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.
[‡]Dept of Computer Science, Rensselaer Polytechnic Institute (RPI), 110 8th St., Troy, NY 12180.

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. The first I-frame of the right sequence is successively coded by H.264 and Slepian-Wolf coding. 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. Experimental results on stereo video sequences using H.264, LDPC codes for Slepian-Wolf coding of the motion vectors and scalar quantization in conjunction with LDPC codes for Wyner-Ziv coding of the residual coefficients show savings in terms of the sum-rate when compared to separate H.264 coding at the same video quality.**

Index terms: multiterminal video coding, stereo matching, H.264 standard.

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] 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 [4]. In earlier works, we proposed a framework for practical MT source coding based on Slepian-Wolf coded quantization [5], which employs the approach of vector quantization followed by Slepian-Wolf coding (SWC) [6]. However, the code designs in [4], [5] are for ideal Gaussian sources assuming *a priori* known correlation. When dealing with practical (e.g., video) sources, correlation modeling is one of the key issues in efficient MT video coding. In this paper, we focus on MT video code design for two correlated video sequences captured by calibrated cameras.

In general, effective coding of a single/monocular video sequence necessitates exploitation of both spatial and temporal redundancies within the sequence. H.264 [7] 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 [8]) in a joint encoding setup.

For MT video coding, although the encoders cannot communicate with each other, the 3D geometric information of the cameras can still help to exploit the binocular correlation between the stereo pair. Many works [9], [10], [11], [12], [13] are done for multiview video coding using this idea. Recently, Song *et al.* [14] designed a model-based coding scheme that combines 3D geometry with distributed coding. A further attempt to obtain pixel-level stereo correspondence leads to *stereo matching*, which is a fundamental problem in stereo vision, and has been extensively studied in the past by many researchers (see [15] and references therein). Assuming knowledge of the camera configurations, stereo matching computes a disparity map from a stereo image pair. It can be formulated as an optimization problem that minimizes the image dissimilarity energy. Quantitative evaluations of different stereo matching 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.

We describe in this paper an MT video coder 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 [5] for direct MT source coding of two Gaussian sources. Specifically, the left video sequence is compressed by the left encoder using H.264 and a reconstructed version is available at the joint decoder. Then, the first I-frame of the right sequence is successively coded: a low-quality version is generated by H.264 and sent to the decoder to obtain a rough disparity map, which is used to compress the refinement bit stream of the right I-frame using SWC with the decoded left I-frame as side information. This way, the low-quality version is thus refined, as well as the disparity map between the I-frames. Using the disparity map as a initial point-to-point correspondence for the remaining P-frames of the right sequence, the joint decoder generates side informations for both the motion vectors and the motion-compensated residual frame by imposing an “identical motion constraint” (which means the corresponding points in the left and right scenes must have identical 3D motions). With side information available at the decoder, we implement SWC of the motion vectors via low-density parity-check (LDPC) coding, and Wyner-Ziv coding (WZC) [17] of the motion-compensated residual frames via Slepian-Wolf coded scalar quantization.

1-4244-1437-7/07/$20.00 ©2007 IEEE

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. 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 paper presents results SWC of the motion vector bits at low rate, and on WZC of the I-frame and the residual P-frames at high rate. These results indicate savings in terms of the sum-rate when compared to separate H.264 coding at the same video quality.

II. MT VIDEO CODING

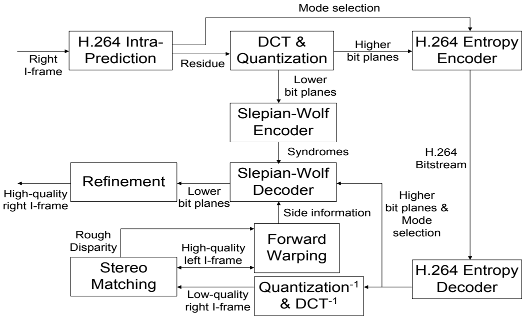

Fig. 1. Multiterminal video encoder-decoder (right I-frame).

