COHERENT VIDEO RECONSTRUCTION WITH MOTION ESTIMATION AT THE DECODER

Claudia Tonoli, Marco Dalai and Riccardo Leonardi

Department of Electronics for Automation - Signal and Communication Laboratory
University of Brescia, Brescia, Italy.
Email: {*name.surname*}@ing.unibs.it

ABSTRACT

In traditional predictive video coding the block matching is performed at the encoder. The obtained motion field is then transmitted to the decoder, together with the prediction residue. Nevertheless, if the motion field is not provided it can be reconstructed, as long as the decoder manages to exploit some correlated information. This paper presents an algorithm for the motion estimation at the decoder side, given the prediction residue only. The main novelty of this algorithm relies on the contextual reconstruction of a frame region composed of several blocks. Simulation results show that taking into account a whole row can improve significantly the results obtained with an algorithm that reconstructs each block separately.

1. INTRODUCTION

In predictive video coding schemes very high compression efficiency is obtained thanks to motion estimation. The basic idea of this approach is to exploit the temporal redundancy across frames, estimated using the motion information. Usually, motion estimation is performed at the encoder side, and then the motion field is transmitted to the decoder, together with the compressed prediction error.

The decoder reconstructs the frame by means of a motion compensated prediction, using the received motion field. The received prediction residue is then added to the compensated frame. In this scheme, the decoder doesn't take into account any additional information aside from the received encoded stream, even if such information could carry important knowledge about the signal to be decoded. On the contrary, if the decoder could exploit further correlated information, rate savings could be achieved.

The possibility of omitting the transmission of motion information has recently attracted an increasing interest. For example in [1] an algorithm for motion derivation at the decoder side for the H.264/AVC codec is presented. In [2] a block-wise algorithm, based on LSE prediction is presented. In this paper we propose a motion compensated prediction of the current frame at the decoder side, performed in absence of motion information. The proposed method is based on DCT energy properties and image spatial continuity. The main novelty of this algorithm relies on the contextual decoding of a region composed of several blocks. Preliminary simulation results seem to be very promising.

The remainder of the paper is structured as follows. In Section 2 motion estimation principles are recalled and decoder based motion estimation is introduced. In Section 3 a model based on side information and the spatial coherence principle is introduced and a parameter for the evaluation of spatial coherence is described. In Section 4 a block-wise motion estimation algorithm is briefly summarized, whereas in Section 5 the proposed region-based algorithm is presented. Simulation results are presented and discussed in Section 6. Finally, concluding remarks are given in Section 7.

2. MOTION ESTIMATION AT THE DECODER

Predictive video coding is based on motion estimation at the encoder and motion compensation at the decoder. For a complete description of the topics related to motion estimation and its applications in state-of-art video coding, we refer the reader to [3], [4], and [5]. In this section, instead, the basic ideas of motion estimation coding are briefly recalled, while introducing the notation that will be used throughout the paper. Then, the idea of motion estimation at the decoder side is formalized.

In the following, the notation is referred to the scheme in Fig. 1. Let $X_{m,n}$, with $0 \leq m < M$ and $0 \leq n < N$, be the $B \times B$ block having $X(mB, nB)$ as the top-left pixel. Let W be the search window in the reference frame X^{ref}, i.e., the previous frame. W is centered in (mB, nB). A predictor for $X_{m,n}$ is searched for among all the blocks contained in W. Each possible predictor $\hat{X}_i$ is identified through the displacement from (m, n), i.e., its motion vector $\mathbf{v_i} = (v_y, v_x)$. Precisely, $\hat{X}_{m,n,\mathbf{v_i}}$ is the $B \times B$ block having $X^{ref}(mB + v_y, nB + v_x)$ as the top-left pixel. The selected predictor $\hat{X}_{m,n,\overline{\mathbf{v}}}$ is the most similar to $X_{m,n}$, according to a given distance measure, i.e., it leads to the residue $R_{m,n,\overline{\mathbf{v}}}$ such that

$$||R_{m,n,\overline{\mathbf{v}}}|| = \left|\left|X_{m,n} - \hat{X}_{m,n,\overline{\mathbf{v}}}\right|\right| = \min_i \left|\left|X_{m,n} - \hat{X}_{m,n,\mathbf{v_i}}\right|\right|$$

The motion field $\mathbf{v}$ represents the motion vectors associated to the predictor wich has been chosen for each block.

