CLASSIFIED PATCH LEARNING FOR SPATIALLY SCALABLE VIDEO CODING

Xiaoyan Sun and Feng Wu

Microsoft Research Asia

{xysun, fengwu}@microsoft.com

ABSTRACT

This paper proposes an advanced spatially scalable video coding approach that exploits the inter layer correlation between different resolution layers by classified patch learning. The novelty of our proposed scheme is twofold. First, the correlation between low and high resolution frames is explored at patch level with regard to image features. Patches extracted from the previous coded frame are classified into structural and textural sets according to the gradient information. Then the inter layer correlation is separately studied for the two sets, resulting in two databases containing pairs of patches at different resolutions. Second, our proposed patch-based compensation manages to simultaneously exploit the spatial and temporal redundancies without overhead bit for motion. Based on the two databases, a high resolution prediction is derived from the current low resolution reconstruction at structural and textural regions, respectively. Experimental results show that our proposed approach improves the performance of H.264/MPEG spatially scalable coding up to 1.9dB and significantly enhances the subjective quality, especially at low bit rates.

***Index Terms*—** Scalable video coding, spatially scalable, inter layer correlation, classified patch learning

1. INTRODUCTION

Spatially scalable coding provides adaptation to the diversity of user devices as well as the heterogeneity on network infrastructures by representing video signals in one bit stream but serving different displaying resolutions. In a pyramidal layered spatially scalable video coding scheme, a base layer bit stream is generated by coding the lowest resolution version of an input video. Then the enhancement layer bit streams are produced by taking advantage of the correlations across neighboring layers as well as that between adjacent frames.

One way to exploit the inter layer correlation between different resolutions is to up-sample the reconstructions at low resolution to predict the frames at high resolution [2][3]. In addition to pixel values, the motion vectors and modes at low resolution layers can also be utilized in the mode/motion estimation at high resolutions [4]. In fact, some of these schemes have been adopted in the current video coding standards, such as MPEG-2 and H.264/MPEG spatially scalable coding (SVC in short) [1].

Recently, learning-based approaches have shown their potential in studying the relationship of image features at different resolutions. Databases consisting of co-occurrence image patches at two different resolutions are introduced as priors for image recovery, such as image hallucination [5][6]. Moreover, it has been extended to image compression, where the database is regarded as a codebook and the indices are embedded in the coded low resolution image [7][8].

In this paper, we propose to exploit the inter layer correlation in spatially scalable video coding by classified patch learning. In the proposed scheme, pairs of reference patches are extracted from low and high resolution reconstructions at the previous frame. Gradient information is involved to classify the reference pairs into two sets, which are then clustered separately to form the structural and textural databases. During compensation, patches extracted from the current low resolution reconstruction are also classified. By finding their matches in the corresponding database, our proposed patch-based compensation scheme conducts a high resolution prediction based on both the current low resolution reconstruction and the two databases. The present scheme can be readily integrated with the current SVC coding scheme. Experimental results demonstrate the effectiveness of our proposed spatially scalable video coding method.

The rest of this paper is organized as follows. The framework of our proposed spatially scalable coding scheme with classified patch learning is introduced in Section II. Then the classified patch learning as well as compensation is described in detail in Section III. Performance of our proposed coding scheme is evaluated in Section IV. Finally, Section V concludes this paper.

2. FRAMEWORK OF OUR PROPOSED CODING SCHEME

Let $F^i = \{f_t^i\}$ represents an input video, where the superscript i indicates the spatial resolution layer and the subscript t is the frame index. The superscript i equals to zero at base layer resolution. Given a low-pass filter, a low resolution video is generated from an original high resolution video via a down-sampling process $\mathcal{D}(\cdot)$

$$f_t^i = \mathcal{D}(f_t^{i+1}), i = 0,1 \text{ in a 2-layer system.} \quad (1)$$

Also, a down-sampled frame can be converted back to high resolution via an up-sampling process $\mathcal{U}(\cdot)$.

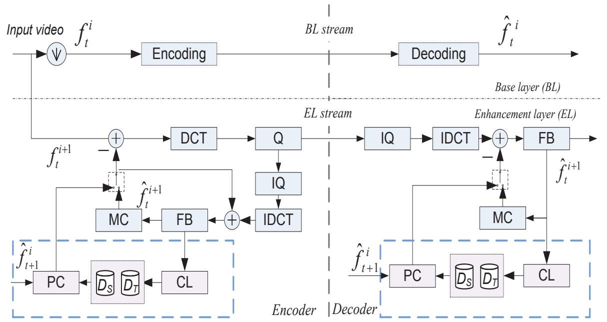


Fig. 1. Framework of our proposed spatially scalable coding scheme with classified patch learning.