Our proposed MT video coding scheme is depicted in Fig. 1 (right I-frame) and Fig. 2 (right P-frame). 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. Then the first frame R_1 of right sequence is intra-coded using a large quantization parameter (QP) to produce a low-quality reconstruction R_1^d at the decoder. A rough disparity map $\tilde{\mathcal{D}}_1$ between R_1^d and the decoded left I-frame L_1^D is generated, and is used to produce a side information R_1^w by warping L_1^D. Now the encoder re-quantizes the residual I-frame using a small QP, and the refining lower bitplanes are compressed by SWC with the syndromes sent to the decoder. Using both the side information R_1^w and the decoded higher bitplanes (from H.264), the refinement bitplanes are decoded and hence the final decoded I-frame R_1^D and disparity $\mathcal{D}_0$ are generated. 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. We denote the reconstructed version of the left and right sequences as $\mathcal{L}^D = \{L_1^D, ..., L_n^D\}$ 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 $\mathcal{M}_k^R$. Hence SWC can be employed to code M_k^R with $\mathcal{M}_k^R$ as decoder side information.

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

$$X_k = R_k - \text{warp}(R_{k-1}^D, M_k^R) = R_k - R_k^M; \quad (1)$$

$$Y_k = \text{warp}(L_k^D, \mathcal{D}_k) - \text{warp}\big(R_{k-1}^D, M_k^R\big), \quad (2)$$

respectively. Finally, WZC is employed to explore the remaining correlation between X_k and Y_k and and joint decoder reconstructs $\mathcal{R}^D = \{R_1^D, R_2^D, ..., R_n^D\}$ using a total transmission rate of $\mathfrak{R}_Y = \sum_{i=1}^n \mathfrak{R}_Y^i$ bps.

III. I-FRAME COMPRESSION

The right I-frame is first intra-coded with $QP = P_l$. The resulting H.264 bitstream is directly sent to the decoder, and a rough disparity map $\tilde{\mathcal{D}}_1$ is generated by matching the reconstructed right I-frame R_1^d ($QP = P_l$) with L_1^D ($QP = P_h = P_l - 6k$). Denote the intra-predicted right I-frame as R_1^P ($QP = P_l$). Then the residual frame $R_1 - R_1^P$ is transformed and re-quantized with $QP = P_h$. Since quantization step size doubles for every increment of 6 in QP [7], all but the lower k bitplanes of the re-quantized coefficients are already transmitted by H.264. The rest k *refinement bitplanes* are compressed by Slepian-Wolf Encoder using multilevel Slepian-Wolf decoding [5] with side information $R_1^w = \text{warp}(L_1^D, \tilde{\mathcal{D}}_1)$ and the decoded higher bitplanes.

IV. MOTION FIELD ESTIMATION AND MOTION FUSION

Although originally designed for stereo matching, the BP based algorithm [15], [16] 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: all scalar disparities d_s become vector disparities $\boldsymbol{d}_s$; the Birchfield and Tomasi's pixel dissimilarity $|F(s, \boldsymbol{d}_s, I)|$ [15] is changed to

$$F(s, \boldsymbol{d}_s, I) = \min\{\bar{d}(s, s', I)/\sigma_f, \bar{d}(s', s, I)/\sigma_f\}, \quad (3)$$

