ELSEVIER

Signal Processing: *Image Communication* 16 (2001) 611–626

SIGNAL PROCESSING: IMAGE COMMUNICATION

www.elsevier.nl/locate/image

Content-based and perceptual bit-allocation using matching pursuits

C. De Vleeschouwer[a,*,1], B. Macq[a]

[a]*Laboratoire de Télécommunications et Télédétection, Université catholique de Louvain, Bâtiment Stévin - 2, place du Levant, B-1348 Louvain-la-Neuve, Belgium*

Received 22 March 1999; received in revised form 17 February 2000; accepted 30 May 2000

Abstract

When communicating at a very-low bit-rate, video coders are unable to preserve high visual quality for all images. A selection of key regions according to human viewing and understanding may therefore be useful: it allows extracting essential features and to code them with a high quality, while the remainder of the image is coarsely transmitted. Such an approach requires a signal decomposition allowing an adaptive spatial variant bit allocation. Matching pursuits (MP) explicitly select the information to be transmitted among a large and overcomplete set of functions and quantize it according to a fixed and constant step. For MP, bit allocation is equivalent to function selection. This makes matching pursuits well suited for spatial variant bit allocation, since functions matched to important objects can be chosen. In this paper, we first show the abilities of matching pursuits to allocate an available bit-budget according to a semantic understanding of the scene. This semantic information can be provided either automatically or interactively. Second, the possibility to incorporate perceptive criteria within the MP coding algorithm is investigated.

Keywords: Content-based scalability; Video coding; Bit-rate allocation; Perceptual coding

1. Introduction

Efficient video coders need techniques to deal with the high temporal redundancy of the three-dimensional video signal [4]. Motion models allow a compact description of moving images and motion-driven prediction leads to a high compaction of the video sequence [30]. In the most commonly used algorithms (MPEG [20], H263 [36], …), a frame is predicted from the previous frame, using local motion information. That is, a particular block of the current frame to be coded is predicted as a displaced block of the previous reconstructed frame [15]. Then, the prediction error, also named displaced frame difference (DFD), is compressed using techniques involving a number of signal expansion methods. Typically, the discrete cosine transform [1] is used, but other orthogonal transforms are possible [32]. Compression results from the quantization of the transformed coefficients. Incorporating the human visual system (HVS) into the quantizer design permits minimizing visible distortion throughout

[1] C. De Vleeschouwer's research is supported by the Belgian NSF.

* Corresponding author. Address: IMEC/DESICS, Kepeldreef 75, 3001, Leuven. Tel.: + 32-16-28-15-70.

E-mail address: vleesch@imec.be (C. De Vleeschouwer).

0923-5965/01/$ - see front matter
PII: S0923-5965(00)00039-4

the picture while drastically reducing the amount of bits that have to be transmitted.

This paper focuses on a recent signal representation technique that has been successfully applied in the framework of DFD coding [2,28]. Matching pursuits, an algorithm proposed by Mallat and Zhang [25] decomposes a signal into a linear expansion of waveforms selected from a redundant dictionary of functions. A major feature that distinguishes matching pursuits from the block-based signal representation techniques is that MP explicitly select the information to transmit among a large and overcomplete set of functions. The atom selection procedure of MP is equivalent to information selection and permits bit-allocation. Actually, MP allows information and details distribution on the video frame.

In this paper, by incorporating perceptive, subjective or semantic criteria throughout the selection procedure, we will explore and demonstrate the abilities of video coding algorithms based on matching pursuits, to efficiently achieve spatial variant bit-allocation. Section 2 proposes a methodology to achieve spatial variant bit-allocation using MP. Section 3 briefly reviews the main results achieved in the context of semantic video sequence interpretation, and incorporates semantic knowledge within the matching pursuits video coding algorithm. Results achieved when using the proposed content-driven MP coding strategy are presented. In Section 4, we show that the MP representation is well adapted to the use of spatially located visual masking criteria. Section 5 concludes.

2. Matching pursuits video coding bit-allocation

Incorporated within a motion compensated video codec, Matching pursuits is an iterative procedure that allows for adaptive representation of the displaced frame difference [2,28]. Bit-allocation procedures for matching pursuits differ from the ones used for DCT- or other transform-based signal representations. For transform-based methods, the signal is first projected on a complete set of orthogonal base functions. Coefficients are then quantized. Quantization may be adapted according to the frequency of the coefficient or according to the activity of the block (spatial masking). The quantization selects which information is transmitted and the selection is constrained by the chosen transform. On the contrary, MP explicitly select the information to be transmitted among a dictionary, which is a large and overcomplete set of functions.

In practice, a set of 2-D functions is fixed and the center of each function is translated into a set of pixel positions. Each step of the iterative procedure selects the function that best matches the signal, i.e. the one that maximizes the inner product between the dictionary function and the signal to represent. As stated by Neff and Zakhor [28], assuming that the DFD is sparse with pockets of energy, where motion prediction was inadequate, the search can be limited around these high-energy pockets. The DFD is divided into 12×12 overlapping blocks located on a 8×8 grid. For each block, the square of all pixel intensities is computed, providing a *block energy* value. The inner product search is then performed in an $S \times S$ search window around the center of the block with the largest energy value.

In the MP representation procedure, the search window selects the spatial area in which the signal is improved by the new atom. It performs spatial-variant bit-allocation. A selection based on the block-energy value is appropriate to maximize the PSNR of the reconstructed signal. In Sections 3 and 4, the window selection takes into account other semantic and perceptive criteria. The objective is to more often select the search window in areas that have to be displayed with a good visual fidelity. Atoms are preferably selected where human visual perception is sensitive. It performs a spatial-variant bit-allocation without transmission of any "side information". Atoms are preferably searched for around interesting blocks but they are located anywhere in the search area. The block decomposition does not appear explicitly in the picture representation (no blocking artifacts). The drawbacks of the block-based analysis procedure, e.g. due to the fact that the position of the objects changes in reference to the fixed grid decomposing the picture into blocks, are thus circumvented.

3. RoI-based bit-allocation

3.1. Video Sequence Semantic Analysis: state of the question

In this section, we focus on the semantic analysis of video sequences and on the possibility it offers to balance the quality across the picture according to the viewer's preferences.

For video-phone or video-conferencing applications, the face is segmented and tracked along the sequence, Robust real-time face tracking has been studied in much recent research [5,10,12,14,16,29,34,37]. Most of this work mentions the fact that the image quality in the face area is more important for a human observer than the image quality in other picture areas.