The framework of our proposed spatially scalable video coding with classified patch learning is illustrated in Fig.1. Some modules, such as entropy coding, are omitted for simplicity. The classified patch learning along with the patch compensation is exhibited by the dashed blocks.

In this framework, the base layer encoding as well decoding is traditional, while the enhancement layer coding employs our proposed classified patch learning in the inter layer prediction. As shown in Fig. 1, at enhancement layer, the high resolution reconstruction $\hat{f}_t^{i+1}$ stored in the frame buffer (FB) is input into the classified patch learning (CL) module. Two databases, the structural database (D_S) and the textural database (D_T), are generated. In addition to the traditional motion compensated prediction, the patch compensation module (PC) produces another prediction from the base layer reconstruction $\hat{f}_{t+1}^i$ and the two databases derived from $\hat{f}_t^{i+1}$. These two predictions are selective utilized at macroblock level by a rate distortion optimal selection.

The corresponding decoding process is exhibited on the right side in Fig. 1. The classified patch learning as well as compensation is as same as that in decoding. It can be observed that the learning and compensation can be performed on-line at the decoder side so that no additional motion bits are required.

3. CLASSIFIED PATCH LEARNING AND COMPENSATION

The essential part of our proposed scheme is classified patch learning and compensation. Supposing the current frame is f_{t+1}^{i+1}, the available reconstructions for predicting f_{t+1}^{i+1} are $\hat{f}_t^i$, $\hat{f}_t^{i+1}$, and $\hat{f}_{t+1}^i$. Fig. 2 illustrates our classified patch learning and compensation process, where the solid lines indicate the learning process, while the dashed lines exhibit the compensation steps.

3.1 Classified patch learning

Classified patch learning can be summarized as the following steps.

Step 1: *Simulation of low resolution image*. Images at different resolutions are required in our scheme to study the inter layer correlation. In the CL module, a low resolution image $\breve{f}_t^i$ is obtained by down-sampling the high resolution image $\hat{f}_t^{i+1}$ (as shown in Fig.2 (a)).

$$\breve{f}_t^i = \mathcal{D}\left(\hat{f}_t^{i+1}\right) \tag{2}$$

Step 2: *Generation of references*. There are two references involved in patch learning. First, a interpolated high resolution frame $\breve{f}_t^{i+1}$ ($\breve{f}_t^{i+1} = \mathcal{U}(\breve{f}_t^i)$) is subtracted from $\hat{f}_t^{i+1}$, resulting in the fine reference h_t^{i+1}, as shown in Fig.2 (b) .

$$h_t^{i+1} = \hat{f}_t^{i+1} - \breve{f}_t^{i+1}. \tag{3}$$

Second, the high frequency component of $\breve{f}_t^{i+1}$ brings in the simplified reference $\tilde{h}_t^{i+1}$ (as shown in Fig.2 (d)) which is the counterpart of the fine signal h_t^{i+1}.

$$\tilde{h}_t^{i+1} = \breve{f}_t^{i+1} * G_H, \tag{4}$$

where G_H is a high pass filter. The two references, h_t^{i+1} and $\tilde{h}_t^{i+1}$, are inputs for classified patch learning.

Step 3: *Classification of patches*. Given a reference image, a rotated match filtering is introduced that involves the computation of two factors – magnitude and orientation – at every point. Mathematically, the orientation is estimated as

$$\theta^*(f) = \arg\max_\theta |f * \boldsymbol{h}_\theta * G_\mathrm{D}|\,, \tag{5}$$

and the magnitude is calculated by

$$M(f) = |f * \boldsymbol{h}_{\theta^*} * G_\mathrm{D}|\,, \tag{6}$$

where G_D stands for derivative Gaussian filtering and $\boldsymbol{h}_\theta$ is a rotation matrix. Then the structural points are identified by checking out the local maximum in the magnitude spectrum (as indicated in Fig.2 (c) by the white points). The other points in the image are regarded as textural ones.

Accordingly, two kinds of patches, the structural and textural patches, are extracted from images at integer structural and textural positions, respectively. The patch size is *n*×*n*.