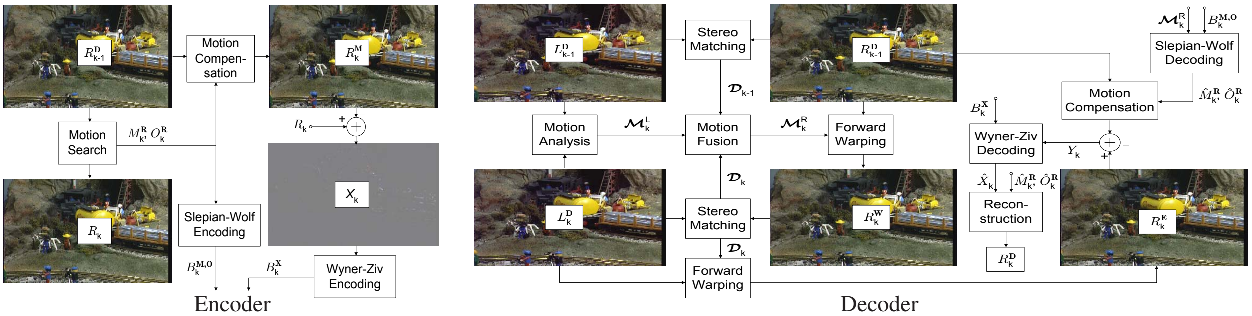
$R^{\mathbf{D}}_{\text{k-1}}$
Motion Compen- sation
$R^{\mathbf{M}}_{\text{k}}$
Motion Search
$M^{\mathbf{R}}_{\text{k}}, O^{\mathbf{R}}_{\text{k}}$
R_{k}
+
−
R_{k}
X_{k}
Slepian-Wolf Encoding
$B^{\mathbf{M,O}}_{\text{k}}$
$B^{\mathbf{X}}_{\text{k}}$
Wyner-Ziv Encoding
Encoder
$L^{\mathbf{D}}_{\text{k-1}}$
Stereo Matching
$R^{\mathbf{D}}_{\text{k-1}}$
$\mathcal{D}_{\text{k-1}}$
Motion Analysis
$\mathcal{M}^{\text{L}}_{\text{k}}$
Motion Fusion
$\mathcal{M}^{\text{R}}_{\text{k}}$
Forward Warping
$\mathcal{D}_{\text{k}}$
$L^{\mathbf{D}}_{\text{k}}$
Stereo Matching
$R^{\mathbf{W}}_{\text{k}}$
$\mathcal{D}_{\text{k}}$
Forward Warping
Decoder
$\mathcal{M}^{\text{R}}_{\text{k}}$
$B^{\mathbf{M,O}}_{\text{k}}$
Slepian-Wolf Decoding
Motion Compensation
$\hat{M}^{\mathbf{R}}_{\text{k}}, \hat{O}^{\mathbf{R}}_{\text{k}}$
$B^{\mathbf{X}}_{\text{k}}$
Wyner-Ziv Decoding
Y_{k}
+
−
$\hat{X}_{\text{k}}$
$\hat{M}^{\mathbf{R}}_{\text{k}}, \hat{O}^{\mathbf{R}}_{\text{k}}$
Recon- struction
$R^{\mathbf{D}}_{\text{k}}$
$R^{\mathbf{E}}_{\text{k}}$

Fig. 2. Multiterminal video encoder-decoder (right P-frames).

where $\bar{d}(s, s', I) = \min\{|I_L(s) - I_R(s')|, |I_L(s) - I_R^{\leftarrow}(s')|, |I_L(s) - I_R^{\rightarrow}(s')|, |I_L(s) - I_R^{\uparrow}(s')|, |I_L(s) - I_R^{\downarrow}(s')|\}$, s' is the matching pixel of s with disparity $\boldsymbol{d}_s$, and $\{I_R^{\leftarrow}(s'), I_R^{\rightarrow}(s'), I_R^{\uparrow}(s'), I_R^{\downarrow}(s')\}$ 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.

The next step is to fuse the disparity map $\mathcal{D}$ and the left motion field $\mathcal{M}^L$ to estimate the right motion field $\mathcal{M}^R$. As shown in Fig. 3 (b), 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, S as the pixel size in the imaging plane, and D as the convergence distance. The stereo scene geometry is illustrated in Fig. 3 (a). The stereo motion fusion algorithm has the following steps (see block diagram in Fig. 3 (c)).

