SIDE INFORMATION GENERATION FOR DISTRIBUTED VIDEO CODING

Ligang Lu, Dake He, Ashish Jagmohan

IBM T. J. Watson Research Center
Emails: lul@us.ibm.com, dakehe@us.ibm.com, ashishja@us.ibm.com

ABSTRACT

Side information(SI) generation is one of the key components of a Wyner-Ziv coder. In this paper we present a novel multi-frame SI generation approach which uses adaptive temporal filtering to estimate the pixel values for SI and motion vector filtering for refinement. For temporal filtering, we derive the optimal mean squared error temporal filter when the noise can be evaluated, and propose a similarity weighted temporal filter when the knowledge of the noise is not available. The temporal filter adapts on the quality of the motion estimation. The quality of SI generation is further improved by using motion vector filtering to reduce the noise effect from motion estimation. Experimental results indicate that the proposed SI generation approach yields good performance in terms of SI quality and conditional entropy.

Index Terms— Wyner-Ziv coding, Slepian-Wolf decoding, distributed video coding, side information generation, adaptive temporal filtering, motion vector filtering.

1. INTRODUCTION

Significant research efforts have been devoted to develop practical distributed video coding (DVC) systems for emerging applications, such as distributed video surveillance network, mobile visual communications, etc. Video encoders for such applications have highly restricted computation and power resources and thus need to adopt low complexity algorithms, while decoders at a central location may have the resources for sophisticated signal processing tasks. Therefore current video coding standards, such as MPEG-x and H.26x, which have been developed for the traditional complex encoder-simple decoder paradigm, are not suitable for these applications. The work by Slepian-Wolf [1] and Wyner-Ziv [2] has laid the theoretical ground for a new video coding paradigm, wherein a low complexity encoding and high complexity decoding system using distributed coding principles may approach the operational rate-distortion performance achieved by traditional systems. In [1] Slepian and Wolf analyzed the lossless coding case and showed that, given a source X and correlated *decoder-only* SI Y, X can be compressed to the the conditional entropy $H(X|Y)$.[1] In [2] Wyner and Ziv analyzed the lossy compression case and derived the corresponding rate-distortion bound.

In recent years, several papers have proposed distributed video communication systems based on the Wyner-Ziv theorem [3, 4]. Figure 1 depicts a general DVC system, wherein the source video sequence is split into two sub sources. The frames from one sub source are encoded by a Wyner-Ziv (W-Z) encoder. The frames of the other sub source are encoded by a traditional encoder, e.g., an H.264 encoder. At the decoder, previously decoded frames are used as reference frames to generate the SI Y. Then the Wyner-Ziv encoded source frame X is decoded by exploiting the correlation between X and Y

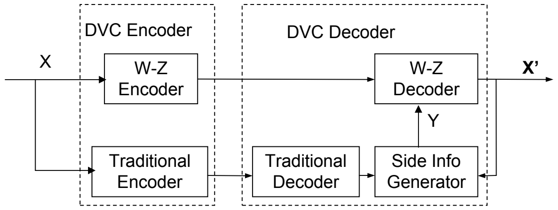

Fig. 1. Block diagram of DVC system.

In order to achieve efficient coding performance, a central issue in a DVC system is the generation of high-quality SI at the decoder. Clearly, the higher the correlation between the source X and the SI Y, the better is the coding performance that can be achieved. Recently, several papers have studied schemes for SI generation. Li and Delp [5] compared the rate-distortion effectiveness of conventional motion compensated prediction to that of motion compensated side estimation, and proposed a multi-reference scheme for SI estimation. Klomp et al. [6] showed that sub-pel motion compensation could be used to yield improved SI interpolation for Wyner-Ziv decoding. Natario et al. [7] proposed a scheme to improve SI interpolation by motion smoothing for pixel domain W-Z video coding. While these schemes reported improvements in SI generation, there is a need for improved SI generation, in order to close the relatively large gap in rate-distortion performance between DVC and traditional video coding schemes.

