_{M \cap O_i^C} D_i(x)\, dx + \int_{O_i \cap M^C} D_i(x)\, dx \qquad (15)$$

where C stands for the set complement and $D_i(x)$ is the Chamfer distance, i.e., the distance from x to the boundary of the object O_i [48]. This distance progressively penalizes unmatched pixels that are far from the object boundary. The

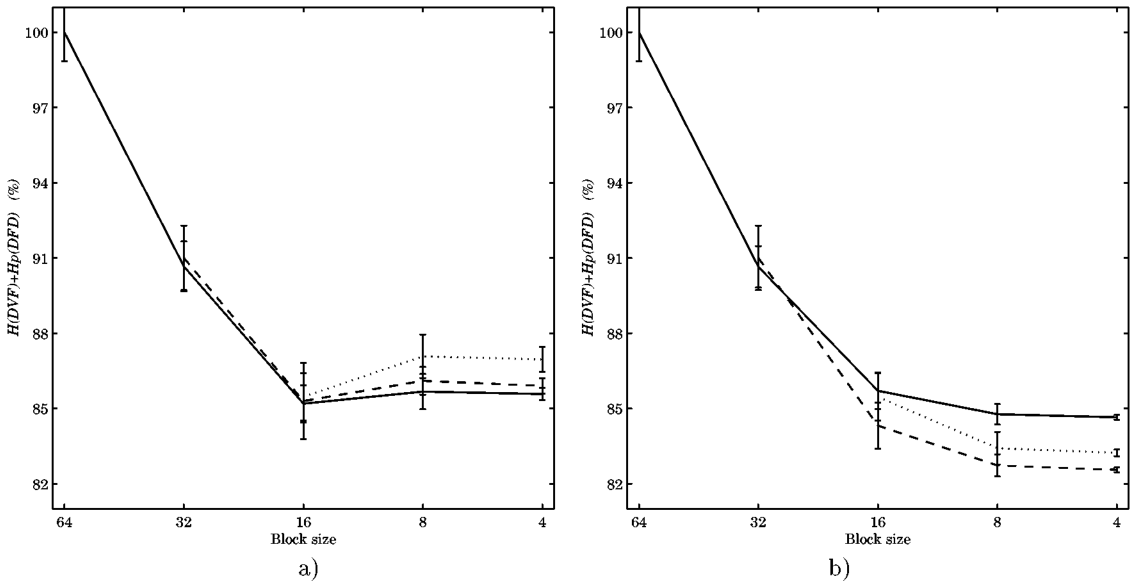

Fig. 7. Bit-rate of DVF and DFD while refining the motion estimate. The values are given as a percentage of the total entropy at the lowest resolution level. The curves represent the average percentage of the bit-rate across the frames and the error bars represent the standard deviation. The different lines indicate different starting resolutions (initial block sizes): solid 64×64, dashed 32×32, and dotted 16×16. (a) Unweighted spatial entropy splitting criterion. (b) Perceptually weighted splitting criterion.

TABLE I
PERCENTAGE OF THE TOTAL BIT-RATE USED FOR THE MOTION FLOW (DVF)

	TAXI	RUBIK	YOSEMITE	TREES	Average
FSBMA	12.57	21.71	46.19	39.21	30 ± 8
Unweight. VSBMA	1.89	3.29	8.69	4.97	4.7 ± 1.5
Weighted VSBMA	0.94	1.87	4.32	3.24	2.6 ± 0.7

segmentation error for a frame ϵ is the sum of the individual segmentation errors ϵ_i. It has been empirically found that this error measure adequately describes the intuitive quality of the segmentation (Figs. 10 and 11 and their particular errors illustrate this point).

Several experiments were carried out using the layer identification algorithm [17], [18] with the three different DVFs: fixed-size BMA, unweighted variable-size BMA and perceptually weighted variable-size BMA. Results regarding two standard sequences, *Taxi* and *Rubik*, are considered here because they represent extreme cases for the segmentation algorithm: small moving objects with simple translations and one large 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 II shows the average segmentation error (normalized by the maximum in each sequence), the number of iterations and the number of identified regions in each case. Figs. 10 and 11 show examples of the motion flows and the segmentations achieved in particular frames of the sequences along with the corresponding segmentation error. The differences in robustness, coherence and meaningfulness between the results of the different algorithms are apparent in the motion flows (Figs. 10 and 11). The quantitative results of the segmentation confirm this intuitive impression. The block appearance of the layers is obviously due to the sparseness of the input flows. The segmentation algorithm is usually initialized with dense flows (one motion vector per pixel) [17], [18], not with sparse (one motion vector per block) flows. However, this worst case situation is appropriate to highlight the usefulness of each flow.

The following trends can be identified from the obtained results: 1) 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* –large object- and *Taxi* –small objects–); however, 2) the segmentation is very sensitive to the robustness and coherence of the sparse flow. This implies that the block size cannot be arbitrarily reduced to improve the segmentation resolution: Despite its higher density, a more noisy flow from a too small block size BMA is worse.

This general behavior 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 (six data) are needed to segment a region. In real (noisy) situations, more vectors per region will be needed. Consequently, the segmentations of large objects are comparatively better. On the other hand, as the number of motion measurements is reduced when using sparse flows (compared to dense flows), their robustness becomes critical. This is why the results based on fixed-size BMA (sparse and noisy) 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