Step 4: *Design of patch databases*. There are two databases generated in the classified patch learning. The structure database $D_S = \{(v_L^m, v_H^m)\}$ consists of pairs of collocated structural patches extracted from the simplified

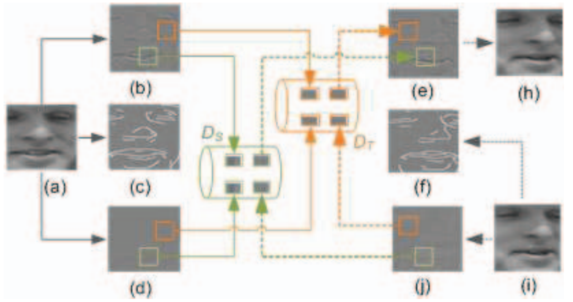


Fig. 2. Illustration of our classified patch learning and compensation. The solid lines indicate the learning process and the dashed lines exhibit the compensation process. Green and orange lines present procedures for structural and textural patches, respectively. Images involved in learning process are (a) high resolution reconstruction $\hat{f}_t^{i+1}$, (b) fine reference h_t^{i+1}, (c) structural map of (a), (d) simplified reference $\tilde{h}_t^{i+1}$. Images involved in the compensation process are (e) fine predication P_{t+1}^{i+1}, (f) structural map of (i), (j) simplified reference $\tilde{h}_{t+1}^{i+1}$, (h) patch compensated prediction B_{t+1}^{i+1}, (i) up-sampled version of low resolution reconstruction $\breve{f}_{t+1}^{i+1}$.

and find references, and the textural database $D_T = \{(u_L^n, u_H^n)\}$ is composed of collocated textural patches as well. Here the superscript m and n are the patch indices, and subscript L and H indicate the simplified and fine patches that extracted from simplified and fine references, respectively. The databases can be further clustered by using a k-mean clustering method.

3.2 Classified patch compensation

The classified patch compensation is illustrated in Fig. 2 by the dashed lines. During compensation, a simplified reference (Fig.2 (j)) is generated by (4) from the up-sampled low resolution reconstruction (Fig.2 (i)) at the current frame. Similar to the learning process, each patch extracted from the simplified reference is classified into either the structural or textural set and accordingly retrieves its fine patch from the corresponding database. The structure map is denoted in Fig.2 (f).

Taking structural patches as example, for each input patch v, an approximate nearest neighbor (ANN) search [9] is used to retrieve a candidate fine patch v_H^j subject to

$$v_L^j \;= \text{argmin}_{v_L' \in D_s} d(v_L', v), \qquad (7)$$

where $d(\cdot)$ stands for the Euclidean distance. In other words, a candidate fine patch is selected when its coupled simplified patch is the most similar one to the input reference patch.

As patches can be overlapped in the fractional compensation, pixel values in the overlapped regions are determined by an average operator. Thus a fine prediction Fig.2 (e)) is obtained by

$$P_{t+1}^{i+1}(x,y) = \frac{1}{R}\sum_{r=1}^{R} v_H^r(x,y)\ , \qquad (8)$$

where R is the number of overlapped patches at (x,y). Together with the up-sampled frame $\check{f}_{t+1}^{i+1}(x,y)$ ($\check{f}_{t+1}^{i+1}(x,y) = \mathcal{U}(\hat{f}_{t+1}^{i}(x,y))$ as exhibited in Fig.2 (i)), the final blended prediction $B_{t+1}^{i+1}(x,y)$ (Fig.2 (h)) is achieved by

$$B_{t+1}^{i+1}(x,y) = \alpha \cdot \tilde{f}_{t+1}^{i+1}(x,y) + \beta \cdot P_{t+1}^{i+1}(x,y)\ , \qquad (9)$$

where α and β are weighted factors.

3.3 Why classification?

Here we used Receiver Operating Characteristics (ROC) curves to demonstrate the effectiveness of the classification. A ROC curve presents the relationship between hit rate h and match error e. Let p denote a test patch and p' be its nearest sample in the database. The match error is defined as

$$e(p) = \|p - p'\| / \|p\|. \qquad (9)$$

For a given match error e, the hit rate h represents the percentage of test data whose match errors are less than e. Clearly, at a given match error, the higher the hit rate is, the better the mapping efficiency.

Fig.3 shows the ROC curves. Here, we perform three tests: (a) testing on general patches (without classification), (b) testing on structural patches only, and (c) testing on textural patches only. In each test, 100 empirical images (from [10] and [11]) are used. These images are equally divided into a test image set and a training image set. 10^5 test patches are uniformly selected from training as well as mapping.

