Optimized Inpainting-based Macroblock Prediction in Video Compression

Yang Xu, Hongkai Xiong
Department of Electronic engineering, Shanghai Jiao Tong University, Shanghai 200240, China
{xuyangz, xionghongkai}@sjtu.edu.cn

Abstract— **In this paper, we propose an optimized inpainting-based macroblock prediction mode in the state-of-the-art H.264/AVC video engine. Parallel with the existing intra- and inter- modes, it is regularized by the global tempo-spatial consistency between the current MB and the co-located decoded region. The target is formulated as a local optimization problem by minimizing the energy of Markov Random Field (MRF). An ordered belief propagation (BP) is developed to solve the optimization problem with spatio-temporal consistency regularity and achieve the patch assignment of largest probability. It ensures a stable marginal belief distribution through updating local messages via iterative forward and backward process, to impose a prioritized inference on the structure. Rate-distortion optimization is used to evaluate the mode selection within the inpainting-based prediction mode, intra- and inter modes. It has been implemented into H.264/AVC, and achieves bit-rate saving and higher PSNR performance, especially in low bit-rate.**

I. INTRODUCTION

The state-of-art video compression schemes such as H.264/AVC have achieved vital efficiency in image and video compression. These mainstream signal processing based compression schemes only exploit statistical redundancy among pixels through intra and inter prediction, followed by entropy coding [1]. To achieve better performance, more intra prediction methods, e.g. template matching [10], have been noticed. Recently, High-performance Video Coding (HVC) has been put forward to develop the next-generation video coding standard. In the Key Technology Area (KTA) software [2], an enhancement to H.264 intra coding was developed, where simplified bidirectional intra prediction (BIP) is used in addition to the existing nine intra prediction modes, and separable directional transforms were also absorbed besides DCT-like integer transform [3]. Various modes have been advocated to suit regions of different properties.

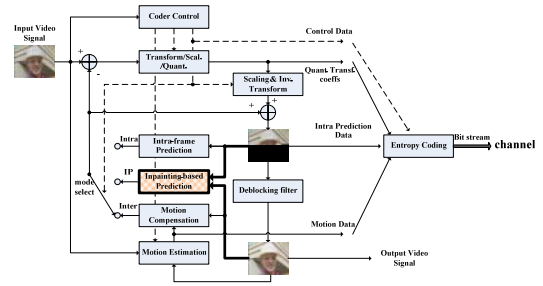
Parallel with traditional prediction track, computer vision technologies have shown remarkable progress in dealing with images of good perceptual quality. Typically, either image inpainting [4] or texture synthesis [5] has been playing a leading effort to exploit visual redundancy by restore images with inferable information around the missing region. Those can be treated in a unified manner under the framework of Markov Random Field (MRF), and optimization algorithms, e.g. belief propagation (BP) [6], are applied. These approaches solve a wide class of problems in image processing and computer vision but are space and time consuming. Assistant information, such as edges, is taken into account to guide and help restoration process. In this context, the structure propagation has ever been advanced as a global optimization problem [7], which preserves important structure in condition of sharp curves outlined by the user earlier.

Based on the fact that missing regions can be restored through image completion approaches, original images are analyzed at the encoder side [8] to skip some regions and extract assistant information (e.g. edges). In the decoder side, the delivered assistant information plays a key role to guide image completion process and restore the skipped regions accurately. Naturally, this effort has been extended to video area with special and temporal texture synthesis and edge-based inpainting [9]. Despite those claim a bit-rate saving at similar visual quality levels compared with the traditional H.264/AVC codec, the passive method of throwing all the restoration burden to decoder fails to ensure pixel-wise fidelity.

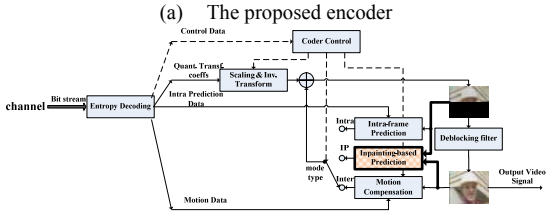
In this paper, we propose an optimized inpainting-based macroblock prediction mode in the state-of-the-art H.264/AVC video engine. It does not only ensure the visual quality of the decoded result, but also be competitive in objective measure, e.g. PSNR and bit-rate. Parallel with the existing intra and inter modes, it is regularized by the global tempo-spatial consistency between the current MB and the co-located decoded region. The target is formulated as a local optimization problem by minimizing the energy of Markov Random Field (MRF). An ordered belief propagation (BP) is developed to solve the optimization problem with spatio-temporal consistency regularity and achieve the patch assignment of largest probability. It ensures a stable marginal belief distribution through updating local messages via iterative forward and backward process, to impose a prioritized inference on the structure. Each MB could be predicted via intra-, inter- and inpainting-based prediction modes, and rate-distortion optimization is calculated to decide optimal mode.

