Projet PERSEE
« Schémas Perceptuels et Codage vidéo 2D et 3D »
n° ANR-09-BLAN-0170

Livrable **D3.1** *30/09/2010*

State of the art in 2D content representation and compression

Vincent	RICORDEL	IRCCyN
Angélique	DRÉMEAU	IRISA
Christine	GUILLEMOT	IRISA
Marco	CAGNAZZO	LTCI
Erica	D'ACUNTO	LTCI
Béatrice	PESQUET-POPESCU	LTCI

Contents

1 Spatial and temporal prediction

The main goal in classical video coding techniques is the minimization of the bit-rate required to obtain a certain quality of the video. Good spatial and temporal prediction techniques can provide a high data compression balanced by a loss of information that slightly affects the video quality. The spatial prediction techniques are mostly inherited from images compression techniques. On the other hand the temporal prediction can be obtained by using motion estimation techniques.

1.1 Temporal prediction

As mentioned in the article by Dufaux and Moscheni [28], we can distinguish four main groups of motion estimation techniques:

- gradient techniques
- pel-recursive techniques
- block matching techniques
- frequency-domain techniques

Gradient techniques have been developed for image sequence analysis applications. They solve the optical flow and results in a dense motion field.

Pel-recursive techniques can be considered as a subset of gradient techniques. The recursion is usually carried out on a pel-by-pel basis leading to a dense motion vector field. This approach was proposed by Cafforio and Rocca in [8].

Frequency domain techniques are based on the relationship between transformed coefficients of shifted images. This technique never reached a widespread use.

As explained in the paper by Jain [46] block matching algorithms are based on matching of blocks between two images, the aim being to minimize a disparity measure. Specifically developed for image sequence coding, they are widely used. In block matching motion estimation, the image is partitioned into blocks and the same displacement vector is assigned to all pixel within a block.

For each block, the displacement vector is evaluated by matching the content of a block of pixel with a corresponding block within a certain search range, placed in the previous frame (or in any refererence frame), and by searching the spatial location minimizing the matching criterion (e.g. MSE).

Block matching was initially designed to estimate displacements with a precision of one pixel in the H.261 standard, but since MPEG-1 half-pixel precision begun to be used, and nowadays a fourth-pixel precision can be achieved, as it happens in the most widely used standard H264[1].

The importance of tepmoral and spatial prediction is witnessed by the fact that the best performing video standard, H.264 [89], heaviliy relies on prediction to achieve its high compression performance.

[1] Eighth-pixel precision is possible using the Key Technical Area (KTA) tools.

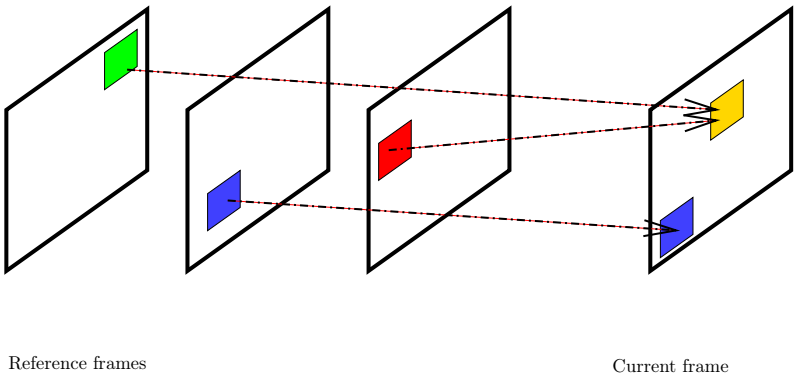

Figure 1: Single or multiple references in H.264 temporal prediction of the current block.

Temporal prediction is realized with a block-matching based technique. The motion estimation and motion compensation techniques are realized in order to provide both predicted (P) and bi-predicted (B) frames. For the first ones the motion estimation is performed referring to the previous I or P frame, while for te second ones, the prediction comes from both direction, from the previous and the following I or P frame. In the paper by Wiegand, Zhang and Girod [90], the prediction is enhanced by using up to 16 reference frames. This is in contrast to prior standards, where the limit was typically one or, in the case of B pictures, two. Moreover for the prediction of B-frames is possible to use any macroblock type, including I-macroblocks. This standard also introduces variable block-size motion compensation with block sizes as large as 16×16 and as small as 4×4, enabling precise segmentation of moving regions. The ability to use multiple motion vectors per macroblock was developed and the motion vectors for each 8×8 or larger partition region can point to different reference pictures. Motion prediction has also been improved by introducing quarter pixel precision and weighted prediction for both B and P frames. Temporal prediction in H.264 is depicted in Fig. 1.

Recently research in video compression has showed several deficiency in block based models: first of all it fails to capture the true motion in natural video as explained in the paper by Han and Podilchuck [42]. Another intrinsic problem with existing motion-compensated predictive coders is the coding of the residual frame or displaced frame difference. Typically, it is encoded by applying transform coding techniques (such as the discrete cosine transform) which work well on still images. However, such methods are quite inefficient on the displaced frame difference (DFD), which consists predominantly of high-frequency data.

In the same paper, a motion estimation method that exploits a dense motion vector field is proposed taking into account the problem of coding both the motion and the subsequent DFD frame. First a dense motion field estimation is computed during the encoding process. It is possible to take advantage of the dense motion information to

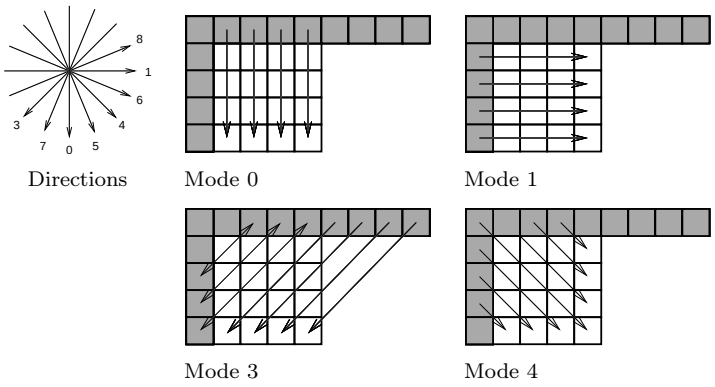

Figure 2: H.264 spatial prediction, INTRA 4×4.

model and encode the DFD frame: the dense motion field is a much more accurate representation of the true motion. Since it is impractical to transmit one unique motion vector for every single pixel in an image, a variable-depth motion field is found from the original motion field. The DFD encoding scheme utilizes the dense motion field in predicting where the DFD energy will be significant, and only coding the DFD values in these regions.

1.2 Spatial prediction

Intra prediction is an effective method for reducing the coded information of an image or an intra frame within a video sequence by exploiting spatial correlation witin a picture. The conventional method today is to create a sample predictor block by extrapolating the reconstructed pixels surrounding the target block to be coded. The sample predictor block is subtracted from the target block and the resulting residual coded using transformation, quantization and entropy coding. This is an effective method for sample predictor block creation in most sequences. However the extrapolation method is not able to represent sample prediction blocks with complex texture. Furthermore, pixels that are far from the surrounding pixels are usually badly predicted.

In H.264 [88], the spatial prediction is improved with a new technique of extrapolating the edges of the previously-decoded parts of the current picture. It's applied in regions of pictures that are coded as intra. This improves the quality of the prediction signal, and also allows prediction from neighboring areas that were not coded using intra coding.

In particular, two classes of intra coding types are supported, which are denoted as INTRA-4×4 and INTRA-16×16. When using the INTRA-4×4 mode, each block of the luminance component can choice one out of nine prediction modes, as shown in Fig. 1.1. Beside DC prediction, eight directional prediction modes are specified. When utilizing the INTRA-16×16 mode, (see Fig. 3) which is well suited for smooth

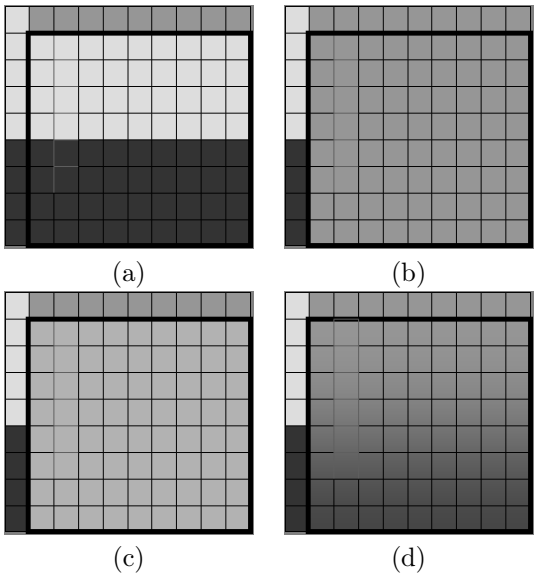

Figure 3: H.264 spatial prediction, INTRA 16×16. (a) Horizontal prediction; (b) Vertical preditction; (c) Average prediction; (a) Planar prediction.

image areas, a uniform prediction is performed for the whole luminance component of a macroblock. Four prediction modes are supported: horizontal, vertical, average and planar.