In traditional coding, both the motion field $\mathbf{v}$ and the block residues $R_{m,n}$ are suitably entropy encoded, and they are transmitted to the decoder. In this case, at the decoder each predictor can be identified very easily, by the use of the motion vector as an index. So each block is reconstructed as:

$$\overline{X}_{m,n} = \hat{X}_{m,n,\mathbf{v}} + R_{m,n} \tag{1}$$

The decoder reconstructs each frame operating in a strictly block-wise mode, since each block is reconstructed independently from its neighbors. It is worth remarking that the frame encoding order is such that the each frame is always decoded after its reference frame.

In the scheme considered in this paper, depicted in Fig. 2, $R_{m,n}$ only is known at the decoder, whereas $\mathbf{v}$ is not transmitted. Therefore, in this scenario the main challenge is to select a proper candidate C for each $X_{m,n}$, given $R_{m,n}$. When C has been chosen, $\hat{X}_{m,n,\mathbf{v}}$ is determined, contextually obtaining the estimated $\mathbf{v}$.

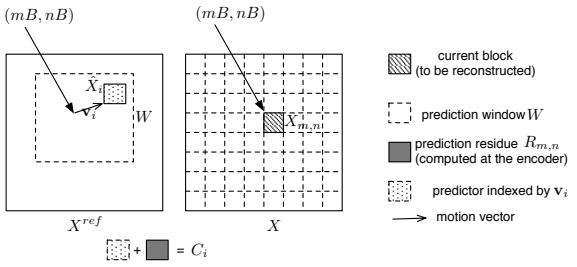

Fig. 1. Candidate set generation.

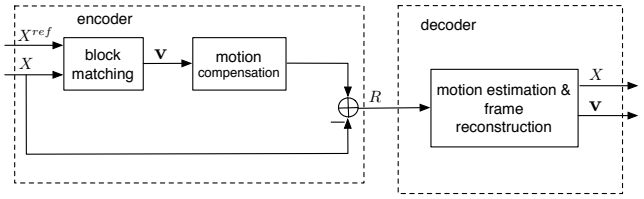

Fig. 2. Scheme of a coding system based on motion estimation at the decoder side

Since the decoder is assumed to know the search window W, it can generate the candidate set associated to the block $X_{m,n}$, as:

$$\mathcal{C}_{m,n} = \left\{ X^{ref}_{i,j} + R_{m,n} \quad | \; X^{ref}_{i,j} \in W \right\} \tag{2a}$$

$$= \left\{ \hat{X}_{m,n,\mathbf{v}_k} + R_{m,n} \right\} \tag{2b}$$

i.e, each block in the search window W is added the known residue. In the latter formulation the correspondence between each candidate and its associated motion vector is explicitated. In the following, the generic element of the candidate set $\mathcal{C}_{m,n}$ is referred to as $C_i^{(m,n)}$, whereas K is the cardinality of $\mathcal{C}_{m,n}$.

In order to determine the best candidate $\overline{C}_{m,n}$ and the correspondent $\overline{\mathbf{v}}_{\mathbf{m},\mathbf{n}}$, a model based on spatial coherence has been formulate, which involves the use of a spatial coherence parameter.

3. SIDE INFORMATION AND SPATIAL COHERENCE

In video coding, the phrase *side information* refers, to a general extent, to pieces of information correlated to the signal to be trasmitted. For the purposes of the model presented in this paper, we define as *side information* the information that the decoder knows and which is correlated to the part of the frame currently being decoded.

