Improved video coding efficiency exploiting tree-based pixelwise coding dependencies

Giuseppe Valenzise[a] and Antonio Ortega[b]

[a]Dipartimento di Elettronica e Informazione, Politecnico di Milano;
[b]Signal and Image Processing Institute, University of Southern California

ABSTRACT

In a conventional hybrid video coding scheme, the choice of encoding parameters (motion vectors, quantization parameters, etc.) is carried out by optimizing frame by frame the output distortion for a given rate budget. While it is well known that motion estimation naturally induces a chain of dependencies among pixels, this is usually not explicitly exploited in the coding process in order to improve overall coding efficiency. Specifically, when considering a group of pictures with an IPPP... structure, each pixel of the first frame can be thought of as the root of a tree whose children are the pixels of the subsequent frames predicted by it. In this work, we demonstrate the advantages of such a representation by showing that, in some situations, the best motion vector is not the one that minimizes the energy of the prediction residual, but the one that produces a better tree structure, e.g., one that can be globally more favorable from a rate-distortion perspective. In this new structure, pixel with a larger descendance are allocated extra rate to produce higher quality predictors. As a proof of concept, we verify this assertion by assigning the quantization parameter in a video sequence in such a way that pixels with a larger number of descendants are coded with a higher quality. In this way we are able to improve RD performance by nearly 1 dB. Our preliminary results suggest that a deeper understanding of the temporal dependencies can potentially lead to substantial gains in coding performance.

Keywords: Tree-based representation, Rate-distortion optimization, QP selection, motion vectors

1. INTRODUCTION

Motion is an intrinsic characteristic of video that induces a series of temporal dependencies between frames. Hybrid video codecs, such as H.264/AVC,[1,2] recognize this fact by employing motion-compensated prediction (MCP) in order to exploit temporal redundancies and reduce significantly the bit rate. The encoding of a frame generally entails a number of decisions, including identifying a good predictor in the previous frame and coding it by means of a motion vector (MV), as well as selecting the coding mode and the quantization step of each macroblock. There has been a substantial amount of work on selecting optimal parameters in video coding, with a majority of the proposed techniques making use of Lagrangian optimization[3,4] in order to find optimal tradeoffs between rate and distortion. Many conventional coding schemes tend to make these choices *frame by frame*. However, it is well known that what is considered optimal from a rate-distortion point of view for a frame, may be potentially suboptimal in view of coding the remaining frames of the sequence, due to the chains of predictions created by motion compensation. There is substantial evidence[5–7] that an optimization that involves a temporal horizon longer than one frame may lead to considerable bit savings.

In the past years, this observation has driven a substantial research effort towards defining strategies to enlarge the optimization scope from single frames to groups of frames. We can basically classify these approaches in three categories: methods that include long-term memory in the motion prediction step; methods that code different portions of the frame separately according to their semantics and, thus, temporal evolution; and finally, techniques that optimize bit allocation at different coding units considering the temporal relations between them. The first kind of approaches includes the long-term memory motion-compensated prediction by Wiegand et al.[8] and the dual frame motion compensation schemes.[9] The idea of these papers is to include, in the search space

Further author information:
G. Valenzise: E-mail: valenzise@elet.polimi.it,
A. Ortega: E-mail: ortega@sipi.usc.edu

of motion vectors, long-term reference frames (i.e., a MV can point to a frame far away in time). In the second class of techniques fall all the methods that segment video frames into objects and code them separately. An example of that is MPEG4 Visual Objects.[10,11] These methods do not aim directly at exploiting the temporal dependencies between frames, but in fact can leverage the different behaviors along time of the different objects and code them in a different way. For instance, many approaches extract the background, which typically does not vary significantly over time, and code it differently from the foreground,[11,12] possibly using it as long-term predictor.[13]