Automatic segmentation of moving objects along video sequences has also been studied in a more general context. In some research, this segmentation is directly integrated within a complete region-based coding scheme [3,13]. In other studies, segmentation attempts to partition the frames into objects that are semantically meaningful to the human observer. This step happens prior to the encoding of the objects, e.g. using MPEG-4 [11,18,35]. Nevertheless, decomposing a video sequence into video objects is very difficult. We refer to the comprehensive review proposed by Meier and Ngan [27]. It summarizes the main motion segmentation and video object generation techniques that have been developed up to now. It appears that existing techniques are promising but as yet unable to accurately locate the boundaries of moving objects in generic video sequences.

In a block-based coding scheme for which the quantization is tuned to some local interest, the analysis and interpretation inaccuracies result in flickering effects among blocks located at the frontiers of objects. To tackle this problem, MPEG-4 enables quality distribution among a set of extracted objects. But the use of this functionality relies on the accurate extraction of the video objects. Inaccuracy results in inefficiency as it burdens the coding cost by reducing the temporal and spatial redundancies among objects. As explained in the next section, matching pursuits coding approach does not suffer from these drawbacks.

3.2. RoI allocation

The aim of this section is to investigate how the MP coding algorithm is able to take some semantic information into account while distributing some available bit-rate along the sequence. The assumption is made that a level of interest is attributed to each region of each frame in the sequence. This can be done by a fully automatic or by a user-driven mechanism. The goal is to bind the visual quality of the coded sequence to the given local interest.

As stated in Section 2, the search for the best matching 2-D dictionary function is performed in a window around a chosen 8×8 block. To maximize PSNR value, one chooses the block with the maximum signal energy. So, atoms are located where the DFD signal is significant. When it can be assumed that the viewer's interest is not uniform on the whole picture, atoms should be located preferably around areas captivating the user's attention. This goal is naturally reached by performing the search around interesting areas. Search window does not have to be selected any more to maximize a global PSNR value, but rather to maximize the user's satisfaction, i.e. to reach the optimal quality while distributing it on the sequence according to his/her preferences. For each 8×8 block, both the DFD energy signal and the local viewer interest I are computed (or interactively provided). Note that the interest value may be computed or provided on a pixel basis. In these cases, the mean interest value is computed on each block. The search window is now chosen around the block that maximizes an increasing function of these two factors. In practice, the product of the DFD block energy with an increasing function of interest I is maximized.

In [9], bit-rate had been shared between segmented video objects in order to adapt their quality to the level of interest. The aim was to reach maximal global quality while keeping VOs quality as close as possible to a linear function of the level of interest.

The extension of this constraint in the context of MP coding leads to atoms spreading on the picture so that local PSNR values are a linear function of the local interest. Considering two 8×8 blocks

labeled i and j, this means that

$$\mathrm{PSNR}_i - \mathrm{PSNR}_j = C(I_i - I_j), \tag{1}$$

where PSNR_i and PSNR_j refer to the PSNR (dB) values computed on 8×8 blocks i and j. I_i and I_j refer to the level of interest attributed to these blocks while C is a user-defined parameter expressing his/her requirement for unequal quality distribution. Thus, we have

$$-10 \log\left(\frac{E_i}{E_j}\right) = C(I_i - I_j), \tag{2}$$

Fig. 1. Interest mask extracted for three successive coded frames of Akiyo (13, 17, 21). Lightning is proportional to the level of interest.

Fig. 2. Content-based bit-allocation for Akiyo video sequence. Frames 21 (left) and 77 (right): original (top), classic encoding (middle), subjective encoding (bottom).

where E_i and E_j are the sum of the DFD squared values on blocks i and j. This means that, when the linear constraint is satisfied, one has

$$E_i \times 10^{(C.I_i/10)} = E_j \times 10^{(C.I_j/10)}.$$

MP is an iterative algorithm for which E_i decreases each time a block i is selected. At each step, *selecting the block i* that *maximizes $E_i \times 10^{(C.I_i/10)}$ progressively decreases the coding error while balancing the quality according to the interest constraint.*

The advantage of MP compared to other coding algorithms is that it does not require accurate prior analysis. Local temporal or spatial instabilities of the analysis system affect the atom selection procedure but these oscillations are absorbed and everything happens as if the quality share-out was bound to a smoothened version of the analytic process. In particular, MP bit allocation strategies and MPEG-4 object-based functionalities can be considered as complementary. Even if it enables quality distribution between a set of extracted video objects, MPEG-4 relies on the accurate extraction of these objects, which remains an unsolved problem for generic and natural sequences. This makes the use of this functionality rather unpractical.

3.3. Allocation results based on automatic RoI selection

In this section, like in [9], we have used a completely automatic fuzzy logic analysis system to extract subjectively interesting areas of a video sequence. The method described in Section 3.2 is used to guide the search process of the MP coding algorithm. Our goal is to show that the proposed method permits to distribute the quality on the picture even with an inaccurate analysis system. The "Akiyo" video sequence has been encoded at 10 kbits/s with a frame rate of 7.5 frames/s. As can be seen from Figs. 1 and 2, despite the instabilities and inaccuracies of the extracted level of interest, the pictures displayed do not suffer from any disturbing flickering effect or sudden spatial quality jump. It is worth noting that no smoothing was introduced on the interest used to guide the coding procedure, as would have been the case with the introduction of the spatial local sensitivity used in [5].

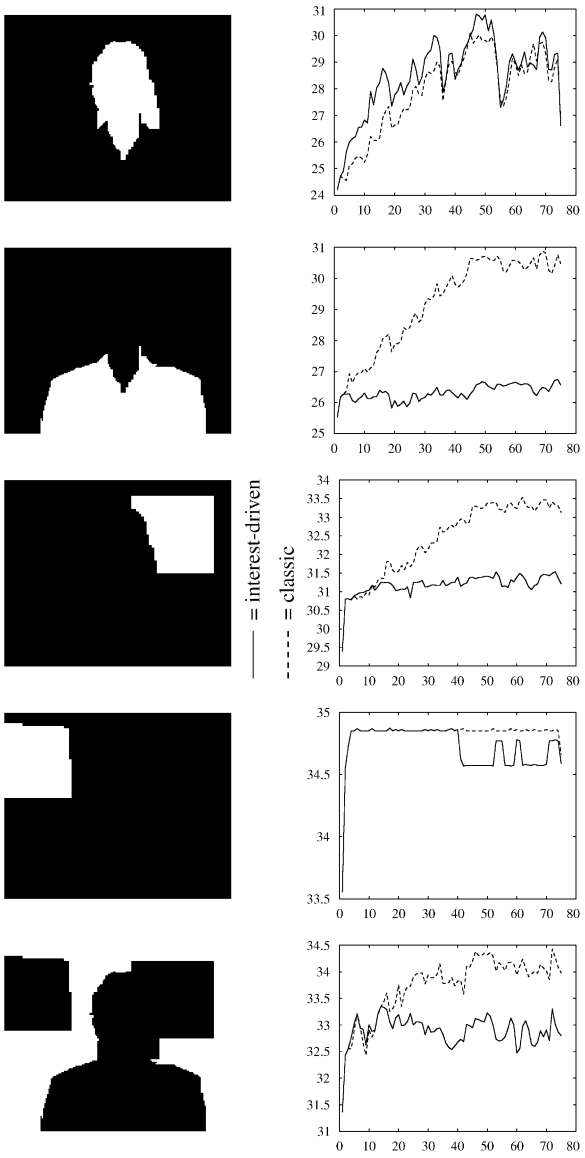

Fig. 3. PSNR curves for the interest-driven coding algorithm and for the classic frame-based MP algorithm. VOP shapes are only used to compute PSNR values. They are not used to drive the allocation process.