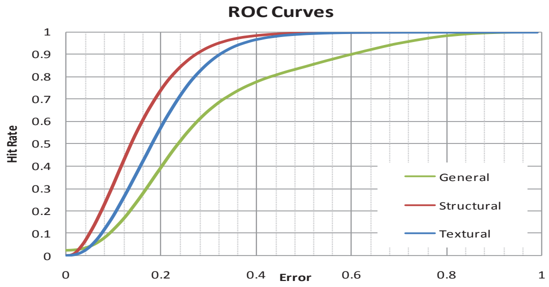


Fig.3 ROC curves of classified patch mapping.

It can be observed that, compared with the general patch case, the classified patches result in a higher hit rate under the same match error and same training set size. As shown in Fig. 3, more than 70% of structural patches and nearly 60% of textural patches have a match error less than 0.2, while only 40% of general patches can fall within the same error range. It indicates that classified patches are relatively low dimensional. The classification facilitates the learning process and enhances the effectiveness of compensation, which enables the high compensation performance in our proposed coding scheme.

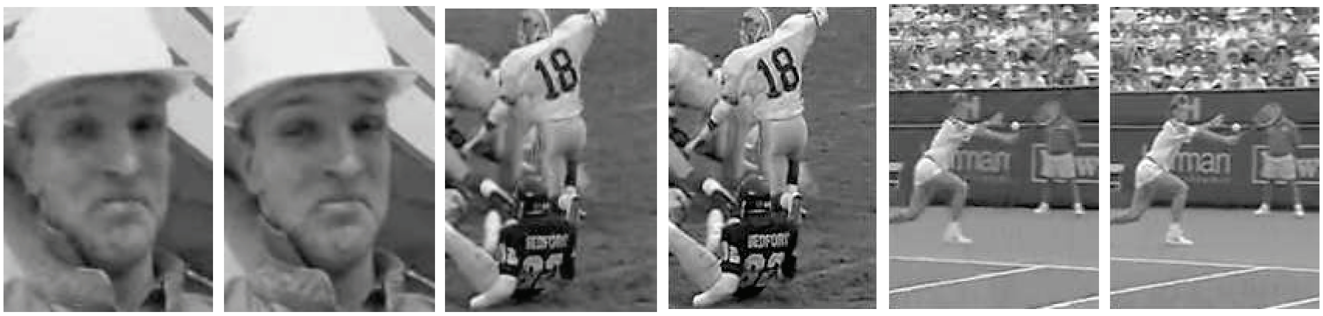
4. EXPERIMENTAL RESULTS

Before evaluation, we would like to point out that our proposed classified patch learning is only used in the coding of luminance component in the current scheme.

In our experiments, the patch size is 11×11. The high pass filter in (4) is performed as subtracting the low-frequency component which is calculated by convolution with a Gaussian kernel, from the original signal. α and β in (9) are 1.0. The simulation to evaluate our scheme is implemented with JSVM 10 [12]. For each sequence, only the first frame is coded as I frame; the others are coded as P frame. Two scalable layers, the QCIF base layer and the CIF spatial enhancement layer, are generated at frame rate 15. In the tests, the base layer quantization parameter (QP_b) is 30, while the enhancement layer quantization parameter (QP_e) changes from 27 to 45 at intervals of 3.

The coding performance of our proposed spatially scalable coding scheme is evaluated in Fig. 5 in terms of PSNR. Compared with the current JSVM scheme (denoted as JSVM), our approach (denoted as CL) is able to achieve 1.9dB gain. For the Football sequence, our scheme averagely outperforms JSVM by more than 1.2dB over the tested bit rates. Also, the comparison results of the Foreman and Stefan sequences shows that our scheme achieves more than 1.8dB and 1.9dB improvements over JSVM at low bit rate, respectively,

We also test on the subjective quality of our proposed scheme in comparison with that of JSVM. As shown in Fig. 4, our scheme significantly enhances the perceptual quality



(a) JSVM (b)CL (c)JSVM (d)CL (e)JSVM (f)CL

Fig.4. Visual quality comparison (QP_e=42). (a) and (b) are Foreman 35th frame, (c) and (d) are Football 115th frame, (e) and (f) are Stefan 92nd frame.

of the reconstructed frames, which presents vivid details (faces, grasses, and characters) and clear structures (walls numbers, and lines).

Here we would like to point out that multi-loop decoding is enabled in our proposed spatially scalable coding approach, while the current JSVM 10 is a single-loop decoding scheme. It has been reported that the rate-distortion penalty of single loop restriction in JSVM is found for most sequences to be small while only a few sequences are found with PSNR losses up to 0.7dB [13]. In contrast, our approach significantly improves the coding performance by up to 1.9dB, which obviously makes use of the inter-layer correlation in a much more efficient way.