The third category of methods, finally, includes methods for dependent coding, and is in fact the most relevant for our work. Dependent coding techniques typically address temporal dependencies in a more theoretical way, treating them as a resource allocation problem. Early work in this direction dates back to Uz, Shapiro and Czigler,[14] who extended the classical DPCM coding theory[15] to analyze the effect of quantizer feedback. The basic result shown there is that frames that are good predictors for subsequent pictures are to be allocated a higher bit rate than normal, and frames which are well predicted are not to be allocated (as an independent optimization approach would suggest) a lower bit budget, because they suffer from error quantization feedback of previous frames. Since we use the results of Uz, Shapiro and Czigler[14] to establish some of the theoretical foundations in our work, we will introduce this technique in more detail in Section 2. The work of Uz et al.[14] is theoretical in nature, and does not provide any explicit algorithm nor experimental results to match with. Ramchandran et al.[6] provided similar insights within an operational rate-distortion framework, and proposed to optimally allocate the bit rate among dependent quantizers by finding the mimimum cost path in a trellis which links the possible quantizer states of the coding units (frames) at each time instant. This method requires to evaluate a set of operational RD points for each frame, which may be prohibitive in many applications; in addition, the choice of quantization parameters is conducted at the frame level, i.e. smaller coded units such as macroblocks are not considered. Chellappa et al.[5] show that the insertion of periodical high-quality long-term reference frames can outperform a normal dual frame motion compensation scheme in which the long-term reference frames are not allocated any extra rate. Even if high-quality frames may be deemed to be a waste of rate when looked at locally, their usefulness as predictors pays off in later frames. Recently, papers by Wiegand and co-authors[7,16] have proposed to jointly optimize transform coefficients and motion vectors by modeling the decoding process as a linear system and solving a quadratic problem. At the expense of a high computational demand, this approach is able to show improvements of nearly 1 dB in rate-distortion curves. The authors observed that in some cases reducing the rate used for frames close to the "root" of the prediction tree (e.g., initial P frames in a group of pictures) may in fact lead to better overall RD performance. However, no clear explanation for this phenomenon is provided, and this fact limits the possibilities to find simple (and faster) heuristics that enable to achieve near-optimal points at less computational costs. This paper can be seen as a first effort to optimize prediction structure as previously proposed,[7,16] but under a more constrained scenario that can make it practical without a very high computational cost. For various reasons, in particular concerns about delay and complexity, techniques that exploit the temporal dimension of video coding (e.g., by introducing temporal masking, structuring mode decisions over time or considering the impact of temporal dependencies) have only received somewhat limited attention.[17]

In this paper we show that taking into account temporal dependencies in a video sequence may lead to a gain in RD performance even using a simple and easy to implement heuristic approach. In order to do so, we propose to model a video sequence as an ensemble of pixel trees, which represent the temporal dependencies induced by motion prediction. Specifically, when considering an IPPP... group of pictures (GOP), each pixel of the first frame can be thought as the root of a tree whose children are the pixels predicted by it in the subsequent frames. In addition to its intuitiveness, this representation is also relatively simple to construct, as it just requires a double pass encoding on each GOP and, as that, has a delay that depends on the adopted GOP size. Furthermore, while all the techniques described above use coding units at a coarse (frame) to medium (macroblock) granularity, the proposed method is conceived to operate, by construction, at the pixel level. The effectiveness of fine granularity distortion estimation methods such as ROPE[18] gives clues that working at the pixel level may be beneficial in terms of performance. To the best of our knowledge, we are the first to consider pixel-wise coding dependencies in a such a tree structured form.

In the rest of the paper we describe this tree-based representation and give preliminary results of its use to set quantization parameters (QP) for each macroblock of a frame. In Section 2 we detail the tree construction

procedure, and we show that the theoretical results of Uz et al.[14] may help to characterize the tradeoff existing between the topology of the trees and bit allocation. In Section 3 we construct a simple, yet effective, two-pass encoding algorithm, where the temporal extent of the dependencies induced by a pixel is heuristically mapped to a corresponding reduction in its quantization parameter. As a proof of concept, this simple algorithm is evaluated on H.264/AVC encoded video sequences: it turns out that we can obtain about 1 dB improvement in RD performance. We want to point out that the proposed scheme is *flexible*, in that it naturally fits to a set of optimization tasks. It is our aim to use this structure to perform mode decision and motion estimation in a future work.

2. TREE-BASED REPRESENTATION

Motion vectors induce a chain of predictions between frames of a video sequence. These predictions can be represented by means of trees. Let us consider an IPPP... GOP: each pixel in a P frame is nothing but a node of a tree having its root in a pixel of the I frame. To formalize better this concept, let $x_j(t)$ denote the jth pixel of the frame at time t, $1 \leq t \leq F$, where F is the number of frames in the GOP. Without loss of generality we vectorize each $m \times n$ frame t in such a way that index j takes values on $1 \dots N$, where $N = mn$. A schematic illustration of this construction is given in Figure 1(a). To build such a representation, motion vectors for the GOP have to be available. We assume that these vectors are gathered and stored by running the video encoder a first time on the GOP.

