FRACTIONAL COMPENSATION FOR SPATIAL SCALABLE VIDEO CODING

Xiaoyan Sun and Feng Wu

Microsoft Research Asia
{xysun,fengwu}@microsoft.com

ABSTRACT

This paper proposes a novel fractional compensation approach for spatial scalable video coding. It simultaneously exploits inter layer correlation and intra layer correlation by learning-based mapping. Instead of using an enhancement layer reconstruction as an entire reference, a set of reference pairs are generated from high-frequency components of both base layer and enhancement layer reconstructions at previous frame. The reference set, which consists of low-resolution and high-resolution patches, can be generated in both encoder and decoder by on-line learning. During the encoding of enhancement layer, a prediction is first gotten from base layer, from which low-resolution patches are extracted. These patches are then used as indices to find the matched high-resolution patches from the reference set. Finally, the prediction enhanced by the high-resolution patches is used for coding. The proposed approach does not need any motion bits. With our proposed FC approach, the performance of H.264 SVC can be improved up to 2.4dB in spatial scalable coding.

Index Terms— video coding, motion estimation, scalable coding, spatial scalability

1. INTRODUCTION

Spatial scalable video coding (SSVC) has been investigated for decades. Among various SSVC coding approaches, the pyramidal layered schemes are well accepted by standards, such as MPEG-2 and H.264/MPEG-4 SVC [1], and extensively investigated by researchers. In a pyramidal layered approach, a base layer bit stream is generated by coding the lowest resolution version of an input video. In addition to the traditional motion prediction within each layer, frames coded at lower resolutions can be up-sampled to form an inter-layer prediction for enhancement layer coding. This inter-layer correlation across neighboring resolutions is expected to be fully exploited to facilitate scalable video coding.

In the state-of-the-art SVC standard [1], inter-layer prediction has been proposed to make use of pixel value, motion and mode information at base layer to predict those at enhancement layer. Subsequently, progresses have been reported by using improved up-sampled filters [3] or subdivided blocks [4] for pixel value or mode prediction. Displacement information has also been introduced in [5] to minimize the difference between up-sampled base layer and enhancement layer at block level. So far, all these schemes focus only on the spatial correlation between two layers. Inter-layer spatial correlation and intra-layer temporal correlation can be exploited only alternatively.

On the other hand, we find a new way to exploit the correlation between different resolution layers. It is enlightened by the image hallucination schemes [8][9] in which databases consisting of co-occurrence image patches at two different resolutions are introduced as priors for image recovery. This idea has been extended to image compression in [6]. It has also been related to the classic vector quantization in [7], where the database is regarded as a codebook and the indices are embedded in the coded low resolution image.

In this paper, we propose a novel fractional compensation approach for spatial scalable video coding in which learning-based mapping is first introduced in inter layer compensation. Reference pairs are extracted from low resolution (LR) and high resolution (HR) reconstruction at previous frame and clustered to form a reference database. This database, rather than a complete reconstructed frame, is then used to enhance a HR prediction from LR reconstruction at current frame by patch mapping. In this way, both the temporal intra-layer correlation and the spatial inter-layer correlation are utilized. This compensation can be performed at fraction level because the reference database as well as the compensation requires only the reconstructed frames and no motion bits are needed. This is the reason that we name this method as fractional compensation. Experimental results demonstrate the effectiveness of the fractional compensation.

The rest of this paper is organized as follows. Framework of the fractional compensation based SSVC is introduced in Section II. Then, our proposed fractional compensation is described in details in Section III. Performance of the proposed approach is evaluated in Section IV. At last, Section V concludes this paper.

2. FRAMEWORK OF CODING SCHEME

The framework of fractional compensation (FC) based SSVC scheme is illustrated in Fig.1. Some modules, such as entropy coding are omitted for simplicity. Our proposed FC is exhibited by the blue-dashed blocks, which consists of three key modules, pair extraction (PE), database generate (DB) and pair compensation (PC).

