RATE/DISTORTION–BASED COMBINATION OF MULTIPLE DCT TRANSFORMS FOR VIDEO CODING

Stéphane Valente

Philips Research France / Video and Communication Group
51, rue Carnot, B.P. 301 — 92156 Suresnes Cedex — France
stephane.valente@philips.com

ABSTRACT

This paper proposes some early results concerning the combination of Discrete Cosine Transforms (DCTs) of multiple sizes in a video encoder. The main difficulty consists in finding a way to obtain a fair competition between several transforms. For that purpose, we derive that the quantization of all transformed coefficients regardless of the DCT size with the same uniform linear quantizer allows to obtain an equivalent level of distortion among all transforms. The selection of the best transform on a particular image area is then made by a rate/distortion framework, which jointly optimizes motion and residual signal coding. Systematic objective and subjective improvements are observed when compared to using a single DCT size.

1. INTRODUCTION

A first attempt to combine DCT transforms on various block sizes can be traced back to the early 90’s, with Qualcomm’s Adaptive Block Size DCT (ABSDCT) aimed at high quality motion image compression for Digital Cinema solutions [1]. Each picture is adaptively divided into 2x2, 4x4, 8x8 and 16x16 blocks. The 2D DCT coefficients of the subblocks are then quantized and entropy coded. The picture divisions are based on block activity and local contrast criterions to favor the image perceived quality, but no clear rate/distortion consideration is involved. Another point worth noting is that no motion compensation is used to avoid motion coding artifacts, therefore reducing the coding efficiency of the overall solution.

In the last few semesters, the H.26L working group has realized significant advances in the field of video compression, by developing new coding tools. Among them, 4x4 and 16x16 DCTs were proposed instead of the commonly used 8x8 DCT, adaptively chosen by rate/distortion measures. The 4x4 DCT is used in conjunction with pixel–prediction algorithms designed for detailed regions, whereas the 16x16 DCT is specialized for large flat areas. Recently, a new ad hoc working group was also created within the H.26L committee to further investigate the gains obtained by adaptively changing the size of the transform with the Adaptive Block Transform (ABT) tool introduced in [2] for motion residual coding, confirming that the debate about the “killer” transform is not over yet.

Combining several transforms within a coder seems to be an interesting idea. However, two fundamental issues need to be solved prior to any attempt to use multiple transforms: the first one is how to set up a fair competition between each transform in a rate/distortion sense. Indeed, in a multiple choice framework, if a particular transform constantly has a prohibitive rate/distortion ratio compared to the other available transforms, it should not be used by the encoder. Similarly, if a transform constantly outperforms the others, a single choice framework will have the same performance for a much smaller complexity. The second issue is how to make the right decisions when several coding possibilities are competing.

Our contribution proposes to answer the above general questions in the case of DCT transforms of different sizes. In section 2, we provide an answer for the fair competition which is conceptually simple: the theoretical normalization of the DCT basis functions is such that all coefficients can be quantized by the same uniform quantizer while producing the same amount of coding noise. In section 3, we introduce a rate/distortion (R/D) framework that jointly optimizes coding mode decisions, motion vectors and transform sizes in a video encoder, as opposed to the separate optimization of a motion field followed by the coding of a residual picture. Finally, in section 4, we report some numerical results that demonstrate the coding gains over a single–transform coder.

2. THEORETICAL ASPECTS OF DCT TRANSFORMS

2.1. Quantization Noise of DCT Transformed Blocks

We derive the theoretical results for the one–dimensional length–N DCT transform. The 2D DCT being simply the separable extension of the 1D DCT, the results can easily be extended to the 2D case. The 1D length–N DCT can be represented by the matrix $H = (H[k, n])_{0 \leq k, n < N}$ with

$$H[k, n] = \Lambda(k) \cos \left( \frac{(2n + 1) k \pi}{2N} \right)$$

where $\Lambda(k) = \sqrt{\frac{2}{N}}$ for $k = 0$ and $\Lambda(k) = \sqrt{\frac{2}{N}}$ otherwise.