We use this tree-based modeling in order to evaluate the impact of each pixel on future frames. We start by recalling some theoretical results provided by Uz et al.[14] Suppose that the operational rate-distortion function of a pixel of the *first* (I) frame has the exponential form

$$D_j(1) = e^{-\alpha R_j(1)} X_j(1) \tag{1}$$

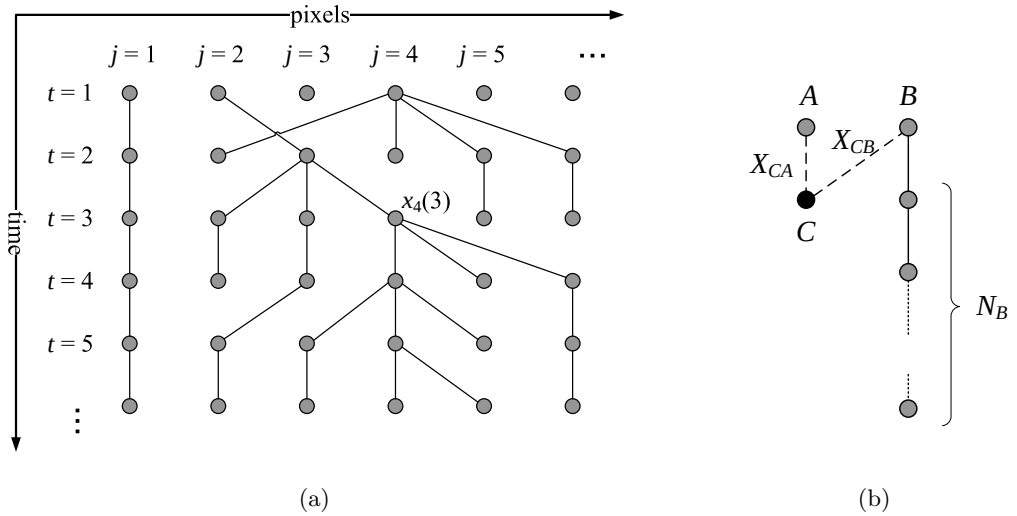
where $D_j(1) = E[(\tilde{x}_j(1) - x_j(1))^2]$ is the quantization distortion between the original pixel j and its quantized version $\tilde{x}_j(1)$, R is the rate and X is a complexity parameter (which basically quantifies how "difficult" it is to code a given source). Pixels of the I frame ($t = 1$) are coded without any kind of temporal prediction; instead, pixels $x_j(t)$ of frames $t \geq 2$ are predicted by some coded pixel $\tilde{x}_k(t-1)$ of the previous frame. If $x_j(t)$ were predicted by the unquantized reference $x_k(t-1)$, the rate-distortion function would depend only on the complexity of the *innovation*, i.e. the difference $x_j(t) - x_k(t-1)$, according to the same kind of exponential law of (1) (in this case, X would be the complexity of the residual). However, since the predictor uses *noisy* samples, a quantizer feedback term has to be added to the rate-distortion function for pixels in the P frames:

$$D_j(t) = e^{-\alpha R_j(t)} (X_j(t) + \rho_j(t) D_k(t-1)), \tag{2}$$

for $t > 1$. Here $X_j(t)$ is the complexity of the innovation, i.e. the complexity of prediction residual $x_j(t) - x_k(t-1)$ when the predictor $x_k(t-1)$ is *not* quantized. Note that (2) decouples the effect on distortion due the complexity *per se* of predicting pixel j from a pixel of the previous frame, from the additional variance produced by the fact that the predictor has been quantized. The propagation of quantization noise is governed by the quantizer feedback ρ, which depends on the kind of prediction adopted. For pixels in video data, the predictor corresponds just to another pixel in the previous frame. Thus it is reasonable to assume $\rho_j(t) = 1$ for $t > 1$, as all the quantization error variance is transferred from the predictor to the predicted pixel. Of course, pixels in the I frames do not have a temporal predictor, so we may see their RD function as (2) with $\rho_j(t) = 0$. Uz et al[14] show that, at optimality, the rates for a chain of dependent pixels $x_j(1) \to x_k(2) \dots \to x_l(t) \dots$ are as follows:

$$R_j(t) = \frac{1}{\alpha} \log X_j(t) + \frac{1}{\alpha} \log \left[\left(\frac{1}{4} + \frac{\lambda \rho_j(t)}{X_j(t)} \right)^{\frac{1}{2}} + \frac{1}{2} \right] + \frac{1}{\alpha} \log \left[\left(\frac{1}{4} + \frac{\lambda \rho_k(t+1)}{X_k(t+1)} \right)^{\frac{1}{2}} + \frac{1}{2} \right] - \frac{1}{\alpha} \log \lambda \tag{3}$$

where λ is a Lagrange multiplier that can be found as described in Uz's paper. In the equation above, rate is assigned according to the importance of a pixel as a predictor (measured through the $X_j(t)$ terms). Equation (3) can be adapted to our proposed tree structure discussed above, with a main caveat. Differently from the case of a linear chain of dependencies, in fact, a tree is endowed with branches; this observation can be incorporated



(a) (b)

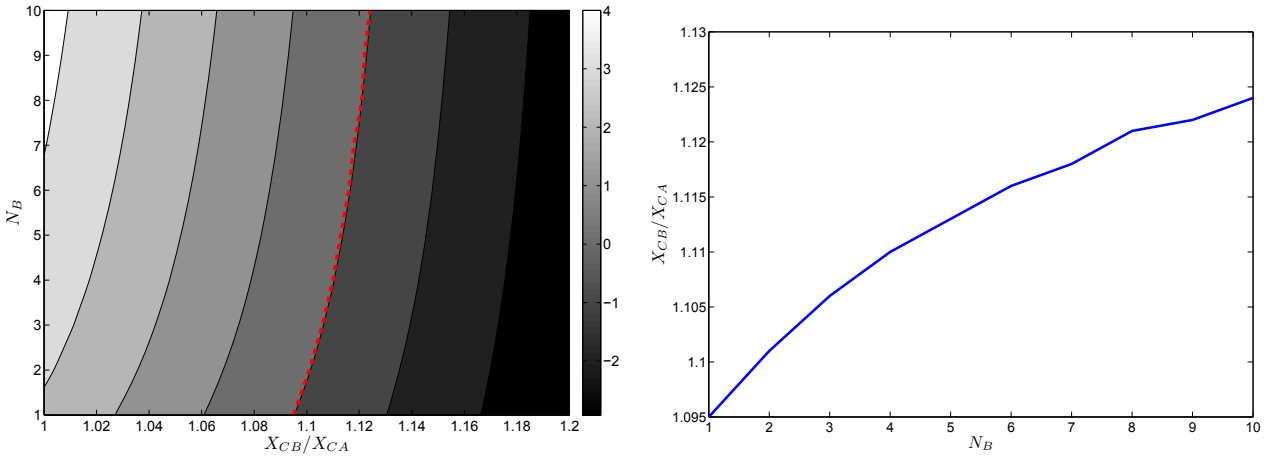
Figure 1. Tree-based representation of video. (a) A schematic representation of the trees of pixels induced by motion-compensated prediction. (b) An example of tradeoff induced by dependent coding. The best predictor for pixel C is A, which does not have any other children. Also, pixel C may be predicted by B, which would entail a higher complexity of the prediction $X_{CB} > X_{CA}$. However, an optimal bit allocation considering the N_B descendants of B would assign a higher bitrate to B, thus reducing its contribution to quantization error. In some cases, C may benefit from this fact and elect B as its parent.

in the formulation above by considering that more "branching" in the tree implies a larger impact of distortion on future pixels. Thus, (3) becomes:

$$R_j(t) = \frac{1}{\alpha}\log X_j(t) + \frac{1}{\alpha}\log\left[\left(\frac{1}{4} + \frac{\lambda\rho_j(t)}{X_j(t)}\right)^{\frac{1}{2}} + \frac{1}{2}\right] + \frac{1}{\alpha}\log\left[\left(\frac{1}{4} + \sum_{k\in\mathcal{C}(j,t)}\frac{\lambda\rho_k(t+1)}{X_k(t+1)}\right)^{\frac{1}{2}} + \frac{1}{2}\right] - \frac{1}{\alpha}\log\lambda, \quad (4)$$

where $\mathcal{C}(j,t)$ are the children of pixel j in the next frame $t+1$. For example, in Figure 1(a) the tree rooted at $x_1(1)$ has only one child in the second frame, while the tree from $x_4(1)$ has four of them. The presence of several "parallel" trees in our structure is instead a minor concern, as this is equivalent to carrying out independent bit allocation across the trees.