2 Transformation

2.1 Anisotropic transforms

The classical scheme of image compression is based on three steps: transform, quantization and lossless coding. Recently, research efforts have focused on the choice of the transform that best represents a natural image. As a matter of fact, in spite of its great success, wavelet transform is not the optimal basis for an image. Indeed, it is very effective in representing smooth signals with pointwise discontinuities (like an image row), but fails in representing discontinuities along curves, like the contours between neighboring visual objects, which typically characterize images. We analyze in the followint three main approaches for the new generation image coding scheme: the object-based paradigm, the directional transforms, and the adaptive lifting schemes.

The **object-based paradigm** is a first tentative solution to this problem. To begin with, considering an image as composed by objects, and not by pixels, is more intuitive and natural. Object-based coding offers a large number of high level functionalities, for example, the user can choose to decode only objects of interest, or to assign them

different coding resources and different error-protection levels. Furthermore an object-based description can be used for subsequent classification tasks, and it is more suited to quality-driven compression schemes. An example of oject based approach is the shape-adaptive wavelet transform, proposed by Li and Li [54], and adopted in MPEG-4 for the arbitrarly-shaped object coding. The main assumption is that, with an object-based approach, the wavelet works only on the interior of the objects, that is, almost stationary signals, and can therefore provide near-optimal performance. This approach requires an adapted coding algorithm, like the shape-adaptive version [12] of the well-knonw SPIHT algorithm [69], and it has shown good performances when the segmentation task is relatively easy [13], like for satellite multispectral and hyperspectral images [10, 11]. On the other hand, the object-based approach has mixed results when applied to natural images [9]: in this case it is profitable only if the segmentation is very accurate, and if the coding cost of the segmentation map (which has to be sent along with transform coefficients) is not too high.

New **directional transforms** represent a more direct solution to wavelet inefficiency on image contours. While in object-based coding the transform remains the wavelet and the intelligence is put on the scheme, here the wits is in the transform. Recent studies have shown that wavelet's inability to adequately describe image contours is due to its separability which (while allowing for a simple implementation) cuts it away from two fundamental properties: directionality and anisotropy [26]. The new directional transforms try to overcome these limits by adding these characteristics to that of wavelet transform, such as multiresolution, localization and critical sampling. Many transforms have been proposed in the last few years, and we show here the most relevant of them. **Curvelets**, introduced by Candès and Donoho [14], provide stable, efficient, and near-optimal representation of otherwise smooth objects having discontinuities along smooth curves. By applying naive thresholding to the curvelet transform of such an object, one can form approximations with rate rivaling the rate obtainable by complex adaptive schemes which attempt to 'track' the discontinuity set. The **contourlets** have been proposed by Do and Vetterli in 2005 [26]. The contourlets framework has many desirable characteristics, such as directionality, anisotropy, an almost optimal NLA[2] behavior for simple classes of images and, unlike other directional transforms, it is easily implemented by a filter bank. Its main drawback is a slight redundancy which, however, is not really a problem in the context of low bit-rate coding. **Bandelets**, introduced by Le Pennec and Mallat [65], decompose the image along multiscale vectors that are elongated in the direction of a geometric flow. This geometric flow indicates directions in which the image grey levels have regular variations. The image decomposition in a bandelet basis is implemented with a fast subband filtering algorithm. Bandelet bases lead to optimal approximation rates for geometrically regular images. For image compression and noise removal applications, the geometric flow is optimized with fast algorithms, so that the resulting bandelet basis produces a minimum distortion. Comparisons with wavelet image compression algorithms show PSNR improvements from 0.5dB to 1.5dB with respect to classical Daubechies filters [3]. **Directionlets** are a a lattice-based perfect-reconstruction and critically sampled

[2]Non-linear approximation; see the extensive review by DeVore for more information [23].

anisotropic multi-directional wavelet transform [81]. The transform retains the separable filtering and subsampling and the simplicity of computations and filter design from the standard two-dimensional WT, unlike in the case of some other directional transform constructions (e.g. curvelets, contourlets or edgelets). The corresponding anisotropic basis functions (directionlets) have directional vanishing moments along any two directions with rational slopes. Furthermore, this novel transform provides an efficient tool for nonlinear approximation of images, achieving the approximation power $O(N^{-1.55})$, which, while slower than the optimal rate $O(N^{-2})$, is much better than $O(N^{-1})$ achieved with wavelets, but at similar complexity.

Another possible approach for new generation transforms are the **adaptive lifting schemes**. The lifting scheme [22] is an efficient and flexible implementation of the wavelet transform. One of the main advantages of the lifting structure is to provide a totally time domain interpretation of the wavelet transform and this feature makes simpler to design new wavelets and content-adaptive wavelets. Adaptive lifting schemes can be use to deal with the problem of contour representation, for example, by constructing directional wavelets, with the filtering direction chosen according to the local orientation of image edges [35, 15, 25], or changing the filters according to the regularity of input signal [34, 21, 43] in order to utilize different and more fit filters when contours or singularities are encountered. A major problem of adaptive lifting schemes is that they are strongly non-isometric transforms, which bars from computing the distortion in the transform domain. On the other hand, this is would be highly desirable in order to perform efficient resource allocation [79]. This problem has been recently adressed by Parrilli *et al.*, both in the case of update-adaptive [63] and prediction-adaptive [64] lifting scheme. The strategy adopted is based on the observation that, although adaptive lifting schemes are nonlinear operators, they can be considered equivalent to suitable time-varying linear filters, which eventually allows us to generalize the traditional distortion computation methods.

3 Oriented Wavelets

This section presents an adaptive oriented wavelet transform introduced in [17], in which the lifting steps of a 1D wavelet are oriented along a discrete set of orientations. The geometry of the image is explicitly described by an orientation map. The orientations are chosen to minimize the energy of the wavelet coefficients in the high-frequency subbands [17], so as to pack the energy of the image as much as possible in the low-frequency subbands. Each level of decomposition consists in splitting the sampling grid in two complementary quincunx cosets and applying the lifting steps along the chosen orientations. The orientation map is coded using two independent interleaved quad-trees. One is used to encode the horizontal and vertical orientations across even levels, while the other is used to encode the diagonal and antidiagonal orientations across odd levels.

3.1 Oriented Wavelets on Quincunx Grid

3.1.1 Quincunx sampling

Let us now consider a 2D lattice defined from an integer matrix M as

$$L(M) = \{m \in \mathbb{Z}^2, m = Mn, n \in \mathbb{Z}^2\}.$$

A quincunx lattice $L^0_{\mathcal{L}}$ can be generated from the matrix

$$M = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

By translating this lattice with the vector $e = \begin{pmatrix} 0 & 1 \end{pmatrix}^\top$, another complementary quincunx lattice $L^0_{\mathcal{H}}$ is obtained. The square sampling grid $\mathbb{Z}^2$ can then be decomposed in these two quincunx lattices as

$$\mathbb{Z}^2 = L^0_{\mathcal{L}} \cup L^0_{\mathcal{H}}.$$

By iterating this decomposition on the $L^0_{\mathcal{L}}$ grid (Fig. , the following multiresolution structure is obtained:

$$\begin{cases} L^k_{\mathcal{L}} & = L(M^{k+1}), \\ L^k_{\mathcal{H}} & = L(M^{k+1}) + M^k e. \end{cases}$$

The lattices $L^k_{\mathcal{L}}$ and $L^k_{\mathcal{H}}$ are either square or quincunx, for even or odd levels k respectively ($L(M^2) = L(2I) = 2\mathbb{Z}^2$). In the following, we will call *quincunx level* a level for which the grids L^k are quincunx grids (k even), whereas a *square level* will denote a level for which the grids L^k are square grids (k odd). Moreover, a quincunx lattice $L(M^k)$ for k odd can be seen as a square lattice rotated by $\frac{\pi}{4}$ where the distance between samples is $2^{\frac{k}{2}}$. Thus, this recursive partitioning defines an l-level quincunx sampling pyramid $(L^0_{\mathcal{H}}, ..., L^{l-1}_{\mathcal{H}}, L^{l-1}_{\mathcal{L}})$ where the downsampling factor at each scale is $|\det(M)|^{\frac{1}{2}} = \sqrt{2}$.

This l-level quincunx pyramid gives a multiresolution representation of the 2D signal. This type of representation has been used previously to define quincunx wavelet bases [31] [37]. However, notice that here and unlike the separable case, only one mother wavelet instead of three is needed to represent the signal. Also, as the downsampling factor is $\sqrt{2}$ instead of 2, twice the number of decomposition levels are necessary to obtain the same low-frequency subband resolution as in a separable decomposition.

3.1.2 Oriented lifting

Rather than using quincunx wavelets, our approach consists in applying a 1D wavelet transform along directions selected adaptively for each wavelet coefficient according to an orientation map. In this section, we assume that the map is known and only the adaptation of the lifting steps is presented. In Section IV.A and V we will present how the map is obtained depending on the application.