Let $z = (z_n)_{0 \leq n < N}$ be the length–N column vector of a signal to be transformed. The DCT of $z$ is then $y = H z$, and the inverse DCT of $y$ is $z = H^T y$. Let $\tilde{y}$ denote the quantized DCT coefficients. The quantization error in the DCT domain is the vector $\tilde{y} - y$, and $\hat{z} - z$ with $\hat{z} = H^T \tilde{y}$ in the pixel domain.

By considering the signal samples and DCT coefficients as random variables, we compute the distortion $D$ in the pixel domain as the mean squared error

$$D = E \left( \frac{1}{N} \sum_{i=0}^{N-1} (\hat{z}_i - z_i)^2 \right) = \frac{1}{N} E \left( \| \hat{z} - z \|^2 \right)$$
where $E(.)$ denotes the mathematical expectation. Because $H$ is a unitary transform, we also have

$$D = \frac{1}{N} E \left( ||\hat{y} - y||^2 \right) = \frac{1}{N} \sum_{i=0}^{N-1} E \left( (\hat{y}_i - y_i)^2 \right). \quad (3)$$

Now let us quantize all DCT coefficients $y_i$ by a uniform scalar quantizer of step $q$. If we assume that the quantization error $\hat{y}_i - y_i$ of each coefficient has a uniform distribution, a well–known result recalled in [3] relates the quantization noise to the step size

$$E \left( (\hat{y}_i - y_i)^2 \right) = \frac{q^2}{12}. \quad (4)$$

From (3), this result extends to the distortion of DCT transformed signals

$$D = \frac{1}{N} \sum_{i=0}^{N-1} \frac{q^2}{12} = \frac{q^2}{12}. \quad (5)$$

In other words, the mean distortion by sample does not depend on the DCT size, but only on the quantization step $q$.

As a conclusion, as long as we keep the same uniform scalar quantizer for all coefficients across all DCT sizes, no transform will be better than the other from the distortion point of view.

### 2.2. Coefficient Dynamics

Even if the distortion is kept constant across all transforms, the coding cost of their coefficients differs, partly because of their different dynamics.

Let us explicit this with a simple example, with an 8 point constant signal $z = (1, \ldots, 1)^T$. Applying expression (1), the two length–4 DCTs of $z$ are each $(\sqrt{2}, 0, \ldots, 0)^T$, and the length–8 DCT is $(\sqrt{\frac{2}{2}}, 0, \ldots, 0)^T$. We can see that the first coefficient of the length–8 DCT is $\sqrt{2}$ times larger than its length–4 counterparts, which are likely to be encoded in lower indexes by the uniform quantizer. The larger the DCT size, the larger the coefficient dynamics and quantized values: this is the price for keeping the distortion comparable across all transform sizes.

In general, we have no answer to say which transform size is better to encode a given signal. Larger transforms will probably have less non–zero coefficients than smaller transforms (definitely less DC coefficients), but will have wider statistical distributions, which are costlier for an entropy coder. The answer depends on the signal to be encoded, and requires a rate/distortion framework to find the optimal trade–off among several coding possibilities.

### 3. RATE/DISTORTION FRAMEWORK

#### 3.1. Principle

Since our coding scheme uses block transforms of different sizes, we use a quadtree (QT) structure to encode the image. This structure corresponds to a tree, where each block (called a node) can be further divided into four subblocks. A node can be a leaf (i.e. it is not further divided) or a branch (it is divided into smaller nodes), and multiple coding decisions can be associated to a node. In addition to offer an efficient representation of the partition of the image into blocks, the QT structure can be optimized by several algorithms readily usable to find optimal rate/distortion decompositions [4, 5, 6].