In particular, our scheme inherits the decoding order from predictive coding. The decoding order is such that, when decoding a given frame, its reference frame has been decoded previously, and hence it is completely known. Besides the reference frames, which can be referred to as *inter-frame* side information, our model relies on another type of side information, i.e., *intra-frame* side information. Intra-frame side information is composed of the already decoded blocks in the current frame. For example, let us suppose that, when $X_{m,n}$ is being decoded, the row above is known, which is the case considered in Section 5. With this hypothesis, the intra-frame side information for $X_{m,n}$ is

$$\mathcal{I}_{m,n} = \{ X_{i,j} \quad | \quad 0 \leq i < m, \forall n \} \tag{3}$$

The key idea underlying this work is the practical exploitation of the intra-frame side information, through the introduction of an assumption on the frame structure. We assume that the frame content is characterized by edges that preserve their continuity across the block boundaries and we define this property *spatial coherence*. The introduced hypothesis is loose, as the block boundaries are decided for coding convenience, regardless of the frame content. In the following, a coding scheme based on this model is presented.

3.1. Spatial Coherence Parameter

As pointed out before, hypotheses about the spatial coherence of the frame need to be formalized, so as to allow a proper evaluation of the candidates and a correct choice of the best fitting one.

In this section, a Spatial Coherence Parameter (SCP) based on the properties of the Discrete Cosine Transform is presented. Frequency domain techniques are used in error concealment methods (see for example [6]), which, like in the case considered in this paper, exploit the already decoded neighborhood to predict a missing block.

Let us consider a $2B \times 2B$ macroblock composed of four neighboring blocks. In the following, some remarks about how its spatial coherence properties reflect in the frequency domain are presented. Then, a parameter "measuring" the level of coherence of the macroblock is introduced.

The presence of a spatial discontinuity, i.e., a non matching edge across a block boundary, introduces high frequencies, that do not belong to the original image. Thus, when this happens, in the DCT domain the energy is distributed on a large number of coefficients, some of them located in the higher frequency range. On the contrary, when edges match properly, in the DCT domain the energy should be very concentrated on few, low frequency located coefficients.

Given a macroblock Y, composed of 4 $B \times B$ blocks as described above, its DCT Spatial Coherence Parameter $p(Y)$ is computed according to the following steps:

- $Z_Y = DCT(Y)$ is computed and normalized in order to have unitary energy.
- DCT coefficients are sorted in descending square modulus magnitude order; let $\tilde{Z}_Y(k)$ be the $k-th$ coefficient in such order.
- given a fixed threshold T, let $p(Y)$ be the minimum value of k that verifies the following condition: $\sum_{j=1}^{k} |\tilde{Z}_Y(j)|^2 \geq T$

4. BLOCK-WISE SELECTION ALGORITHM

The Spatial Coherence Parameter can be used as a test to determine whether a block fits the given neighborhood. In [2] an algorithm for motion estimation at the decoder based on this principle, but using a different SCP, is described. This algorithm, which operates on each block separately, in the following is referred to as block-wise algorithm and it is used as a reference for the performance assessment of the row-wise algorithm presented in this paper. According to this algorithm, first the candidate set for the current block $X_{m,n}$ is generated, as described in Section 2. Then, a ranking of the candidates is obtained, to determine which candidate fits best the known causal neighborhood, i.e., the upper-left , upper and left neighbors. For each $C_i^{(m,n)}$, the macroblock $Y_{C_i}^{(m,n)}$, composed of the current candidate $C_i^{(m,n)}$ and the three known causal neighbors, is constructed as:

$$Y_{C_i}^{(m,n)} = \begin{bmatrix} \overline{X}_{m-1,n-1} & \overline{X}_{m-1,n} \\ \overline{X}_{m,n-1} & C_i^{(m,n)} \end{bmatrix} \tag{4}$$

