Multiterminal Video Coding

Yang Yang, Vladimir Stanković†, Wei Zhao, and Zixiang Xiong
Dept of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843.
†Dept of Communication Systems, Lancaster University, Lancaster, LA1 4WA, UK.

***Abstract*— Following recent works on the rate region of the quadratic Gaussian two-terminal source coding problem and limit-approaching code designs, this paper examines multiterminal source coding of two correlated video sequences to save the sum rate over independent coding. Specifically, the first video sequence is coded by H.264 and used at the joint decoder to facilitate Wyner-Ziv coding of the second video sequence. An efficient stereo matching algorithm based on loopy belief propagation is then adopted at the decoder to produce pixel-level disparity maps between the corresponding frames of the two decoded video sequences on the fly. Based on the disparity maps, side information for both motion vectors and motion-compensated residual frames of the second sequence are generated at the decoder before Wyner-Ziv encoding. Preliminary results on stereo video sequences using H.264 in conjunction with LDPC codes for Slepian-Wolf coding of the motion vectors show savings in terms of the sum rate when compared to separate coding at the same video quality.**

I. Introduction

Multiterminal (MT) source coding [1] is gaining research interest lately due to its potential applications in distributed sensor networks and distributed multiview video coding. Theoretical limit of MT source coding of jointly Gaussian sources was given recently in [2] for the direct setting (with two encoders) where the encoders directly observe the sources, and in [3], [4] for the indirect/CEO setting where the encoders observe independently corrupted versions of the same source. Practical MT code designs based on generalized coset codes were provided by Pradhan and Ramchandran in [5]. In earlier works, we proposed a framework for practical MT source coding based on Slepian-Wolf coded quantization [6], [7], which employs the optimal approach of vector quantization followed by Slepian-Wolf coding (SWC) [8]. However, the code designs in [5], [6], [7] are for ideal Gaussian sources assuming *a priori* known correlation. When dealing with practical sources (e.g., video), correlation modeling is one of the key issues in efficient MT video coding. In this work, we focus on MT video code design for two correlated video sequences captured by calibrated stereo cameras.

In general, effective coding of a single/monocular video sequence necessitates exploitation of both spatial and temporal redundancies within the sequence. H.264 [9] provides the currently most efficient solution by using motion estimation/compensation to strip off the temporal redundancy between frames, the DCT of the resulting motion-compensated residual frames for energy compaction and decorrelation, and variable-length coding for compression.

For stereo video sequences synchronously captured by two calibrated video cameras, the compression efficiency can be further improved by exploiting the inter-sequence correlation (as done in the MPEG-2 stereo video coding standard [10]) in a joint encoding setup.

For MT video coding, although the encoders cannot communicate with each other, the binocular correlation between the stereo pair can still be extracted from the 3D geometric information of the cameras. This leads to *stereo matching*, which is a fundamental problem in stereo vision, and has been extensively studied in the past by many researchers. Assuming knowledge of the stereo camera configuration, classical stereo matching attempts to compute a disparity/depth map from a stereo image pair.

In general, stereo matching can be formulated as an optimization problem that minimizes the image dissimilarity energy, e.g., squared intensity difference, absolute intensity difference, and shift absolute difference [11]. Boykov *et al.* [12] and Kolmogorov and Zabih [13] presented efficient graph-cut based stereo algorithms, which find a smooth disparity map that is consistent with the image intensities. Geiger *et al.* [14] derived an occlusion process and a disparity field using dynamic programming. Based on Markov random fields, Sun *et al.* [15] proposed a stereo algorithm using belief propagation (BP), which considers three coupled Markov random fields: a smooth disparity field, a spatial line process, and a binary occlusion process. Quantitative evaluations of different stereo algorithms in terms of "bad" pixel percentage (available at http://cat.middlebury.edu/stereo) showed that the BP based algorithm [15] is among the most efficient.

According to the theory [2], a technique that integrates the best stereo matching algorithm (that handles the binocular redundancy) with the most efficient H.264 monocular video compression standard (that removes spatial and temporal redundancies) is potentially the best solution for MT video coding. With the powerful H.264, one approach to MT video coding is to use the disparity maps generated by the stereo matching algorithm to exploit the redundancy in each part of the H.264 bitstream (e.g., overhead bits, motion vector bits, and texture bits for DCT coefficients). Such an MT video coder will perform no worse than separate H.264 coding of the two video sequences.