Quadtree are typically constructed by top–down or bottom–up methods [4]. The top–down method (used in [1]) consists in starting from an entire block, and deciding whether it must be represented by a single leaf or divided into four subblocks, and so on. The bottom–up method starts from the smallest possible blocks in the QT, decides whether to merge them into larger blocks, and recursively proceeds to the largest possible blocks. To minimize rate as well as distortion, the split or merge decisions are made by locally minimizing a Lagrangian cost function in the form $d + \lambda r$, where $d$, $r$ and $\lambda$ are respectively the distortion, the rate and the Lagrangian multiplier.

Other optimization algorithms exist, and our implementation is largely inspired from [6]. The QT structure is sequentially built by a Viterbi algorithm parsing all possible transitions between a node at a given QT level and its spatial neighbors (possibly at different QT levels). The optimal QT structure corresponds to the best path arriving at the last QT node minimizing the Lagrangian cost function on the whole image. Compared to top–down or bottom–up methods, the advantage of this approach is that the optimal QT decomposition can be found even when there are coding dependencies between the leaves. This is often the case in practical video coding systems, for instance when motion vectors are differentially encoded (i.e. the coding cost of a particular motion vector depends on the last coded one, that is to say from the former QT nodes in the QT structure). This is also true in future systems like H.26L, where arithmetic coders constantly update statistics tables in function of past encoded symbols.

#### 3.2. Implementation Details

The encoding stage proceeds in two passes. The first pass parses the QT trellis to find the optimal path. At each node, only the best incoming path is kept, along with its Lagrangian cost $d_k + \lambda r_k$. All transitions to the neighboring nodes, which may account for different block sizes and different coding modes, are evaluated and the corresponding Lagrangian cost is updated, thus propagating the best path. This cost is computed by using the true coding distortion and the true coding cost obtained by simulating the binarization of output symbols.

When the termination node of the QT is reached, the second pass back–tracks the best path, producing the QT segmentation, the sequence of coding mode decisions and the actual symbol bits.

Our current implementation defines two types of pictures, INTRA and INTER. INTRA pictures are encoded using DCT transforms of multiple sizes. No advanced INTRA prediction mode is implemented, only DC coefficients are predicted from the previously coded blocks. AC coefficients are encoded by non–zero coefficients, runs of zero coefficients and end–of–block symbols. INTER pictures are also partitioned into blocks of different sizes, but three coding modes are available to encode each block:

- **INTRA block**: the block is encoded by a DCT transform like in INTRA pictures;
- **MV block**: a motion vector for this block is encoded in the bitstream. The block is then reconstructed by motion–compensating the corresponding block in the last decoded picture. The motion vectors have pixel–resolution, and the maximum displacement value is 16 pixels. It only exploits the temporal redundancy of coded pictures;
**MV+T block:** a motion vector and a residual texture are encoded by a DCT transform for this block, exploiting the temporal and spatial redundancies of coded pictures.

Entropy coding is performed by arithmetic coders using adaptive first-order statistics models. Five contextual models are defined, depending on the type of data they handle (QT segmentation bits, coding mode decisions, motion vectors, DC coefficients, non-zero AC DCT coefficients and runs of zero AC DCT coefficients).

As far as the rate control is concerned in the next experiments, we use the DCT quantizer steps proposed by H.26L (adapted to take into account the size-dependent normalization factors of the DCT transforms), along with the recommended quantizer-dependent Lagrangian multipliers.

The major advance of this framework is that it jointly optimizes motion and texture encoding in finding the best coding possibility among all available coding decisions. Usually, a classical video coder would do a preliminary motion field search minimizing the distortion of the motion-compensated signal. Then, it would decide the block coding modes by choosing the INTRA mode, the MV mode or the MV+T mode. The motion field and texture encoding would therefore be separately optimized. On the contrary, our best path search tries several motion vectors for each possible block, and decides the best coding mode for a given motion vector. Comparing the different possibilities at each node transition, the Viterbi algorithm simultaneously optimizes the block sizes, the motion vectors and the coding modes over the whole image.