In this paper we present a novel multi-frame SI generation approach for Wyner-Ziv coding. The proposed approach is based on the use of temporal and motion vector filtering to yield more accurate SI. The organization of this paper is as follows. In the next section, we discuss SI generation using multi-frame adaptive temporal filtering. Also in Section 2 we describe a motion vector refinement filtering process which further improves the quality of the generated SI. Finally, in Section 3, we present experimental results to evaluate the proposed approach.

2. SIDE INFORMATION GENERATION FOR DVC

In this section, we propose an efficient SI generation scheme for DVC. The diagram of our scheme is shown in Figure 2. As shown in Figure 2, our SI generator consists of a similarity estimator, a motion estimator, an adaptive temporal filter, and a motion vector filter.

This work was supported in part by the National Aeronautics and Space Administration of the United States of America under Grant NNS05AA75C.

[1] This is the same rate as in the case where Y is available both at the encoder and decoder.

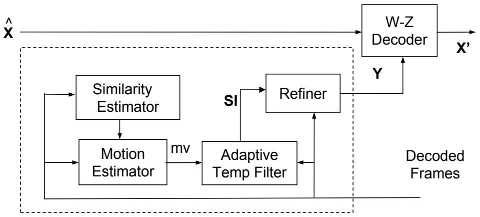

Fig. 2. Block diagram of proposed SI generation system.

The similarity estimator measures the similarity of local statistical features, and produces a similarity measure. The similarity measure is then used in motion estimation and motion compensated temporal filtering, as well as resolving empty mapping or multiple mappings in SI generation. The principles and exemplary schemes of using the similarity measure in SI extrapolation and interpolation have been described in [8]. Our motion estimator employs a new minimization scheme by introducing a similarity constraint in the objective distance function to reduce the search algorithm's sensitivity to various noises and slow pixel intensity changes. It also effectively trades off minimizing the pixel difference with the motion smoothness in adjacent similar neighborhood [9]. In this paper, our focus is on the new developments in SI generation using multi-frame adaptive temporal filtering and motion vector refinement filtering.

2.1. Multi-frame motion compensated adaptive temporal filtering

To generate the SI $\mathbf{Y}_N$ for decoding the current W-Z frame $\mathbf{X}_N$, we use M previous decoded frames $\mathbf{X}'_{N-M}, \mathbf{X}'_{N-M+1}, \cdots, \mathbf{X}'_{N-1}$, and one decoded future reference frame $\mathbf{X}'_{N+1}$ in motion compensated adaptive temporal filtering. Note that the decoded frame can either be a decoded reference frame or a decoded W-Z frame. Figure 3 describes the case when $M = 2$. As shown in the figure, we use motion estimate to find the best matching blocks between the reference frames $N-1$ and $N-2$ and between $N-1$ and $N+1$. The pixel values of SI are estimated using the multi-frame adaptive temporal filter described below.

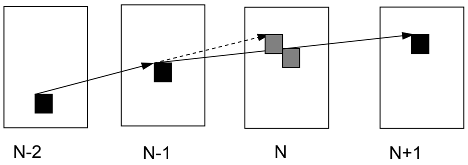

Fig. 3. Multi-frame temporal filtering.

A. Ideal case: Minimum mean squared error filter

We wish to estimate the present value X_n of a discrete-time process $X = \{X_i\}_{i=1}^{\infty}$ in terms of the observations of the sum

$$X'_i = X_i + Z_i \tag{1}$$

where $Z = \{Z_i\}_{i=1}^{\infty}$ is a discrete-time noise process independent of X. The non-causal estimate Y_n is the output of a linear time-invariant non-causal system [10]

$$Y_n = \sum_{k=1}^{M} h_k X'_{n-k} + h_{-1} X'_{n+1} \tag{2}$$

with the input X'_n and delta response h_n. Using the orthogonality principle that X_n is orthogonal to the estimation error, the optimal mean squared estimation yields