We describe in this paper an MT video coder (without allowing collaboration among the two encoders) that is capable of outperforming separate H.264 coding of two stereo video sequences. Our coder shares the basic structure of Slepian-Wolf coded quantization [6] for direct MT source coding of two Gaussian sources. Specifically, the left video sequence is compressed by the first encoder using H.264 and a recon-

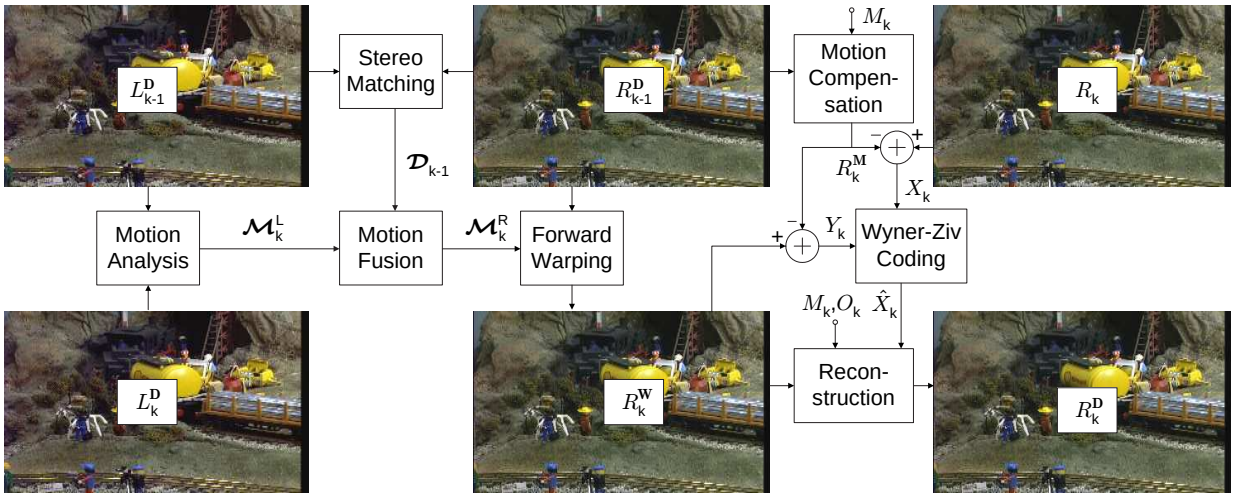

Fig. 1. Multiterminal video coding framework.

structed version is available at the joint decoder. Then, the first frame of the right sequence is intra-coded by H.264 and its reconstruction is used to generate the first disparity map at the joint decoder by employing the BP based stereo matching algorithm [16]. Knowing the disparity map as a point-to-point correspondence between the first pair of stereo video frames, we impose an "identical motion constraint" (which means the corresponding points in the left and right scenes must have identical 3D motions) to explore the redundancy in the next frame. Under this constraint, we devise a novel algorithm that incorporates a 3D geometric model to quantitatively fuse the left motion field with the disparity map, producing an estimation of the right motion field. With this estimated motion field, the joint decoder not only generates the side information for the motion vectors in H.264, but also warps the reconstructed first frame of the right sequence to form an estimation of its second frame. Finally, side information for the H.264 motion-compensated residual frame is obtained by taking the difference between the warped version and the H.264 motion-compensated version of the second frame. With side information available at the decoder, we apply SWC implemented via low-density parity-check (LDPC) code to the motion vectors, and Wyner-Ziv coding (WZC) [17] of the motion-compensated residual frames by using Slepian-Wolf coded scalar quantization.

As mentioned earlier, H.264 bitstream consists of header bits, motion vector bits, and texture bits. In the low-rate regime, most the of the rate budget is spent on the former two; and there is not much room for further savings in the texture bits from WZC in this scenario. This paper focuses on the low-rate regime and presents some initial results on SWC of the motion vector bits that indicate savings in terms of the sum rate when compared to separate H.264 coding at the same video quality.

In the high-rate regime, additional WZC of the motion-compensated residual frames is a must, but it is more challenging because the "bad" matching pixels in the disparity map and motion field will introduce much more noise to the side information of residual frame pixels than to that of the motion vectors (which are generated at macroblock level instead of pixel level). This is part of our ongoing research.