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 $V_h^l = V_h^r$)
$$\frac{v_h^r}{v_h^l} = \frac{r_p L_p}{l_p R_p} = \frac{\sin(\alpha + \frac{\theta}{2})\sin(\beta)}{\sin(\beta + \frac{\theta}{2})\sin(\alpha)}. \quad (4)$$
3) Estimating the right vertical motion vector using
$$\frac{v_v^r}{v_v^l} = \frac{v_h^r}{v_h^l} = \frac{\sin(\alpha + \frac{\theta}{2})\sin(\beta)}{\sin(\beta + \frac{\theta}{2})\sin(\alpha)}. \quad (5)$$

V. SWC of Motion Vectors and WZC of Residual Coefficients

In SWC of M_k^R and WZC of X_k based on the decoder side informations $\mathcal{M}_k^R$ and Y_k, respectively, the key requirement is the correlation model. However, unlike ideal sources (e.g., i.i.d. jointly Gaussian), this correlation is not available *a priori*. As in other works on distributed video coding in the literature (e.g., [18]), we collect joint statistics from training video sequences between each source-side information pair to build a generic correlation model.

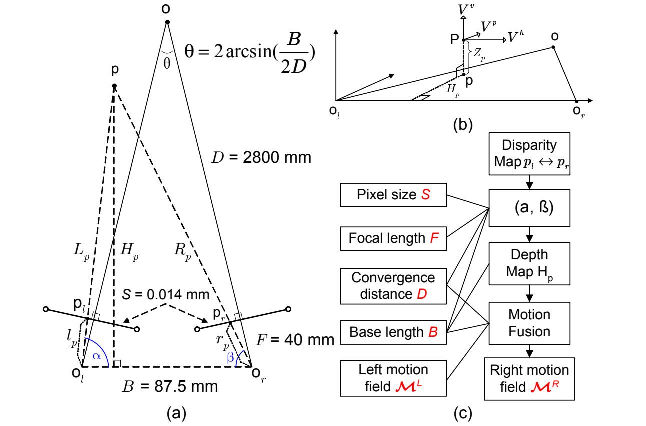

Fig. 3. Stereo motion fusion (a) 3D geometry; (b) motion decomposition; (c) block diagram.

Then the Wyner-Ziv encoder quantizes X_k using scalar quantization. The resulting quantization levels and the motion vectors M_k^R are then coded by two separate Slepian-Wolf encoders, which send the syndrome bits for each bit-plane of the two sources. Finally, the joint decoder uses the syndrome bits and the log-likelihood ratios (computed using the correlation model and the side information) to reconstruct $\hat{X}_k$ and $\hat{M}_k^R$. Detailed encoding/decoding algorithms can be found in [5].

VI. Simulation Results

In our simulations, we use the Y-component of the 720 × 288 "tunnel" stereo video sequences. Both the left and right sequences are coded by H.264 standard, coding parameters and the statistics of the resulting bitstream for both the low-rate case and the high-rate case are given in Table I.

TABLE I
H.264 compression parameters and statistics.

Parameters	Low-rate regime	High-rate regime
QP I frame	35	22
QP P frame	33	20
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	Low-rate regime	High-rate regime
Bit rate	866.3 Kbps	6.630 Mbps
Average SNR	31.15 dB	40.59 dB

The disparity maps and motion fields are generated in half-pel precision by the modified stereo matching algorithm described in Section IV. The parameter values 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].

In low-rate case, only the motion vectors for the inter-coded blocks are Slepian-Wolf coded based on the side information generated at the decoder. Using the joint statistics collected from all 20 frames of “tunnel” sequence as generic correlation model, and 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. 4 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.

