WYNER-ZIV CODING OF MULTIVIEW IMAGES WITH UNSUPERVISED LEARNING OF TWO DISPARITIES

David Chen, David Varodayan, Markus Flierl, Bernd Girod

Information Systems Laboratory, Stanford University, Stanford, CA 94305
{dmchen, varodayan, mflierl, bgirod}@stanford.edu

ABSTRACT

Wyner-Ziv coding of multiview images is an attractive solution because it avoids communications between individual cameras. To achieve good rate-distortion performance, the Wyner-Ziv decoder must reliably estimate the disparities between the multiview images. For the scenario where two reference images exist at the decoder, we propose a codec that effectively performs unsupervised learning of the two disparities between an image being Wyner-Ziv coded and the two reference images. The proposed two-disparity decoder disparity-compensates the two references images and generates side information more accurately than an existing one-disparity decoder. Experimental results with real multiview images demonstrate that the proposed codec achieves PSNR gains of 1-5 dB over the one-disparity codec.

Index Terms— image coding, data compression, stereo vision, disparity

1. INTRODUCTION

Multiview images captured by a camera array are very similar. Exploiting these similarities is desirable for compression. The conventional approach requires a joint encoder, but this method is not practical if the cameras do not communicate with one another. Distributed coding has emerged as an alternative, with separate encoders and a joint decoder. The information theoretic Slepian-Wolf and Wyner-Ziv theorems suggest that distributed coding can be as efficient in coding performance as conventional joint compression [1][2].

A common way to relate multiview images is through disparities. Distributed coding must solve the challenge of estimating the disparities only at the decoder. A similar issue arises in distributed video coding (DVC), in which the motion between neighboring frames must be estimated at the decoder [3][4]. In that situation, motion-compensated temporal interpolation (MCTI) of two intra-coded key frames is commonly used [3]. MCTI assumes, however, symmetric motion vectors, which would be an invalid assumption for disparities between multiview images in general.

A Wyner-Ziv stereoscopic image codec in [5] proposes unsupervised learning of the disparity at the decoder using the Expectation Maximization (EM) algorithm [6]. It is a lossy extension of a simpler lossless Slepian-Wolf codec reported in [7]. The decoder iteratively learns by EM the unknown disparity between an image being transmitted X and a previously received image Y. A disparity-compensated version of Y serves as the side information. When two previously received images Y_1 and Y_2 reside at the decoder, however, the system still only uses one reference image rather than using both simultaneously.

In this paper, we generalize the one-disparity codec in [5] to perform unsupervised learning of two disparities, one between X and Y_1 and one between X and Y_2. Disparity-compensated versions of Y_1 and Y_2 are used to generate higher-quality side information, leading to significantly improved rate-distortion efficiency in coding X. Unlike previous multiple-reference codecs in [8] or [9] which make hard decisions about disparities, our EM-based codec makes soft decisions in estimating the disparities. In Section 2, the one-disparity codec is reviewed. In Section 3, the proposed two-disparity codec is presented with a generalized EM algorithm and an improved reconstruction method. In Section 4, experimental results for multiview image sets demonstrate that the two-disparity codec achieves PSNR gains of 1-5 dB over the one-disparity codec.

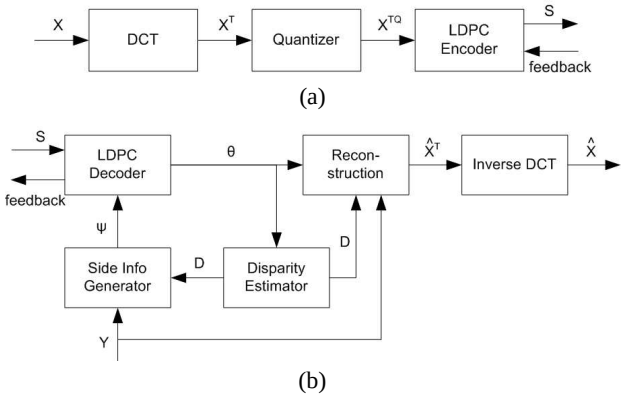

Fig. 1. One-disparity Wyner-Ziv image (a) encoder and (b) decoder.