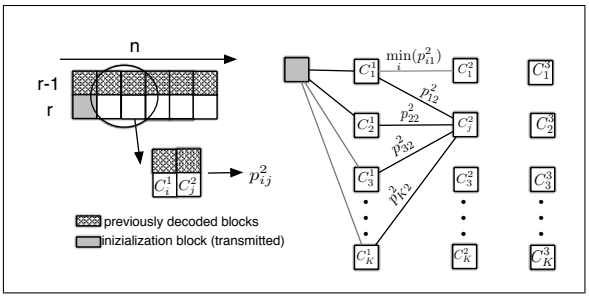

Fig. 3. Computation of the p_{ij}^n parameters and trellis construction

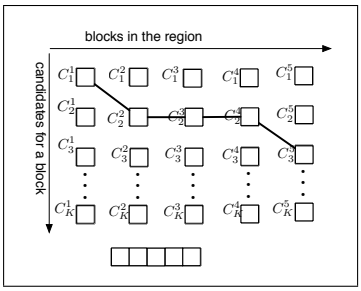

Fig. 4. Trellis contruction: the path represents the combination $(C_1^1, C_2^2, C_2^3, C_2^4, C_3^5)$

Then, the SCP $p(Y_{C_i}^{(m,n)})$ is computed. The selected predictor $\overline{C}^{(m,n)}$ is the one that minimizes p:

$$\overline{C}^{(m,n)} = \underset{C_i \in \mathcal{C}_{mn}}{\operatorname{argmin}}\, p(Y_{C_i}^{(m,n)}) \tag{5}$$

Hence, the estimated motion vector is the one associated with $\overline{C}^{(m,n)}$. This algorithms works well as long as the hypothesis that the true original block is always the most coherent one (in terms of DCT SCP) holds.

5. REGION BASED MODEL FOR MOTION ESTIMATION AT THE DECODER

In this section a novel algorithm for motion estimation at the decoder is introduced. This algorithm is based on the contextual decoding of several blocks belonging to a given region of the frame. All the possible combinations of blocks for the region (or at least the more likely ones, depending on whether any optimization is carried out, see Section 5.1) are generated, combining the candidates of each block in the region.

The algorithm presented in this paper is designed to overcome one of the main drawbacks of the block-wise algorithm, i.e., the impossibility of detecting when the Spatial Coherence Parameter fails to identify the correct original block. In fact, errors occur when one or more candidates happen to have a SCP smaller than the original block SCP, i.e., when the condition (5) does not provide the correct candidate. A wrong block is likely to induce an error on the next block when it is used as a neighbor for the latter, and so on, allowing the error to propagate across the frame. A global algorithm, taking into account a region of adjacent blocks, can exploit the propagation of a wrong choice. Indeed, the main feature of this algorithm is an average of the SCP over the blocks of the entire region. Let us consider the case in which the correct block is not the minimum SCP block, so a wrong candidate is chosen. Since the neighbor of the next block contains such wrong candidate, all the computed parameters will be high, because no good match with a wrong neighbor can be found. So the averaging can be thought of as a balancing of the presence of a not minimal SCP correct block with the effect of the error propagation on other blocks. In the following, instead of an average a non-normalized sum will be performed; this does not affect the just discussed property.

For the sake of simplicity, the algorithm has been implemented taking as a region a single row of blocks. This special region shape simplifies the interdependency of unknown blocks, allowing for a Viterbi-based minimization of the SCP sum.