To compare the performance of the interest-driven MP coding algorithm with a classic frame-based MP algorithm, an accurate segmentation of the video sequence is provided (see Fig. 3). A PSNR value is computed on each object of the sequence for both coding scenarios. One may note that the PSNR increases faster in the face area for the

interest-driven case. This proves that first coded atoms are preferably located in this area. After some delay, quality of the classic MP coding scheme joins the one of the interest-driven scheme. This can be explained by the nature of the sequence studied. The head moves on a still background. In the classic coding scheme, once the level of quality of the background is sufficient, no more atoms are put on it. All atoms are spread on the face, just as in the interest-driven version.

This example illustrates the efficiency of the allocation procedure. We can also notice that an unbalanced allocation between two picture areas is only relevant when both areas contain significant information (objects in motion, etc). In these cases the generation of accurate segmentation is still an unsolved problem. Nevertheless, as our scheme is supposed to deal with analysis inaccuracies, a coarse automatic or user-assisted segmentation may be enough to distribute the quality around some selected objects and along all the sequences. This is investigated in the following section.

3.4. Allocation results based on interactive RoI selection

Here, the viewer has coarsely selected regions of interest in the video sequence "news" encoded at

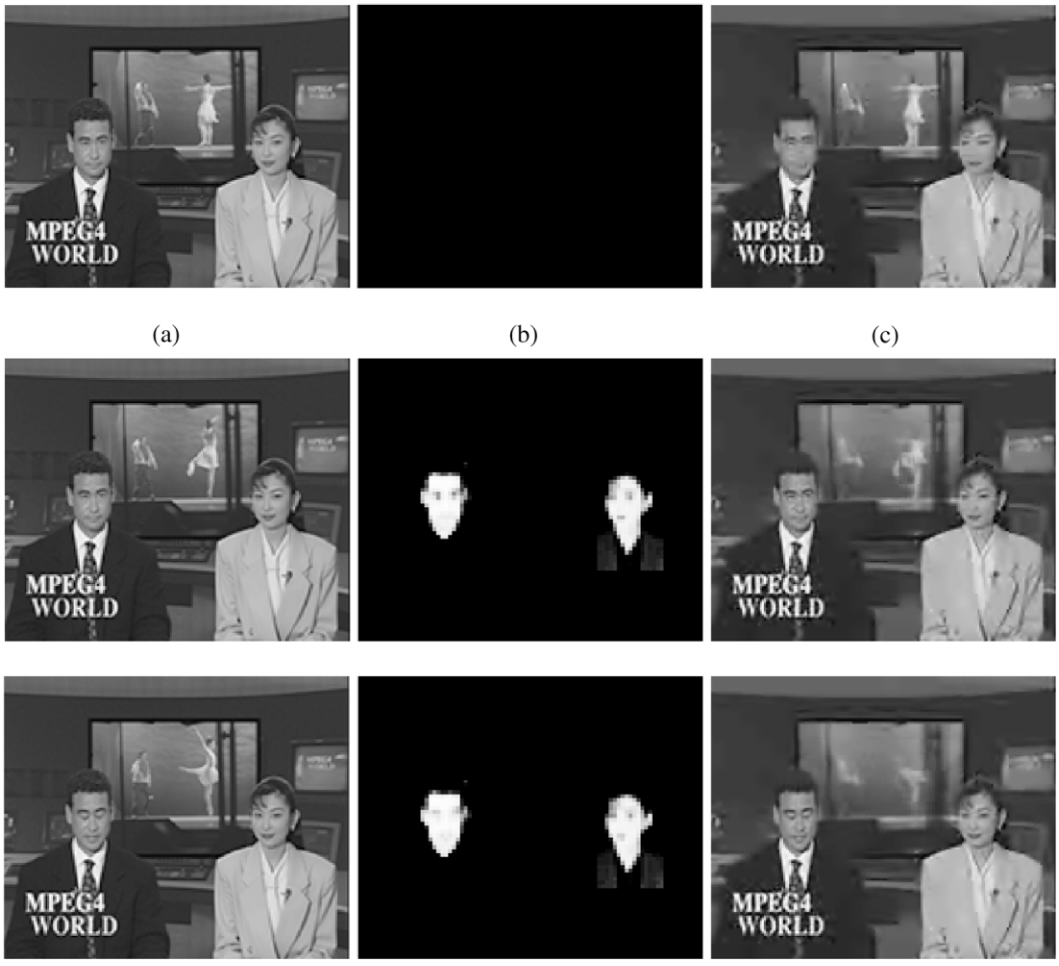

Fig. 4. The "news" video sequence encoded according to the user's interest. Frames 1 (top), 7 (middle), 13 (bottom): (a) original, (b) interest mask, (c) subjective encoded picture.

24 kbits/s and 10 frames/s. A watershed algorithm and a basic tracking algorithm has been used to distinguish complete objects from the kernels provided by the user. Along the sequence display, the user interest goes from the head to the background and then comes back to the head. From Figs. 4 and 5, one may note that the quality is distributed in accordance with the users' requirements along the sequence, even if the provided segmentation is only user-assisted and rather inaccurate. Moreover, changes of interest result in fast transitions in the distribution of quality on the picture.