II. MT VIDEO CODING

Our proposed MT video coding scheme is depicted in Fig. 1. Let $\mathcal{L} = \{L_1, L_2, ..., L_n\}$ and $\mathcal{R} = \{R_1, R_2, ..., R_n\}$ be the left and right stereo video sequences, respectively. First, the left sequence $\mathcal{L}$ is compressed at Encoder 1 by H.264 and transmitted to the joint decoder, using a transmission rate of $\mathfrak{R}_L$ bits per second (bps). Assume that only the first frame L_1 is intra-coded I-frame and all the other frames $L_2, ..., L_n$ are inter-coded P-frames. Similarly, the right sequence $\mathcal{R}$ is also compressed by H.264 at Encoder 2, but only the first frame is directly transmitted. Denote $\mathfrak{R}_R^1$ as the bit rate (in bps) that Encoder 2 spent on coding R_1. The coded bitstream for the k-th inter-coded frame R_k ($k = 2, 3, ..., n$) consists of three parts, namely, the overhead bits O_k^R, the motion vector bits M_k^R, and texture bits C_k^R for the DCT coefficients. Let the reconstructed version of left and right sequences be $\mathcal{L}^D = \{L_1^D, ..., L_n^D\}$ and $\mathcal{R}^D = \{R_1^D, ..., R_n^D\}$, respectively.

Before compressing R_k for $k = 2, ..., n$ at Encoder 2, we assume that the joint decoder has access to the reconstructions $\{L_1^D, ..., L_{k-1}^D, L_k^D\}$ and $\{R_1^D, ..., R_{k-1}^D\}$. We first employ stereo matching to generate disparity map $\mathcal{D}_{k-1}$ between L_{k-1}^D and R_{k-1}^D. Using a slightly modified stereo matching algorithm (by allowing vertical disparities), we also obtain a forward motion field $\mathcal{M}_k^L$ from L_{k-1}^D to L_k^D. Then, assume that the 3D stereo camera settings are known, and follow the "identical motion constraint" we apply a novel motion fusing algorithm to produce the right forward motion field $\mathcal{M}_k^R$ based on the known information $\mathcal{D}_{k-1}$ and $\mathcal{M}_k^L$. Clearly, the motion vectors M_k^R in the H.264 bitstream are correlated to the motion field $\mathcal{M}_k^R$ as shown in Fig. 2. Hence SWC can be employed to code M_k^R with $\mathcal{M}_k^R$ as decoder side information.

Next, R^D_{k-1} is warped according to the right motion field $\mathcal{M}^R_k$, generating an estimate R^W_k of the k-th frame R_k. Assume ideal Slepian-Wolf decoding, such that M^R_k is perfectly reconstructed at the decoder, then exactly the same motion compensated frame R^M_k at the encoder can be formed by warping R^D_{k-1} according to M^R_k. Consequently, the *source* and the *side information* for WZC can be computed as

$$X_k = R_k - R^M_k; \qquad Y_k = R^W_k - R^M_k, \tag{1}$$

respectively.

Finally, WZC is employed to explore the remaining correlation between X_k and Y_k and and joint decoder reconstructs $\mathcal{R}^D = \{R^D_1, R^D_2, ..., R^D_n\}$ using a total transmission rate of $\mathfrak{R}_Y = \sum_{i=1}^n \mathfrak{R}^i_Y$ bps.

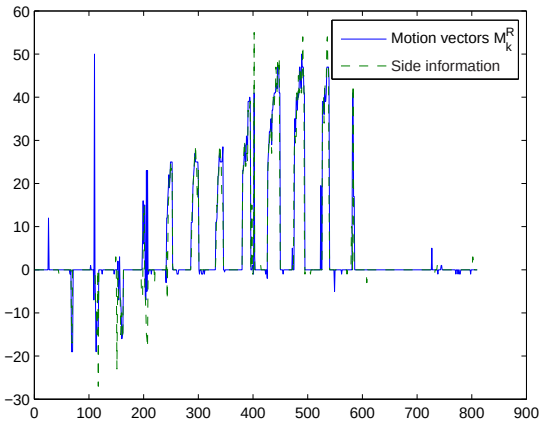

Fig. 2. Comparison of motion vectors M^R_k and their hard-decision side information.