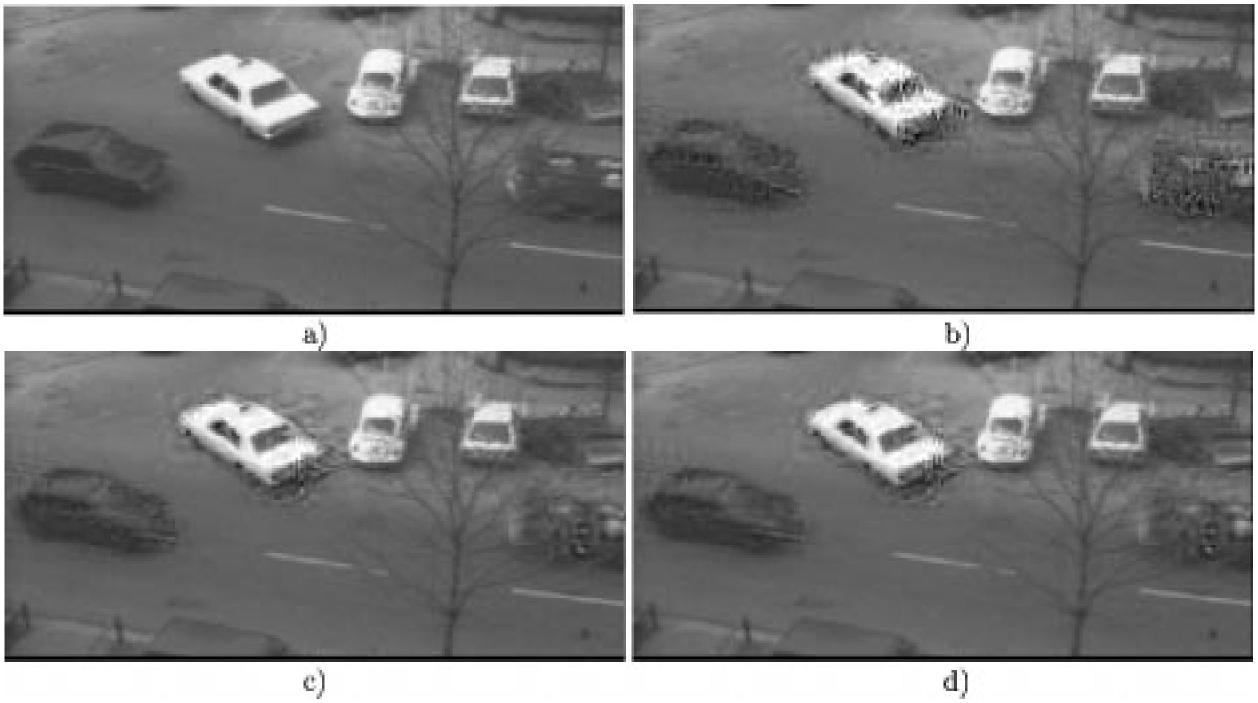

Fig. 8. Reconstruction results with different motion estimations and a fixed MPEG-like quantization. (a) Original (detail of frame 7 of the *Taxi* sequence). (b) Fixed-size BMA. (c) Unweighted variable-size BMA. (d) Weighted variable-size BMA.

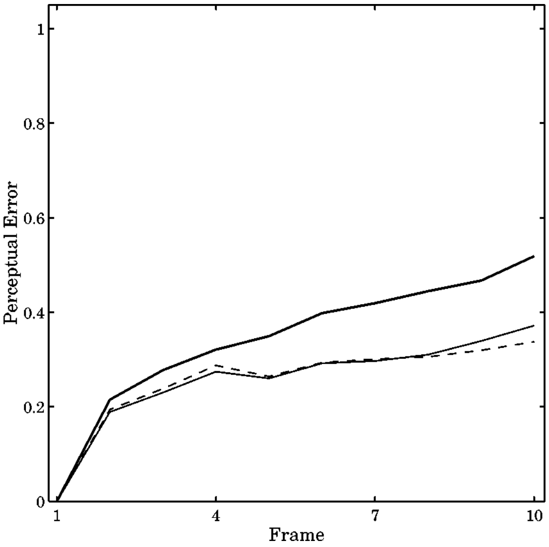

Fig. 9. Perceptual distortion measures for the frames of the *Taxi* sequence using different motion estimations with the same (MPEG-like) quantizer. The thick solid line corresponds to the fixed-size BMA. The thin solid line corresponds to the unweighted variable-size BMA and the thin dashed line corresponds to the weighted variable-size BMA.

segmentation error (see Table II). Moreover, a closer estimation of the actual number of moving objects is obtained due to the smoothness of the flow within moving regions, which prevents false splittings.

To summarize, despite the fact that a better bit allocation between DVF and DFD (Table I and Fig. 7) does not improve the quality of the decoded sequence (Figs. 8 and 9), the segmentation results (Table II and Figs. 10 and 11) show that perceptual weighting gives rise to a more meaningful flow: it makes the sparse flow more suitable for obtaining robust rough segmentations.

C. Experiment 3: Relative Relevance of the Improvements in the Motion Estimation and the Quantization

To study the combined effect and the relative advantages of the proposed improvements, the four possible combinations of motion estimation and quantization algorithms were compared at a fixed bit-rate. The perceptually weighted and the unweighted (suboptimal) variable-size BMA were combined with the the 2-D linear (MPEG-like) and the 2-D nonlinear MPE quantizers. Fig. 12 shows an example of a reconstructed frame using the four different approaches considered. Fig. 13 shows the corresponding distortion for each frame of the *Rubik* sequence. These reconstruction examples and error results confirm what expected from the separate analysis of the previous sections: The quality improvement due to the use of the appropriate entropy measure in the motion estimation is negligible compared to the benefits of a better quantization.

In this case, the simple increase in the quantizer bandpass that comes from considering the visual nonlinearities (Fig. 2) may explain the enhancement. Due to the noisy (high-frequency) nature of the error signal, wide band quantizers may be better than narrower band (CSF-based) quantizers. In fact, the benefits of this enhanced quantizer are more apparent on video frames (Figs. 5 and 12) than in still images [13], [14]. The interesting constant distortion result reported in Section V-A (Fig. 6) is also reproduced here: While the perceptual distortion increases quickly when using the linear quantizer, it remains constant with the nonlinear MPE quantizer regardless the motion estimation algorithm.