We use this general model to show that the temporal impact of a pixel (measured, for example, as the depth of the tree rooted in it), plays a key role in determining rate-distortion tradeoffs between local greedy decisions and global optima. In order to do so, we consider a simplified situation (see Figure 1(b)), where a node C has two candidate parents in the previous frame: A, which happens to be the best predictor for C (i.e. the one that has minimum complexity X_{CA}); and B which is not the best predictor but which has a larger number of descendants than A. For simplicity, we assume that A has no other children. This situation may occur, for example, in motion estimation, whenever pixel C is searching for the best motion vector among the two candidates A and B. Conventional approaches would select A, since it minimizes the complexity of the innovation, i.e. a measure of the residual prediction energy (such as MAD, SSD or a Lagrangian cost involving MAD or SSD). However, due to (2), the quantization error of the first pixel in the tree propagates down the tree. Therefore it is reasonable that, in the situation of Figure 1(b), pixel B should be coded with a higher quality when the number of its descendants is larger, in order to reduce the impact of quantization errors in future pixels. Thus, if B is a sufficiently good predictor for C, changing the parent of C from A to B may lead to a global gain in the rate-distortion sense, *even though A may be a better predictor*. To illustrate that, we simulated the scenario of Figure 1(b) using the exponential rate-distortion functions (1)-(2) suggested by Uz et al. with $\alpha = 1$, and varying the number of children of B and the ratio X_{CB}/X_{CA}, for a total rate constraint of 1 bit per pixel. In the simulation X_A and X_B were set to 100, X_{CA} and all the complexities between the descendants of B were put to 70. Without loss of generality, we suppose that the structure of the tree of B is linear (i.e. without branches), so that the only



(a) (b)

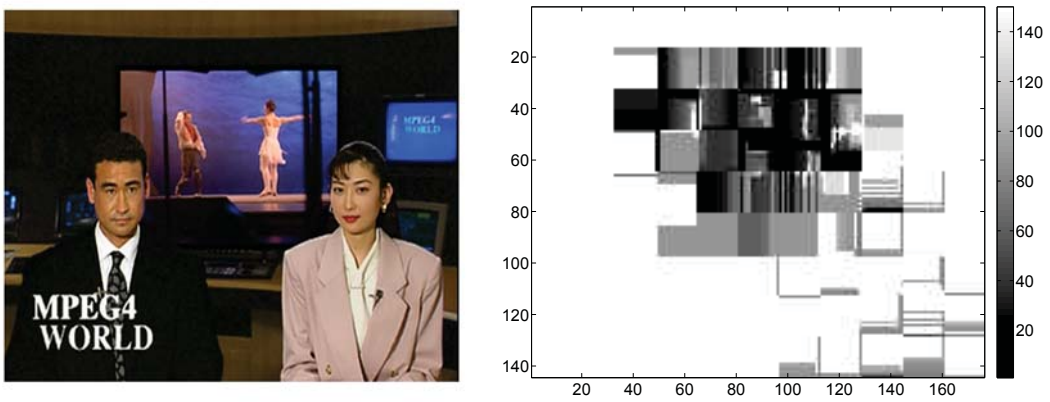
Figure 2. Tradeoff between depth of the tree springing from B and the worsening of complexity incurred by switching from the best predictor to a suboptimal one in Figure 1(b). $X_A = X_B = 100$, $X_{CA} = 70$ (a) The surface of D_{gain}; highlighted is the frontier that separates positive and negative gains. For positive gains it is more convenient to make pixel C depend from B, even if the complexity of the innovation increases. (b) the frontier plotted as a function of N_b.