2. ONE-DISPARITY CODEC

The Wyner-Ziv image codec from [5] can be summarized by the block diagram in Fig. 1. One image Y is transmitted by conventional coding, such as JPEG. Another image X must be encoded independently of Y but decoded using Y as side information. To exploit spatial correlation, the Wyner-Ziv encoder transforms X into 8-by-8 blockwise discrete cosine transform (DCT) coefficients X^T and quantizes X^T by a JPEG-recommended quantization table. Quantized coefficients X^{TQ} are losslessly communicated using a rate-adaptive low-density parity-check (LDPC) code [10]. The rate-adaptive LDPC code enables small portions of the syndrome S to be incrementally sent, assuming a feedback channel is available.

The Slepian-Wolf decoder within the larger Wyner-Ziv decoder in Fig. 1 is the loop formed by the LDPC decoder, the disparity estimator, and the side information generator. This loop is an instance of the EM algorithm, as detailed thoroughly in [7]. Using the received portions of S and the reference image Y, the Slepian-

Wolf decoder iteratively estimates the quantized transform coefficients X^{TQ}. When the disparity D between X and Y can be accurately estimated, the Slepian-Wolf decoder provides significant bit rate saving over conventional lossless transmission of X^{TQ} . Thus, reliable disparity estimation is a critical step in efficient coding.

As part of the decoder's EM algorithm, the disparity estimator iteratively calculates an *a posteriori* distribution of D having observed Y and S. On iteration t, the update is

$$P^{(t)}\{D\} = P^{(t-1)}\{D\}P\left\{Y, S|D; \theta^{(t)}\right\} \quad D = d_1, ..., d_M \quad (1)$$

where D ranges over M values with nonzero probability and $\theta^{(t)}$ is the current statistical estimate of X^{TQ}. $\theta^{(t)}$ is determined by the LDPC decoder through joint bitplane decoding, as described in [7]. The probability $P^{(t)}\{D\}$ indicates how likely the image Y_D, which is Y shifted by D, will match X.

Similarly, the side information generator iteratively updates the distribution of disparity-compensated side information ψ, a soft estimate of X^{TQ} having observed Y_D for multiple values of D. Each shifted image Y_D must be transformed and quantized to produce side information Y_D^{TQ} in the transform domain that is directly comparable to X^{TQ}. Thus, ψ conveys probabilities of quantized transform coefficients. On iteration t, the update is

$$\begin{aligned}\psi^{(t)}(w) &= \sum_{d=d_1}^{d_M} P^{(t)}\{D=d\}P\left\{X^{TQ}=w|Y_d\right\} \\ &= \sum_{d=d_1}^{d_M} P^{(t)}\{D=d\}P_N\left(w - Y_d^{TQ}\right) \quad (2)\end{aligned}$$

where w indicates quantized coefficient values. In the second equation, $P_N(n)$ is the distribution of the noise that remains between quantized transform coefficients after optimal disparity compensation. A Laplacian distribution is an accurate model for N. Eq. (2) can be interpreted as blending together probabilities of quantized transform coefficients for different shifted versions of Y.

After the coefficients X^{TQ} are recovered by the Slepian-Wolf decoder, the Wyner-Ziv decoder proceeds to reconstruct the actual image. The existing one-disparity system uses nearest-neighbor reconstruction [3]. A comparison of this and other reconstruction methods, including an optimal reconstruction scheme, will be given in Section 3.

3. TWO-DISPARITY CODEC

When two images Y_1 and Y_2 are available at the decoder, the one-disparity codec described in Section 2 performs suboptimally. Intuitively, image X will be better matched to Y_1 in some regions and to Y_2 in some other regions. Also, in general, the union of Y_1 and Y_2 provides a larger field-of-view than either image alone. A two-disparity codec can use Y_1 and Y_2 to generate higher-quality side information, which leads to both lower bit rate in Slepian-Wolf decoding of the quantized transform coefficients X^{TQ} and higher-quality in reconstructing X itself.

A diagram for the improved codec is shown in Fig. 2. Only the decoder is depicted because the encoder is the same as in the one-disparity system of Fig. 1. The two decoders are structurally similar. In the EM algorithm, the main difference is that the disparity estimator now estimates two disparities and the side information generator selectively chooses the best content from Y_1 and Y_2 to match X. In the reconstruction, the additional reference image leads to a lower error in estimating transform coefficients.