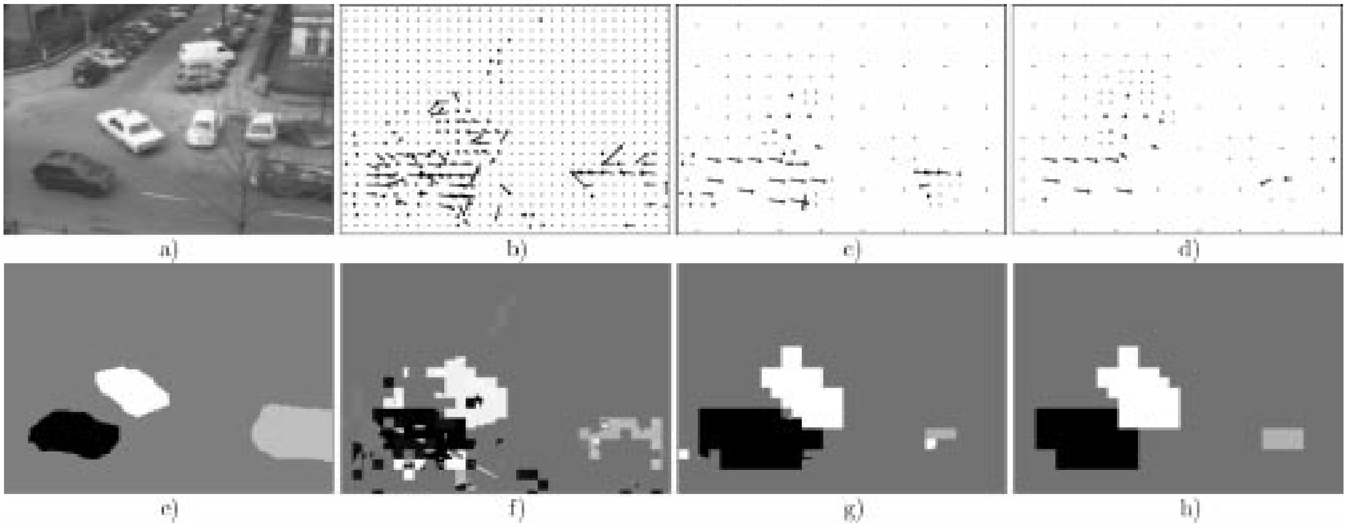

Fig. 10. DVFs and motion-based segmentation results (*Taxi*, frame 9). (a) Original frame. (b) Fixed-size BMA flow. (c) Unweighted variable-size BMA flow. (d) Perceptually weighted variable-size BMA flow. (e) Ideal segmentation. (f) Segmentation with fixed-size BMA ($\epsilon = 1.02$). (g) Segmentation with unweighted variable-size BMA, ($\epsilon = 0.49$). (h) Segmentation with perceptually weighted variable-size BMA, ($\epsilon = 0.35$).

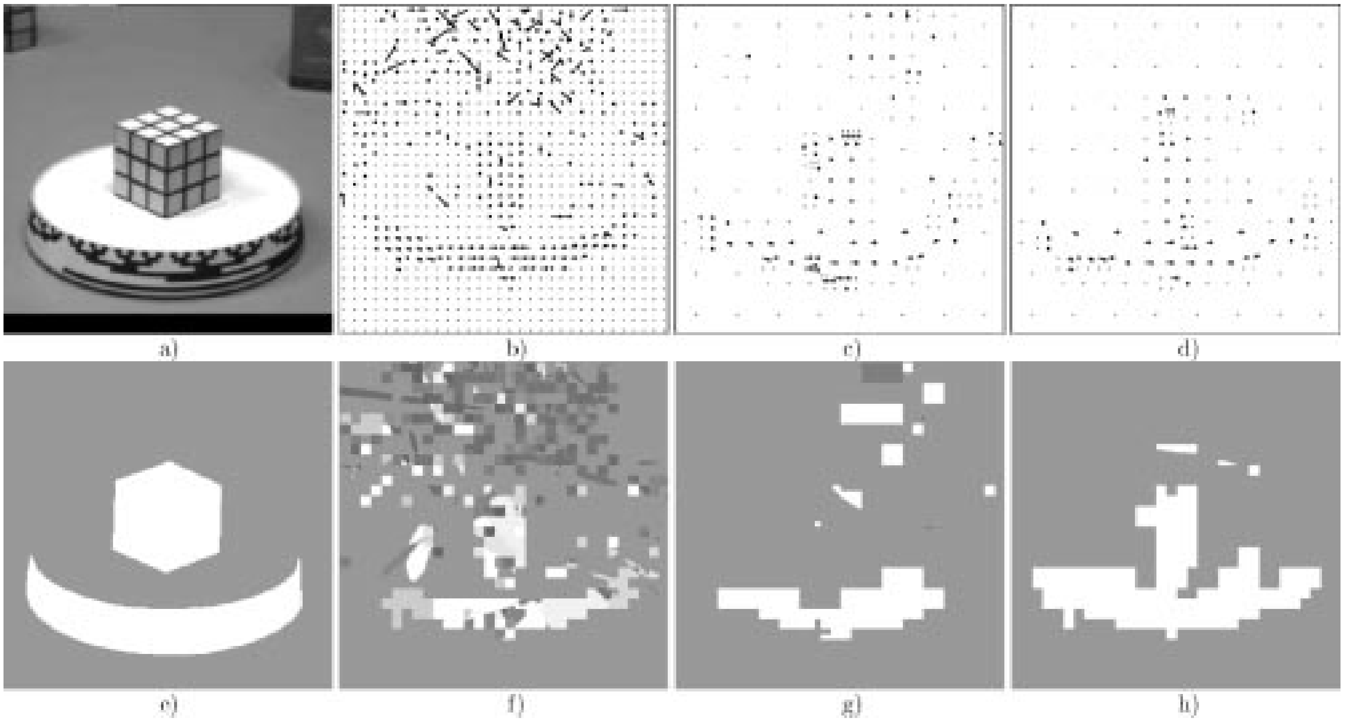

Fig. 11. DVFs and motion-based segmentation results (*Rubik*, frame 3). (a) Original frame. (b) Fixed-size BMA flow. (c) Unweighted variable-size BMA flow. (d) Perceptually weighted variable-size BMA flow. (e) Ideal segmentation. (f) Segmentation with fixed-size BMA ($\epsilon = 0.98$). (g) Segmentation with unweighted variable-size BMA, ($\epsilon = 0.51$). (h) Segmentation with perceptually weighted variable-size BMA, ($\epsilon = 0.25$).