The wavelet coefficients, corresponding to the prediction errors at the successive levels, are computed on the $L^k_{\mathcal{H}}$ grids. The approximation images are computed on the $L^k_{\mathcal{L}}$ grids on which the decomposition is iterated. Each point in $L^k_{\mathcal{H}}$, has four neighbors in $L^k_{\mathcal{L}}$. Therefore, a point $n \in L^k_{\mathcal{H}}$ can be predicted from any combination of these neighbors. For instance, in the case of lifted quincunx wavelets, the sample $S[n]$ is predicted from an average of all four neighbors. Here, this same sample is instead predicted from either its neighbors in the same line *or* in the same row . This requires the knowledge of the orientation map which defines which direction of prediction is used at location n. Since only two choices are allowed, this map is binary. Generally, the orientation which minimizes the prediction error is chosen, defining a binary map $m_{L^k_{\mathcal{H}}}$ on $L^k_{\mathcal{H}}$.

To compute the wavelet coefficient at location n, the predict steps of a 1D biorthogonal wavelet are applied in the orientation chosen in n. Thus, the predict steps of the 1D wavelet defined in Eq. 1

$$P(\alpha_i) : S[n_1] \leftarrow S[n_1] + \alpha_i(S[n_1 - 1] + S[n_1 + 1]), n_1 \text{ odd}, \tag{1}$$

$$P(\alpha_i) : S[n_1] \leftarrow S[n_1] + \alpha_i \sum_{n \in \mathcal{R}_{n_1}} S[n], \tag{2}$$

where $\mathcal{R}_{n_1}$ is the set of the two horizontal neighbors of n_1 in $L^k_{\mathcal{L}}$, or:

$$P(\alpha_i) : S[n_1] \leftarrow S[n_1] + \alpha_i \sum_{n \in \mathcal{C}_{n_1}} S[n], \tag{3}$$

where $\mathcal{C}_{n_1}$ is the set of the two vertical neighbors of n_1 in $L^k_{\mathcal{L}}$.

Note that more orientations could be defined by either interpolating samples in $L^k_{\mathcal{L}}$ (as done e.g. in [24] for the separable wavelet) or by considering further neighbors. However, the increased flexibility in orientation comes at the expense of an increased cost for the orientation map. Having a larger set of orientations would also increase the complexity of the transform as all modes of coding have to be tested. The problem of finding the optimal balance between the number of allowed orientations and the cost of the geometric side information is a complex question, which is not addressed here and left open for future work.

The update steps have to be modified according to the predict steps. A sample at location n_2 in $L^k_{\mathcal{L}}$ may indeed be used zero to four times to predict its neighbors in $L^k_{\mathcal{H}}$, unlike in the 1D case where it is used exactly twice. The factors β^*_i used in the modified update steps are obtained by weighting the original factors β_i of the 1D wavelet given in Eq. 4. Since the orthogonality property is lost due to this varying number of predictors, the aim of the update step is rather to ensure that some statistical properties of the original signal are preserved in the low frequency band. Another criterion for determining the proper β^*_i could be to minimize globally the scalar product between the wavelet basis functions, so as to obtain a close to orthogonal transform. This possibility has not been investigated here. Instead, the following empirical modification is proposed. For the 5/3 wavelet, this modification ensures the mean of the original

image is preserved in the low-frequency band. Depending on the number v of neighbors predicted from a sample at location n_2, the update factors β_i^* are defined as

$$\beta_i^* = \begin{cases} \frac{2}{v}\beta_i & \text{if } v \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the update step is modified as follows:

$$U(\beta_i) : S[n_2] \leftarrow S[n_2] + \beta_i^* \sum_{n \in \mathcal{U}_{n_2}} S[n], \tag{4}$$

where $\mathcal{U}_{n_2} \subset L^k_{\mathcal{H}}$ is the set of neighbors of n_2 using the sample $S[n_2]$ as a predictor. When the direction of prediction is the same for all points in $L^k_{\mathcal{H}}$, the decomposition is equivalent to the 1D wavelet applied along that direction on $L^k_{\mathcal{H}}$.

All these modifications translate directly in square levels (k odd) when viewed as quincunx levels (k even) rotated by $\frac{\pi}{4}$. The orientations are then diagonal or antidiagonal instead of horizontal or vertical, but the lifting steps apply similarly

3.1.3 Representation of the Orientation Map with Quad-Trees

In order to pack the energy as much as possible in the lower frequency subbands, the orientation map is chosen so as to minimize the prediction error at each level. Without entropy coding, the orientation map would cost slightly less than 1 bpp, which is prohibitive. However, this binary information on the filtering direction is not always relevant. Indeed, when the distortion obtained by predicting in either one or the other direction is similar, the choice of the proper orientation does not impact the distortion significantly. This happens mainly in uniform regions, where both predictors are similar, or in textured regions where the pixels are less correlated, hence where the prediction fails. Thus, the orientation information is only important on edges, which concerns a small proportion of the pixels in natural images. It is therefore possible to propagate the orientation information from edges to other regions to reduce the entropy of the map substantially with a negligible impact on the overall distortion.

In order to do so, the orientation map is coded using two independent interleaved quad-trees. One is used to encode the horizontal and vertical orientations across even levels, while the other is used to encode the diagonal and antidiagonal orientations

3.2 Edge driven oriented wavelets transform

In the context of still image compression, G. Jeannic[47] also developed a new representations based on the 2D discrete wavelet transform. This transform belongs to those that adapt the wavelet basis to the local content of the image, and that are implemented via oriented lifting schemes.

In his thesis G. Jeannic compared the ability of three of these oriented representations at minimizing the energy of the reconstructed high frequencies sub-band along the filtering orientation. Two successive oriented lifting schemes are applied on the image. The first one along the local direction of regularity, the second one along the

further horizontal or vertical orientation. In order to match the image content, the first direction has to be estimated. He proposed two methods. The first one extracts the significant edges of the image. The second one maximizes the mean probability of the gradient measure. Along with those methods, the regions of the images are classified into four structural categories: the uniform areas, the mono-oriented areas where only one dominant orientation is detected, the multi-oriented areas where more than one dominant orientation is detected, and the isotropic textured areas where no dominant orientation is detected.

He thus proposed a structural representation based on oriented lifting schemes which decomposes the image on sub-bands that can be interpreted in different manners depending on the structural classes of its elements. For example for mono-oriented areas, the high frequencies along the direction of regularity do not have the same meaning than the high frequencies across it.

The geometric information needs to be coded, along with the oriented wavelet coefficients, for the synthesis (image reconstruction). The reduced representation of the geometrical features in the images can be represented by the extracted edges for the first method of estimation, and by quad-trees for both methods. For chain coding, he proposed a new method that exploits the gradient orientation of the previous resolution decoded image, and shows improvement over a markovian approach using the past coded edge elements at the current resolution. However the quad-trees cost, with adaptive arithmetic coding and using the spatial coded/decoded context, is less expensive than the associated extracted edges one.

An adaptive quantization is performed taking into account the structural properties of the image, and by exploiting some proprieties of the human visual system On one hand the elements of the isotropic textured areas can be roughly quantized because of induced masking effect, while the opposite case appears for the uniform areas. On the other hand the sub-bands of the mono-oriented and multi-oriented areas are similarly quantized. The high frequencies along the direction of regularity can be neglected in comparison with the high frequencies across it, exploiting the anisotropy of the structural representation.

3.3 Dictionaries adapted to sparse representations

The use of sparse dictionaries can help in image and video compression. The *quality* of the sparse expansion, in term of sparsity and approximation, depends on the algorithm used to create it. It is also related to the dictionary in which the expansion is performed. The more the dictionary will be adapted to the characteristics of the signal and promote sparsity of the expansion, the "best" the sparse expansion will be. The definition of dictionaries constitutes thus an important issue and is the subject of numerous contributions.

Two different types of dictionaries can be distinguished:

- dictionaries made up of a predefined set of functions,
- dictionaries learned from a set of signals.

Predefined dictionaries have the advantage to be simple to use. However their "success" depends on their well-adaptation to a sparse description of the considered signals. For example a dictionary well-adapted to textured-image signals do probably not encourage a good sparse decomposition of homogenous-image signals.

Learning enables to define a dictionary particular to a certain type of given signals (the training set) under some criteria we can control. Particularized to sparse decompositions, the problem can be formalized as follows. Given a set of training signals $\{\mathbf{y}_j\}_{j=1}^K$, we want to find $\mathbf{D}^\star$ which leads to the best distortion-sparsity compromise:

$$\mathbf{D}^\star = \arg\min_{\mathbf{D}} \left\{ \sum_j \min_{\mathbf{x}_j} \|\mathbf{y}_j - \mathbf{D}\mathbf{x}_j\|_2^2 + \lambda \|\mathbf{x}_j\|_0 \right\},$$

where $\mathbf{x}_j$ is the sparse decomposition of signal $\mathbf{y}_j$ in dictionary $\mathbf{D}$.

The approaches to dictionary design that have been proposed so far have a common two-step process:

- Find the sparse coefficients given the dictionary,
- Update the dictionary assuming known and fixed sparse vectors.

Their main differences rely on the methods used to successively estimate the sparse vectors and the dictionary. We present some of them in the next subsection.