$$R_{XX'}(m) = \sum_{k=1}^{M} h_k R_{X'X'}(m-k) + h_{-1} R_{X'X'}(m+1) \tag{3}$$

where $R_{XX'}$ and $R_{X'X'}$ are the cross-correlation and autocorrelation functions, respectively. Here we have assumed that the underling processes are joint wide-sense stationary (WSS). If we have the knowledge about the noise v, the coefficients of the MSE optimal filter can be obtained by solving (3). For example, if $M = 2$ and using the property of $R(\tau) = R(-\tau)$ for real correlation function , we have

$$\begin{aligned}
h_2 &= \frac{\Delta_1}{\Delta}, h_1 = \frac{\Delta_2}{\Delta}, h_{-1} = \frac{\Delta_3}{\Delta}, \\
\Delta_1 &= \begin{vmatrix} R_{XX'}(1) & R_{X'X'}(2) & R_{X'X'}(0) \\ R_{XX'}(1) & R_{X'X'}(0) & R_{X'X'}(2) \\ R_{XX'}(2) & R_{X'X'}(2) & R_{X'X'}(3) \end{vmatrix}, \\
\Delta_2 &= \begin{vmatrix} R_{X'X'}(3) & R_{XX'}(1) & R_{X'X'}(0) \\ R_{X'X'}(1) & R_{XX'}(1) & R_{X'X'}(2) \\ R_{X'X'}(1) & R_{XX'}(2) & R_{X'X'}(3) \end{vmatrix}, \\
\Delta_3 &= \begin{vmatrix} R_{X'X'}(3) & R_{X'X'}(2) & R_{XX'}(1) \\ R_{X'X'}(1) & R_{X'X'}(0) & R_{XX'}(1) \\ R_{X'X'}(1) & R_{X'X'}(2) & R_{XX'}(2) \end{vmatrix}, \\
\Delta &= \begin{vmatrix} R_{X'X'}(3) & R_{X'X'}(2) & R_{X'X'}(0) \\ R_{X'X'}(1) & R_{X'X'}(0) & R_{X'X'}(2) \\ R_{X'X'}(1) & R_{X'X'}(2) & R_{X'X'}(3) \end{vmatrix}.
\end{aligned}$$

B. Empirical case: Similarity weighted temporal filter

It is usually nontrivial to evaluate the noise since it is a combination of multiple kinds of noises, including channel noise and source coding noise. We present here an empirical filter evaluated in our practice. After the motion estimator has found the best matching blocks B_{N-2}, B_{N-1}, and B_{N+1} in the reference frames $N-2$, $N-1$, and $N+1$, respectively, the SI estimate is given by

$$B_N = \begin{cases} \frac{a_2 \gamma_{N-2,N-1} B_{N-2}}{a_2 \gamma_{N-2,N-1} + (a_1 + a_{-1}) \gamma_{N-1,N+1}} + \frac{(a_1 B_{N-1} + a_{-1} B_{N+1}) \gamma_{N-1,N+1}}{a_2 \gamma_{N-2,N-1} + (a_1 + a_{-1}) \gamma_{N-1,N+1}} & \text{if } \gamma_{N-2,N-1} \\ & \gamma_{N-1,N+1} \geq T \\ \frac{a_1 B_{N-1} + a_{-1} B_{N+1}}{a_1 + a_{-1}} & \text{otherwise} \end{cases}$$

where T is a constant threshold; a_{-2}, a_{-1}, and a_1 are weights; $\gamma_{N-2,N-1}$ (or $\gamma_{N-1,N+1}$, respectively) is the cross-correlation coefficient between B_{N-2} and B_{N-1} (or B_{N-1} and B_{N+1}, respectively). Here the adaptation of the filter coefficients are based on the values of the cross-correlation coefficients which in fact measure the similarity between the two matched blocks and indicate the relative quality of the motion vectors. In the next section we will describe our new refinement filtering process.

2.2. Multi-frame motion vector refinement filtering

In the above subsection, it is shown that temporal filtering can be used to reduce the noise effect in estimating the pixel value in X_N for a given motion vector. In this subsection, we propose a refinement scheme that uses filtering to reduce the noise effect from motion estimation in the multi-frame setting. The scheme is conceptually simple, easy to implement, and delivers robust and good performance in practice (see the experimental results in the next section).