possible branch is when C is connected to B. We computed the total distortion for the $N_B + 3$ nodes in two different cases: 1) we set the parent of pixel C to be A: this corresponds to performing independent allocation across the two trees of Figure 1(b); 2) we set B as the parent of pixel C, so that now the total number of children of B is $N_B + 1$. We computed the rates for each node according to (4). The total distortions for both cases (respectively, D_1 and D_2) are computed applying recursively (2). We define the distortion gain $D_{\text{gain}} = D_1 - D_2$ as the difference of the two distortions. This is plotted in Figure 2. It can be observed that there is a gain for some combinations of X_{CB}/X_{CA} and N_B, specifically, when B has a larger number of descendant, it is allocated a higher bit budget, thus its quality improves. Therefore, when the increase in the complexity of the innovation due to passing from predictor A to B is relatively small, the reduction in the quantization error of B is somewhat able to counterbalance the RD loss due to a worse predictor. With the specific parameters used in this simulation, we are able balance the up to about a 12% worse predictor having 8 or more descendants (see Figure 2(b)). While this is a small example based on model of actual dependencies, these are encouraging results, indicative that there is room for a potentially substantial gain deriving from considering temporal dependencies. In the next section we make this intuition more concrete using as a proof of concept H.264/AVC coded video sequences.

3. QP HEURISTIC ASSIGNMENT

The model proposed in the previous section has two strong implications. On the one hand, it suggests that, in some cases, the best motion vector is not the one that minimizes the energy of the prediction residual, but the one that produces a better tree structure which can be globally more favorable from a rate-distortion perspective. At the same time, it entails that a higher bit rate, and thus better quality, should be given to pixels which have a stronger impact on the future, i.e. pixels with a larger number of descendants.

In this section we provide a preliminary proof of concept of this second fact with a H.264/AVC video encoder, while experiments on the first part are deferred to future work. In order to give higher quality to pixels which have a larger descendance, we first build motion trees as explained in Section 2, starting from the motion vectors obtained by coding the group of pictures a first time, with a fixed quantization parameter (QP). Then, for each frame t we compute the temporal extent of the trees that start from that frame on. For example, supposing that the GOP size F is 6 pictures in Figure 1(a), the depth of the tree spreading from pixel 1 of frame 1 is $\text{depth}(x_1(1)) = 6$; similarly we can compute $\text{depth}(x_4(1)) = 3$ and $\text{depth}(x_4(3)) = 4$. In this way we obtain a pixel-level map of the depths of each tree: an example of such a map is illustrated in Figure 3(b) for the first frame of the *News* video sequence.



(a) (b)

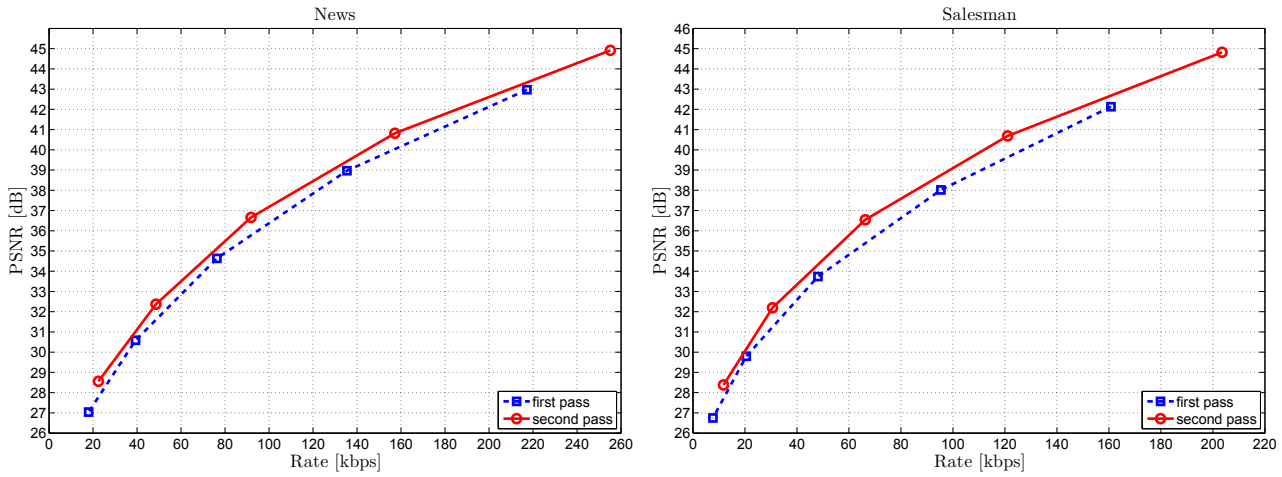
Figure 3. Depth of pixel trees for the *News* video sequence. (b) First frame of the *News* video sequence. (c) A pixel level map of the trees' depth. Bright areas correspond to a larger depth in the trees.