3.3.1 Dictionary learning

The learning methods for dictionary adapted to sparse expansions can be divided into 4 classes: the Bayesian approaches, the Method of Optimal Directions (MOD), the K-SVD algorithm and the learning of structured dictionaries, such as unions of bases.

Bayesian approaches

The Bayesian approaches place the optimization of dictionaries into a probabilistic framework. Each training signal $\mathbf{y}_j$ is seen as a noisy combination of atoms chosen from a dictionary $\mathbf{D}$.

$$\mathbf{y}_j = \mathbf{D}\mathbf{x}_j + \mathbf{n}, \tag{5}$$

where $\mathbf{n}$ is a white Gaussian noise.

Two approaches can then be distinguished.

The first one considers the following ML estimation problem:

$$\mathbf{D}^\star = \arg\max_{\mathbf{D}} \sum_{j=1}^{K} \log p(\mathbf{y}_j|\mathbf{D}), \tag{6}$$

where

$$p(\mathbf{y}_j|\mathbf{D}) = \int_{\mathcal{X}^M} p(\mathbf{y}_j, \mathbf{x}_j|\mathbf{D}) d\mathbf{x}_j = \int_{\mathcal{X}^M} p(\mathbf{y}_j|\mathbf{x}_j, \mathbf{D}) p(\mathbf{x}_j) d\mathbf{x}_j. \tag{7}$$

Several different probability distributions are proposed in the literature. We find thus Cauchy distributions and Laplace distributions ([58] et [53]), supposed to encourage sparsity (the Laplace distribution actually implements the ℓ_1-norm measure).

The marginalization (7) is untractable. Olshausen and Field propose in [58] to replace it by a maximization:

$$\mathbf{D}^\star = \arg\max_{\mathbf{D}} \sum_{j=1}^{K} \max_{\mathbf{x}_j} \log p(\mathbf{y}_j, \mathbf{x}_j | \mathbf{D}). \tag{8}$$

A gradient descent is then use to estimate the sparse vectors $\mathbf{x}_j$ and the dictionary $\mathbf{D}$. This solution tends to increase the values of the dictionary atoms, the authors propose thus to constraint the ℓ_2-norm of the atoms.

Confronted to the same integration (7), Lewicki and Sejnowski ([53]) choose to approximate the joint probability distribution $p(\mathbf{y}_j|\mathbf{D})$ instead of maximizing on the variables $\mathbf{x}_j$. They resort to a Laplace approximation, which approximates a complex probability distribution with a Gaussian. This technique enables to solve analytically the integration (7) and thus to take into account the uncertainties we have on the probability distribution of the $\mathbf{x}_j$'s. It presents also the advantage to avoid the definition of constraints of the norms of the dictionary atoms. A simple gradient descent can then be used to estimate the dictionary without any other consideration.

The second approach considers the following MAP estimation problem:

$$(\mathbf{D}^\star, \{\mathbf{x}_j^\star\}) = \arg\max_{\mathbf{D}} \sum_{j=1}^{K} \log p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}). \tag{9}$$

This is the approach adopted by Murray and Kreutz-Delgado dans [57] and Kreutz-Delgado and Rao dans [50]. A gradient descent is used to estimate the dictionary. Confronted to the same problem of increasing values of the atoms as Olshausen and Field, the authors choose a prior which constraints the dictionary to have a unitary Frobenius norm. But the main contribution of their work is the use of the FOCUSS algorithm to perform the estimation of the sparse vectors. This choice improves dramatically the performance of the learning algorithm in comparison with other previous Bayesian algorithms.

Method of Optimal Directions (MOD)

The Method of Optimal Directions, introduced by Engan *et al.* in [29], is explicitely inspired by the Lloyd-Max algorithm used for the quantization dictionary learning ([36]). The estimation of the sparse vectors is performed, as in the algorithm proposed by Kreutz-Delgado and Rao, by a sparse decomposition algorithm (OMP is here prefered to a ℓ_1-norm). The estimation of the dictionary constitutes the main contribution of the MOD method and resides in the minimization of the global approximation error $\sum_j \|\mathbf{y}_j - \mathbf{D}\mathbf{x}_j\|_2^2$. Although resulting in an update equation close to the one proposed by Olshausen and Field in [58], a fast implementation is made possible and the performance is improved. However, in this method, as well as in Olshausen and Field's one, an atom normalization is required.

K-SVD algorithm
Proposed by Aharon *et al.* in [1], this method relies on a singular value decomposition (SVD) to estimate the dictionaty $\mathbf{D}$. After estimating the sparse vectors by a sparse decomposition algorithm like OMP, the dictionary atoms are updated successively. The algorithm proceeds as follows. The contribution of the considered atom, denoted $\mathbf{d}_k$, in the description of the signals is evaluated by a representation error matrix corresponding to the difference between the signals $\mathbf{y}_j$'s "using" the atom $\mathbf{d}_k$ and their "troncated" sparse approximations (*i.e.,* in which we removed the atom $\mathbf{d}_k$). The obtained vectors form a matrix on which a SVD decomposition is applied. The atom $\mathbf{d}_k$ is then estimated as the first singular vector. The K-SVD algorithm achieves good performance in comparison with other algorithms in literature.

Learning of structured dictionaries: union of bases
In the context of transform coding, the use of redundant dictionaries can have important repercussions in the coding cost of the atoms indices chosen for the sparse decomposition. A way to reduce this cost is to introduce some structure in the dictionary. This can be simply done by considering a set of bases. Several contributions deal thus with learning of unions of bases. We present here two of them, which propose different approaches although based on same techniques: the one optimizes the union of bases together, the other considers each basis independently.

A first method is introduced by Lesage *et al.* in [52]. Based on a SVD as Aharon's algorithm, the algorithm proposes another use of it. It proceeds by estimating at the same time all atoms of one basis and not successively as K-SVD does. The sparse vectors are also estimated in a different way, using the Block Coordinate Relaxation method (BCR) introduced in [70]. This method extends the soft-thresholding presented previously to union of orthonormal bases.

Another algorithm is proposed by Sezer *et al.* in [72]. Instead of considering the entire dictionary, the authors assume that each signal has a sparse decomposition in a unique basis. The algorithm presents thus an additional step of classification: each signal is classified according to the basis which minimizes the approximation error, resulting in several "training subsets". The algorithm proceeds then in a classical way by estimating the sparse vectors and the bases successively on each corresponding subset. The sparse vectors are calculated by a hard-thresholding (presented previously). The bases are updated with a SVD-based method similar to the one used by Lesage *et al.* .

4 Rate-quality optimization

The rate-quality optimization (RQO) problem consists essentially in finding the best coding technique for each set of data into which the input signal can be partitioned. This approach is a underlying common element to many compression techniques, from EBCOT [76] to the video coding techniques [74].

In the RQO context therefore, one usually chooses the coding technique (or "mode")

maximazing the quality for a given rate. However, while defining and computing the coding rate is rather easy (it is defined in terms of number of bits per pixel or per second, and usually computed by actually performing the encoding operation), the *quality* definition and computation is much more vague.

Broadly speaking, there are two classes of quality measures for visual data (images and video): subjective and objective measures. The first ones are obtained by interviewing a set of people looking at the decoded images (and possibly at the original one) and judging about the perceived quality. This kind of measure is supposed to be the most representative one, but of course it cannot be integrated into a compression algorithm.

On the other hand, objective measures can be computed as a mathematical function of the original image, say $f(n,m)$ and the decoded one, say $\widehat{f}(n,m)$, an operation that can be performed by an automatic system (a coding algorithm). A very common class of objective measures is based on the mean square error (MSE):

$$\mathrm{MSE} = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\left[f(n,m) - \widehat{f}(n,m)\right]^2,$$

like the peak signal-to-noise ratio (PSNR):

$$\mathrm{PSNR} = 10\log_{10}\frac{255^2}{\mathrm{MSE}}.$$

However these measures are commonly regarded as not accurately representative of perceived quality [30]. For this reason, a huge quantity of perceptual objective quality measures have flourished in recent years. A common factor to these measures is the attempt to model the complex human visual system (HVS).

In the following, we review QRO methods using objective measures, both non-perceptual (*i.e.*, MSE-based) and perceptual. The first have the merit to be very simple to compute and to analyze; the second can gain from a more accurate model of HSV.

4.1 The optimization problem

As told in the previous paragraph, the optimization problem ends up in finding the best coding mode for each data block, within a given set of coding technique [74]. Sometimes this problem is referred to as *operational control*. In the case of video this means a jointly optimal choice of

- the so-called coding mode, which can be chosen among compensated and non-compensated ones;
- for the motion-compensated modes, the motion information, i.e. all the information needed in order to compute the motion compensated prediction of the current block (motion vector(s) and segmentation information if present);
- the quantization step.

This choice is performed with a Lagrangian technique. The three free parameters related to modes, quantization, and motion information are tied together by experimental relationship which nevertheless depend on the video coder model, and which allow to perform the optimization over a single parameter.

To better illustrate the problem, let us consider a source producing K samples $S_1, S_2, \ldots, S_K$. Each sample can be a scalar or a vector, for example it can be a macroblock (MB) in a video sequence. Let us suppose that there exist several ways to encode each sample. Let M be the set of all possible coding modes.