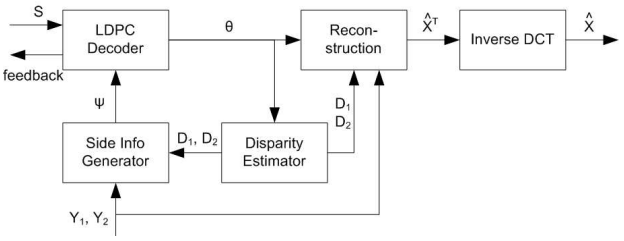

Fig. 2. Two-disparity Wyner-Ziv image decoder. Syndrome S is generated by the encoder in Fig. 1(a).

3.1. Generalized EM Learning Algorithm

The EM algorithm of the one-disparity system can be generalized to simultaneously estimate two disparities. For D_1 between X and Y_1 and D_2 between X and Y_2, the proposed update step becomes

$$\begin{aligned}&P^{(t)}\{D_k\} = P^{(t-1)}\{D_k\}\,P\left\{Y_k, S|D_k; \theta^{(t)}\right\} \\ &D_k = d_1, ..., d_M \qquad k = 1, 2 \quad (3)\end{aligned}$$

which reduces to (1) if either $P^{(t)}\{D_1\} = 0$ or $P^{(t)}\{D_2\} = 0$. The two distributions on D_1 and D_2 are also normalized such that their joint sum is unity. Effectively, the search space for disparity-compensation candidates has been doubled, from M different candidates for a single D to $2M$ different candidates for D_1 and D_2.

The side information generator selectively blends together probabilities for the shifted reference images as before, using twice as many candidates as in (2). On iteration t, the new update is

$$\begin{aligned}\psi^{(t)}(w) &= \sum_{k=1}^{2}\sum_{d=d_1}^{d_M} P^{(t)}\{D_k=d\}P\left\{X^{TQ}=w|Y_{k,d}\right\} \\ &= \sum_{k=1}^{2}\sum_{d=d_1}^{d_M} P^{(t)}\{D_k=d\}P_{N_k}\left(w - Y_{k,d}^{TQ}\right) \quad (4)\end{aligned}$$

so that the weighted sum includes shifted versions of Y_1 and Y_2. N_1 and N_2 represent two different Laplacian noise models. Since full distributions of D_1 and D_2 are maintained, the blending operation allows all shifted candidates to contribute partially to the side information. Other non-EM codecs which disparity-compensate two reference images restrict D_1 and D_2 to be deterministic and thus only permit two shifted candidates to be blended together [8].

The proposed method can be easily generalized to the case of K reference images. In (3) and (4), index k would range over $\{1, \ldots, K\}$ instead of just $\{1, 2\}$.

3.2. Transform Coefficient Reconstruction

Once the quantized transform coefficients X^{TQ} are Slepian-Wolf decoded, the reconstruction block in Fig. 2 calculates the dequantized transform coefficients X^T. Side-information-assisted reconstruction (SIAR) gives better results than choosing the centroid of the quantization bin $[Z_i, Z_{i+1}]$ determined by X^{TQ}. Several SIAR methods are compared in this section, including the optimal two-disparity method used in our codec.

3.2.1. Nearest-Neighbor Reconstruction

This first method has been widely used for DVC [3]. It assumes only one reference image Y is available. For our two-disparity codec, we

can generate a single Y using a weighted sum of the most likely shifted versions of Y_1 and Y_2, in the form

$$\begin{aligned} Y &= P^{(\infty)}\{D_1 = d_1^*\} \cdot Y_{1,d_1^*} + P^{(\infty)}\{D_2 = d_2^*\} \cdot Y_{2,d_2^*} \\ d_k^* &= \underset{d \in \{d_1,\ldots,d_M\}}{\operatorname{argmax}} P^{(\infty)}\{D_k = d\} \qquad k = 1,2 \end{aligned} \tag{5}$$

where $P^{(\infty)}$ denotes the final probability after convergence of EM. If $Y^T \in [Z_i, Z_{i+1}]$, the value is used to estimate X^T. Otherwise, Y^T falls outside the correct bin and the estimate of X^T is clipped to the boundary of the nearest neighboring bin.