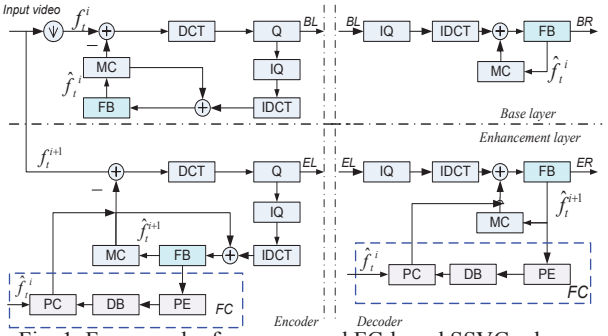


Fig. 1. Framework of our proposed FC-based SSVC scheme

Here we denote an input video by $F^i = \{f_t^i\}$, 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 LR video is generated from the original HR video via a down-sampling process $D(\cdot)$

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

As shown in Fig. 1, the base layer encoding is traditional, only the input frame f_t^i is a down-sampled version of the original one. Notice that the reconstruction $\hat{f}_t^i$ stored in the base layer frame buffer (FB) is input into PC module for enhancement layer coding.

At enhancement layer, a HR reconstruction $\hat{f}_t^{i+1}$is input into PE module for pair extraction, the resulting patch pairs are then clustered and stored in DB. In PC module, patches extracted from LR reconstruct are used as indices for retrieving their HR reference patches. At last, the reference patches are integrated with LR reconstruction in PC to present the FC prediction for enhancement layer coding.

The corresponding decoding process is exhibited on the right side in Fig. 1. It can be observed that the FC can be online performed at the decoder side so that no additional motion bits are required.

Here we would like to point out the different between block and patch. Blocks are not overlapped with each other but patches will. In traditional SSVC schemes, blocks are commonly of size $n \times n$ $(n = 4,8, \text{or} 16)$. In our proposed FC, the patch size is not restricted and can be flexible.

3. FRACTIONAL COMPENSATION

Fig.2 illustrates the basic idea of our proposed FC approach. Assuming the current frame is f_{t+1}^{i+1}, there are three reconstructions, $\hat{f}_t^i$, $\hat{f}_t^{i+1}$, and $\hat{f}_{t+1}^i$, available for predicting f_{t+1}^{i+1}. At time t, an additional reference $\tilde{f}_t^{i+1}$ is generated by either up-sampling $\hat{f}_t^i$ or low-pass filtering $\hat{f}_t^{i+1}$. It is then high-pass filtered to form the simplified reference $\tilde{h}_t^{i+1}$. The difference between HR reconstruction $\hat{f}_t^{i+1}$ and reference $\tilde{f}_t^{i+1}$ is stored in h_t^{i+1}as a sort of fine reference. After patch extraction, we get a reference database composing of pairs of reference patches extracted from simplified reference and fine reference at same positions, respectively.

Similarly, another simplified reference $\tilde{h}_{t+1}^{i+1}$ at time t+1 is generated by up-sampling the base layer reconstruction $\hat{f}_{t+1}^i$ followed by a high-pass filtering. Based on the patches extracted from $\tilde{h}_{t+1}^{i+1}$ and those inside database, the patch compensation is then employed to generate the compensated signal $\hat{h}_{t+1}^{i+1}$ and the prediction p_{t+1}^{i+1} as well.

In the following sections, technologies in FC, including PE, DB and PC, are introduced in details.

3.1. Patch Extraction

In our FC scheme, patches are extracted from high frequency components of images for conducting the database and predictions. In specific, patches in our scheme are centered at edge regions. The reason is twofold. First, smooth region in a HR image can be well approximated by the up-sampled version of its LR image. Second, it has been shown that patch primitives in contour regions are in low dimensionality [8]. Extracting patches at only edge regions can greatly facilitate the following learning and mapping process.

In addition, the dimensionality of patches can be further decreased by removing the effect of remaining low frequency part. A contract map for image f is first calculated as a set of scaling factors for normalization.

$$|f| = E\left(g(f)\right), \quad (2)$$

where $g(\cdot)$ is a low pass filter and $E(\cdot)$ is an energy function.

For each patch $q(x, y)$ centered at position (x, y), its normalized version $\overline{\boldsymbol{q}}$ is obtained and used as reference patch for learning and compensation.

$$\overline{\boldsymbol{q}} = \boldsymbol{q}/|f(x, y)|, \quad (3)$$

where $f(x, y)$ is the value in f at position (x, y).