It can be observed that, for this sequence, the static portions of the frame are the ones with a larger number of descendants, as they are the ones which are easily predicted in successive frames. In general, for the case of video, predictable motion corresponds to *physical* motion, and it is observed in many cases that pixels with many branches do not necessarily lead to many good predictions. In other words, this means that often the complexity of the innovation X_j for these pixels is relatively high, and therefore in (2) it would be the dominant factor (since the effect of quantization feedback would be relatively less important, reducing quantization error would not be very beneficial). In addition, we have observed that, often, trees with high "branching" do not have a large temporal extent. So, we opted for a using depth of the trees as importance weight: the greater the depth of the trees, the more important the pixel. In a practical video coding scenario, the smallest coding unit is represented by macroblocks (MB): therefore, we pool the pixels' weights at the MB level by simply averaging them. In a second coding pass, these macroblock-level weights are used to reduce, accordingly, the QP of each block from a predefined $\mathrm{QP}_{\mathrm{target}}$ as follows. Once a weight for each MB has been computed, this is translated into a $\Delta\mathrm{QP}_i(t)$, $0 \leq \Delta\mathrm{QP}_i(t) \leq \Delta_{\max}(t)$, for the MB i of frame t, in such a way that pixels with no descendants are not allocated any extra quality ($\mathrm{QP}_i(t) = \mathrm{QP}_{\mathrm{target}}$), while the QP for pixels with the maximum number of descendance levels (i.e. $F-t+1$) is reduced by $\Delta_{\max}(t)$. A simple mapping as the following one can be adopted:

$$\mathrm{QP}_i(t) = \mathrm{round}\left(\mathrm{QP}_{\mathrm{target}} - \Delta\mathrm{QP}_i(t)\right), \quad \text{where } \Delta\mathrm{QP}_i(t) = \frac{\mathrm{depth}_i(t)}{\Delta_{\max}(t)}, \tag{5}$$

with $\mathrm{depth}_i(t)$ being the average depth from frame t of pixel trees in macroblock i. The maximum QP reduction $\Delta_{\max}(t)$ starts with some value Δ_1 for the first frame of the GOP and progressively goes to zero as the frame index t approaches the last picture, since the last frames have lower temporal dependencies and, consequently, their quality should not uselessly increased too much. The heuristic QP selection algorithm described above requires the construction of the tree structure from each frame to the end of the GOP, therefore its complexity scales as $O(NF)$, where as before N is the number of pixels in a frame.

We performed preliminary tests to assess the validity of the proposed heuristic method using a H.264/AVC video encoder to encode two QCIF sequences, *News* and *Salesman*. We code 150 frames of both sequences ($F = 150$), where the first frame is Intra coded, and the remaining ones are all P pictures. In order to build the motion trees as explained in Section 2, we disabled the use of Intra macroblocks in P pictures and put the maximum number of previous reference frames for motion search equal to one. Each sequences is first coded with a fixed QP to extract motion vectors; these are used to build motion trees and to assign a QP to each MB accordingly, as specified in (5). We set a $\mathrm{QP}_{\mathrm{target}}$ equal to the QP of the first pass, and reuse the same motion vectors and the same coding modes extracted in the first pass to re-encode the group of pictures with the new QPs. Note that this is somewhat suboptimal, and an iterative approach alternating between motion estimation (and therefore tree construction) and QP refinement is to be preferred.[7] Although suboptimal, our approach



(a) (b)

Figure 4. RD curves for two QCIF video sequences. The number of frames in the group of pictures is 150.

has the advantage of being fast, while still producing good RD gains, as shown in Figure 4. In this picture, RD curves for the first and second pass coding have been traced out by trying different QPs (respectively, $\mathrm{QP}_{\mathrm{target}}$). Even if the proposed implementation is still preliminary, we are able to show nearly 1 dB improvement in PSNR in both sequences, by simply taking into account a pixel-level map of the depths of dependencies at the time of deciding the quantization parameter of each MB.