TABLE II
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
Weighted VSBMA	0.27 ± 0.03	7 ± 6	2.1 ± 0.3	0.49 ± 0.14	4 ± 1	4.3 ± 0.5

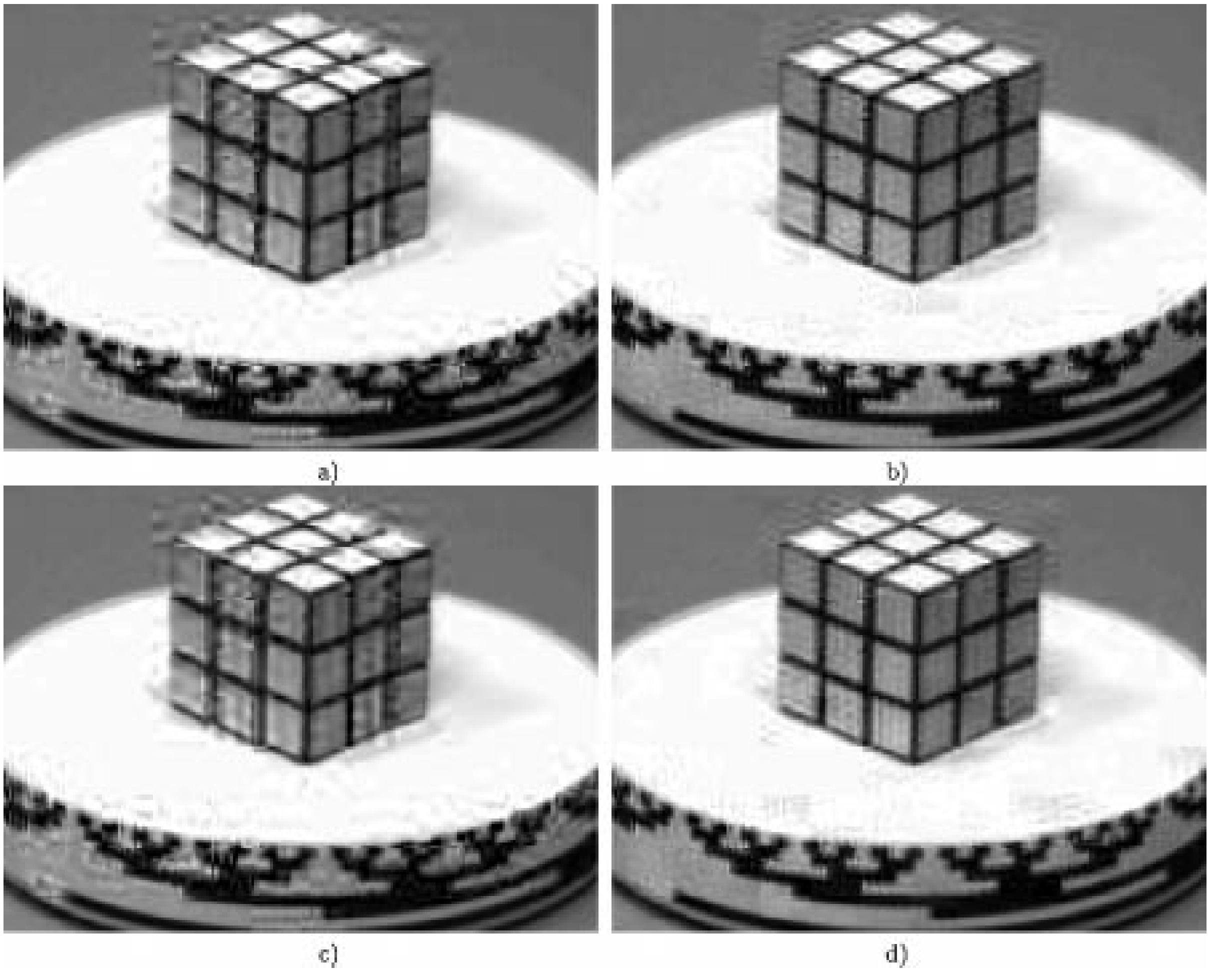

Fig. 12. Decoded results using different combinations of quantizers and motion estimations. (a) Unweighted variable-size BMA and 2-D linear MPE, MPEG-like, quantization. (b) Unweighted variable-size BMA and 2-D nonlinear MPE quantization. (c) Perceptually weighted variable-size BMA and 2-D linear MPE, MPEG-like, quantization. (d) Perceptually weighted variable-size BMA and 2-D nonlinear MPE quantization.

D. Experiment 4: The Proposed Scheme Versus Previous Comparable Schemes

In this section, the considered elements of the motion compensated video coder were combined to simulate and compare previously reported schemes (MPEG-1 or H.261 and H.263) and the proposed ones. MPEG-1 and H.261 use a fixed-size BMA and a linear (CSF-based) MPE quantizer. A regular implementation of H.263 use an unweighted splitting criterion variable-size BMA and a linear MPE quantizer. The video coder scheme proposed here use a perceptually weighted variable-size BMA and a nonlinear MPE quantizer (either 2-D or 3-D).

Examples of the decoded frames at a fixed bit-rate are shown in Figs. 14 and 15. The perceptual distortion for each frame of the reconstructed sequences is shown in Fig. 16. Both frames and distortions confirm the improvements of the proposed schemes compared to the previous similar schemes in terms of subjective quality at a fixed bit-rate.

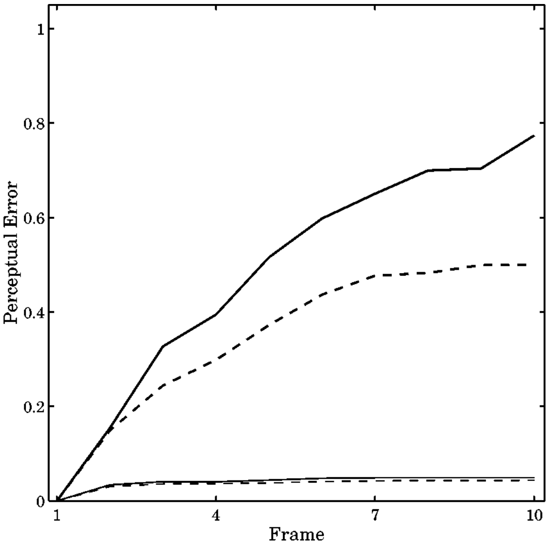

Fig. 13. Perceptual distortion measures for the frames of the *Rubik* sequence using different combinations of quantizers and motion estimations. The thick lines correspond to the approaches using the linear MPE quantizer. The thin lines correspond to the approaches using the 2-D nonlinear MPE quantizer. The dashed lines correspond to the approaches using perceptually weighted variable-size BMA and the solid lines correspond to the approaches using unweighted variable-size BMA.