Notice that, in our proposed FC, there are three reference patch sets, fine patch set P_F, previous simplified patch set P_{PS}, and current simplified patch set P_{CS}, that are extracted from the three references, h_t^{i+1}, $\tilde{h}_t^{i+1}$, and $\tilde{h}_{t+1}^{i+1}$, respectively.

3.2. Database Design

Let $\{\boldsymbol{v}_j, \boldsymbol{u}_j\}$, j= 1, … , J, be the sequence of patch pairs composing the reference database, where $\boldsymbol{v}_j$ and $\boldsymbol{u}_j$ are patches belong to P_F and P_{PS}, respectively. A clustering methods can be employed to optimize the partition cells of $\boldsymbol{v}_j$ and $\boldsymbol{u}_j$ simultaneously. For simplicity, the proposed database is designed on optimizing the partition cells of $\boldsymbol{u}_j$ only, and all the simplified patches are divided into K clusters $\boldsymbol{S}_k, k = 1,2,..,K$, by minimizing the distortion

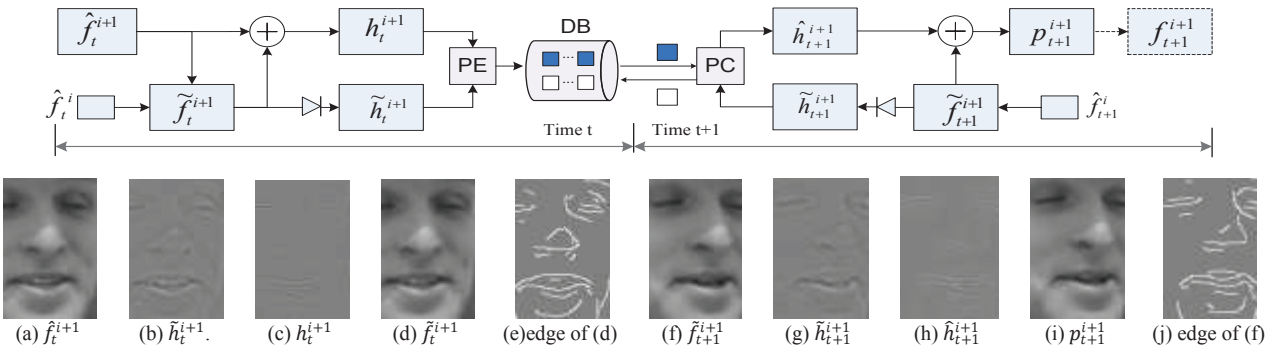
$$d_k = \sum_{k=0}^{K} \sum_{all\ \boldsymbol{u}_i \in \boldsymbol{S}_k} d(\boldsymbol{u}_i, \boldsymbol{u}_k^*), \quad (4)$$

where $\boldsymbol{u}_k^*$ is the centroid point of all the simplified patch $\boldsymbol{u}_i$ belonging to $\boldsymbol{S}_k$, and $d(\cdot)$ stands for the Euclidean distance. The $\boldsymbol{u}_k^*$ is calculated by the nearest neighboring principle

$$\boldsymbol{u}_k^* = \boldsymbol{u}_i, s.t. \min_{\boldsymbol{u}_i \in \boldsymbol{S}_k} \sum_{all\ \boldsymbol{u}_j \in \boldsymbol{S}_k} d(\boldsymbol{u}_i, \boldsymbol{u}_j). \quad (5)$$

Giving a centroid point $\boldsymbol{u}_k^*$, the corresponding fine patch centroid of $\boldsymbol{S}_k$ is determined as

$$\boldsymbol{v}_k^* = \boldsymbol{v}_i, \ if \ \boldsymbol{u}_k^* = \boldsymbol{u}_i. \quad (6)$$



(a) $\hat{f}_t^{i+1}$ (b) $\tilde{h}_t^{i+1}$. (c) h_t^{i+1} (d) $\tilde{f}_t^{i+1}$ (e)edge of (d) (f) $\tilde{f}_{t+1}^{i+1}$ (g) $\tilde{h}_{t+1}^{i+1}$ (h) $\hat{h}_{t+1}^{i+1}$ (i) p_{t+1}^{i+1} (j) edge of (f)

Fig. 2. Illustration of our proposed FC. (a)-(e) are at time t and (f)-(j) are at time t+1. (a) HR reconstruction, (b) simplified reference, (c) fine reference, (d) additional reference, (e) edge of (d), (f) additional reference, (g) simplified reference, (h) compensated reference, (i) resulting prediction, (j) edge of (f).