In Fig. 4, the first frame has been intra-encoded without reference to the user's interest. For the two next frames, a mask selecting the speaker's head has been used to allocate preferably in these areas the available bit-budget. One may notice that the background quality decreases rapidly while the heads are rather accurately encoded. Actually, the background information is not transmitted any more, After some delay, the objects present in the background disappear. This is noticeable on the first frame of Fig. 5. In this figure, the user interest is extended to incorporate the background. Rapidly, the background quality is

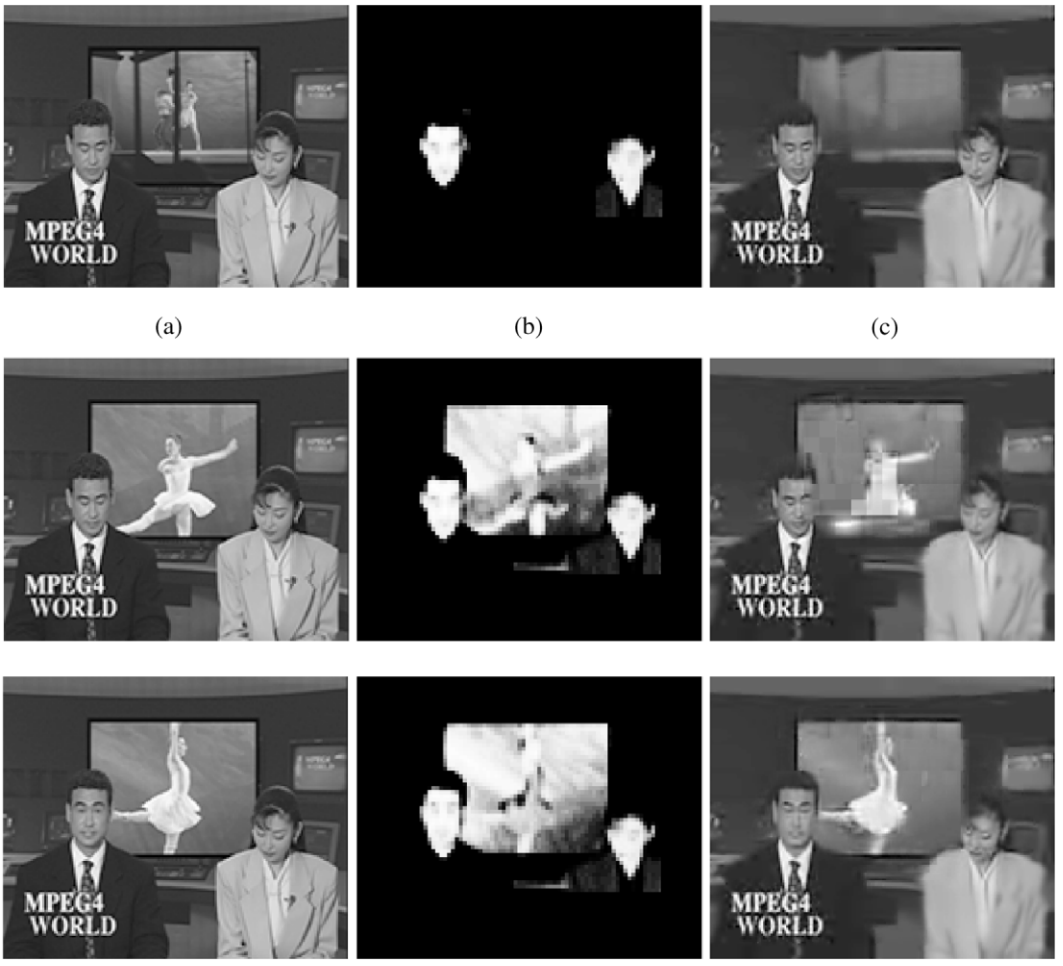

Fig. 5. The "news" video sequence encoded according to the user's interest. Frames 235 (top), 241 (middle), 268 (bottom): (a) original, (b) interest mask, (c) subjectively encoded picture.

improved. Of course, this is done in return for a quality decrease on the head. This may be noted when comparing the heads on frames 235 and 268.

These results demonstrate that the MP coding scheme should be able to distribute quality on complex video sequences, only requiring limited user assistance. For some specific applications (e.g. video-conferencing), a priori knowledge is available and permits the development of efficient tracking algorithms (see Section 3.1). Automatic application-dedicated and intelligent codec could thus be developed.

4. Perceptive bit-allocation

It is now admitted that the retina of the eye splits the visual stimulus, decomposing an image into components that are characterized by their spatial frequency and their location in the visual field. Sensitivity of the HVS to a component depends on these characteristics. Moreover, interactions exist between components with similar features (masking). The relevance of the HVS models in regards with the MP signal representation is discussed in this section. MP represents a signal as the sum of precise spatially localized functions. It makes local