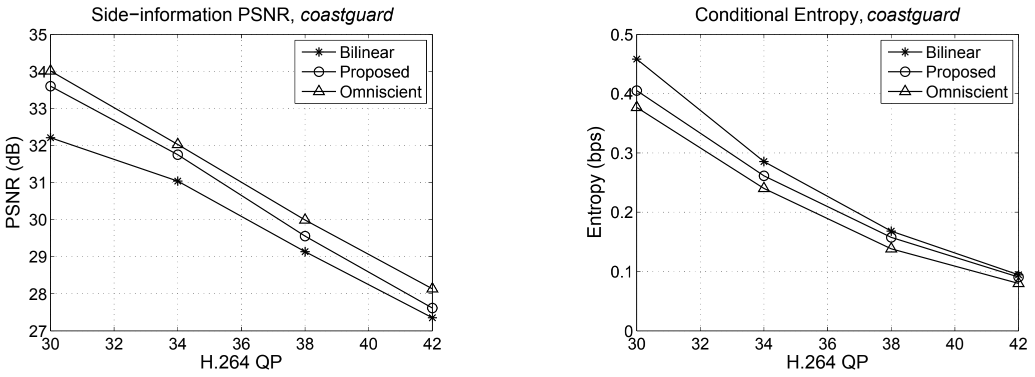

Fig. 4. Side information comparison for CIF sized *coastguard* sequence. (a) Side information PSNR vs. Quantization parameter (QP) of H.264 coded frames. (b) Conditional entropy of quantized source given SI vs. H.264 QP.

Before describing our filtering scheme for motion vectors, we observe that in motion estimation, a motion vector is identified by searching for a pair of image blocks in a predefined range that minimizes an objective distance function. In H.264, for example, a motion vector is identified for a given image block B in the current frame X_N by using the following formula,

$$\vec{v^*}(B) \triangleq \arg\min_{\vec{v}\in\mathcal{V}} d(B, T_{\vec{v}}(B, X'_i)), \qquad (4)$$

where $\mathcal{V}$ is a set of predefined motion vectors, d denotes the mean absolute difference between two same-sized image blocks, X'_i, $i \neq N$, is a previously decoded frame, and $T_{\vec{v}}(B, X'_i))$ is a block in X'_i resulting from moving(translating) B onto X'_{N-1} according to the motion vector $\vec{v}$. From a statistical point of view, (4) is equivalent to maximum likelihood(ML) decoding of a motion vector $\vec{V}$. (4) is further equivalent to maximum *a posteriori*(MAP) decoding if the following two conditions are satisfied:

C1) The random variable $\vec{V}$ is uniformly distributed over $\mathcal{V}$;

C2) Given $\vec{V}$ and B, the noise introduced in the motion (translation) is additive, Laplacian with zero mean, and independent of $\vec{V}$ and B (Here B is regarded as a random variable by abusing notation).

In the case of SI generation, the problem of motion estimation get more complicated because the current frame is unknown to the decoder. In order to get robust motion estimation, in [8] and [9], the distance function d in (4) is replaced with a more sophisticated similar measure that takes into account local statistical features in addition to the mean absolute difference. In the following, we further refine the scheme by using filtering to reduce the noise effect from motion estimation.

Without losing generality, consider the case where $M = 1$, and we would like to generate a SI frame Y_N from X'_{N-1} and X'_{N+1}. Using (4), one can estimate a motion vector $\vec{v}_B$ for any block B in X'_{N-1} (corresponding to X'_{N+1}). Projecting B according to $\vec{v}_B$ onto the imaginary plane of X_N, we can estimate a block $\hat{B}$ in Y_N. However, in general the projection of B does not fall into the integer-pel grid, and thus estimating $\hat{B}$ (which is in the integer-pel grid) from the projection will introduce noise. Our refinement scheme is designed to reduce the effect of such quantization noise in generating Y_N.