If there are more than one candidate high frequency patch of $\boldsymbol{v}_k^*$, one of them will be selected as the centroid point using nearest neighboring principle.

At last, the selected simplified patches together with their corresponding high frequency patches form the final database $\boldsymbol{L}$.

3.3. Patch Compensation

During patch compensation, a reference patch $\boldsymbol{u}$ in P_{CS} retrieves a candidate fine patch $\boldsymbol{v_j}$ by an approximate nearest neighbor (ANN) search [11] subject to

$$\boldsymbol{u}_j = \text{argmin}_{\boldsymbol{u}' \in \boldsymbol{L}} d(\boldsymbol{u}', \boldsymbol{u}). \quad (7)$$

In other words, a candidate fine patch is selected when its coupled simplified patch is the most similar one to the input reference patch.

The retrieved patches are then adjusted by multiplying the scaling factors, which are calculated from the simplified reference $\tilde{f}_{t+1}^{i+1}$ as (2), to align with the reconstruction in terms of brightness. The resulting patches are integrated with $\tilde{f}_{t+1}^{i+1}$ to form the final prediction.

Notice that patches can be overlapped in fractional compensation. In this case, the decision of pixel values in the overlapped regions is a problem should be handled during blending. Here we use a straightforward average operator to deal with the problem. For each position (x, y), the blended value of $P_{t+1}^{i+1}(x, y)$ is achieved by

$$P_{t+1}^{i+1}(x, y) = \alpha \cdot \tilde{f}_{t+1}^{i+1}(x, y) + \beta \cdot \frac{1}{R} \sum_{r=1}^{R} v_r(x, y) \quad (8)$$

where R is the number of overlapped patches $\boldsymbol{v}$ at (x, y), and α and β are weighted factors.

Before evaluation, we would like to point out that the proposed FC is only used for the coding of luminance component. Detailed parameters and band-pass filters are discussed in the following section.

4. PERFORMANCE EVALUATION

To test the robustness of the FC, we fix the parameters as follows for all the tests. In our experiments, the patch size is set to 11×11. The high pass filter in Fig.2 is performed as subtracting the low-frequency component, which is calculated by convolution with a Gaussian kernel, from the original signal. The contrast information in (2) is the square root values of the Gaussian smoothed energy signal of pixel intensities. α and β in (8) are 1.0. Edges shown in Fig.2 (e) and (j) are detected by the method proposed in [10]. Patches centered at edge pixels are involved in the fractional compensation.

The simulation to evaluate our scheme is implemented with JSVM 10 [2]. For each sequence, only the first frame is coded as I frame; the others are coded as P frame. All the macroblock modes are enabled. Our proposed FC is treated as the inter-layer intra prediction mode that competes with the other coding modes during enhancement layer coding. Two scalable layers, QCIF base layer and CIF spatial enhancement layer, are generated at frame rate 15. The enhancement layer QP_e changes from 27 to 45 at interval of 3.

The coding performance of our proposed FC-based SSVC scheme is evaluated in Fig. 4. In this test, the base layer QP_b is 30. Compared with the current JSVM scheme, our approach is able to achieve more than 1.8dB gain. We also test on the subjective quality of our proposed scheme in comparison with that of JSVM. As shown in Fig. 5, our scheme significantly enhances the perceptual quality of the reconstructed HR frames, especially at edge regions. When base layer are coded at high bit rate, the SSVC scheme equipped with our proposed fractional compensation is able to achieve 2.4dB gain at shown in Fig. 3.