In high-rate case, we implement the algorithms described in Section III for the I-frame (with $P_l = 34$ and $P_h = 22$) and in Section IV for the residual coefficients of the P-frames. Generic correlation models between the sources and the side informations are generated based on the joint statistics collected from all 20 frames of “tunnel” sequence. Scalar quantization followed by LDPC code based multilevel Slepian-Wolf codes are employed for Wyner-Ziv coding. The total saving is 32,548 bits, which is equivalent to 48.8 Kbps, or 0.75% of the total bit rate. Again, a performance comparison among separate encoding, MT coding, and joint encoding is shown in Fig. 5.

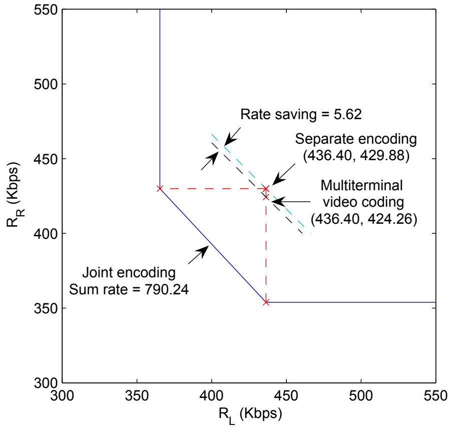

Fig. 4. Comparison between separate H.264 encoding, MT coding, and joint encoding (with same PSNR = 31.15 dB).

However, compared to separate H.264 compression, the computational complexity of our scheme is higher, and most of the computation time is spent on the stereo matching algorithm (which takes around 40 minutes per frame on a Pentium IV 2.0GHz PC).

VII. 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. Results on rate savings for motion vectors in the low-rate regime and for I-frame and residual coefficients in the high-rate regime are given.

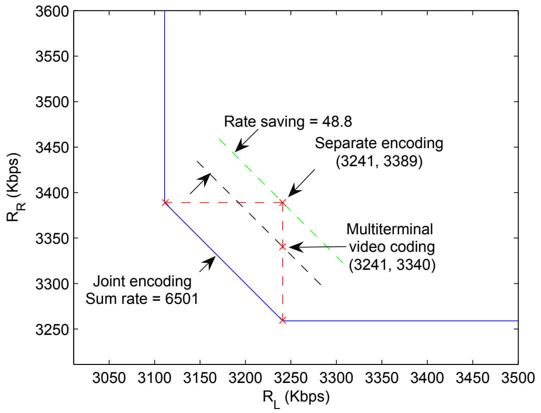

Fig. 5. Comparison between separate H.264 encoding, MT coding, and joint encoding (with same PSNR = 40.59 dB).

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.

[3] Y. Oohama, “The rate-distortion function for the quadratic Gaussian CEO problem,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 1057–1070, May 1998.

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

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

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

[7] 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.

[8] 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

[9] M. Flierl and B. Girod, “Coding of multi-view image sequences with video sensors,” *Proc. ICIP’06*, Atlanta, GA, Oct. 2006.

[10] X. Guo, Y. Lu, F. Wu, W. Gao and S. Li, “Distributed multiview video coding,” *Proc. SPIE*, vol. 6077, San Jose, CA, USA, Jan. 2006.

[11] M. Ouaret, F. Dufaux and T. Ebrahimi, “Fusion-based multiview video coding,” *Proc. ACM International Workshop on Video Surveiliance and Sensor Networks*, Santa Barbara, CA, Oct. 2006.

[12] E. Martinian, A. Behrens, J. Xin and A. Vetro, “View synthesis for multiview video compression,” *Picture Coding Symposium*, Beijing, China, Apr. 2006.

[13] C. Yeo and K. Ramchandran, “Distributed video compression for wireless camera networks,” in *Proc. SPIE Conf. on Video and Image Communications, VCIP’07*, San Jose, California, Feb. 2007.

[14] B. Song, A. K. Roy-Chowdhury, and E. Tuncel, “A Multi-terminal Model-based Video Compression Algorithm,” *Proc. ICIP’06*, Atlanta, GA, Oct. 2006.

[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.