Making the usual continuity assumption that for any B' in a small neighborhood of B, we have $||\vec{v}_B - \vec{v}_{B'}||_1 \leq \delta$, where δ is small real number, and $||\cdot||_1$ denotes the L1 norm between two vectors. In view of this assumption, the quantization noise described above can be effectively reduced if we estimate $\hat{B}$ by using filtering, i.e.,

$$\hat{B} = \sum_{B\in\mathcal{N}(\hat{B})} \frac{1}{|\mathcal{B}|} B, \qquad (5)$$

where $\mathcal{N}(\hat{B})$ denotes the set of blocks B in X_{N-1} such that the projection of B according to $\vec{v}_B$ on X_{N-1} is within sub-pel distance of $\hat{B}$.

Furthermore, one can assume that each $B \in \mathcal{N}(\hat{B})$ is equally likely without additional prior knowledge. Let d_B denote the minimum mean absolute difference at which $\vec{v}_B$ is determined by using (4). Assuming that conditions $C1$ and $C2$ are satisfied, we see that (5) can be refined as follows.

$$\hat{B} = \sum_{B\in\mathcal{N}(\hat{B})} \frac{e^{-d_B}}{\sum_{B\in\mathcal{N}(\hat{B})} e^{-d_B}} B, \qquad (6)$$

where e denotes the base of the natural logarithm function.

Finally, we note that our refinement filtering scheme can be combined with temporal filtering in the interested multi-frame setting. The performance of the combined scheme is experimentally evaluated in the next section.

3. RESULTS

To evaluate the performance of the presented SI generation algorithm, we present results for the luminance component of the standard 352 × 288 (CIF) sized video sequences *coastguard* and *mobile_and_calendar*. The sequences are coded using the GOP format $IWPWP\ldots$, where I and P frames denote H.264 coded intra-predicted and single-list inter-predicted frames respectively, and W frames denote Wyner-Ziv coded frames. While decoding each W frame, the decoder generates SI using previously decoded temporally neighboring H.264 frames as predictors.

The goodness of the generated SI is measured using two key metrics: (1) The PSNR of the SI compared to the original source frame, averaged over all W frames; and (2) The conditional entropy[2]

[2] From the Slepian-Wolf theorem [1], the conditional entropy of a discrete source given the decoder SI is the minimum rate required to correctly reconstruct the source.

of the quantized source frame, conditioned on the generated decoder SI, averaged over all W frames. The conditional entropy is computed using the memoryless Laplacian assumption, as this is the favored assumption in current Wyner-Ziv coders [3]. We compare the described metrics for the following: (1) The proposed SI generation algorithm; (2) A *bilinear* SI generation algorithm which uses a full quarter-pel motion search followed by bilinear interpolation to generate SI; and (3) Side information generated by an *omniscient* H.264 coder which uses the original source frame to perform motion estimation, and uses the same motion estimation algorithm as the JM11 reference encoder [11].

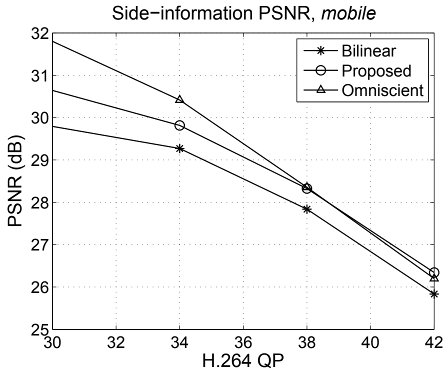

Fig. 5. Side information PSNR comparison for CIF sized *mobile and calendar* sequence.

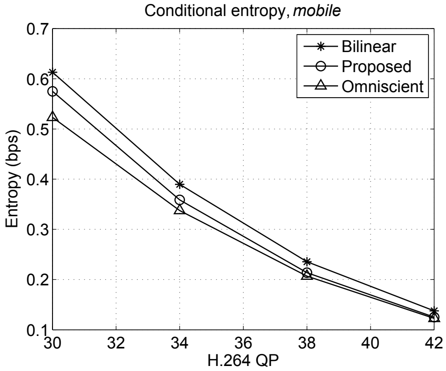

Fig. 6. Comparison of conditional entropy of quantized source given SI for CIF sized *mobile and calendar* sequence.