The operational control of the encoder consists in choosing the best set of encoding modes according to a cost function and given a rate constraint. Let $\mathbf{I} = I_1, I_2, \ldots, I_K$ be the coding modes of all the source samples. Let $D(\mathbf{S}, \mathbf{I})$ and $R(\mathbf{S}, \mathbf{I})$ respectively the distortion associated to the coding of $\mathbf{S}$ in mode $\mathbf{I}$. The target of the RD optimization is to minimize $D(\mathbf{S}, \mathbf{I})$ given that the rate $R(\mathbf{S}, \mathbf{I})$ is below an assigned value R_c. This constrained problem can be changed into an unconstrained one by the use of a Lagrangian parameter:

$$\mathbf{I}^* = \arg\min_{\mathbf{I}} J(\mathbf{S}, \mathbf{I}|\lambda_{\text{MODE}}) \tag{10}$$

where

$$J(\mathbf{S}, \mathbf{I}|\lambda_{\text{MODE}}) = D(\mathbf{S}, \mathbf{I}) + \lambda_{\text{MODE}} R(\mathbf{S}, \mathbf{I}) \tag{11}$$

We observe that a rigorous optimization would require a joint minimization of all the source symbols. If we use a simplifying hypothesis, *i.e.* we suppose for the moment that each symbol coding mode can be optimized independently, we arrive to a different formulation. In this case we can obtain the optimal mode for the i-th source symbol just by solving the unconstrained problem:

$$I_i^* = arg\min_{I} J(S_i, I|\lambda_{\text{MODE}}) \tag{12}$$

where

$$J(S, I|\lambda_{\text{MODE}} = D(S, I) + \lambda_{\text{MODE}} R(S, I) \tag{13}$$

In the video coding case, the source symbols S_i are the MB of a video sequence. For the moment, let us suppose that the optimal quantization step Q is given. The criterion to minimize is:

$$J_{\text{MODE}}(S_k, I_k|Q, \lambda_{\text{MODE}}) = D_{\text{REC}}(S_k, I_k|Q) + \lambda_{\text{MODE}} R_{\text{REC}}(S_k, I_k|Q) \tag{14}$$

where D_{REC} and R_{REC} have to be suitably computed according to the coding mode, as shown in the following. It is important to observe that for a given total rate R_c, could give several couples (Q, λ_{rmMODE}) which attain the same value of the criterion J, so a joint optimization would be needed. However, in [74] authors show an experimental relationship between the best values of Q and λ_{MODE}. So an operative strategy could be to tune the rate by choosing Q, and using the value of λ_{MODE} according to Q and to the relationship (which in turns, depends on the model of the video coder).

An important issue is how to compute this criterion for each motion compensated mode. Once the best motion information has been chosen, we can compute the "best" motion-compensated prediction of the current block according to a given mode. Then

we use the same paradigm (transform-quantization-inverse transform) to compute D_{REC}, while in the computation of R_{REC} we have to consider both the cost of the transformed coefficient and that of the motion information, encoded by the techniques that we are introducing. The second issue is how to choose λ_{MODE}.

In facts, some experimental relationships have been found between Q and λ_{MODE} [74, 87]: for the H.263 coder, $\lambda_{\mathrm{MODE}} = 0.85Q^2_{H.263}$; while for H.264, $\lambda_{\mathrm{MODE}} = 0.85 \cdot 2^{(Q_{H.264}-12)/3}$).

Most of the currently used RD optimization problems are based on a classical MSE-based quality measure. This means that the mean square error is used as the function D in the previous equations. Unfortunately as pointed many times in literature (see for example the article by Wang and Bovik [86]), even though the MSE possesses many favorable properties for application and analysis, tests have demonstrated that the main square error cannot easy predict the human perception of image fidelity and quality.

Sometimes the quality perceived by human visual system is badly evaluated by the classical MSE-based quality measures. Moreover, for 2D and most of all for 3D video coding it is not completely known the correlation between the calculated MSE and the perception for human visual system. This issue opens the way to a new research branch for the development of coding algorithms in order to improve perceptual coding. It is possible to introduce perceptually-based block coding or quantization in classical video coding architectures, such as in the H.264 architecture.

A possible perception-based technique for MPEG [59] takes into account the different levels of distortion that are tolerable by viewers in different parts of a picture by segmenting the scene into flat, edge, and textured regions and quantizing these regions differently. The visually important areas are represented by Importance Maps. These maps are generated by combining factors known to influence human visual attention and eye movements. Lower quantization is assigned to visually important regions, while areas classified as being of low visual importance are more harshly quantized. Results indicate a subjective improvement in picture quality.

Another possibility is using a method based on a perceptual quality measure called the MND (maximum noticeable distortion) and computes the quantization matrix depending on specific statistics of the picture and the viewing condition as showed in the paper by Chen and Challapali [18].

It is known that human eyes cannot sense any changes below the just noticeable distortion (JND) threshold around a pixel due to their underlying spatial/temporal sensitivity and masking properties. Obviously, any un-noticeable signal difference need not to be coded in the bitstream and reflected in the distortion measure. An appropriate (even imperfect) JND model can significantly help to improve the performance of video coding algorithms.

In the paper by Yang, Ling, Lu, Ong and Yao [91] it is explained a new perceptually-adaptive video coding scheme for hybrid video compression, in order to achieve better perceptual coding quality and operational efficiency. A new estimator for color video is first devised in the image domain. How to efficiently integrate masking effects together is a key issue of JND modeling. Spatial masking factors are integrated with the nonlinear additivity model for masking (NAMM). The JND estimator applies to

all color components and accounts for the compound impact of luminance masking, texture masking and temporal masking. Extensive subjective viewing confirms that it is capable of determining a more accurate visibility threshold that is close to the actual JND bound in human eyes. Secondly, the image-domain JND profile is incorporated into hybrid video encoding via the JND-adaptive motion estimation and residue filtering process. The scheme works with any prevalent video coding standards and various motion estimation strategies. To demonstrate the effectiveness of the proposed scheme, it has been implemented in the MPEG-2 TM5 coder and demonstrated to achieve average improvement of over 18% in motion estimation efficiency, 0.6 dB in average peak signal-to perceptual-noise ratio (PSPNR) and most remarkably, 0.17 dB in the objective coding quality measure (PSNR) on average. Theoretical explanation is presented for the improvement on the objective coding quality measure. With the JND-based motion estimation and residue filtering process, hybrid video encoding can be more efficient and the use of bits is optimized for visual quality.

Recently another metric has been proposed by Wang and Bovik [5] in order to achieve a better correlation between the percieved quality and the rate distortion optimization: the structural similarity index (SSIM). A formulation for the SSIM index can be:

$$\mathrm{SSIM}(x,y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_1}\right) \tag{15}$$

For example, in the paper [82] by Wang, Ma and Gao has been proposed a SSIM based Lagrangian perceptual distortion rate optimization method. This method is then followed by a dynamic adaptive Lagrangian multiplier selection scheme based on the proprieties of the input sequences. Experiments demonstrate that this approach can achieve a better perceptual distortion rate performances and better visual quality compared to classical SAD/SSD based RDO coding.

The same SSIM index is exploited also in the paper [61] by Ou, Huang and Chen in order to provide a perceptually optimized bit allocation for H.264. Here the authors propose a novel rate distortion model to characterize the relationship between rate and structural similarity index. The distortion metric is defined as: $D_{\mathrm{SSIM}}(R) = 1 - \mathrm{SSIM}(R)$

Experiments show that an exponential function can describe the relationship between D_{SSIM} and bit allocation; so it is possible to characterize the relationship between rate and SSIM index this way:

$$D(R) = \alpha e^{-\beta R} \tag{16}$$

where α and β are positive parameters. Optimum bit allocation is then possible by solving the following constrained problem:

$$\min_{r_i} \sum_{i=1}^{N_b} \alpha_i e^{-\beta_i r_i}, \tag{17}$$

$$\sum_{i=1}^{N_p} r_i \leq T_0, \tag{18}$$

$$L_i \leq r_i \leq U_i, i = 1, ..., N_b \tag{19}$$

4.2 Use of saliency maps

In the context of lossy video coding context, the bit budget has to be shared according to a rate vs. distortion tradeoff. An appropriate distribution of the resources improves the overall perceived quality. The idea is simple, it consists in promoting the quality of the regions of interest (ROI) in the videos (namely the regions that are more important visually), this method is also known as selective compression. This implies to have (a priori) informations about the scene to be coded, these ROI can be obtained by modeling the visual attention and by computing saliency maps. There are two ways to achieve a selective compression.

The first one is called indirect because it consists in filtering the video before its encoding in order to reduce the amount of information of the regions with lesser visual interest So the filter smoothes these regions with less interest, but the ROI are unchanged. The choice of the pre-processing is important, because it has to be complementary to the coding, and in particular to the quantization step, a non-linear filtering seems to be more efficient. In his thesis O. Le Meur[51] proposed to use a filter called "leveling"[55], it is the combination of two morphological operators (an opening and a closing by reconstruction), they have the advantage of preserving the spatial structures of the video.