III. Efficient stereo matching algorithm

Let $I = \{L, R\}$ be a pair of stereo images. Denote $\mathcal{D}$ as the smooth disparity field, $\mathcal{N}$ as a spatial line process, and $\mathcal{O}$ as a binary process indicating the occlusion regions. Then, according to [15],

$$P(\mathcal{D}, \mathcal{N}, \mathcal{O}|I) \propto \prod_{s \notin \mathcal{O}} \exp(-F(s, d_s, I)) \prod_s \exp(-\eta_c(o_s)) \prod_s \prod_{t \in N(s)} \exp(-\phi(d_s, d_t)(1 - l_{s,t}) + \gamma(l_{s,t})), \tag{2}$$

where s, t represent pixels in the reference frame, $N(s)$ denotes the set of neighboring pixels of s that is larger than s, and d_s, o_s are the disparity and occlusion of pixel s, respectively. We can further write the posterior probability over $\mathcal{D}$ as [15]

$$P(\mathcal{D}|I) \propto \prod_s \exp(-\rho_d(d_s)) \prod_s \prod_{t \in N(s)} \exp(-\rho_p(d_s, d_t)), \tag{3}$$

where

$$\rho_d(d_s) = -\ln\Big((1 - e_d)\exp(-\frac{|F(s, d_s, I)|}{\sigma_d}) + e_d\Big), \tag{4}$$

$$\rho_p(d_s, d_t) = -\ln\Big((1 - e_p)\exp(-\frac{|d_s - d_t|}{\sigma_p}) + e_p\Big), \tag{5}$$

and $|F(s, d_s, I)|$ is Birchfield and Tomasi's pixel dissimilarity [11]. Hence the standard "max-product" algorithm can be implemented by the following steps [15].

1) Initialize all messages $m_{st}(x_t)$ as uniform distributions and messages $m_s(x_s) = \exp(-\rho_d(x_s))$.
2) Update messages $m_{st}(x_t)$ iteratively for $i = 1, 2, ..., T$

$$m^{i+1}_{st}(x_t) \leftarrow \kappa \max_{x_s} \exp(-\rho_p(x_s, x_t)) m^i_s(x_s) \prod_{x_k \in N(x_s) \setminus x_s} m^i_{ks}(x_s). \tag{6}$$

3) Compute beliefs

$$b_s(x_s) \leftarrow \kappa m_s(x_s) \prod_{x_k \in N(x_s)} m_{ks}(x_s), \qquad x^{MAP}_s = \arg\max_{x_k} b_s(x_k). \tag{7}$$

In our experiments, we use a modified BP algorithm described in [16], which is more time efficient without sacrificing the quality of matching results.

IV. Motion field estimation

Although originally designed for stereo matching, the BP based algorithm can also be applied for motion field estimation. Since most stereo cameras are aligned such that no vertical disparity exists between corresponding pixels, the algorithm in [15] only allows horizontal disparities, which are clearly not enough for motion field. Hence we slightly modify the above algorithm by allowing vertical disparities. First, all d_s's in equations (2) - (7) become vectors $\boldsymbol{d}_s$, and absolute value "$|\cdot|$" becomes L^1-norm "$\|\cdot\|$". Also, the Birchfield and Tomasi's pixel dissimilarity $|F(s, \boldsymbol{d}_s, I)|$ is changed to

$$F(s, \boldsymbol{d}_s, I) = \min\{\bar{d}(s, s', I)/\sigma_f, \bar{d}(s', s, I)/\sigma_f\}, \tag{8}$$

where $\bar{d}(s, s', I) = \min\{|I_L(s) - I_R(s')|, |I_L(s) - I^{\leftarrow}_R(s')|, |I_L(s) - I^{\rightarrow}_R(s')|, |I_L(s) - I^{\uparrow}_R(s')|, |I_L(s) - I^{\downarrow}_R(s')|\}$, s' is the matching pixel of s with disparity $\boldsymbol{d}_s$, and $\{I^{\leftarrow}_R(s'), I^{\rightarrow}_R(s'), I^{\uparrow}_R(s'), I^{\downarrow}_R(s')\}$ are the linearly interpolated intensity halfway between s' and its neighboring pixel to the left, right, top and bottom, respectively, and σ_f is the image noise variance that depends on the quality of input pictures.

V. Motion fusion

In this section, we describe the algorithm used to fuse the disparity map $\mathcal{D}$ and the left motion field $\mathcal{M}^X$ to estimate the right motion field $\mathcal{M}^Y$. As shown in Fig. 3, the 3D motion vector can be decomposed into three components: horizontal motion V_h that is parallel to $o_l o_r$, vertical motion V_v that is perpendicular to the $oo_l o_r$ plane, and parallel motion V_p that is perpendicular to both V_h and V_v (which is ignored in the motion fusion algorithm). Denote F as the focal length of both cameras, B as the base line distance $o_l o_r$ between two cameras, and D as the convergence distance. The stereo scene geometry is illustrated in Fig. 4. The stereo motion fusion algorithm has the following steps.