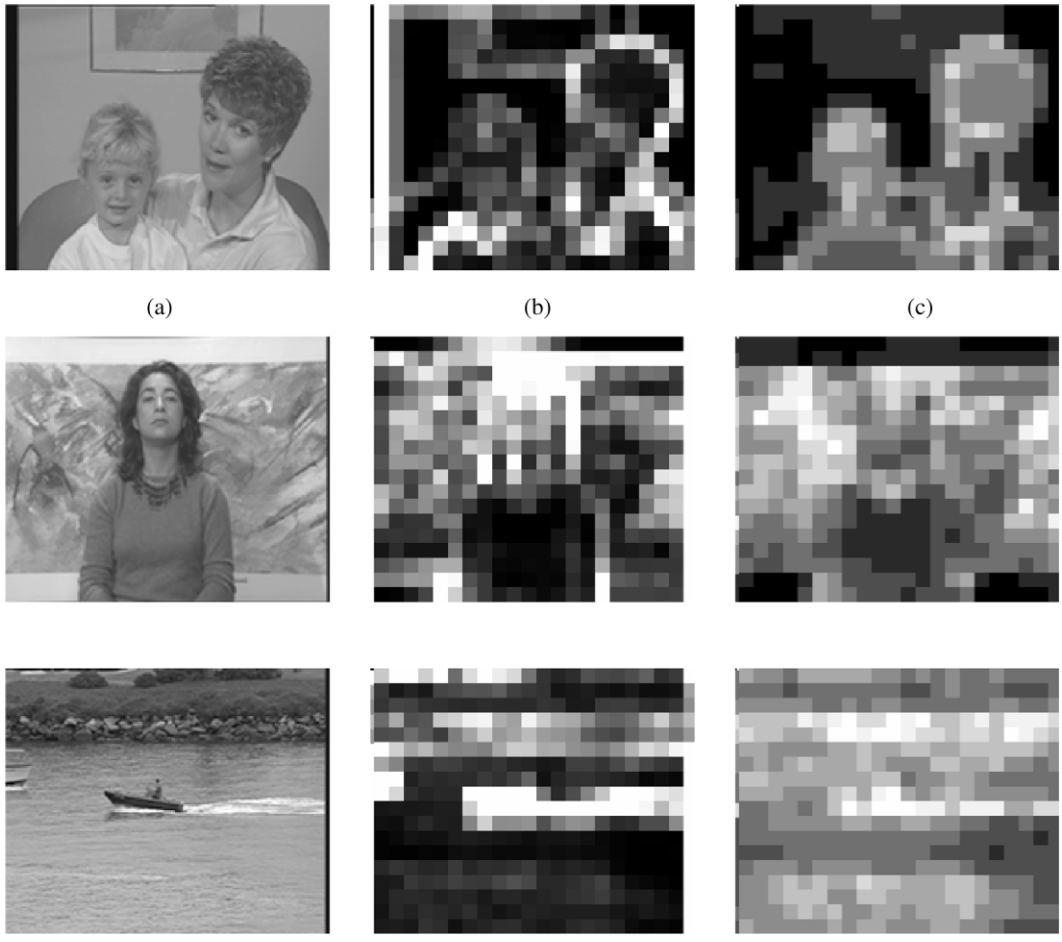

Fig. 6. Original image (a), local activity (b) and threshold extracted from the distance histogram (c) for "mother–daughter" (top), "silent" (middle) and "coastguard" (bottom). Block lightning is proportional to the activity or threshold measure. Local activity overestimates masking capabilities around edges.

spatial modeling of the HVS more suitable than models based on a frequency analysis of the signal.

4.1. Sensitivity of the HVS to spatial frequencies: discussion

Many studies on eye sensitivity have shown that the perception of noise depends on its frequency distribution. A contrast sensitivity function (CSF) has been build to express this dependency [26]. Nevertheless, it is based on assumptions that makes its use for natural image processing applications questionable:

- The CSF results from studies performed on gratings, i.e. monochromatic signals of one single frequency and orientation. Extension to random real pictures is risky.
- The model validity for low frequencies is subject to the assumption that curves can be extrapolated to very low frequencies. Indeed, sinusoidal gratings at zero frequency have no sense. This is disturbing in the framework of picture coding where the DC component is very important.

For DCT or wavelet-based coding methods, the eye sensitivity has been taken into account for the design of quantization matrix [17,22,24,31]. In an MP representation framework, in order to minimize visible distortions on the reconstructed picture, the sensitivity of the HVS to each dictionary function could also be estimated. Following a similar methodology as the one presented in [24] and using the sensitivity curve of Mannos and Sakrison, one may express the weighted noise power reduction resulting from the selection of a single atom g_γ at each step of the algorithm as

$$P_{\mathrm{N}} = W_{\mathrm{F}}(g_\gamma) \cdot |\langle f(x, y), g_\gamma(x, y)\rangle|^2, \tag{4}$$

where

$$W_{\mathrm{F}}(g_\gamma) = \iint |G_\gamma(u, v)|^2 \cdot |\mathrm{CSF}(f)|^2 \, \mathrm{d}u \, \mathrm{d}v, \tag{5}$$

where $G_\gamma(u, v)$ is the Fourier transform of the selected dictionary function. u and v are normalized horizontal and vertical frequencies. f is the spatial frequency in cycles/degree. i.e. $f = \pi.L.N.\sqrt{v^2 + u^2}/180$, where L is the distance to the screen expressed in screen height and N is the number of rows on the screen.

In order to minimize the weighted noise power, the atom should of course be chosen as the one that maximizes P_{N}. While computing the weighting factor, we observe that the dictionary functions have a wide spectrum with many spectral components located around the maximum of the frequency sensitivity curve. Considering that the model used is derived from gratings, the wide extent of the spectrum of the dictionary functions suggests that frequency weighting should not be used. This has been confirmed by experiments. Selecting other factors than unity degrades quality. Further examination of the phenomenon has revealed that, on a particular frame, only small perceptive improvements may result from the selection of adapted weighting factors. This is obtained as a counterpart of an objective PSNR quality reduction. Incorporated within a video codec, this PSNR reduction is amplified by motion compensation, the efficiency of which also decreases. Degradation of motion compensation effectiveness completely outstrips visual improvements that may be obtained on a particular frame.

Normalized luminance / 40% Treshold	<128	<154	<192	<224	>224
0-1	3.4	3.2	3.1	3.1	3
2-3	2.3	2.2	2.1	2.1	2
4-5-6	1.8	1.7	1.6	1.6	1.5
6-...-10	1.3	1.2	1.1	1.1	1
>10	0.3	0.3	0.3	0.3	0.3

Fig. 7. Block energy weighting factor as a function of the block mean luminance and the threshold extracted from the "distance histogram".

4.2. Local interaction between components: spatial masking

4.2.1. Models involved

The MP basis functions' support is often very compact, so that their frequential contents is not correctly modeled by gratings. On the contrary, they are spatially located. Models expressing local noise visibility are thus better suited than frequency-based models. Weber's law [33] states that the visibility threshold of a noise is larger for bright areas than for dark ones. Other models express the masking phenomenon. Masking is the reduction of visibility of noise by a stimulus at the location where the noise is introduced. In the MP representation framework, the understanding of masking phenomenon may prevent the selection of an atom in areas where it is masked by the local signal.

Rather than complex and computationally heavy masking models [7,8,19,21,38–40], practical tools for subjective visual quality measurement are based on the concept of *spatial activity* [23]. The use of spatial activity relies on the fact that noise visibility decreases in areas with sharp luminosity variations.