The second one is called direct because it adapts directly the core of the encoder according to the knowing of the ROI. In the case of a classical encoding (block based) the goal is then to control the distribution of the bit budget according to the visual interest of each encoded macro-block, in order to improve the overall visual quality. A. Bradley[6] has shown that selective compression of still images is able to increase the subjective quality when in one hand the ROI are relatively small, and when in the other hand the rate constraint causes artifacts on the salient areas. In most cases, the encoding parameter that we control is the quantization step. Typically a macro-block of low visual interest will be quantized more coarsely than a macro-block visually important. W. Osberg[60] proposed a method to control the quantization in a MPEG coder using saliency maps. It also used a model of spatial masking spatial to redistribute the coding errors. Then the ROI and the areas where the artifacts are easily visible are finely quantized, while visually less important regions and areas capable of spatial masking are coarsely quantized. The method improves the perceived subjective quality of the decoded video compared to a classical coding method. O. Le MeurLemeur2005 also proposed a method of direct selective compression in two steps. First he determines the quantization step of each macro block satisfying a minimal cost, this stage should provide a decoded image with an homogeneous quality. Secondly he redistributes the exceeded bits to the salient areas. So the method improves the PSNR of the ROI. He observed the same results as those given by A. Bradley[6] and L. Huguenel[45] i.e the direct selective compression approach is relevant for ROI with small sizes and when the background has a visual masking capabilities of the quantization noise. O. Brouard[7] proposed a method of pre-analysis of the HD video before its encoding. The pre-analysis module incorporates a visual attention model.

This model aims at analyzing the video by taking into account its high level informations in order to transmit to the encoder an optimal set of parameters and to exploit efficiently the coding tools. He illustrated the method through two coding applications. The first one proposes to adapt the GOP structure according to the spatio-temporal content of the video. The second is a compression scheme with a visual differentiated quality guided by the saliency maps.

5 Multiple description coding

Multiple description (MD) coding involves the transmission of several correlated representations of the source signal over independent channels. In the simplest case, two representations (called *descriptions*) are generated by the MD encoder and sent over (logically) different channels. The MD decoder handles two different situations: in the first one, errors have occurred on one of the channels and the decoder ignores the data coming from it, delivering an approximate version of the orignal signal using the other channel output (side decoding); in the second situation both channels were unaffected by errors and a decoder produces a better version of the original signal (central decoding). The question is now: what should these two different representations of the signal be and how can the reconstructions (central and side reconstructions) be best obtained?

The beginnings of this new coding strategy date as far as 1979 when El Gabel, Gersho, Ozarow formulated the following question: *What are the achievable distortions for a memoryless source at fixed given transmission bitrates when this source is described by several bitstreams [62],[33]?*. This problem was tackled from an information theoretic point of view. In this period the problem was mostly cast in the literature as a source coding technique, probably because of the rate-distortion optimization philosophy. Later on, Goyal, Vaishampayan and others included MDC in the class of joint source-channel coding but consensus has not yet been completely reached among researchers. Nowadays, we are leaning toward the latter classification, since the MDC strategy takes simultaneously into account the possibility of losses in the encoding process.

MDC has known a spectacular regain of interest when researchers such as Vaishampayan, [80], Wang, Orchard, Reibman [83], Kovacevic and Goyal [39] proposed viable methods for error resilience via multiple description coding. These works were motivated by the important advances in multimedia communications.

In the following we will consider different approaches to MDC: MD by quantization, by correlating transforms, by filter banks and by unequal error protection. An excellent survey on MDC issues can be found in the paper by Goyal [38].

5.1 MD by quantization

The first practical approach to Multiple Description Coding is proposed by Vaishampayan, [80]. This technique relies on quantization and the idea is to build two discrete descriptions for a source, each of them belonging to a certain dictionary of symbols.

The imposed criterion is that the resulting quadratic distortion when both channels work correctly is smaller than the individual side distortions. The solution is based on scalar quantization. Two uniform quantizers are involved and the second one is shifted by half a quantization interval with respect to the first one. Thus if one description is lost the source is recovered from a description quantized with the original step size, whereas if both descriptions are received the resulting quantization step is halved.

Besides this scalar quantization approach, vector quantization solutions have been proposed, using lattice vectors. The lattice vector quantizers are quite similar to scalar quantizers, but the source is split into vectors of length L. A first lattice leads to a finely quantized vector which is subdivided into two descriptions, each in a coarser sublattice. Such a representation will be decoded simply by inversing the indexation at the central decoder. As in the scalar case, the index assignment is obtained by solving the rate-distortion optimization of the central distortion under side rate constraints [49].

5.2 MD by correlating transforms

Multiple Description Correlating Transform (MDCT) has been introduced in the literature by Wang, Orchard and Reibman in [83] for two variables. The results have been generalized later on by Goyal and Kovacevic to the case of n variables [39]. The latter introduced in [39] the notion of Multiple Description Transform Coding (MDTC), which is an extension of the Wang, Orchard and Reibman's work. In their method the source vector is first quantized with a uniform scalar quantizer. The obtained vector is then transformed with a discrete invertible transform and the coefficients are independently entropy coded after being grouped into $m \leq n$ subsets to be sent over the m channels.

A second method that generates correlation into the transmitted signal was also proposed by Goyal *et al.* [41]. In this case the signal is expanded by the means of a frame decomposition.

5.3 Filter banks

Another case of MD methods with transform coding, which can be viewed as a particular case of discrete-time frame decomposition is based on filter banks.

The first application of filter banks to multiple descriptions is proposed by Yang and Ramchandran in [92]. Here, the analysis filters are orthonormal. At the reconstruction, the associated synthesis filters are used. The filtered signals obtained at the analysis stage are decimated by a factor of two, quantized and entropy coded for transmission over each of the channels.

The same problem of designing optimal filter banks for MDC has also been approached by Dragotti *et al.* in [27]. The difference between the two approaches resides in the place of the quantizer in the transmission chain. The advantage of this approach is that the quantization cells are not changing shape and the quantization error is not increased by the use of non-orthogonal transforms.

5.4 Unequal Error Protection (UEP)

Researchers propose to transform a scalable source bitstream into an M-description packet stream in which each packet contains approximately the same amount of information. A strategy for prioritized encoding mainly designed for video conferencing-type applications over lossy packet networks is given in [2] and serves as a starting point in this new class of MD methods.

Puri and Ramchandran, [67], combine these considerations with Forward Error Correction in order to add redundancy to a given source. They propose to split the information bitstream into several layers in decreasing order of importance and each of those layers is further protected by progressively weaker channel codes.

An important issue for this MD encoder is the optimal partitioning of the bitstream into layers. A general framework for a variety of transmission scenarios in the packetized media streaming context is proposed in [19]. The combination of MD with layered coding is also reported for correlating transforms MDC in [85].

5.5 Image coding

The first applications of multiple description coding to images are investigated by Wang *et al.* [83], Goyal *et al.* [40], Jiang and Ortega [48], Servetto *et al.* [71] and they are closely related to the general methods we have already enumerated.

The pairwise correlating transform is applied to different blocks of an image, [83], after classifying these blocks into four classes in order to ensure similar statistical properties of the transformed coefficients. The selection is made upon geometrical/image regularity considerations such as smoothness, edge orientation for the first three classes, whereas the fourth is assigned to what is left after this classification. This is due to the fact that real images are not statistically stationary, therefore the correlating transform applied globally could introduce large estimation errors.

The transform coding approach to images of Goyal *et al.* [40] uses the generalized multiple description method proposed before and applies it to a four-channel coding scenario similar to JPEG.

Another technique for MD image transmission was proposed by Jiang and Ortega [48] and it uses the polyphase transform for description generation followed by selective quantization in order to introduce the desired amount of redundancy.

A similar method to Jiang and Ortega's is introduced by Miguel *et al.* in [56]. The evolution of the PSNR with the description number is studied, aiming to prove that the descriptions are balanced.

Methods that also use wavelet transforms when building MD schemes are those by Servetto *et al.* [71], Pereira *et al.* [66], Channappayya *et al.* [16], Tillo *et al.* [78]). Other applications are based on lapped orthogonal transforms in [20].

Matching-pursuit like methods have been proposed by Radulovic and Frossard [68]. They are building multiple descriptions based on redundant dictionaries.

5.6 Video coding

Building MD schemes for the transmission of video sequences offers more degrees of freedom as compared with images since the source has an additional dimension in this case, which is given by the temporal axis (the "frame" direction). Several directions have been investigated for video MDC.

Before going into the details of these practical schemes, we present two very simple strategies that are more or less the founding stones to the further considered directions when it comes to dealing with video sequences. The *temporal splitting* involves, in some sense, reducing the frame-rate of the video source by a factor of two in each description. A second simple technique for building two or more descriptions involves the partitioning of each individual frame in the video sequence, and this has come to be known as *spatial splitting*.

A very good survey on MD for hybrid video coders is provided by Wang, Reibman and Lin in [84]. Video schemes are classified here according to the solution to drift effect they are proposing (or not) and the introduced redundancy. Thus the existing coders involving prediction loops belong to one of the three classes, according to the use of mismatch in side decoders.