VI. Final Remarks

The current motion compensated video coding standards include very basic perceptual information (linear threshold models) only in the quantizer design. In this paper a multigrid motion compensated video coding scheme based on a more

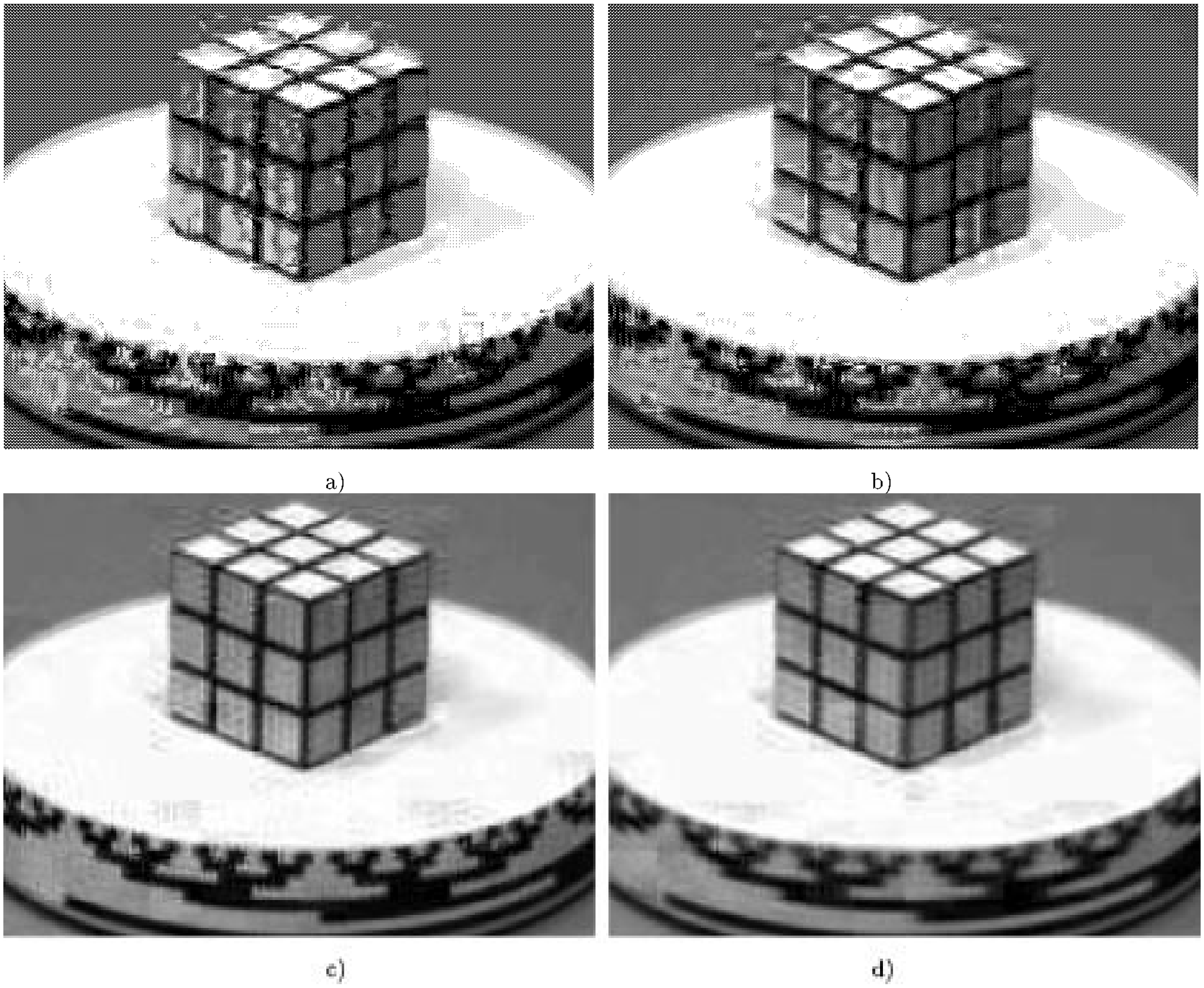

Fig. 14. Reconstruction results with the *Rubik* sequence using previously reported encoding configurations (a-b) and the proposed 2-D or 3-D alternatives (c-d). (a) Fixed size BMA for motion estimation and MPEG-like quantization (linear MPE). (b) Unweighted variable-size BMA and MPEG-like quantization (linear MPE). (c) Perceptually weighted variable-size BMA and 2-D nonlinear MPE quantization. (d) Perceptually weighted variable-size BMA and 2-D nonlinear MPE quantization and temporal filtering.

accurate HVS contrast discrimination model has been presented. The model accounts for the nonuniform nature of the HVS redundancy removal in the frequency domain. Here the basic idea is to design the entire encoding process to preserve no more than the subjectively significant information at a given subjective distortion level. This aim affects not only the quantizer design, but also the motion estimation.

On the one hand, as a result of a more accurate perception model and the MPE restriction criterion, an improved nonlinear quantizer has been proposed. On the other hand, this perceptual quantizer is used here to decide if additional motion information is perceptually significant. This definition of perceptually significant motion information gives rise to an appropriate entropy-constrained BMA (using the actual DFD entropy) in a natural way. In this way, superfluous effort in the motion description (predicting details that are going to be discarded by the quantizer) is avoided and a perceptual feedback is introduced in the motion estimation refinement.

The reconstructed frames and perceptually meaningful distortion measures show that the proposed schemes improve the results of previous comparable schemes such as H.263 with unweighted motion refinement and MPEG-like quantization. In particular, nonlinear MPE quantizers lead to better subjective quality than the linear MPE (CSF-based, MPEG-like) quantizers at the same bit-rates because they more accurately preserve the relevant information of the DFDs. A very interesting consequence is that the (2-D or 3-D) nonlinear quantizers keep the distortion bounded over a large group of frames, so the need to introduce bit-consuming intracoded frames can be reduced. According to the perceptual distortion results at a fixed bit-rate, the benefits of a better bit allocation between DVF and DFD (the benefits of the perceptual weight in motion estimation) are negligible compared to the benefits of perceptually more accurate quantizers. The significant improvement due to this enhancement of the quantizer suggests that quantizer design may be more important than optimal motion estimation from the rate-distortion point of view.