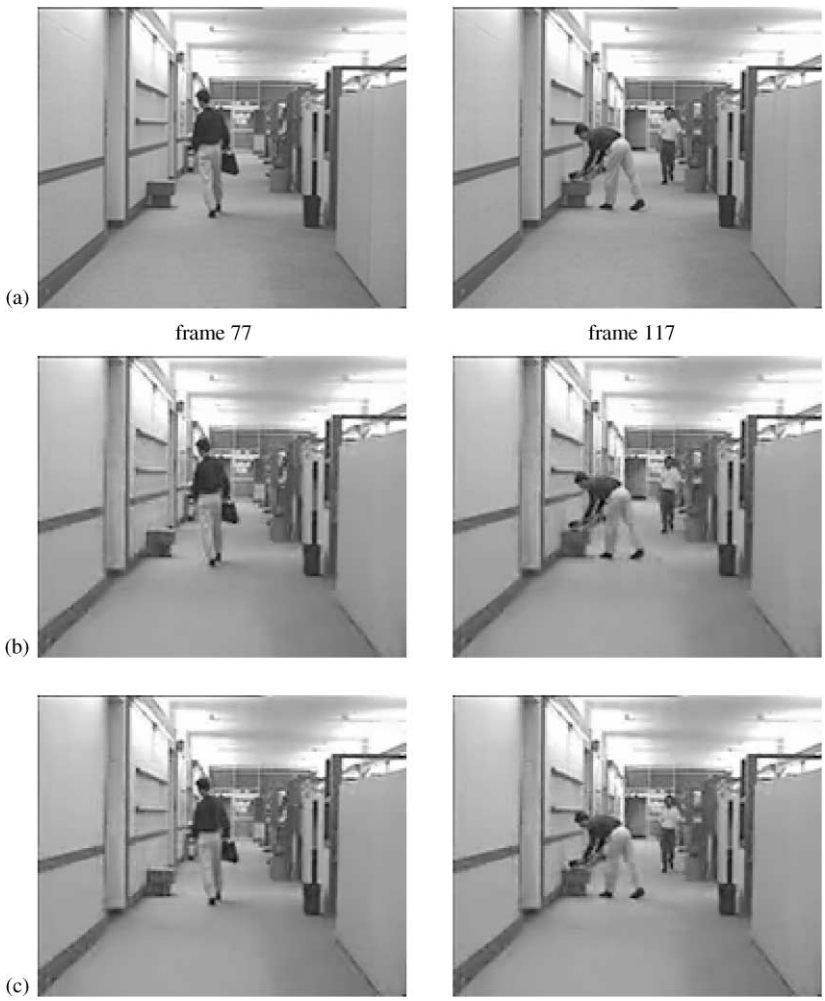

Fig. 8. "Hall-monitor" sequence (15 kbits/s, 7.5 frames/s): (a) original, (b) encoded without perceptive considerations, (c) encoded with block categorization according to masking capabilities during search window selection.

Spatial activity $A_{(i,j)}$ around pixel position (i, j) is defined as the sum of local variations of surrounding pixels.

$$A_{(i,j)} = \sum_{k,l} W(k,l) \cdot |x(i+k, j+l) - \bar{x}|^r, \quad (6)$$

where $W(k, l)$ is a weighting window used for activity localization. $\bar{x}$ is the mean pixel value around position (i, j).

4.2.2. Perceptive allocation using matching pursuits

In this section, a modification of the search window selection procedure is proposed in order to incorporate spatial masking concepts within the MP coding algorithm. The goal is to reduce the chance to select and transmit atoms that are likely to be masked by the local content of the scene.

The perception of noise in some spatial area depends on mean luminance and on the masking capabilities of the area. With the aim of developing a low-cost watermarking algorithm, Darmstaedter et al. [6] have tried to discriminate blocks according to their masking capabilities. Actually, they have categorized picture blocks according to their mean luminance value and their spatial activity. The activity of an 8×8 block is assimilated to the

Fig. 9. "Mother-and-daughter" sequence (28 kbits/s, 10 frames/s): (a) original, (b) without perceptive considerations, (c) with block categorization according to masking capabilities during search window selection.

one of its central pixel, using a power $r = 2$ and a block window as weighting function $W(k, l)$. So, according to Eq. (6), the activity is simply defined as the variance of the block. For a block whose top left corner is located in (i, j),

$$A_{(i,j)} = \frac{1}{64} \sum_{k,l=0}^{7} |x(i+k, j+1) - \bar{x}|^2. \quad (7)$$

The relevance of the categorization based on this measure of spatial activity has been tested on a large set of pictures. It has revealed that local activity overestimates the masking capabilities around edges.

A similar block categorization has been incorporated within the search window selection procedure of the MP coding algorithm. Actually, the *block energy* is weighted by a factor depending on the category the block belongs to. The purpose is to place more atoms around blocks with poor masking capabilities. As contours have to be sharply represented, local activity, which overestimates masking abilities around edges, is not a suitable classification measure. Another measure has been investigated. For each pixel, the distance to each of its neighbours (i.e. the luminance difference) is computed. A histogram of the distances is constructed. P being a percentage, the smallest threshold T, for which $P\%$ of the distances are lower than T, is extracted. This threshold and the mean luminance of the block determine the category of the block. The greater the threshold T for a block, the greater its masking ability. Fig. 6 compares local activity and threshold values for three frames of sequences "Mother-daughter", "silent" and "coastguard": local activity overestimates the masking capabilities around the edges. Fig. 7 summarizes the luminance and the threshold values for each category ($P = 40$). The energy weighting factor that

Fig. 10. Blow-ups of encoded frames. Left: encoded without perceptive consideration. Right: encoded with masking capabilities estimation. Top: "Hall-monitor", frame 77. Bottom: "Mother-and-daughter", frame 28.

has been chosen for each category is also mentioned. These have been empirically chosen. The smaller weighting factors are attributed to the blocks with high masking capabilities (large threshold and large luminance). As a result, these blocks are less often chosen in the search for a new atom.

4.2.3. Perceptive allocation results

Incorporating masking capabilities within the search window selection procedure brings improvements for all sequences. Figs. 8, 9 and 11 present coded frames of the "hall-monitor", "mother-and-daughter" and "coastguard" sequences. One may observe that only small differences exist between the encoding schemes proposed. For "coastguard", one notices that artifacts in the background are reduced. For "hall-monitor" and "mother-and-daughter", improvements mainly appear along the edges surrounded by flat areas. Blow-ups of some frames are provided in Fig. 10. For these sequences, the improvements cannot be considered significant. Nevertheless, one observes from Fig. 12 that, for sequences with complex areas that change in time, e.g. the public in "Stefan", large improvements are achieved for the rest of the picture. Here, Stefan, the