1) Estimating the depth. Calculate angles α and β using the horizontal coordinate of the pixel s. Then the depth of s is $H_p = B/[(\tan(\alpha))^{-1} + (\tan(\beta))^{-1}]$.

2) Estimating the right horizontal motion vector $v_h^r = V_h^r r_p / R_p$ based on the depth H_p and the left horizonal motion vector $v_h^l = V_h^l l_p / L_p$ using (note that $V_h^l = V_h^r$)

$$\frac{v_h^r}{v_h^l} = \frac{r_p L_p}{l_p R_p} = \frac{\sin(\alpha + \frac{\theta}{2})\sin(\beta)}{\sin(\beta + \frac{\theta}{2})\sin(\alpha)}. \quad (9)$$

3) Estimating the right vertical motion vector using

$$\frac{v_v^r}{v_v^l} = \frac{v_h^r}{v_h^l} = \frac{\sin(\alpha + \frac{\theta}{2})\sin(\beta)}{\sin(\beta + \frac{\theta}{2})\sin(\alpha)}. \quad (10)$$

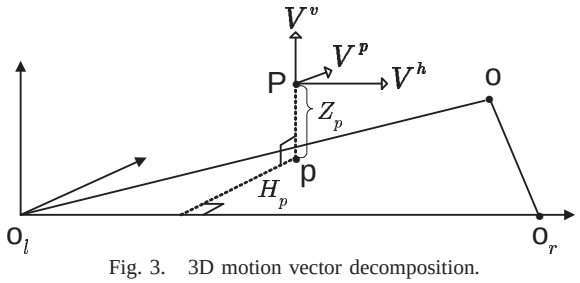

Fig. 3. 3D motion vector decomposition.

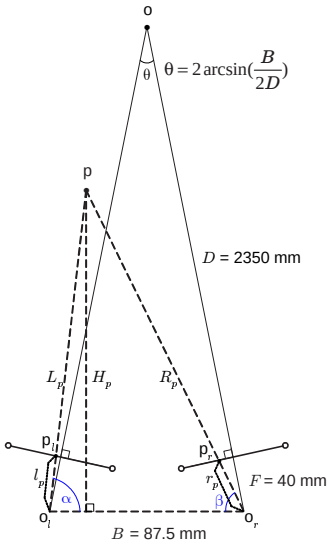

Fig. 4. 3D geometry for stereo video pair.

VI. WZC of Motion Compensated Residual Coefficients

In WZC of X_k based on the decoder side information Y_k, the key assumption is that we know the correlation between the target frame X_k and the side information Y_k. However, unlike ideal sources (e.g., i.i.d. jointly Gaussian), the correlation between X_k and Y_k is not available *a priori*. As in other works on distributed video coding in the literature (e.g., [18]), we use training video sequences to find an average difference distribution between X_k and Y_k, and assume an additive noise correlation model. Moreover, classification of H.264 macroblocks (e.g., between occluded and non-occluded ones) may also help to build different correlation models for different classes.

Then the encoder quantizes X_k using scalar quantization. Knowing the averaged additive noise model, the encoder can also compute the required transmission rates for each bit-plane of the quantization indices, and send the corresponding syndrome bits for Slepian-Wolf compression. Finally, the joint decoder uses the syndrome bits and the log-likelihood ratios to reconstruct $\hat{X}_k$. Detailed encoding/decoding algorithms can be found in [7].

VII. Simulation Results

In our simulations, we use the Y-component of the 720 $\times$ 288 "tunnel" stereo video sequences downloaded from "http://www.tnt.uni-hannover.de/project/eu/distima/images". Both the left and right sequences are coded by H.264 standard, coding parameters and some statistics of the resulting bitstream are given in Table I.

TABLE I

H.264 compression parameters and statistics.