However, a side effect of the perceptual control of the motion estimation is a scale-dependent refinement strategy that gives rise to more robust and meaningful motion flows compared to unweighted refinement criteria. This is confirmed by the segmentation results that show how the perceptual weighting

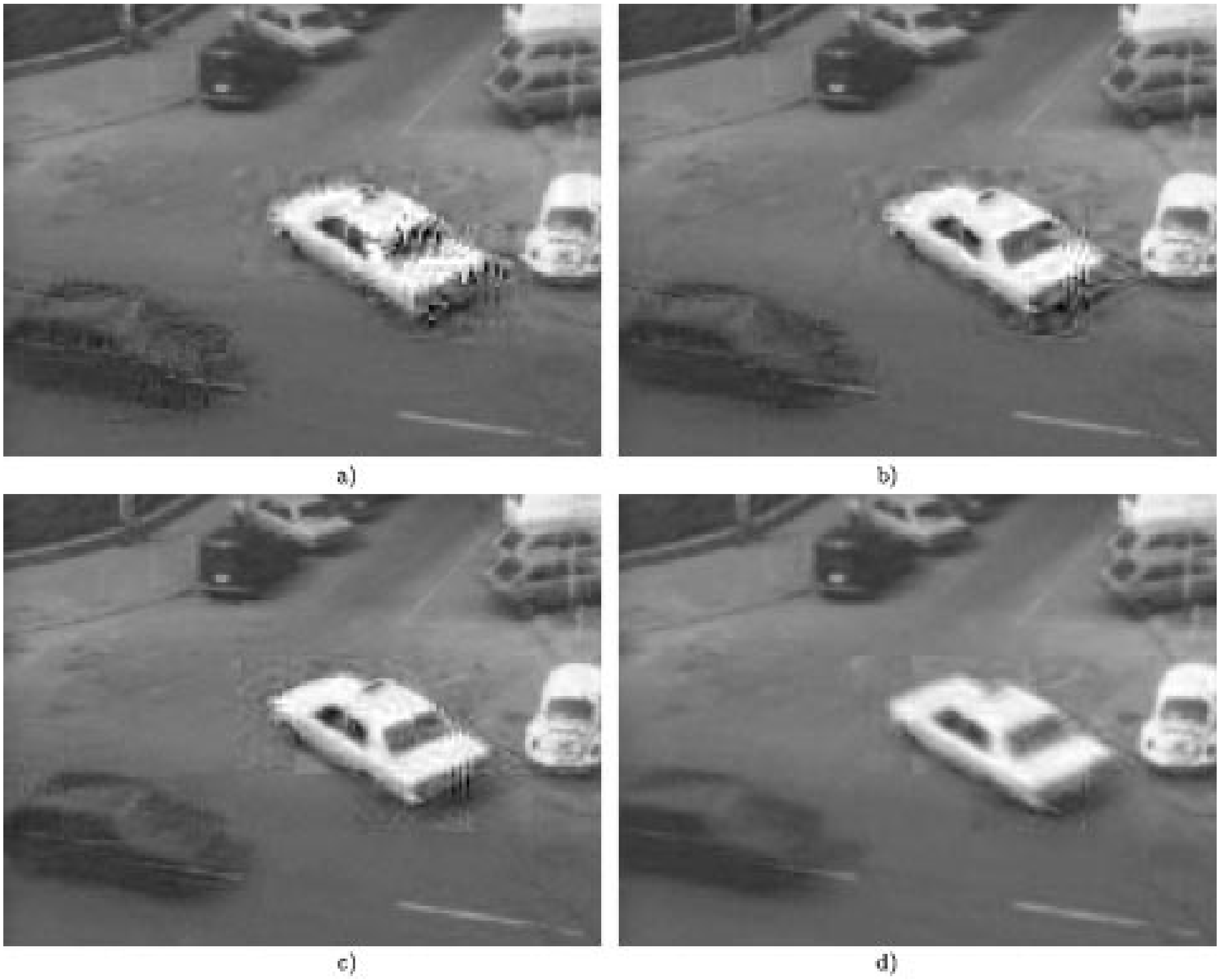

Fig. 15. Reconstruction results with the *Taxi* sequence using previously reported encoding configurations (a-b) and the proposed 2-D or 3-D alternatives (c-d). (a) Fixed size BMA for motion estimation and MPEG-like quantization (linear MPE). (b) Unweighted variable-size BMA and MPEG-like quantization (linear MPE). (c) Perceptually weighted variable-size BMA and 2-D nonlinear MPE quantization. (d) Perceptually weighted variable-size BMA and 2-D nonlinear MPE quantization and temporal filtering.

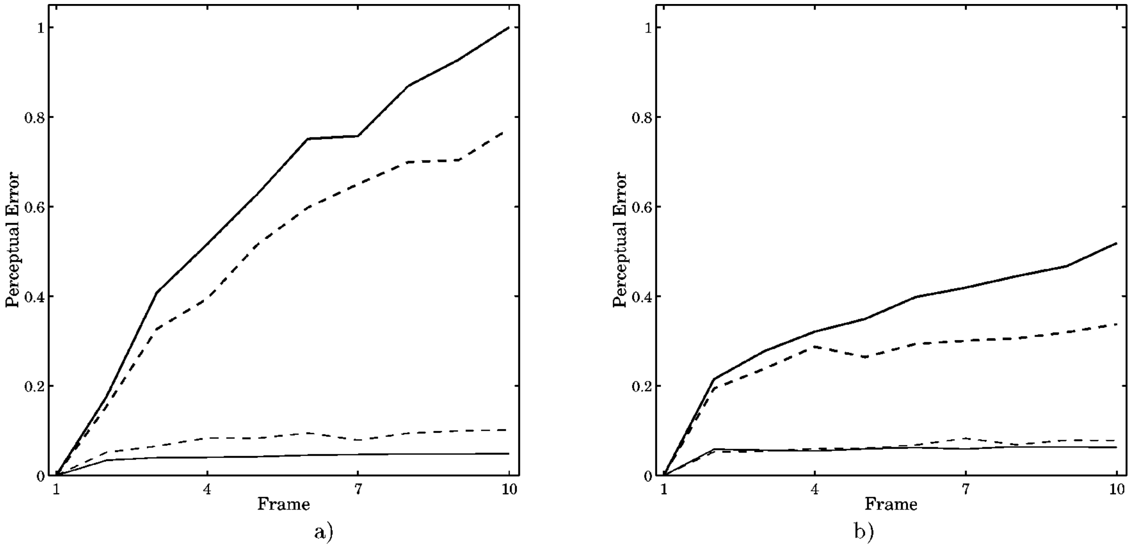

Fig. 16. Perceptual distortion measures for the frames of the (a) *Rubik* and (b) *Taxi* sequence using previous and the proposed schemes. The thick lines correspond to the previous apporaches and thin lines correspond to the proposed schemes. Solid thick line corresponds to fixed-size BMA and linear MPE quantizer (H.261, MPEG1). Dashed thick line corresponds to unweighted variable-size BMA and linear MPE quantizer (H.263). The solid thin line corresponds to perceptually weighted variable-size BMA and 2-D nonlinear MPE and the dashed thin line corresponds to perceptually weighted variable-size BMA and 3-D nonlinear MPE. As in Fig. 6, the frame-by-frame implementation of the perceptual distortion measure may slightly overestimate the visual effect of the frame blurring introduced by the temporal filter.

makes a sparse flow suitable to obtain rough segmentations of the moving objects. These results suggest that the perceptual quantizer should not to be taken into account in the motion estimation due to rate-distortion reasons, but to obtain more meaningful flows that may be of interest for higher-level video coding.