3.2.2. Optimal One-Disparity Reconstruction

The assignment in nearest-neighbor reconstruction only approximates the optimal one-disparity reconstruction, which is the conditional expectation $E\left[X^T | X^T \in [Z_i, Z_{i+1}], Y^T\right]$. This expectation has been analytically evaluated for a Laplacian noise model [9].

3.2.3. Optimal Two-Disparity Reconstruction

For two reference images, the two previous SIAR methods are suboptimal, and the optimal two-disparity reconstruction is the conditional expectation $E\left[X^T | X^T \in [Z_i, Z_{i+1}], Y_1^T, Y_2^T\right]$. A closed-form solution for the two-disparity case is also given in [9] for the Laplacian noise model. The authors of [9] calculate the optimal shifts d_1^* and d_2^* using MCTI and then assume for simplicity the two disparities are equally probable. Since our two-disparity codec learns the probabilities of the disparities, a more general reconstruction formula is proposed here. Breaking the interval $[Z_i, Z_{i+1}]$ into J non-overlapping subintervals $Z_i = q_0 < q_1 < \ldots < q_J = Z_{i+1}$ so that simple integrals can be used, the solution is derived to be

$$\begin{aligned} \hat{X}^T &= \frac{\sum_{k=1}^{2} P^{(\infty)}\{D_k = d_k^*\} \sum_{j=0}^{J-1} E_1^{k,j}}{\sum_{k=1}^{2} P^{(\infty)}\{D_k = d_k^*\} \sum_{j=0}^{J-1} E_0^{k,j}} \\ E_1^{k,j} &= \frac{\lambda_k}{2} \int_{q_j}^{q_{j+1}} X^T \exp\left(-\lambda_k \left| X^T - Y_{k,d_k^*}^T \right|\right) dX^T \\ E_0^{k,j} &= \frac{\lambda_k}{2} \int_{q_j}^{q_{j+1}} \exp\left(-\lambda_k \left| X^T - Y_{k,d_k^*}^T \right|\right) dX^T \end{aligned} \tag{6}$$

where λ_k $(k = 1, 2)$ are the parameters of the Laplacian noise terms N_k $(k = 1, 2)$ from (4) and d_k^* is calculated from (5). In Section 4, experimental results will show that two-disparity reconstruction outperforms the other two methods.

4. EXPERIMENTAL RESULTS

The performance of the one-disparity and two-disparity codecs are evaluated using two sets of multiview images from [11], which we call *Teddy* and *Barn*. One image from each set is shown in Fig. 3(a-b) and takes the role of the image X which is Wyner-Ziv coded. Two other images in each set are chosen to serve as the reference images Y_1 and Y_2. We assume high-quality versions of Y_1 and Y_2 reside at the decoder, having been previously transmitted using conventional coding. For simplicity, disparities are only estimated at block resolution, and the search space is constrained to horizontal integer-pel shifts of blocks in Y_1 or Y_2. Fig. 3(d-g) shows the minimum mean-squared-error disparities D_1 and D_2, for a block size of 8.

The rate-distortion performance of several decoders are compared. First, an impractical decoder called an oracle receives the blockwise disparities from Fig. 3(d-g). Disparity values are obtained

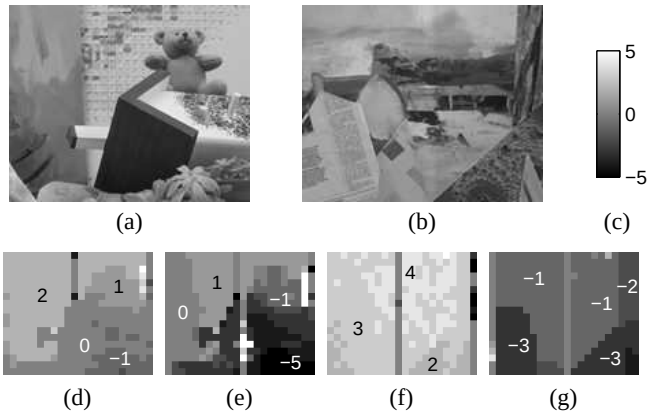

Fig. 3. Image X from multiview sets (a) *Teddy* and (b) *Barn*, each of size 176-by-144 and bit depth 8. (c) Disparity legend. Horizontal disparites from (d) Y_1 to X for *Teddy*, (e) Y_2 to X for *Teddy*, (f) Y_1 to X for *Barn*, and (g) Y_2 to X for *Barn*.