Parameters	Left sequence $\mathcal{L}$	Right sequence $\mathcal{R}$
QP I frame	35	35
QP P frame	33	33
Total frames	20	20
Inter-search mode	16×16,16×8,8×16	16×16,16×8,8×16
Motion precision	quarter-pel	quarter-pel
Statistics	**Left bitstream**	**Right bitstream**
I-frame	95,112 bits	94,448 bits
Overhead	16,994 bits	16,822 bits
Motion vectors	39,494 bits	38,970 bits
Coefficients	139,160 bits	136,108 bits
Total	287,248 bits	286,584 bits
Bit rate	436.40 kbps	429.88 kbps
Average SNR	31.11 dB	31.18 dB

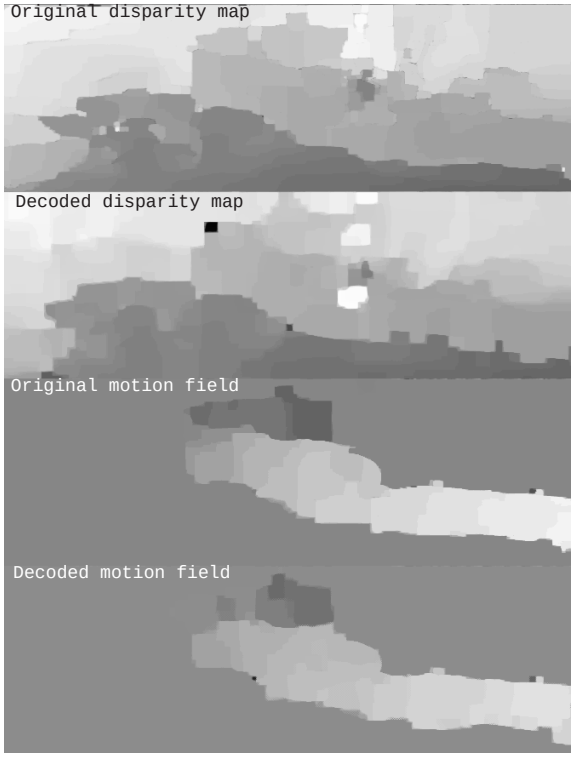

Fig. 5. Generated disparity maps (top two) and motion fields (bottom two).

Fig. 5 shows the disparity maps and motion fields generated by the modified stereo matching algorithm described in Sec-

tions III and IV. The parameter values in (2) - (7) are consistent with those in [15]: $e_d = 0.01, \sigma_d = 8, e_p = 0.05, \sigma_p = 0.6$. We also incorporate segmentation results produced by the mean-shift algorithm [19].

Since the sum rate is low (866.28 kbps at a frame rate of 30 frames/sec), we can see that the disparity map and the motion field generated by the decoded frames are not very reliable compared to those from the originals. Hence only the motion vectors for the inter-coded blocks are Slepian-Wolf coded based on the side information generated at the decoder. Using a multilevel Slepian-Wolf code implemented by LDPC codes, we are able to save 3,747 bits from the 38,970 motion vector bits in the right bitstream. All the other components are directly transmitted to the decoder. Figs. 6 and 7 compare the rate-distortion performance for separate encoding, MT coding, and joint encoding of "tunnel" stereo video sequences, where in the joint encoding case we interleave the left and right stereo video sequences and use H.264 to code the interleaved sequence with two reference frames in motion estimation, to generate a benchmark for MT video coding.

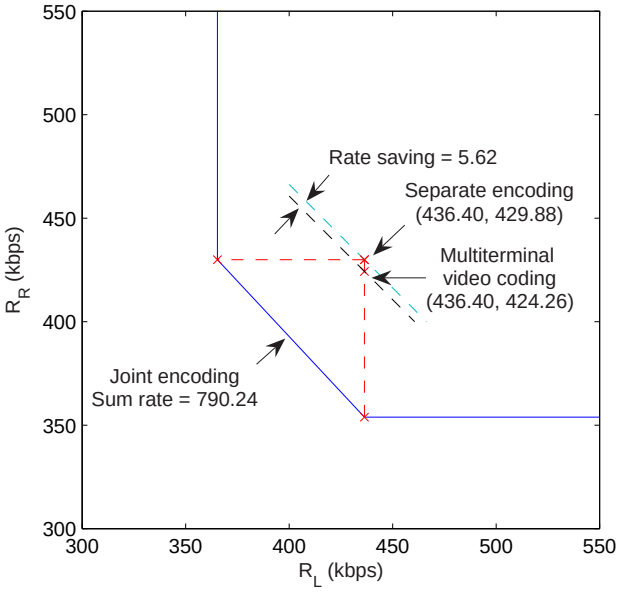

Fig. 6. Comparison between separate H.264 encoding, MT coding, and joint encoding (with same PSNR = 31.14 dB).