II. THE PROPOSED SCHEME

In the proposed video compression engine, the inpainting-based prediction mode is parallel with the existing intra- and inter- modes. No assistant information needs to be extracted and encapsulated. As shown in Fig. 1, each MB could be coded via intra-, inter-, or inpainting-based prediction mode. In the inpainting-based prediction mode, the prediction is constructed through exampalar based inpainting method, where sample patches in both the current and previous frames compose a dictionary of candidates. We choose suitable patches from the dictionary to paste on the MB region by the optimization criterion.



(a) The proposed encoder



(b) The proposed decoder

Fig. 1. The framework of the proposed codec.

In order to select an optimal mode in a rate-distortion (RD) sense, Lagrangian minimization is applied to mode selection problem. We use the general form of cost function in Lagrangian R-D optimized (RDO) mode selection:

$$J = D + \lambda \tilde{R} \tag{1}$$

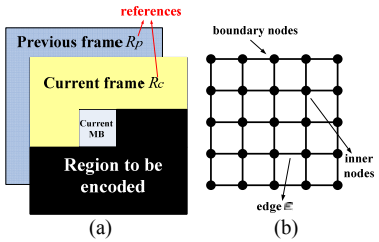
A. *Image Inpainting as a Discrete Optimization Problem*

In the proposed method, the target region to be inpainted is the current macroblock M, and the candidate patches are acquired from the constructed region in both the current and previous frames $S = R_p \cup R_c n$, as depicted in Fig. 2(a). The candidates dictionary C consist of all $wl \times w$ patches from the source region S. The nodes of the MRF are de-sampled from the pixels in Mn, and the edges ε of the MRF make up a 4-neighborhood system. The nodes consist of boundary nodes, whose neighborhood intersects the source region S, and inner nodes, as depicted in Fig. 2(b).

The predicted MB is constructed by copying suitable candidate patches over the nodes' position. To this end, we define the energy of MRF to turn the task into a discrete optimization problem of minimizing the energy of MRF. We define the single node potential $V_s(c_s n)$ and pairwise potential $V_{mn}(c_m, c_n)$ as displayed in Fig. 3. The single node potential $V_s(c_s)$, or called data cost, for placing patch $c_s n$ over node s, presents how well the patch agrees with the source region around node s:

$$V_s(c_s) = \sum_{(x,y)\in R_s \cap R_c} |c_s(x,y) - R_c(x,y)|^2 + \gamma \sum_{(x,y)\in R_s \cap R_p} |c_s(x,y) - R_p(x,y)|^2 \tag{2}$$

In (2), the first term expresses the distortion between the candidate patch and the source region in the current frame, and the second term describes the mismatch between the candidate patch and the source region in the previous frame, which is decayed by a coefficient γ, for motion change between the current and previous frames. If the current frame is IDR frame and there is no previous reference frame, the second term is zero. In a similar fashion, the smoothness potential $V_{mn}(c_m, c_n n)$, due to placing patches $c_m n$, c_n over neighbor nodes m and n, measures how well these patches agree at the region of overlap:



(a) (b)

Fig. 2. (a) Candidate patches are acquired from the constructed region in both the current and previous frames; (b) The nodes and edges of the MRF.

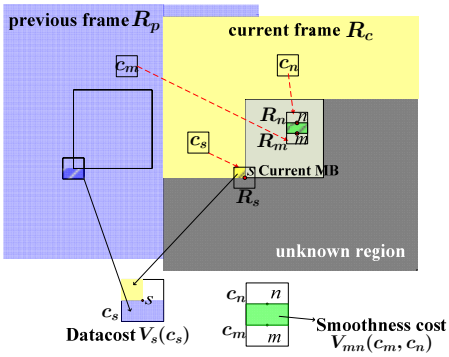


Fig. 3. Compute data cost $V_s(c_s)$ and smoothness cost $V_{mn}(c_m, c_n)$.

$$V_{mn}(c_m, c_n) = \sum_{(x,y)\in R_m \cap R_n} |c_m(x,y) - c_n(x,y)|^2 \tag{3}$$

In our test, the overlapping region of the two neighboring nodes is half the area of the patch in order to obtain smooth result and make best use of dependency between the neighbors.