by block matching using either reference image Y_1 or Y_2. The oracle resembles a conventional coding scheme in which disparity is calculated at the encoder and transmitted to the decoder. It is meant to measure an upper performance bound for practical disparity learning. Second, the one-disparity decoder is tested using Y_1 or Y_2 individually. Third, the proposed two-disparity decoder is tested using Y_1 and Y_2 together. For all decoders, the EM algorithm is permitted to run for 50 iterations at each incremental rate of the rate-adaptive LDPC code. If after 50 iterations convergence is not achieved, the LDPC decoder advances to the next higher incremental rate.

Rate-PSNR plots for *Teddy* and *Barn* are presented in Fig. 4. For *Teddy*, the two-disparity decoder achieves up to 0.9 dB and 2.3 dB increases over one-disparity decoding with Y_1 and Y_2, respectively. For *Barn*, the two-disparity decoder again performs better, producing gains of up to 5.3 dB and 2.1 dB over one-disparity decoding with Y_1 and Y_2, respectively. Gaps between two-disparity decoding and the oracle reflect the inefficiency of decoder-side disparity learning relative to traditional encoder-side disparity search.

The key reason for the two-disparity decoder's superior performance is that it can select, on a blockwise basis, the reference image that minimizes the mismatch with X in the transform domain. Let d_1^* and d_2^* be defined as in (5). Fig. 5 shows the fraction of blocks for which $P\{D_1 = d_1^*\} \gg P\{D_2 = d_2^*\}$, the fraction for which $P\{D_1 = d_1^*\} \ll P\{D_2 = d_2^*\}$, and the fraction for which $P\{D_1 = d_1^*\} \approx P\{D_2 = d_2^*\}$. The last case, where Y_1 and Y_2 are almost equally good choices, resembles classical two-hypothesis prediction. If either Y_1 or Y_2 is unavailable, as is the case for one-disparity decoding, Fig. 5 shows that the decoder would be forced to make suboptimal choices for a significant fraction of blocks. Fig. 5 also shows that the fractions depend on rate. At low rates, quantization is coarse and the two sets of quantized transform coefficients Y_1^{TQ} and Y_2^{TQ} are similar, so many blocks of X^{TQ} can be accurately compensated using both Y_1^{TQ} and Y_2^{TQ}. At high rates, quantization is finer and the differences between Y_1^{TQ} and Y_2^{TQ} are more pronounced, so most blocks of X^{TQ} match significantly better with the quantized transform coefficients of one reference image than with the other.

The two-disparity decoder uses the optimal two-disparity reconstruction. If nearest-neighbor or optimal one-disparity reconstruction is used instead in the reconstruction block of Fig. 2, the quality

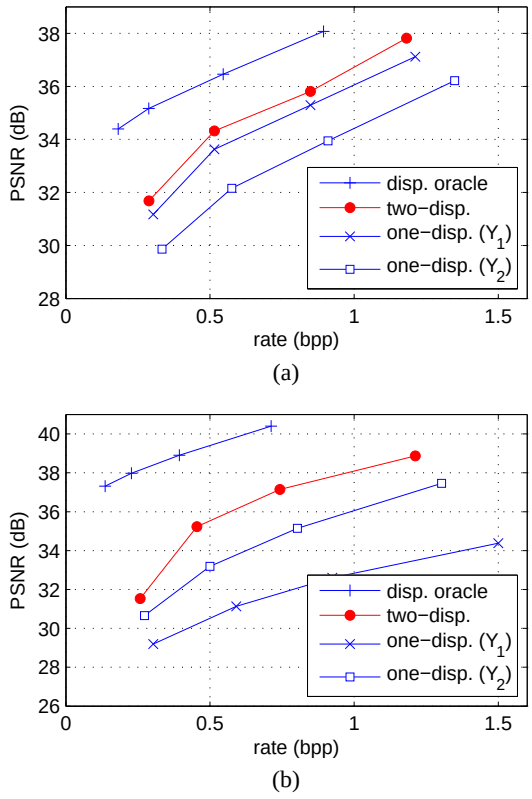

Fig. 4. Rate-PSNR performances for four different Wyner-Ziv decoders, for (a) *Teddy* and (b) *Barn*.