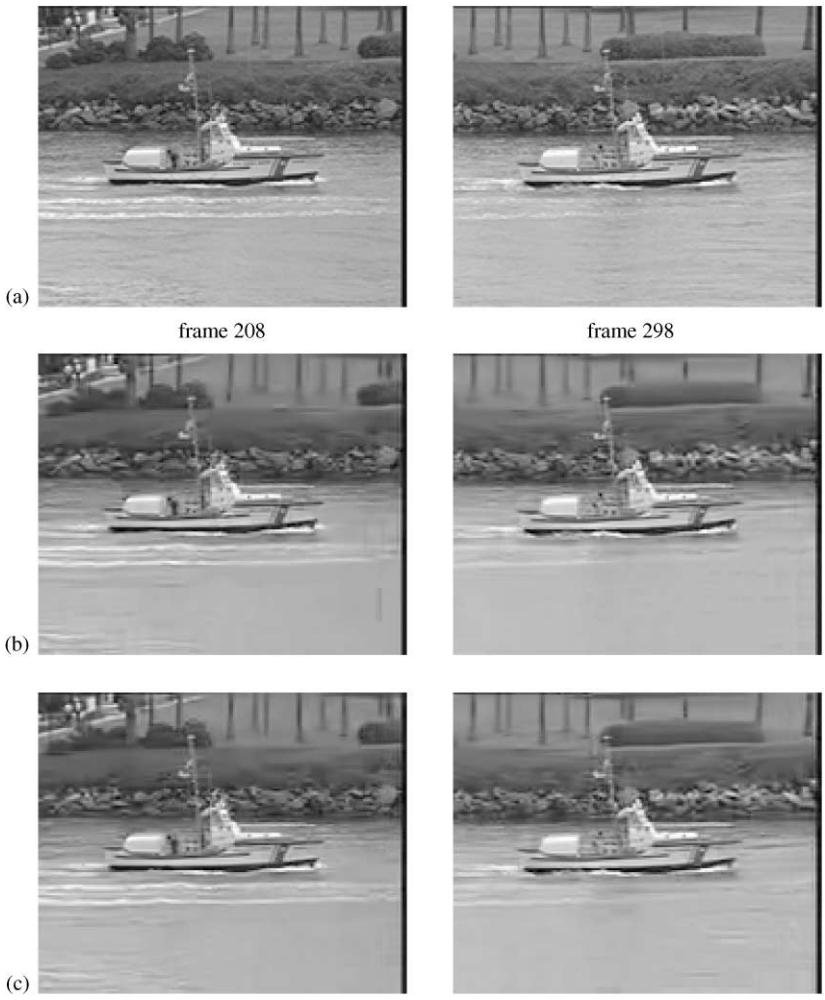

Fig. 11. "Coastguard" sequence (48 kbits/s, 10 frames/s): (a) original, (b) encoded without perceptive considerations, (c) encoded with block categorization according to masking capabilities during search window selection.

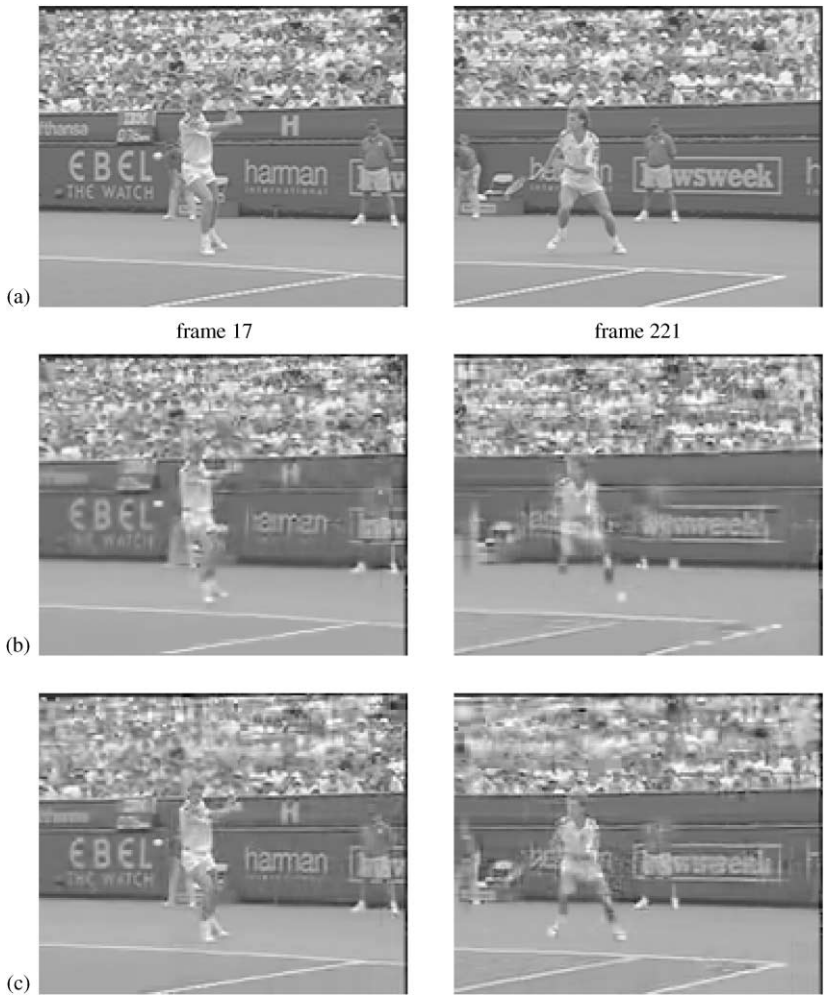

Fig. 12. "Stefan" sequence (75 kbits/s, 7.5 frames/s): (a) original, (b) encoded without perceptive considerations, (c) encoded with block categorization according to masking capabilities during search window selection.

linesman and the billboards are significantly improved compensating for public degradation.

5. Conclusion

In this paper, we have investigated the matching pursuits abilities of incorporating both perceptive and semantic criteria within a matching pursuits video codec.

In the context of very low bit-rate coding, the knowledge we have of the HVS is incomplete and partial. This is because most of the HVS models were built around visual thresholds, i.e. for slightly impaired pictures. Incorporating these models within VLBR coding algorithms did not always provide the expected improvements. From the different perceptive criteria we have investigated, only an empirical measure of the local masking abilities of the picture content has provided significant visual improvements of the displayed sequence.

The main objective of this paper has been to demonstrate that MP inherently permits spatially distributing information on each frame of the picture and adapting this distribution along

time. A significant advantage of MP vis-a-vis object-based methods as the one proposed within MPEG-4 is that it does not require accurate picture segmentation and interpretation. Even if the analysis suffers from inaccuracies or temporal instabilities, the sequence displayed remains pleasant to the eye and the coding efficiency is preserved. These results are promising for the development of intelligent dedicated coding schemes in which an automatic analyzing system provides some information concerning the relevance of the information captured by the camera. It also opens the door to simple user-assisted coding schemes in which the user coarsely selects areas of interest in the sequence.

References

[1] N. Ahmed, K.R. Rao, Miscellaneous orthogonal transforms, in: Orthogonal Transforms for Digital Signal Processing, Springer, New York, 1975, Chapter 7, pp. 169–171, ISBN 0-387-06556-3.

[2] M. Banham, J. Brailean, A selective update approach to matching pursuits video coding, IEEE Trans. Circuits Systems Video Technol. 7 (1) (February 1997) 119–129.

[3] J.R. Casas, L. Torres, A region-based subband coding scheme, Signal Processing: Image Communication 10 (1–3) (1997) 173–200.

[4] C. Chen, Video compression: standards and applications, J. Visual Commun. Image Representat. 4 (2) (June 1993) 103–111.

[5] S. Daly, K. Matthews, J. Ribas-Corbera, Face-based visually-optimized image sequence coding, in: International Conference on Image Processing, Chicago, Illinois, 4–7 October 1998, Vol. 3, pp. 443–447.