One application of MD to video coding is presented by Tang and Zakhor in [75]. Here the structure given by the discrete cosine transform is modified in order to allow the use of matching pursuits and the MD system is built upon a three-loop structure.

Gallant *et al.* have developed a standard compliant MDVC scheme based on spatial oversampling of the video signal by the means of an inverse zero-padded DCT, [32] and a polyphase transform that generates two descriptions. Tillier et al. [77] presented in 2007 a wavelet-based video coder both progressive and MD, using a polyphase redundant decomposition of the video signal. The missing frames at the decoder are recovered by applying different types of linear and non-linear interpolations and a post-processing step is performed in order to eliminate a visual artefact of granularity in the decoded sequence. In these works, H.264/AVC codecs are used.

Starting from the idea that the tradeoff between error resilience and compression efficiency of most existing MDC methods is dependent on the targeted quality, network capabilities as well as the characteristics of the video itself, Heng *et al.* introduced an adaptive multiple description scheme [44]. Different simple MD modes are defined and the system chooses between them based on a rate-distortion optimization. The authors consider the following four modes: single description (SD) coding, temporal splitting (TS), spatial splitting (SS) and repetition coding (RC). Among these modes the most efficient in terms of coding is obviously the SD mode, whereas the most efficient in terms of error resilience is the RC mode.

Aside with techniques which aim to design a MD coder *ex-novo*, Shirani et al. [73] pointed out that an MD coder based only on pre-/post-processing and use of legacy coders reduces significantly the development time, hence the development cost. However, this benefit comes at the price of sub-optimal performance with respect to the from-scratch solutions. The idea of reusing information from the lower fidelity version of a frame in central decoding by means of a convex combination has been originally proposed by Zhu *et al.* [93]; in their work, the lower fidelity frame (*i.e.* the side-decoded

one) was a transmitted B-frame of lower hierarchical level.

Another approach which is not multiple description coding but follows similar guidelines is proposed by Apostolopoulos in [4] for video communication over unreliable networks. In this work multiple states are created at the encoder by temporal splitting. However, there is no explicit redundant coding of video frames, the higher rate resulting when putting together two states encoded individually coming from the fact that the frames are further apart and thus the motion compensation is less effective. The proposed encoding strategy is combined with path diversity which means explicitly sending different packets on different paths. This idea has some important benefits: the burst losses are transformed into individual losses, the outage probability decreases and smaller fluctuations in transmission quality are encountered by averaging the number of paths.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54(11):4311–4322, November 2006.

[2] J. Albanese, A. Blomer, J. Edmonds, M. Luby, and M. Sudan. Priority encoding transmission. *IEEE Transactions on Information Theory*, 42(6):1737–1774, November 1996.

[3] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220, April 1992.

[4] J. Apostolopoulos. Reliable video communication over lossy packet networks using multiple state encoding and path diversity. In *SPIE Visual Communications and Image Processing Conference*, San Jose, CA, USA, January 2001.

[5] A. Bovik, Z. Wang, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[6] A. P. Bradley. Can region of interest coding improve overall perceived image quality ? In *Proceedings of the APRS Workshop on Digital Image Computing (WDIC)*, volume 1, pages 41–44, Février 2003.

[7] O. Brouard. *Pré-analyse de la vidéo pour un codage adapté. Application au codage* TVHD *en flux* H.264. PhD thesis, Ecole Polytechnique de l'Université de Nantes, Juillet 2010.

[8] C. Cafforio and F. Rocca. Methods for measuring small displacements of television images. *IEEE Transactions on Information Theory*, 22(5):573–579, September 1976.

[9] M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva. Cost and advantages of shape adaptive wavelet transform in object-based image coding. *EURASIP Journal of Image and Video Processing*, 2007:1–13, 2007.

[10] M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva. Improved class-based coding of multispectral images with shape-adaptive wavelet transform. *IEEE Geoscience and Remote Sensing Letters*, 4(4):566–570, October 2007.

[11] M. Cagnazzo, G. Poggi, and L. Verdoliva. Region-based transform coding of multispectral images. *IEEE Transactions on Image Processing*, 16(12):2916–2926, December 2007.

[12] M. Cagnazzo, G. Poggi, L. Verdoliva, and A. Zinincola. Region-oriented compression of multispectral images by shape-adaptive wavelet transform and spiht. In *IEEE International Conference on Image Processing*, volume 4, pages 2459–2462, Singapore, October 2004.

[13] M. Cagnazzo, L. Verdoliva, and G. Poggi. Costs and advantages of shape-adaptive wavelet transform for region-based image coding. In *IEEE International Conference on Image Processing*, volume 3, pages 197–200, Genova, Italy, September 2005.

[14] E. Candes and D. Donoho. Curvelets-a surprisingly effective nonadaptive representation for objects with edges. *Curve and Surface Fitting*, January 1999.

[15] C. Chang and B. Girod. Direction-adaptive discrete wavelet transform for image compression. *IEEE Transactions on Image Processing*, 16(5):1289–1302, May 2007.

[16] S. Channappayya, J. Lee, R. Heath Jr., and A. Bovik. Frame-based multiple description coding in the wavelet domain. In *IEEE International Conference on Image Processing*, volume 3, pages 920–923, Genova, Italy, September 2005.

[17] V. Chappelier and C. Guillemot. Oriented wavelet transform for image compression and denoising. *IEEE Transactions on Image Processing*, 15(10):2892–2903, Oct. 2006.

[18] Y. Chen and K. Challapali. Fast computation of perceptually optimal quantization matrices for mpeg-2 intra pictures. In *IEEE International Conference on Image Processing*, volume 3, pages 419–422, Chicago, IL, October 1998.

[19] P. Chou and Z. Miao. Rate-distortion optimized streaming of packetized media. *IEEE Transactions on Multimedia*, 8(2):390–404, April 2006.

[20] D. Chung and Y. Wang. Multiple description image coding using signal decomposition and reconstruction based on lapped orthogonal transforms. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(6):895–908, September 1999.

[21] R. Claypoole, G. Davis, W. Sweldens, and R. Baraniuk. Nonlinear wavelet transforms for image coding via lifting. *IEEE Transactions on Image Processing*, 12(12):1449–1459, December 2003.

[22] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl*, 4(3):245–267, 1998.

[23] R. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[24] W. Ding and F. Wu. Lifting-based wavelet transform with directionally spatial prediction. In *Picture Coding Symposium*, Dec. 2004.

[25] W. Ding, F. Wu, X. Wu, S. Li, and H. Li. Adaptive directional lifting-based wavelet transform for image coding. *IEEE Transactions on Image Processing*, 16(2):416–427, February 2007.

[26] M. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, December 2005.

[27] P. Dragotti, S. Servetto, and M. Vetterli. Optimal filter banks for multiple description coding: analysis and synthesis. *IEEE Transactions on Information Theory*, 48(7):2036–2052, July 2002.

[28] F. Dufaux and F. Moscheni. Motion estimation techniques for digital tv: a review and a new contribution. *Proceedings of the IEEE*, 83(6):858–876, June 1995.

[29] K. Engan, S. O. Aase, and J. H. Husoy. Method of optimal directions for frame design. In IEEE Computer Society, editor, *Proc. IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 2443–2446, 1999.

[30] A. Eskicioglu and P. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, December 1995.

[31] J.C. Feauveau. Analyse multi-résolution avec un facteur de résolution. *IJournal de Traitement du Signal*, 7(2):117–128, Oct. 1990.

[32] M. Gallant, S. Shirani, and F. Kossentini. Standard-compliant multiple description video coding. In *IEEE International Conference on Image Processing*, volume 1, pages 946–949, Thessaloniki, Greece, October 2001.

[33] A. El Gamal and T. Cover. Achievable rates for multiple descriptions. *IEEE Transactions on Information Theory*, 28(6):851–857, November 1982.

[34] O. Gerek and A. Çetin. Adaptive polyphase subband decomposition structures for image compression. *IEEE Transactions on Image Processing*, 9(10):1649–1659, October 2000.

[35] O. Gerek and A. Çetin. A 2d orientation-adaptive prediction filter in lifting structures for image coding. *IEEE Transactions on Image Processing*, 15(1):106–111, January 2006.

[36] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1991.

[37] A. Gouze, M. Antonini, and M. Barlaud. Quincunx lifting scheme for lossy image coding. In *International Conference on Image Processing*, volume 1, pages 665–668, Sept. 2000.

[38] V. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18(5):74–93, September 2001.

[39] V. Goyal and J. Kovacevic. Optimal multiple description transform coding of gaussian vectors. In *Data Compression Conference*, pages 388–397, Snowbird, Utah, USA, March 1998.

[40] V. Goyal, J. Kovacevic, R. Arean, and M. Vetterli. Multiple description transform coding of images. In *IEEE International Conference on Image Processing*, volume 1, pages 674–678, Chicago, IL, October 1998.

[41] V. Goyal, J. Kovacevic, and M. Vetterli. Quantized frame expansions as sourcechannel codes for erasure channels. In *Data Compression Conference*, pages 326–335, Snowbird, Utah, USA, March 1999.