Based on the definition of data cost and smoothness cost, our goal is to minimize the energy of MRF by assigning a patch $\hat{c}_m \in Cnj$ to each node m:

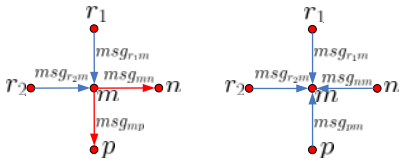
$$\min E(\hat{c}) = \sum_{m\in M} V_m(\hat{c}_m) + \sum_{(m,n)\in\varepsilon} V_{mn}(\hat{c}_m, \hat{c}_n) \tag{4}$$

B. *Ordered belief propagation*

Belief propagation (BP) is an iterative algorithm trying to find a maximum a posteriori (MAP) estimate by iteratively solving a finite set of equations until convergence. Through continuously propagating local messages between the nodes of the MRF, beliefs of every node is updated and optimal output is achieved when all the messages have stabilized finally. The messages are expressions of the node's opinion to its neighbors. Specifically, the set of messages sent from node m to its neighbor n implies the opinion of m about assigning label $c_n \in C$ to node n, and is denoted as $\{msg_{mn}(c_n)_{c_n\in C}\}$:

$$msg_{mn}(c_n) = \min_{c_m\in C}\{V_{mn}(c_m, c_n) + V_m(c_m) + \sum_{r\neq n,(r,m)\in\varepsilon} msg_{rm}(c_m)\} \tag{5}$$

From (5), we can find that if one node m need to send message to its neighbor node n, it must first traverse each one



(a) Calculating messages (b) Calculating beliefs

Fig. 4. (a) If a node m need send messages to its neighbors n and p, it must make use of the messages msg_{r_1m} and msg_{r_2m} from the neighbors that have already sent some messages. (b) If a node m need calculate beliefs $b_m(c_m)$, it should collect the messages coming from all the neighbors.

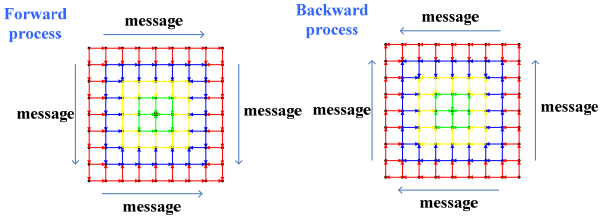


Fig.5. Belief propagation, forward process and backward process, nodes are visited in order from boundary to inner side.

of its own candidates c_m, and decide which one of them provides the greatest support for assigning candidate c_n to node n. The support of candidate c_n to node n is determined by the compatibility between candidate c_m and c_n (smoothness cost $V_{mn}(c_m, c_n)$), and the likelihood of assigning candidate c_m to node m (data cost $V_m(c_m)$), as well as the opinion of its other neighbors about candidate $c_m g$ (sum of messages $\sum_{r\neq n,(r,m)\in\varepsilon} msg_{rm}(c_m)$). As a result, if one node m need to send its messages to its neighbor n, it must first receive the messages from other neighbors(if the neighbors have already sent some messages), and then add its own opinion into the messages sent to node n (see Fig. 4(a)).

To improve BP's convergence and accelerate completion, we consider an ordered BP to ensure the quality of messages, where nodes are arranged in a special order list. In one iteration, each node should not only send messages to its neighbors, but also receive feedbacks. That is, one iteration is decomposed into forward and backward processes where nodes are interleavely visited from head to tail through the order list in the forward process and inversely in the backward process. Once a node is accessed, it sends messages to all its neighbors and updates its beliefs. The problem is further regarded as in which order the nodes should be arranged to effectively reduce the computing complexity and speed up convergence. It is obvious to find that boundary nodes with more information from the surrounding known region is more confident to determine which kind of patches are suitable, while the inner nodes only rely on the opinions from their neighbors to select patches. Moreover, the nodes that have larger intersection area with known region are more confident, e.g. the node in the top left corner is the most confident one in the MB. In this way, nodes are arrayed to the order list according to their confidence decreasingly. As in Fig.5, within forward process, we scan the nodes from the top left to the below right along the boundary, like peeling an onion. Within the backward process, the nodes are visited in the inverse order from inner to outside.