4. CONCLUSIONS AND FUTURE WORK

This paper proposes a tree-based representation of a video sequence that models the temporal dependencies between pixels induced by motion vectors. We build on the theoretical formulation of Uz et al.[14] to assert that the optimal rates in a tree are assigned according to the temporal impact of a pixel on its descendants, i.e. pixel with more descendants are coded with higher quality. We leverage this fact to evaluate a possible tradeoff between selecting, greedily, the best predictor for a pixel, and choosing instead a suboptimal one in terms of prediction residual energy, but which has a larger descendance (and thus, potentially, is coded with higher quality). Our simulation using a simple exponential RD model suggested by Uz et al. reveals that, actually, there exists a tradeoff between the number of descendants of the suboptimal predictor and the increase produced in the innovation complexity. This has two strong implications: on one hand, it corroborates the intuition that pixels with deeper impact on the future have to be allocated extra rate than pixels with fewer descendants. On the other hand, it gives clues that a motion estimation approach which embeds in the Lagrangian search function a weighting term to account for the temporal dependencies may lead to observable gains in the rate-distortion sense.

For now, we have offered a simple proof of concept for the first conclusion, by selecting the quantization parameter of the frames of a video sequence according to the depth of its pixels' trees. Even though the simple preliminary approach used here, the RD curves are improved by about 1 dB in PSNR. In future work, we aim at expanding this understanding in two ways, namely by building a finer model of RD tradeoffs induced by temporal dependencies in our tree structure, and by testing its validity for performing motion estimation and mode decision with real video codecs.

REFERENCES

[1] ITU-T, *Information Technology - Coding of Audio-visual Objects - Part 10: Advanced Video Coding* (May 2003). ISO/IEC International Standard 14496-10:2003.

[2] Wiegand, T., Sullivan, G., Bjontegaard, G., and Luthra, A., "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on circuits and systems for video technology* **13**(7), 560–576 (2003).

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

[4] Ortega, A. and Ramchandran, K., "Rate-distortion techniques in image and video compression," *IEEE Signal Processing Magazine* **15**(6), 23–50 (1998).

[5] Chellappa, V., Cosman, P., and Voelker, G., "Dual Frame Motion Compensation with uneven quality assignment," *IEEE Transactions on Circuits and Systems for Video Technology* **18**(2), 249 (2008).

[6] Ramchandran, K., Ortega, A., and Vetterli, M., "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Transactions on Image Processing* **3**(5), 533–545 (1994).

[7] Schumitsch, B., Schwarz, H., and Wiegand, T., "Optimization of transform coefficient selection and motion vector estimation considering interpicture dependencies in hybrid video coding," in [*Proc. SPIE*], **5685**, 327–334 (2005).

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

[9] Gothe, M. and Vaisey, J., "Improving motion compensation using multiple temporal frames," in [*IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*], **1**, 157–160 vol.1 (May 1993).

[10] Vetro, A., Sun, H., and Wang, Y., "MPEG-4 rate control for multiple video objects," *IEEE Transactions on Circuits and Systems for Video Technology* **9**(1), 186–199 (1999).

[11] Vetro, A., Haga, T., Sumi, K., and Sun, H., "Object-based coding for long-term archive of surveillance video," in [*Proc. IEEE Int. Conf. Multimedia & Expo*], **2**, 417–420 (2003).

[12] Hepper, D., "Efficiency analysis and application of uncovered background prediction in a low bit rate image coder," *IEEE Transactions on Communications* **38**(9), 1578–1584 (1990).

[13] Mavlankar, A. and Girod, B., "Background extraction and long-term memory motion-compensated prediction for spatial-random-access-enabled video coding," in [*Picture Coding Symposium*], (May 2009).

[14] Uz, K., Shapiro, J., and Czigler, M., "Optimal bit allocation in the presence of quantizer feedback," in [*Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*], (1993).

[15] Jayant, N. and Noll, P., [*Digital coding of waveforms: principles and applications to speech and video*], Prentice Hall Professional Technical Reference (1990).

[16] Winken, M., Schwarz, H., Marpe, D., and Wiegand, T., "Joint Optimization of Transform Coefficients for Hierarchical B Picture Coding in H. 264/AVC," in [*IEEE International Conference on Image Processing*], **4** (2007).

[17] Ortega, A., "Video coding: Predictions are hard to make, especially about t he future," in [*Proc. Image Media Processing Symposium*], (2007).

[18] Zhang, R., Regunathan, S., and Rose, K., "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications* **18**(6), 966–976 (2000).