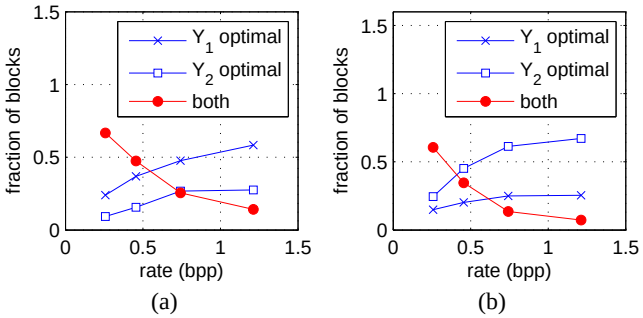

Fig. 5. Fraction of blocks with most probable shift candidate from Y_1, Y_2, or both reference images, for (a) *Teddy* and (b) *Barn*.

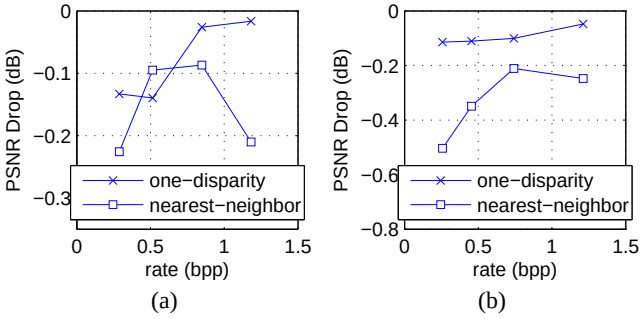

Fig. 6. PSNR drop from two-disparity reconstruction using suboptimal reconstruction for (a) *Teddy* and (b) *Barn*.

of the final image will be reduced. Fig. 6 displays the PSNR drops relative to two-disparity reconstruction resulting from usage of suboptimal reconstructions. The reductions are as large as 0.2 dB and 0.5 dB for *Teddy* and *Barn*, respectively.

5. CONCLUSION

This paper presents a Wyner-Ziv image codec that learns two disparities in an unsupervised fashion at the decoder. The proposed two-disparity decoder generalizes the statistical estimation framework of an existing one-disparity decoder and contains an improved reconstruction step. Significant rate-distortion gains are achieved over one-disparity decoding. Future research should investigate extensions of our framework to disparity compensation with three or more reference images and with fractional-pel accuracy.

6. REFERENCES

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, no. 4, pp. 471–480, July 1973.

[2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Info. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[3] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[4] R. Puri, A. Majumdar, and K. Ramchandran, "PRISM: A video coding paradigm with motion estimation at the decoder," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2436–2448, Oct. 2007.

[5] D. Varodayan, Y.-C. Lin, A. Mavlankar, M. Flierl, and B. Girod, "Wyner-Ziv coding of stereo images with unsupervised learning of disparity," in *Proc. Picture Coding Symposium*, Lisbon, Portugal, Nov. 2007.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the expectation maximization algorithm," *J. Royal Stat. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] D. Varodayan, A. Mavlankar, M. Flierl, and B. Girod, "Distributed grayscale stereo image coding with unsupervised learning of disparity," in *Proc. Data Compr. Conf.*, Snowbird, Utah, March 2007, pp. 143–152.

[8] K. Misra, S. Karande, and H. Radha, "Multi-hypothesis distributed video coding using low-density parity-check codes," in *Proc. Allerton Conf. on Commun., Control, Comput.*, Monticello, IL, Sept. 2005.

[9] D. Kubasov, J. Nayak, and C. Guillemot, "Optimal reconstruction in Wyner-Ziv video coding with multiple side information," in *Proc. Internat. Workshop on Multi. Signal Process.*, Crete, Greece, Oct. 2007.

[10] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive distributed source coding using low-density parity-check codes," in *Proc. Asilomar Conf. on Signals, Systems, Comput.*, Pacific Grove, CA, Nov. 2005, pp. 1203–1207.

[11] D. Scharstein and R. Szeliski, "Middlebury stereo vision page," http://vision.middlebury.edu/stereo.