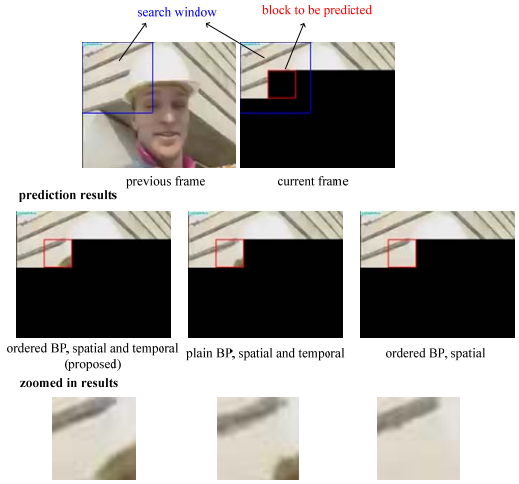


Fig.6. Prediction results of ordered BP, plain BP, spatial-temporal and uniquely spatial reference.

After several iterations, when the messages are asymptotically stable, the final set of beliefs $\{b_m(c_m)\}_{c_m\in C}$ of each node, which expresses the probability of assigning patch c_m to node m, is obtained:

$$b_m(c_m) = -V_m(c_m) - \sum_{r:(r,m)\in\varepsilon} msg_{rm}(c_m) \quad (6)$$

The set of beliefs is related to the node's data cost and the messages from all its neighbors (see Fig. 4(b)). Actually, each belief $b_m(c_m)$ approximates the maximum conditional probability given the fact that node m has been assigned the patch c_m, and it is the evidence to select candidates. Finally, candidate of maximum belief is assigned for each node:

$$\hat{c}_m = \arg\max_{c_m\in C} b_m(c_m) \quad (7)$$

As the order of message propagation is proposed to ensure the confidence of messages, compared with plain BP without the special order, superiority can be found out in Fig. 6 after the same number of iterations. In plain BP, nodes are visited from left to right and from up to down. Besides, in our proposed method, both the spatial and temporal information are considered, we can compare the results with output of uniquely spatial consideration, where there is only the first term when computing data cost in (2).

C. *Composition of final patches*

After BP has completed and the final patches have been selected out for each MRF node, we need to compose them to produce the final result. As there are some overlapping regions between the neighboring nodes, every pixel in the MB region is covered by several patches centered in its surrounding nodes. The final result of the pixel is composed by blending the patches with weights related to the nodes' confidence:

$$I(x,y) = \sum_{i:(x,y)\in R_i} w_i \hat{c}_i(x,y), \quad s.t. w_i \propto \frac{1}{i}, \sum_i w_i = 1 \quad (8)$$

By computing a sum of squared error (SSE) between the prediction MB and original MB, we find that the distortion of SSE is reduced averagely by about 3% through weighted blending of patches, compared with simply paste the patches

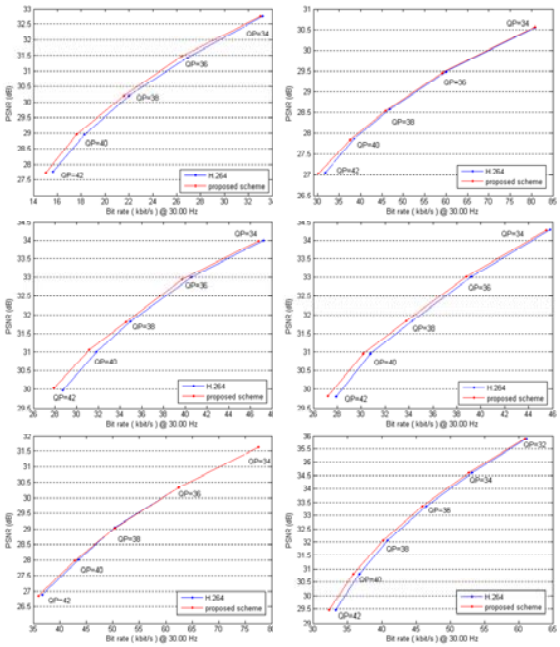


Fig. 7. Performance comparision, from left to right and up to down: "Forman", "Coastguard", "Suzie", "Mother_daughter", "Salesman","Akiyo".

to the position of nodes.