5.1. Row-wise SCP Sum minimization

In this section, the algorithm previously introduced is described in detail. Since we are presenting a preliminary implementation of a novel approach, a simple region shape has been chosen. Indeed the chosen region shape is an entire row of blocks:

$$T_{\overline{m}} = X_{(\overline{m},n)},\ for\ 0 \leq n < N \tag{6}$$

The rows are decoded from the top one (T_1) to the bottom one (T_{N-1}). The first row, T_0, is assumed to be completely transmitted, because it is required for the algorithm inizialization. So, a whole row is decoded at once, under the hypothesis that the decoding of the row above has already been performed. Such hypothesis guarantees that, for each block, its upper and upper-left neighbors are known. On the contrary, the left neighbor is uncertain, since it still belongs to the current row and it is being decoded too. Hence, for each candidate C_k^n of the n-th block, K coherence parameters p_{ij}^n need to be computed, each one using a different candidate C_i^{n-1} as left neighbor. (Note that from now on m is left implicit where it is clear that $m = \overline{m}$, i.e., the current row index).

Let the region predictor $\underline{L}_{\mathbf{q}} = (C_{q_1}^1, C_{q_2}^2, \cdots, C_{q_N}^N)$, indexed by the index vector $\mathbf{q}$, be the row predictor composed of the candidate $C_{q_1}^1$ for the first block, the candidate $C_{q_2}^2$ for the second block, and so on. As depicted in Fig. 3, the parameters p_{ij}^n are the SCP of the macroblock Y_{ij}^n, defined as follows:

$$Y_{ij}^n = \begin{bmatrix} \overline{X}_{\overline{m}-1,n-1} & \overline{X}_{\overline{m}-1,n} \\ C_i^{(\overline{m},n-1)} & C_j^{(\overline{m},n)} \end{bmatrix} \tag{7}$$

The sum $\sigma(L_{\mathbf{q}})$ of the SCPs of all the involved blocks is considered, for each combination $\underline{L}_{\mathbf{q}}$:

$$\sigma(L_{\mathbf{q}}) = \sum_{n=1}^{N} p_{q_{(n-1)}q_n}^n \tag{8}$$

In order to track the combinations with lower SCP Sum σ, a Viterbi-like minimization (see for example [7]) is used: each path on the trellis represents a combination of candidates, i.e., a candidate row $\underline{L}_{\mathbf{q}}$. In Fig. 4 this model is depicted; the highlighted path represent the combination composed of candidate C_1^1 for the first block, C_2^2 for the first block, and so on. Being each candidate univocally indexed by a motion vector, each path represents a row of the motion field, as well. At each step, the selected candidate $\overline{C}_l^{n-1}$ is such that

$$p_{lj}^n = \min_i(p_{ij}^n) \tag{9}$$

for the hypotesis $C^n = C_j^n$. Thanks to the application of the Viterbi algorithm, the combination with higher σ are automatically discarded.

The main drawback of this scheme is that, despite the use of the Viterbi algorithm, the decoding complexity remains quite high. This is important to be noted, especially because the encoder is as complex as in predictive coding. Since this is a preliminary implementation of a novel algorithm, complexity optimizations are still to be studied. In this implementation, we have supposed that the motion field is smooth, so as to discard the combinations leading to a motion field that varies abruptly from one block to the next.

6. EXPERIMENTAL RESULTS

The performance of the proposed algorithm has been evaluated by assessing the improvement with respect to the block-wise algorithm. The percentage of correctly reconstructed blocks, i.e., of correctly estimated motion vectors, vs. the PSNR of the encoded sequence are plotted in Fig. 5, for the first 50 frames of the *Foreman*, *Mobile* and *Highway* sequences.From the obtained results, it emerges that the region-based reconstruction significantly improves the performance of the block-wise selection. There is one exception only, which regards the lossless case. For the *Highway* sequence (plot with square markers), the region-based algorithm leads to slightly worst performance. In all the other cases the percentage of correct blocks significatively increases, even if, at very low qualities, it still remains quite low.