VIII. Conclusion

In this paper, we addressed MT video coding that targets at saving the sum rate over separate monocular video compressions with H.264. The main idea is to explore the binocular redundancy by using disparity maps generated by stereo matching to form side informations in WZC. Preliminary results on rate savings for motion vectors in the low-rate regime are given.

References

[1] T. Berger, "Multiterminal source coding", *The Inform. Theory Approach to Communications*, G. Longo, Ed., New York: Springer-Verlag, 1977.

[2] A. Wagner, S. Tavildar, and P. Viswanath, "The rate region of the quadratic Gaussian two-terminal source-coding problem," submitted to *IEEE Trans. Inform. Theory*, Oct. 2005.

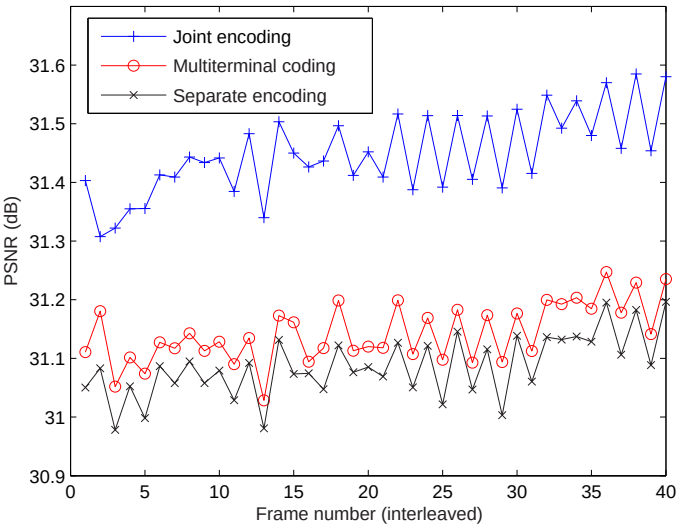

Fig. 7. Comparison (in terms of PSNR vs. frame number) between separate H.264 encoding, MT coding, and joint encoding (with same sum rate of 860.6 kbps). The frame indices are for the interleaved version of the left and right sequences (20 frames each) in each case.

[3] Y. Oohama, "Multiterminal source coding for correlated memoryless Gaussian sources with several side information at the decoder," *Proc. ITW-1999*, Kruger National Park, South Africa, June 1999.

[4] V. Prabhakaran, D. Tse, and K. Ramchandran, "Rate region of the quadratic Gaussian CEO problem," *Proc. ISIT-2004*, Chicago, IL, June 2004.

[5] S. Pradhan and K. Ramchandran, "Generalized coset codes for distributed binning," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3457–3474, Oct. 2005.

[6] Y. Yang, V. Stanković, Z. Xiong, and W. Zhao, "Asymmetric code design for remote multiterminal source coding," *Proc. DCC-2004*, Snowbird, UT, March 2004.

[7] Y. Yang, V. Stanković, Z. Xiong, and W. Zhao, "On multiterminal source code design," submitted to *IEEE Trans. Inform. Theory*, Oct. 2006.

[8] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, July 1973.

[9] T. Wiegand, G. Sullivan, G. Bjintegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 13, pp. 560-576, July 2003.

[10] J.-R. Ohm, "Stereo/multiview encoding using the MPEG family of standards," in *Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems VI*, vol. 3639, pp. 242-253, Jan. 1999

[11] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 401–406, Apr. 1998.

[12] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, Nov. 2001.

[13] V. Kolmogorov and R. Zabih,, "Multi-camera scene reconstruction via graph cuts," in *Proc. European Conf. Computer Vision*, 2002.

[14] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," *Intl J. Computer Vision*, vol. 14, pp. 211–226, 1995.

[15] J. Sun, H. Y. Shum and N. N. Zheng, "Stereo matching using belief propagation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.25, no. 7, pp. 787–800, 2003.

[16] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision", *International Journal of Computer Vision*, vol. 70, no. 1, Oct. 2006.

[17] A. 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, Jan. 1976.

[18] R. Puri, A. Majumbar, P. Ishwar, and K. Ramchandran, "Distributed video coding in wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 23, pp. 94-106, July 2006.

[19] D. Comanicu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 603–619, May 2002.