Figure 4 compares the performance of the three SI generation methods for the CIF-sized *coastguard* sequence. As can be seen from Figure 4(a), the quality of the SI generated using the proposed method, as measured by the PSNR compared to the source, lies between the quality of the SI generated by the bilinear method and the omniscient H.264 coder. For low quantization parameters, the proposed method generates SI which is much better than the bilinear method, and is almost as good as the omniscient H.264 coder. Figure 4(b) shows the conditional entropy for the three methods. Again, as can be seen the proposed method yields SI which is very close to the SI generated by the omniscient H.264 coder, for lower quantization parameters.

Figures 5-6 shows similar results for the CIF-sized *mobile _and _calendar* sequence. The important difference is that, in this case, the performance gap between the proposed method and the omniscient encoder decreases with increasing quantization. We hypothesize that this difference is due to the spatial characteristics of the two sequences; *coastguard* consists of random texture which is especially difficult to interpolate unambiguously from poorly quantized predictors, while *mobile_and_calendar* contains more regular structures which are easier to interpolate. Also, as can be seen, our use of motion vector filtering and temporal filtering allows the generated SI to be even better than the SI generated by the omniscient H.264 coder, in certain cases. Finally, in terms of performance of the entire Wyner-Ziv system, we note that for *mobile_and_calendar*, for instance, the Wyner-Ziv coder using the proposed SI is about 0.5 dB inferior to the Wyner-Ziv coder using the SI generated by the omniscient H.264 coder, for quantization parameter 30. At low rates, the Wyner-Ziv system using the proposed SI generation method is superior to that using the omniscient H.264 SI. Greater details can be found in [12].

4. REFERENCES

[1] D. Slepian and J.K. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inform. Theory*, Vol. 19, pp. 471–480, 1973.

[2] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, Vol. 22, pp. 1–10, 1976.

[3] A. Aaron, E. Setton, and B. Girod, “Toward practical Wyner-Ziv coding of video,” in *Proc. IEEE Intl. Conf. Image Processing*, 2003.

[4] R. Puri and K. Ramchandran, “PRISM: A new robust video coding architecture based on distributed compression principles,” in *Allerton Conference on Communication, Control and Computing*, 2002.

[5] L. Zhen and E. J. Delp, “Wyner-Ziv video side estimator: conventional motion search method revisited.” in *Proc. of ICIP*, pp. 825–828, Genova, Italy, 2005.

[6] S. Klomp, Y. Vatis, and J. Ostermann, “Side information interpolation with sub-pel motion compensation for Wyner-Ziv decoder,” in *Proc. Intl. Conf. on Signal Processing and Multimedia Appl.*, Setubal, Portugal, 2006.

[7] L. Natario, C. Brites, J. Ascenso, and F. Pereira, “Extrapolating side information for low-delay pixel-domain distributed video coding,” in *Proc. Intl. Workshop on Very Low Video Coding*, Sardenha, Italy, 2005.

[8] L. Lu and V. Sheinin, “Side information generation for low complexity video coding systems based on Wyner-Ziv theorem,” in *Proc. Intl. Sym. on Broadband Multimedia Systems and Broadcasting*, Las Vegas, 2006.

[9] L. Lu and V. Sheinin, “Motion estimation with similarity constraint and its application to distributed video coding,” submitted to *ICME*, 2007.

[10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes, 3rd Edition*. McGraw-Hill, 1991.

[11] Joint Video Team Reference Software, Version 11 (JM11), http://iphome.hhi.de/suehring/tml/download.

[12] D. He, A. Jagmohan, L.Lu and V. Sheinin, “An efficient low complexity Wyner-Ziv video codec”, in preparation.