III. EXPERIMENTAL RESULTS

The proposed mode has been implemented in the Joint Model version JM15.1 of H.264/AVC. In the experiments, we used YUV 4:2:0 sequence format with QCIF resolution (176x144). All sequences are tested with 30 Hz frame rate, IPPP… sequence type, and a GOP size of 10 frames. Inpainting mode is enabled only in P slices. We set decay coefficient $\gamma = 0.7$, and inpainting mode is compared with I4x4, I16x16, P16x16, P8x16, P16x8, P8x8, P4x8, P8x4 modes in P slices.

We test six sequences: "Foreman", "Coastguard", "Suzie", "Mother_daughter", "Salesman" and "Akiyo". Fig.7 shows the coding performance comparison between the proposed scheme and H.264/AVC. Table I shows the inpainting mode selecting rate under different QP level, and Fig. 8 gives out some examples of inpainting mode MBs, which are outlined by red rims. From the results, we can observe that the ratio of inpainting-based prediction mode selection is augmenting along with the increase of QP levels. It means the inpainting-based prediction mode is superior in low bit-rate conditions.

IV. CONCLUSION AND FUTURE WORK

In this paper, we present a generic video coding framework by adding an inpainting based prediction mode in the traditional H.264/AVC scheme. Macroblocks are predicted by inpainting mode, besides the existing intra- and inter-modes, and rate distortion optimization is considered to decide which mode the current MB should choose. Under the framework of MRF, we use ordered belief propagation to get the best patch arrangement, and tempo-spatial correlation is considered. The approach has been implemented into

Fig. 8 Examples of inpainting mode MBs

TABLE I. INPAINTING MODE RATIO UNDER DIFFERENT QP LEVELS

Inpainting Mode Ratio		QP LEVEL					
		32	34	36	38	40	42
sequences	Forman	4.55%	7.07%	7.58%	10.4%	12.7%	15.3%
	Coast-guard	1.62%	3.84%	7.68%	14.9%	26.0%	36.6%
	Suzie	7.54%	10.4%	13.0%	15.5%	23.0%	33.3%
	Mother_ daughter	8.96%	11.9%	14.3%	16.5%	20.9%	26.5%
	Salesman	3.13%	3.64%	3.16%	8.18%	12.0%	17.6%
	Akiyo	9.63%	13.0%	17.0%	20.8%	29.6%	34.6%

conventional video coding scheme H.264/AVC and has show improved performance, especially in low bitrate condition.

In the future, more vision based technologies can be introduced into video coding framework. We consider investigating multidimensional feature tensor, e.g. structure, texture, color, etc, which can be extracted and analyzed for matching and completion to achieve higher compression rate and better performance.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits Syst. Video Technol., vol. 13(7), pp. 560-576, Jul.2003.

[2] Key technology area reference software, http://iphome.hhi.de/suehring/tml/download/kta .

[3] Yan Ye, Marta Karczewicz, "Improved intra coding," ITU-Telecommunications standardization Sector, 33rd Meeting, Shenzhen, China, 20 October, 2007.

[4] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," Proceedings of ACM SIGGRAPH, New Orleans, USA, pp.259-263, July 2000.

[5] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," ICCV, Corfu, Greece, pp.1033-1038, September 1999.

[6] Nikos Komodakis and Georgios Tziritas, "Image completion using global optimization," IEEE Computer Society Conference on CVPR, pp.442-452, June 2006.

[7] Jian Sun, Lu Yuan, Jiaya Jia and Heung-Yeung Shum, "Image completion with structure propagation," SIGGRAPH, Vol. 24, pp. 861-868, 2005.

[8] Dong Liu, Xiaoyan Sun, Feng Wu, Shipeng Li, and Ya-Qin Zhang, "Image compression with edge-based inpainting," IEEE Transactions on CSVT, Vol. 17, Issue 10, pp. 1273-1287, Oct 2007.

[9] Chunbo Zhu, Xiaoyan Sun, Feng Wu, and Houqiang Li, "Video coding with spatio-temporal texture synthesis and edge-based inpainting," IEEE International Conference on Multimedia and Expo, Hannover, Germany, pp. 813-816, June 2008.

[10] Thiow Keng Tan, choong Seng Boon, and Yoshinori Suzuki, "Intra prediction by template matching," IEEE International Conference on Image Processing, pp.1693-1696, Oct. 2006.