Systematic psychophysical testing instead of the simple frame-by-frame perceptual measure may be necessary to measure accurately the relevance of the temporal features of the proposed 1-D+2-D quantizer design. Nevertheless, the results presented suggest that when amplitude nonlinearities are taken into account, the consideration of the temporal properties is not so significant.

VII. Acknowledgements

The authors wish to thank the reviewers for their detailed comments on the original manuscript. The authors also wish to thank Dr. J. Soret for valuable discussions on the hardware implementation of the algorithms described in this paper.

References

[1] D. LeGall, "MPEG: A video compression standard for multimedia applications," *Commun. ACM*, vol. 34, no. 4, pp. 47–58, 1991.

[2] G. Tziritas and C. Labit, *Motion Analysis for Image Sequence Coding*, Amsterdam, The Netherlands: Elsevier, 1994.

[3] A. M. Tekalp, *Digital Video Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[4] *ITU-Telecommunication standardization sector*, Draft recommendation H.263, 1994.

[5] *Overview of the MPEG-4 Version 1 Standard*, ISO/IEC JTCI/SC29/WG11 N1909, 1997.

[6] M. Bierling, "Displacement estimation by hierarchical block-matching," *Proc. SPIE, Conf. Vis. Commun. Image Process.*, vol. 1001, pp. 942–951, 1988.

[7] F. Moscheni, F. Dufaux, and H. Nicolas, "Entropy criterion for optimal bit allocation between motion and prediction error information," *Proc. SPIE, Conf. Vis. Commun. Image Process.*, vol. 2094, pp. 235–242, 1993.

[8] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: A review and new contribution," *Proc. IEEE*, vol. 83, no. 6, pp. 858–876, 1995.

[9] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression*. Norwell, MA: Kluwer, 1997.

[10] ——, "A video compression scheme with optimal bit allocation among segmentation, motion and residual error," *IEEE Trans. Image Processing*, vol. 6, no. 11, pp. 1487–1502, 1997.

[11] J. Lee, "Joint optimization of block size and quantization for quadtree-based motion estimation," *IEEE Trans. Image Processing*, vol. 7, pp. 909–912, June 1998.

[12] J. Malo, F. Ferri, J. Albert, and J. M. Artigas, "Splitting criterion for hierarchical motion estimation based on perceptual coding," *Electron. Lett.*, vol. 34, no. 6, pp. 541–543, 1998.

[13] J. Malo, F. Ferri, J. Albert, and J. Soret, "Comparison of perceptually uniform quantization with average error minimization in image transform coding," *Electron. Lett.*, vol. 35, no. 13, pp. 1067–1068, 1999.

[14] J. Malo, F. Ferri, J. Albert, J. Soret, and J. M. Artigas, "The role of perceptual contrast nonlinearities in image transform coding," *Image Vis. Comput.*, vol. 18, no. 3, pp. 233–246, 2000.

[15] J. Malo, A. M. Pons, and J. M. Artigas, "Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain," *Image Vis. Comput.*, vol. 15, no. 7, pp. 535–548, 1997.

[16] A. M. Pons, J. Malo, J. M. Artigas, and P. Capilla, "Image quality metric based on multidimensional contrast perception models," *Displays*, vol. 20, pp. 93–110, 1999.

[17] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Processing*, vol. 3, pp. 625–638, May 1994.

[18] L. Torres, D. García, and A. Mates, "A robust motion estimation and segmentation approach to represent moving images with layers," in *IEEE ICASSP*, Munich, Germany, 1997, pp. 2981–2984.

[19] G. E. Legge, "A power law for contrast discrimination," *Vis. Res.*, vol. 18, pp. 68–91, 1981.

[20] D. H. Kelly, "Motion and vision II: Stabilized spatiotemporal threshold surface," *J. Opt. Soc. Amer.*, vol. 69, no. 10, pp. 1340–1349, 1979.

[21] B. L. Beard, S. Klein, and T. Carney, "Motion thresholds can be predicted from contrast discrimination," *J. Opt. Soc. Amer. A Opt. Image Sci.*, vol. 14, no. 9, pp. 2449–2470, 1997.

[22] E. Martinez-Uriegas, "Color detection and color contrast discrimination thresholds," in *Proc. OSA Annu. Meeting ILS-XIII*, Los Angeles, CA, 1997, p. 81.

[23] B. Girod, "Motion compensation: Visual aspects, accuracy and fundamental limits," in *Motion Analysis and Image Sequence Processing*, M. I. Sezan and R. L. Lagendijk, Eds., 1993.

[24] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.

[25] A. B. Watson, "DCT quantization matrices visually optimized for individual images," *Hum. Vis., Vis. Proc. Digit. Disp. IV*, vol. 1913, 1993.

[26] J. Malo, A. M. Pons, and J. M. Artigas, "Bit allocation algorithm for codebook design in vector quantization fully based on human visual system nonlinearities for suprathreshold contrasts," *Electron. Lett.*, vol. 31, no. 15, pp. 1222–1224, 1995.

[27] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[28] B. Macq, "Weighted optimum bit allocations to orthogonal transforms for picture coding," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 875–883, 1992.

[29] M. Chan, Y. B. Yu, and A. G. Constantinides, "Variable size bock matching motion compensation with applications to video coding," *Proc. IEEE, Vis., Image, Sig. Process.*, vol. 137, no. 4, pp. 205–212, 1990.

[30] E. Reusens *et al.*, "Dynamic approach to visual data compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 197–211, 1997.