5. CONCLUSION

In this paper, we propose to exploit the inter layer correlation in spatially scalable video coding by classified patch learning. Pairs of reference patches are extracted from the previous coded frame at different resolutions and classified according to the gradient information to form the structural and textural databases. Based on the two databases, our proposed patch-based compensation scheme derives a high resolution prediction by patch mapping at structural and textural regions, respectively. Our proposed scheme can be readily integrated with the current SVC coding scheme. Experimental results demonstrate the effectiveness of our proposed classified patch learning for spatially scalable video coding.

6. REFERENCES

[1]. H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable extension of the H.264/MPEG-4 AVC video coding standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 9, pp. 1103–1120, Sep. 2007.

[2]. M. Flierl and P. Vandergheynst, “An improved pyramid for spatially scalable video coding,” in Proc. IEEE ICIP 2005, Genova, Italy, 2005.

[3]. T. Wang; C.-S. Park; J.-H. Kim; M.-S. Yoon; S.-J. Ko, “Improved inter-layer intra prediction for scalable video coding,” I Proc. TENCON 2007, pp. 1-4, 2007

[4]. Y. Liu, G. Rath, and C. Guillemot, “Improved Intra Prediction for H.264/AVC scalable Extension”, IEEE MMSP 2007, pp. 247-250, 2007

[5]. W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” IEEE Computer Graphics and Applications, March/April, 2002

[6]. J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum, “Image Hallucination with Primal Sketch Priors”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2003.

[7]. Y. Li, X. Sun, H. Xiong, and F. Wu, “Incorporating Primal Sketch Based Learning into Low Bit-Rate Image Compression,” IEEE ICIP, 2007, vol. III, pp, 173-176, 2007

[8]. F. Wu, X. Sun, “Image compression by visual pattern vector quantization (VPVQ)”, DCC2008, pp. 282-291, 2008

[9]. D. Mount, and S. A. Ann, “Library for approximation nearest neighbor searching,” http://wwww.cs.umd.edu/mount/ANN

[10]. http://r0k.us/graphics/kodak.

[11]. ftp://ftp.kodak.com/www/images/

[12]. T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien, Joint Draft 10, Joint Video Team, JVT-W201, San Jose, CA, USA, April 2007

[13]. H. Schwarz, T. Hinz, D. Marpe, and T. Wiegand, “Constrained inter-layer prediction for single-loop decoding in spatial scalability”, in Proc. IEEE ICIP2005, pp.870-873, 2005

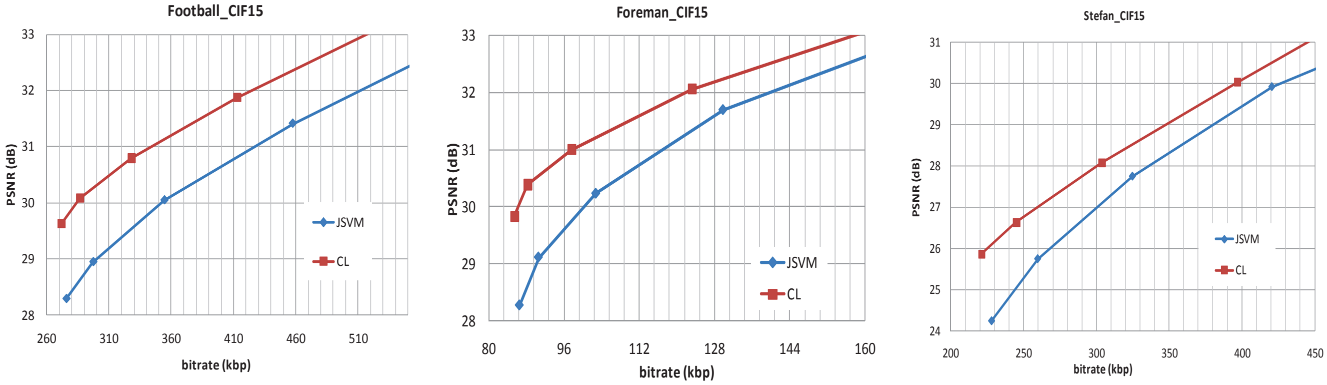


Fig.5. Performance comparison of JSVM and our approach (CL) at 15fps. Test sequences from left to right are Football, Foreman, and Stefan.