[42] S. Han and C. Podilchuk. Video compression with dense motion fields. *IEEE Transactions on Image Processing*, 10(11):1605–1612, November 2001.

[43] H. Heijmans, B. Pesquet-Popescu, and G. Piella. Building nonredundant adaptive wavelets by update lifting. *Applied and Computational Harmonic Analysis*, 18(3):252–281, May 2005.

[44] B. Heng, J. Apostolopoulos, and J. Lim. End-to-end rate-distortion optimized md mode selection for multiple description video coding. *EURASIP Journal on Applied Signal Processing*, 2006:1–12, 2006.

[45] L. Huguenel. Codage par zones d'intérêt dans la cadre d'un codeur MPEG4 AVC. Technical report, Rapport de Diplôme de Recherche Technologique (réalisé chez Thomson, supervisé par l'IRISA), 2005.

[46] J. Jain and A. Jain. Displacement measurement and its application in interframe image coding. *IEEE Transactions on Communications*, 29(12):1799 – 1808, December 1981.

[47] G. Jeannic. *Représentation structurelle d'images par transformées locales en ondelettes orientées et codage.* PhD thesis, Ecole Polytechnique de l'Université de Nantes, Novembre 2008.

[48] W. Jiang and A. Ortega. Multiple description coding via polyphase transform and selective quantization. In *SPIE Visual Communications and Image Processing Conference*, volume 3653, pages 998–1008, San Jose, CA, USA, January 1999.

[49] J. Kelner, V. Goyal, and J. Kovacevic. Multiple description lattice vector quantization: Variations and extensions. In *Data Compression Conference*, volume 1, pages 480–489, Snowbird, Utah, USA, March 2000.

[50] K. Kreutz-Delgado and B. D. Rao. Focuss-based dictionary learning algorithms. In *Proc. SPIE*, volume 4119 : "Wavelet Applications in Signal and Image Processing VIII, July-August 2000.

[51] O. Le Meur. *Attention sélective en visualisation d'images fixes et animées affichées sur écran : Modèles et évaluation de performances - Applications.* PhD thesis, Ecole Polytechnique de l'Université de Nantes., 2005.

[52] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya. Learning unions of orthonormal bases with thresholded singular value decomposition. In *Proc. IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages v293–v296, 18-23 March 2005.

[53] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Comp.*, 12:337–365, 2000.

[54] S. Li and W. Li. Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(5):725–743, May 2000.

[55] F. Meyer. From connected operators to levelings. In *Proceedings of the fourth international symposium on Mathematical morphology and its applications to image and signal processing.*, pages 191–198. Kluwer Academic., 1998.

[56] A. Miguel, A. Mohr, and E. Riskin. Spiht for generalized multiple description coding. In *IEEE International Conference on Image Processing*, volume 3, pages 842–846, Kobe, Japan, October 1999.

[57] J. L. Murray and K. Kreutz-Delgado. An improved focuss-based learning algorithm for solving sparse linear inverse problems. In *Proc. Asilomar Conf. On Signals, Systems and Computers*, volume 1, pages 347 – 351, 2001.

[58] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[59] W. Osberger, A. Maeder, and N. Bergmann. Perceptually based quantization technique for mpeg encoding. In *SPIE Human vision and electronic imaging III*, volume 3299, pages 148–159, San Jose, CA, January 1998.

[60] W. Osberger, A. J. Maeder, and N. W. Bergmann. A perceptually based quantization technique for MPEG encoding. In *Proceedings of SPIE Conference on Human Vision and Electronic Imaging.*, volume 3299, pages 148–159, Janvier 1998.

[61] T. Ou, Y. Huang, and H. Chen. A perceptual-based approach to bit allocation for h.264 encoder. In *SPIE Visual Communications and Image Processing Conference*, volume 7744, July 2010.

[62] L. Ozarow. On a source-coding problem with two channels and three receivers. *Bell Syst. Tech. J.*, 59(10):1909–1921, December 1980.

[63] S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu. Distortion evaluation in transform domain for adaptive lifting schemes. In *IEEE Workshop on Multimedia Signal Processing*, volume 1, pages 200–205, Cairns, Australia, October 2008.

[64] S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu. Estimation of quantization noise for adaptive-prediction lifting schemes. In *IEEE Workshop on Multimedia Signal Processing*, volume 1, pages 1–6, Rio de Janeiro, Brazil, October 2009.

[65] E. Le Pennec and S. Mallat. Sparse geometric image representation with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, April 2005.

[66] M. Pereira, M. Antonini, and M. Barlaud. Channel adapted scan-based multiple description video coding. In *IEEE International Conference On Multimedia and Expo*, volume 2, pages 609–612, August 2002.

[67] R. Puri and K. Ramchandran. Multiple description source coding through forward error correction codes. In *Asilomar Conf. on Signals, Systems and Computers*, volume 1, pages 342–346, Pacific Grove, CA, October 1999.

[68] I. Radulovic and P. Frossard. Multiple description coding with redundant expansions and application to image communications. *EURASIP Journal of Image and Video Processing*, 2007, 2007.

[69] A. Said and W. Pearlman. A new fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

[70] S. Sardy, A. G. Bruce, and P. Tseng. Block coordinate relaxation methods for nonparametric signal denoising with wavelet dictionaries. *Journal of computational and graphical statistics*, 9:361–379, 2000.

[71] S. Servetto, K. Ramchandran, V. Vaishampayan, and K. Nahrstedt. Multiple description wavelet based image coding. *IEEE Transactions on Image Processing*, 9(5):813–826, May 2000.

[72] O. G. Sezer, O. Harmanci, and O. G. Guleryuz. Sparse orthonormal transforms for image compression. In *Proc. IEEE Int'l Conference on Image Processing (ICIP), San Diego, CA.*, October 2008.

[73] S. Shirani, M. Gallant, and F. Kossentini. Multiple description image coding using pre- and post-processing. In *Interrnational Conference on Information Technology: Coding and Computing*, volume 1, pages 35–39, Las Vegas, NV, April 2001.

[74] G. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15(6):74–90, November 1998.

[75] X. Tang and A. Zakhor. Matching pursuits multiple description coding for wireless video. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(6):566–575, June 2002.

[76] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Transactions on Image Processing*, 9(7):1158–1170, July 2000.

[77] C. Tillier, T. Petrisor, and B. Pesquet-Popescu. A motion-compensated overcomplete temporal decomposition for multiple description scalable video coding. *EURASIP Journal of Image and Video Processing*, 2007(1):1–10, 2007.

[78] T. Tillo, M. Grangetto, and G. Olmo. Multiple description image coding based on lagrangian rate allocation. *IEEE Transactions on Image Processing*, 16(3):673–683, March 2007.

[79] B. Usevitch. Optimal bit allocation for biorthogonal wavelet coding. In *Data Compression Conference*, pages 387–395, Snowbird, Utah, USA, March 1996.

[80] V. Vaishampayan. Design of multiple description scalar quantizers. *IEEE Transactions on Image Processing*, 39(3):821–834, May 1993.

[81] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P. Dragotti. Directionlets: anisotropic multidirectional representation with separable filtering. *IEEE Transactions on Image Processing*, 15(7):1916–1933, July 2006.

[82] S. Wang, S. Ma, and W. Gao. Ssim based perceputal distortion rate optimization coding. In *SPIE Visual Communications and Image Processing Conference*, volume 7744, July 2010.

[83] Y. Wang, M. Orchard, and A. Reibman. Multiple description image coding for noisy channels by pairing transform coefficients. In *IEEE Workshop on Multimedia Signal Processing*, volume 1, pages 419–424, Princeton, New Jersey, USA, June 1997.

[84] Y. Wang, A. Reibman, and S. Lin. Multiple description coding for video delivery. *Proceedings of the IEEE*, 93(1):57–60, January 2005.

[85] Y. Wang, A. Reibman, M. Orchard, and H. Jafarkhani. An improvement to multiple description transform coding. *IEEE Transactions on Signal Processing*, 50(11):2843–2854, November 2002.

[86] Z. Wang and A. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, January 2009.

[87] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan. Rateconstrained coder control and comparison of video coding standards. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):688–703, July 2003.

[88] T. Wiegand and G. Sullivan. The h.264/avc video coding standard Standards in a nutshell. *IEEE Signal Processing Magazine*, 24(2):148–153, March 2007.

[89] T. Wiegand, G. Sullivan, G. Bjøntegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, July 2003.

[90] T. Wiegand, X. Zhang, and B. Girod. Long-term memory motion-compensated prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1):70–84, February 1999.

[91] X. Yang, W. Ling, Z. Lu, E. Ong, and S. Yao. Just noticeable distortion model and its applications in video coding. *Elsevier Signal Processing: Image Communication*, 20(7):662–680, August 2005.

[92] X. Yang and K. Ramchandran. Optimal multiple description subband coding. In *IEEE International Conference on Image Processing*, volume 1, pages 684–658, Chicago, IL, October 1998.

[93] C. Zhu and M. Liu. Multiple description video coding based on hierarchical b pictures. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(4):511–521, April 2009.