### 4. EXPERIMENTAL RESULTS

#### 4.1. INTRA Coding

For INTRA pictures, the framework simply optimizes the size of the DCT transforms, since there is a single coding mode, as shown in figure 1. We compared our multiple-transform coder, using DCTs of size 4x4, 8x8 and 16x16, with another coder using only the 4x4 DCT on INTRA images. Both coders were run over standard luminance sequences, comprising CONTAINER, FOREMAN, MOBILE, MOTHER and TABLE, with a fixed quantization parameter. We took the values 12, 16, 20 and 24 (translated into H.26L quantization steps) to obtain final PSNRs between 30 and 40 dB.

**Fig. 1.** First image of the TABLE sequence using 4x4 to 16x16 DCT transforms: intra-coded picture (left) and transform sizes (right).

PSNR improvements between 0.5 and 1.5 dB are systematically obtained over the single 4x4 transform coder on all sequences.

Figure 2 shows the R/D plot averaged over 50 images for the TABLE sequence.

**Fig. 2.** R/D plot for the TABLE sequence in CIF format at 30 Hz (50 images in INTRA mode): 4x4 only versus 4x4 to 16x16 DCT transforms.

#### 4.2. INTER Coding

To assess the added-value of multiple DCT transforms in INTER pictures (which are predominant in coded video sequences), we compare two video coders exploiting blocks from 4x4 to 16x16 pixels to perform the motion-compensation. The first one uses a single DCT size of 4x4 pixels, regardless of the size of the INTRA or motion-compensated coded block. Thus, only the motion compensation module of the coder takes advantage of the various block sizes, as in typical video coding systems. The mode coding decisions and block sizes are still optimized by the R/D framework.

The second coder ties the size of the DCT transform to the size of the coded block. For instance, if a 16x16 block is coded in the MV+T mode, the residual signal will undergo a 16x16 DCT transform, instead of 16 4x4 DCTs. This seems to be a smarter choice because it does not potentially add blocking artifacts at the transform boundaries within a motion-compensated block.

Figure 3 shows the R/D plot for the TABLE sequence in CIF format at 30 Hz averaged over one INTRA and 299 INTER pictures, for the quantization parameters 12, 16, 20 and 24. The PSNR improvement is around 0.5 dB due to highly textured regions. On the other test sequences, the PSNR increase for the multiple-transform coder always exists, typically about 0.25 dB.

Although the coding gain seems limited, the picture quality is subjectively improved, according to human viewers. Figure 4 shows the 170th picture of the TABLE sequence coded in the INTER mode with the quantization parameter set to 20. The improvement is also clearly visible in figure 5, where the table tennis net is better reproduced and the flat areas are smoother (please see the PDF file of the paper for better judgment).

### 5. CONCLUSION

This paper has provided some early results concerning the combination of multiple DCT transforms to encode video pictures in INTRA and INTER modes. The main difficulty consists in making
Fig. 3. R/D plot for the TABLE sequence in CIF format at 30 Hz (300 images in INTER mode): 4x4 only versus 4x4 to 16x16 DCT transforms.

Fig. 4. Inter-coded image 170 of the TABLE sequences at 372 kbit/s in CIF resolution: with 4x4 DCT only (top) and with 4x4 to 16x16 DCT transforms (bottom).

Fig. 5. Magnified crop of inter-coded image 170: with 4x4 DCT only (top left), with 4x4 to 16x16 DCT transforms (top right) and original image (bottom).

sure to obtain a fair competition between several transforms. For that purpose, we showed that the quantization of all transformed coefficients regardless of the transform size with the same uniform linear quantizer allows to obtain an equivalent level of distortion among all transforms. Then, the selection of the best transform on a particular image area is made by a rate/distortion framework, which jointly optimizes motion and residual signal coding, going further than the sequential optimization of both entities. Systematic objective and subjective improvements are obtained in comparison with using a single DCT size.

6. REFERENCES