Here we would like to point out that multi-loop decoding is enabled in our FC based SSVC 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 [12]. Differently, our FC approach significantly improves the coding performance up to 2.4dB, which obviously make use of the inter-layer correlation in a much more efficient way.

5. CONCLUSION

This paper proposes a novel fractional compensation approach for spatial scalable video coding. It makes use of the inter-layer correlation at previous coded frames for current

HR frame coding. Rather than a complete reconstructed frame, a set of reference pairs are extracted at fraction level from low and high resolution at same positions and clustered to form a reference database. This reference database is then used to conduct a HR prediction from the current LR reconstruction by find the best match patch from the reference database. Experimental results show the effectiveness of our proposed fractional compensation.

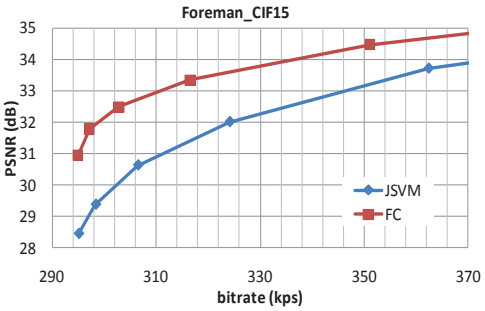


Fig. 3. PSNR comparison of Foreman sequence (QP_b = 20)

5. 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]. 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

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

[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]. 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

[6]. 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

[7]. F. Wu, X. Sun, "Image compression by visual pattern vector quantization (VPVQ)", Data Compression Conference 2008, pp. 282-291, 2008

[8]. 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.

[9]. W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning Low-level Vision," International journal of Computer Vision, 40(1), pp25-47, 2000.

[10]. M. Jacob, M. Unser, "Design of Steerable Filers for Feature Detection Using Canny-Like Criteria," IEEE Trans. of Pattern Analysis and Machine Intelligence, Vol. 26, no. 8, pp. 1007-1019, Aug. 2004

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

[12]. 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

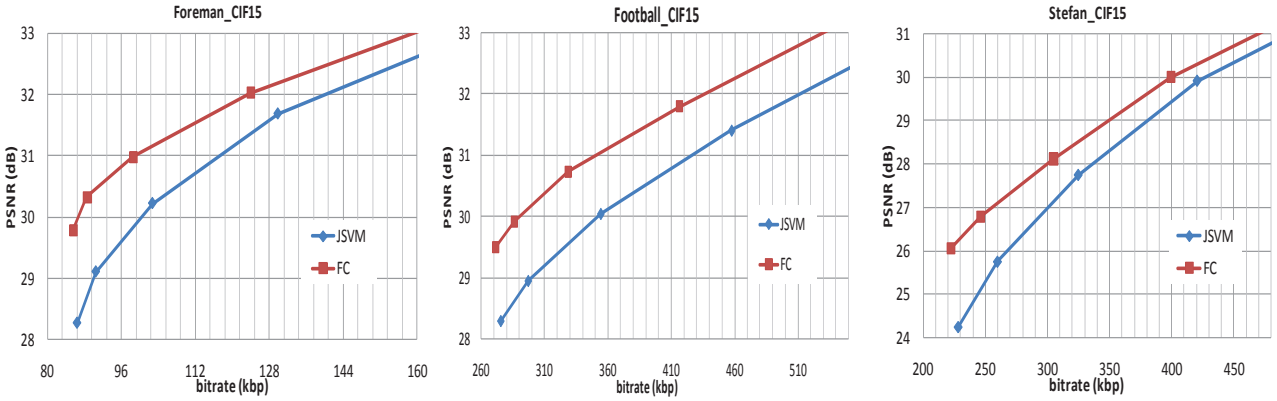


Fig. 4. Performance comparison of JSVM and our FC-based approach at 15fps, QP_b=30. Test sequences from left to right are Foreman, Football and Stefan.

(a)JSVM (28.01dB) (b) FC (30.02dB) (c) JSVM (27.96dB) (d)FC (29.39dB)

Fig. 3. Visual quality comparison. (QP_b=30, QP_e =45) (a) and (b) are Foreman 66th frame; (c)and (d) are Football 31st frame.