[31] F. Dufaux and F. Moscheni, "Segmentation-based motion estimation for second generation video coding techniques," in *Video Coding: A Second Generation Approach*, L. Torres and M. Kunt, Eds. Norwell, MA: Kluwer, 1996.

[32] F. Dufaux, I. Moccagatta, B. Rouchouze, T. Ebrahimi, and M. Kunt, "Motion-compensated generic coding of video based on multiresolution data structure," *Opt. Eng.*, vol. 32, no. 7, pp. 1559–1570, 1993.

[33] J. Li and X. Lin, "Sequantial image coding based on multiresolution tree architecture," *Electron. Lett.*, vol. 29, no. 17, pp. 1545–1547, 1993.

[34] A. B. Watson and J. A. Solomon, "A model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A Opt. Image Sci.*, vol. 14, pp. 2379–2391, 1997.

[35] H. R. Wilson, "Pattern discrimination, visual filters and spatial sampling irregularities," in *Computational Models of Visual Processing*, M. S. Landy and J. A. Movshon, Eds. Cambridge, MA: MIT Press, 1991, pp. 153–168.

[36] A. B. Watson, "Efficiency of a model human image code," *J. Opt. Soc. Amer. A Opt. Image Sci.*, vol. 4, no. 12, pp. 2401–2417, 1987.

[37] J. G. Daugman, "Entropy reduction and decorrelation in visual coding by oriented neural receptive fields," *IEEE Trans. Biomed. Eng.*, vol. 36, pp. 107–114, 1989.

[38] A. J. Ahumada and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," *Proc. SPIE*, vol. 1666, pp. 365–374, 1992.

[39] J. Solomon, A. Watson, and A. Ahumada, "Visibility of DCT basis functions: effects of contrast masking," in *Proc. Data Compress. Conf.*, Snowbird, UT, 1994, pp. 361–370.

[40] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 127–135, 1982.

[41] K. N. Nygan, H. C. Koh, and W. C. Wong, "Hybrid image coding scheme incorporating human visual system characteristics," *Opt. Eng.*, vol. 30, no. 7, pp. 940–946, 1991.

[42] D. J. Heeger, "Model for the extraction of image flow," *J. Opt. Soc. Amer. A Opt. Image Sci.*, vol. 4, pp. 1455–1471, 1987.

[43] A. B. Watson and A. J. Ahumada, "Model of human visual motion sensing," *J. Opt. Soc. Amer. A Opt. Image Sci.*, vol. 2, pp. 322–342, 1985.

[44] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.

[45] W. Niehsen and M. Brünig, "Covariance analysis of motion-compensated frame differences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 536–539, 1999.

[46] S. Winkler, "Issues on vision modeling for video quality assessment," *Signal Process.*, vol. 78, no. 2, pp. 231–252, 1999.

[47] A. B. Watson, J. Hu, J. F. McGowan, and J. B. Mulligan, "Design and performance of a digital video quality metric," *Proc. SPIE, Hum. Vis., Vis. Proc. Digit. Disp. IX*, vol. 3644, pp. 168–174, 1999.

[48] G. Borgefors, "Distance transformations in digital images," *Comput. Vis. Graph. Image Process.*, vol. 34, pp. 344–371, 1986.

Jesús Malo was born in 1970. He received the M.Sc. degree in physics in 1995 and the Ph.D. degree in physics in 1999, both from Universitat de València, València, Spain.

Since 1994, he has been with the Vision Group of Universitat de València. In the Fall of 1999, he worked with the Image and Vision Group at the Institute of Optics, CSIC. In 2000 and 2001, he worked as Fulbright Postdoctoral Fellow at the Vision Group of the NASA Ames Research Center, Moffett Field, CA, and at the Laboratory of Computational Vision of the Center for Neural Science, New York University. He is interested in models of low-level human vision, their relations with information theory and their applications to computer vision, image processing, and vision science experimentation.

Dr. Malo was the recipient of the Vistakon European Research Award in 1994.

Juan Gutiérrez was born in 1971. He received the Licenciado degree in physics (electricity, electronics, and computer science) in 1995 from the Universtitat de València, València, Spain, where he is currently pursuing the Ph.D. degree in motion estimation and segmentation.

Since 1997, he has been with the Computer Science and Electronics Department, Universtitat de València, València, Spain, as a Teacher and a Research Fellow. He has performed two research stays, one at the Digital Imaging Research Centre at Kingston University (U.K.) for seven months working on multi-object tracking, and the other at the Department of Informatics and Mathematical Modeling, Technical University of Denmark for two months, working on optical flow regularization. His current research interests include image analysis, motion understanding, and regularization theory.

I. Epifanio, photograph and biography not available at time of publication.

Francesc J. Ferri was born in 1964. He received the Licenciado degree in physics (electricity, electronics, and computer science) in 1987 and the Ph.D. degree in pattern recognition in 1993, both from Universitat de València, València, Spain.

He has been with the Computer Science and Electronics Department, Universtitat de València, since 1986; first as a Research Fellow, and since 1988, as a Teacher of computer science and pattern recognition. He has been involved in a number of scientific and technical projects on computer vision and pattern recognition. During a sabbatical in 1993, he joined the Vision, Speech, and Signal Processing Research group at the University of Surrey (U.K.), where he was working in feature selection and statistical pattern recognition methodology. He has authored or coauthored more than 90 technical papers in international conferences and well-established journals in his fields of interest. He has also helped in refereeing several of these publications. His current research interests include feature selection, nonparametric classification methods, inductive learning, computational geometry, image analysis, and motion understanding.

Dr. Ferri is a member of the ACM and the AERFAI (Spanish Association for Pattern Recognition and Image Analysis, affilate society of IAPR), where he has helped several times in organizing the biennal Spanish Conference on Pattern Recognition and Image Analysis.

José M. Artigas received the M.S. degree in physical sciences in 1974 and the Ph.D. degree in 1979, both from Universitat de València, València, Spain.

From 1985 to 1992, he worked in collaboration with the group of Prof. F. W. Campbell of Cambridge University (U.K.). Since 1984, he has been on the staff of the Faculty of Physics, Universitat de València, where he has held the position of Associate Profesor of Optics. His two main research interests have been in physiological optics (spatial vision) and color vision. He has advised nine Ph.D. dissertations. He is currently developing a project concern a new noninvasive methods for early detection of visual pathologies.

Dr. Artigas is a member of the Spanish Optical Society (SEDO) and the American Optical Society (OSA).