An approximation of the rate-distortion curve for the proposed algorithm, compared with the block-wise algorithm, is also reported in Fig. 6, for the *Foreman* sequence. In order to give an idea of how the presented algorithm could perform in a realistic scenario, it has been applied to a lossy codec. In more detail, the rate and PSNR values for the case of transmission of the whole motion field have been obtained using a simplified H.264 codec. The block size has been set to 16 and the considered prediction mode is P, i.e., monodirectional prediction, with a single reference picture. An important remark about the rate estimation must be given: the coding efficiency in modern predictive codecs, such as the H.264 codec, depends heavily on how arithmetic coding is performed. Since our methods has not been really implemented in H.264 yet, it is impossible to measure exactly the rate savings. In order to produce a reliable estimate, the bits devoted to the motion transmission for each block have been computed, and, for the correctly predicted block, the result has been subtracted from the overall bit-rate. A signalling overhead has also been taken into account. The three plots appear to be almost overlapping: in the considered case the coding gain seems to be limited by both the signalling overhead and the fact that the motion represents a small amount of the total rate.

7. CONCLUSIONS

In this paper a novel algorithm for motion estimation at the decoder side is presented. The contextual decoding of a region of blocks is formalized, considering a simple case of a row of blocks. Thus, the blocks belonging to a whole row are decoded contextually, exploiting the property of spatial coherence of the frame. In order to evaluate the spatial coherence, a DCT-based Spatial Coherence Parameter is presented. The sum of the SCP for each candidate row is to be computed and minimized. In order to perform the minimization and to reconstruct the row, a model based on the Viterbi algorithm is introduced. Simulation results show that the proposed approach outperforms the block-wise algorithm that employs the same parameter. Moreover, the preliminary analysis of the rate-distortion performance seems to be promising.

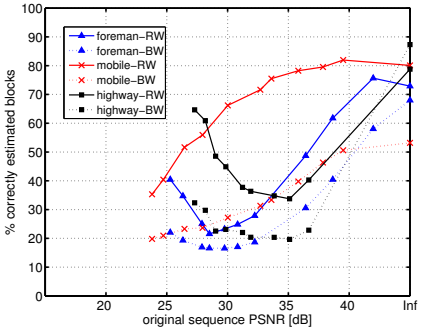

Fig. 5. Performance assessment for the row-wise (RW) and block-wise (BW) algorithm

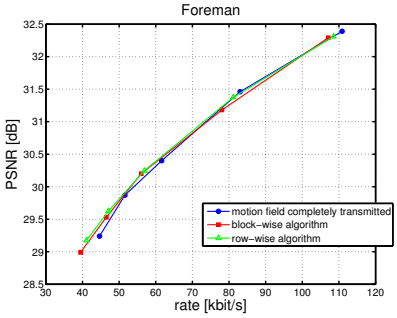

Fig. 6. Estimated rate-distortion curves for the *Foreman* sequence

8. REFERENCES

[1] S. Kamp, M. Evertz, and M. Wien, "Decoder side motion vector derivation for inter frame video coding," in *Proc. of ICIP08*. IEEE, San Diego, CA, 12-15 Oct. 2008, pp. 1120–1123.

[2] C. Tonoli, P. Migliorati, and R. Leonardi, "Video coding with motion estimation at the decoder," in *Proc. of Thyrrenian Workhshop on Digital Communication*, Pula, Sardinia, Italy, 2-4 Sept. 2009.

[3] T. Wiegand, G. J. Sullivan, and G. Bjontegaard, "Overview of the h.264/avc video coding standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, July 2003.

[4] T. Wiegand, G.J. Sullivan, and A. Luthra, "Draft ITU-T recommendation and final draft international standard of joint video specification," Joint Video Team Doc. JVT-G050r1, June 2003.

[5] S. Kappagantula and K.R. Rao, "Motion compensated interframe image prediction," *IEEE Trans. on Communications*, vol. 33, pp. 1011–1015, 1985.

[6] J.W. Park, G. J.W. Kim, and S.U. Lee, "DCT coefficients recovery-based error concealment technique and its application to the mpeg-2 bit stream error," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, pp. 845–854, December 1997.

[7] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on Information Theory*, vol. 13, pp. 260–269, April 1967.