[6] V. Darmstaedter, J.F. Delaigle, J.J. Quisquater, B. Macq, Low cost spatial watermarking, Comput. and Graphics (Special issue on data Security in Image Communications and Networking) 22 (4) (July/August 1998) 417–424.

[7] J. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Opt. Soc. Am. A 2 (7) (July 1985) 1160–1169.

[8] R.J. De Valois, D.G. Albrecht, L.G. Thorell, Spatial frequency selectivity of cells in Macaque visual cortex, Vision Res. 22 (1982) 545–559.

[9] C. De Vleeschouwer, T. Delmot, X. Marichal, B. Macq, A fuzzy logic system for content-based bitrate allocation, Signal Processing: Image Communication 10 (1–3) (1997) 115–141.

[10] N. Doulamis, A. Doulamis, D. Kalogeras, S. Kollias, Low bit-rate coding of image sequences using adaptive regions of interest, IEEE Trans. Circuits Systems Video Technol. 8 (8) (December 1998) 928–934.

[11] T. Ebrahimi, MPEG-4 video verification model: A video encoding/decoding algorithm based on content representation, Signal Processing: Image Communication 9 (4) (May 1997) 367–384.

[12] A. Eleftheriadis, A. Jacquin, Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates, Signal Processing: Image Communication 7 (3) (September 1995) 231–248.

[13] S.-C. Han, J.W. Woods, Adaptative coding of moving objects for very low bit rates, IEEE J. Selected Areas Commun. 16 (1) (January 1998) 56–70.

[14] J. Hartung, A. Jacquin, J. Pawlyk, J. Rosenberg, H. Okada, P.E. Crouch, Object-oriented H.263 compatible video coding platform for conferencing applications, IEEE J. Selected Areas Commun. 16 (1) (January 1998) 42–55.

[15] J.R. Jain, A.K. Jain, Displacement measurement and its application in interframe image coding, IEEE Trans. Commun. 29 (12) (December 1981) 1799–1806.

[16] M. Kampmann, Segmentation of a head into face, ears, neck and hair for knowledge-based analysis-synthesis coding of videophone sequences, in: International Conference on Image Processing, Chicago, Illinois, 4–7 October 1998, Vol. 2, pp. 876–880.

[17] S.A. Klein, A.D. Silverstein, T. Carney, Relevance of human vision to JPEG-DCT compression, in: Human Vision, Visual Processing and digital display III, SPIE Electronic Imaging: Science and Technology, 1992, Vol. 1666, pp. 200–213.

[18] R. Koenen, Overview of the MPEG-4 standard (ISO/IEC JTC 1 SC29/WG11 N1730), available on the MPEG home page http://drogo.cselt.stet.it/mpeg/. Stockholm, July 1997.

[19] J.J. Kulikowski, P.O. Bishop, Fourier analysis and spatial representation in the visual cortex, Experientia 37 (1981) 160–163.

[20] D. Le Gall, MPEG: a video compression standard for multimedia applications, Commun. ACM 34 (4) (April 1991) 46–58.

[21] G.E. Legge, Spatial frequency masking in human vision: binocular interactions, J. Opt. Soc. Am. A 69 (6) (June 1979) 838–847.

[22] Y.-C. Li, T.-H. Wu, Y.-C. Chen, A scene adaptative hybrid video coding scheme based on the LOT, IEEE Trans. Circuits Systems Video Technol. 8 (1) (February 1998) 92–103.

[23] J.O. Limb, Distortion criteria of the human viewer, IEEE Trans. Systems Man Cybernet. 9 (12) (December 1979) 778–793.

[24] B. Macq, Weighted optimum bit allocations to orthogonal transforms for picture coding, IEEE J. Selected Areas Commun. 10 (5) (June 1992) 875–883.

[25] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (December 1993) 3397–3415.

[26] J.L. Mannos, D.J. Sakrison, The effects of a visual fidelity criterion on the encoding of images, IEEE Trans. Inform. Theory 20 (4) (June 1974) 525–536.

[27] T. Meier, K.N. Ngan, Automatic segmentation of moving objects for video object plane generation, IEEE Trans. Circuits Systems Video Technol. 8 (5) (September 1998) 525–538.
[28] R. Neff, A. Zakhor, Very low bit rate video coding based on matching pursuits, IEEE Trans. Circuits Systems Video Technol. 7 (1) (February 1997) 158–171.
[29] A.V. Nefian, M.H. Hayes, Face detection and recognition using hidden Markov models, in: International Conference on Image Processing, Chicago, Illinois, 4–7 October 1998, Vol. 1, pp. 141–145.
[30] A.N. Netravali, J.D. Robbins, Motion-compensated television coding: Part I, BELL System Tech. J 58 (3) (March 1979) 631–670.
[31] N.B. Nill, A visual model weighted cosine transform for image compression and quality assessment. IEEE Trans. Commun. 33 (6) (June 1985) 551–557.
[32] M. Ohta, S. Nogaki, Hybrid picture coding with wavelet transform and overlapped motion-compensated interframe prediction coding. IEEE Trans. Signal Process. 41 (12) (December 1993) 3416–3424.
[33] L.A. Olzak, J.P. Thomas, Handbook of Perception and Human Performance, Vol. 1: Sensory Processes and Perception, Wiley, University of California, Los Angeles, CA, 1986 (Chapter 7: Seeing Spatial Patterns).
[34] R.J. Qian, M.I. Sezan, K.E. Matthews, A robust real-time face tracking algorithm, in: International Conference on Image Processing, Chicago, Illinois, 4–7 October 1998, Vol. 1, pp. 131–135.
[35] T. Sikora, The MPEG-4 video standard verification model. IEEE Trans. Circuits Systems Video Technol. 7 (1) (February 1997) 19–31.
[36] Telecommunication Standardization Sector of ITU, Draft ITU-T Recommendation H.263, ITU Recommendations, July 1995.
[37] H. Wang, S.-F. Chang, A highly efficient system for automatic face region detection in MPEG video, IEEE Trans. Circuits Systems Video Technol. 7 (4) (August 1997) 615–628.
[38] M.A. Webster, R.L. De Valois, Relationship between spatial frequency and orientation tuning of striate cortex cells, J. Opt. Soc. Am. A 2 (7) (July 1985).
[39] H.R. Wilson, D.K. McFarlane, G.C. Phillips, Spatial frequency tuning of orientation selective units estimated by oblique masking, Vision Res. 23 (9) (1983) 873–874.
[40] H.R. Wilson, G.C. Phillips, Orientation bandwiths of spatial mechanisms measured by masking, J. Opt. Soc. Am. A 1 (2) (February 1984) 226–232.