Distributed Compression of Stereoscopic Images with Unsupervised Learning of Disparity

David Varodayan and Aditya Mavlankar
Information Systems Laboratory, Department of Electrical Engineering
Stanford University, Stanford, CA 94305
Email: {varodayan, maditya}@stanford.edu

***Abstract*— Distributed compression is particularly attractive for stereoscopic images since it avoids communication between cameras. Since compression performance depends on exploiting the redundancy between images, knowing the disparity is important at the decoder. Unfortunately, distributed encoders cannot calculate this disparity and communicate it. In this paper, we propose an Expectation Maximization algorithm to perform unsupervised learning of disparity during the decoding procedure. Experimental results show that this performs nearly as well as a system which knows the disparity through an oracle.**

I. Introduction

Colocated pixels from pairs of stereoscopic images are very statistically dependent after compensation for disparity induced by the geometry of the scene. Much of the disparity between these images can be characterized as shifts of foreground objects relative to the background. Assuming that the disparity information and occlusions can be coded compactly, joint lossless compression is much more efficient than separate lossless encoding and decoding. Surprisingly, distributed lossless encoding combined with joint decoding can be just as efficient as the wholly joint system, according to the Slepian-Wolf theorem [1]. Distributed compression is preferred because it avoids communication between the stereo cameras. The difficulty, however, lies in discovering and exploiting the scene-dependent disparity at the decoder, while keeping the transmission rate low.

A similar problem arises in the area of low complexity encoding of video captured by a single camera [2] [3]. These systems encode frames of video separately and decode them jointly, so discovering the motion between successive frames at the decoder is helpful. One very computationally burdensome way to learn the motion is to run the decoding algorithm with every motion realization [3]. Another approach requires the encoder to transmit additional hashed information, so the decoder can select a good motion configuration before running the decoding algorithm [4]. Since the encoder transmits the hashes at a constant rate, it wastes bits when there is little motion. On the other hand, if there is too much change between frames, the fixed-rate hash may be insufficient for reliable motion search. Due to the drawbacks of excessive computation and difficulty of rate allocation for the hash, we use neither of these approaches for the compression of stereoscopic images.

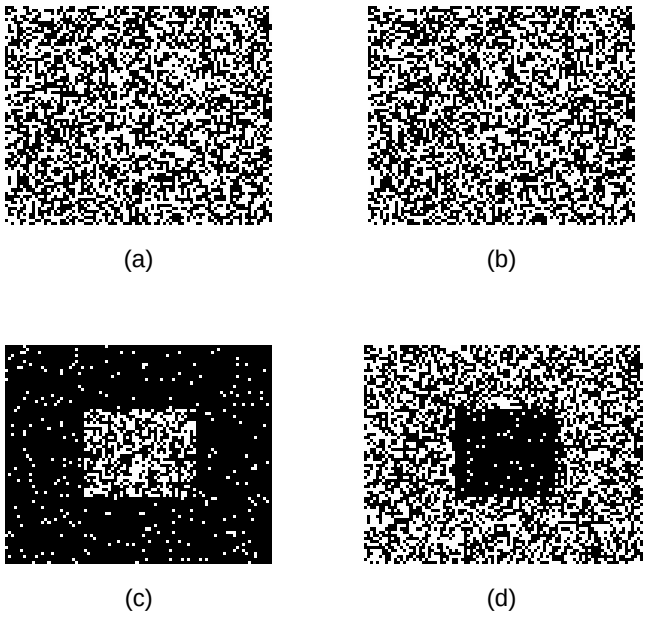

Fig. 1. (a) Source image X (b) Source image Y (c) Sum of X and Y modulo 2 (d) Sum of X and Y modulo 2 (shifted to realign the shifted rectangle)

In Section II, we consider a toy version of the problem and propose a novel decoding algorithm, which learns disparity unsupervised. We describe the algorithm formally within the framework of Expectation Maximization (EM) in Section III. Section IV reports our simulation results.

II. Random Dot Stereogram Compression

We model stereoscopic images X and Y as binary random dot stereograms [5]. The disparity information D governs the relationship between X and Y. We generate Y by copying X and shifting an arbitrary rectangular region of it horizontally. The newly revealed area is filled randomly. Finally, independent identically distributed (i.i.d.) binary noise is added (modulo 2) to Y to mimic camera noise. Thus, D consists of the boundaries of the shifted rectangle as well as the magnitude and direction of the shift. Fig. 1 shows sample realizations of X and Y and their sums under different shifts. The interesting observation of [5] is that stereoscopic viewing of X and Y as a

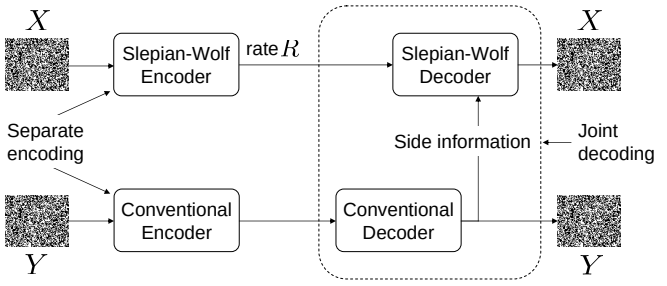

Fig. 2. Distributed compression: separate encoding and joint decoding

single image creates an illusion of depth; the shifted rectangle appears on a different plane compared to the rest of the image.

Our compression setup is shown in Fig. 2. Images X and Y are encoded separately and decoded jointly. For simplicity, we assume that Y is conventionally coded and is available at the decoder. The challenge is to encode X efficiently in the absence of Y so that it can be reliably decoded in the presence of Y. The Slepian-Wolf theorem states that X can be communicated losslessly to the decoder using R bits on average as long as $R > H(X|Y)$ [1].

Fig. 3 depicts three compression systems that can be applied to this problem. The system in Fig. 3(a) performs compression of X with respect to the colocated pixels of Y under the assumption of no disparity [6]. The encoder computes the syndrome S (of length R bits) of X with respect to a Low-Density Parity-Check (LDPC) code [7]. The decoder initially estimates X statistically using the colocated pixels of Y and refines these estimates using S via an iterative belief propagation algorithm. When disparity is introduced between X and Y, this scheme performs badly because the estimates of X are poor in the shifted region. For comparison, Fig. 3(b) shows an impractical scheme in which the decoder is endowed with a disparity oracle. The oracle informs the decoder which pixels of Y should be used to inform the estimates of the pixels of X during LDPC decoding. Finally, Fig. 3(c) depicts our proposed practical decoder that learns disparity D via EM. The disparity oracle of Fig. 3(b) is replaced by a disparity estimator, which maintains an *a posteriori* probability distribution on D. Every iteration of LDPC decoding sends the disparity estimator a soft estimate of X (denoted by θ) in order to refine the distribution on D. In return, the disparity estimator updates the side information ψ for the LDPC decoder by blending information from the pixels of Y according to the refined distribution on D. The following section formalizes the process in terms of EM.

III. Expectation Maximization Algorithm

Let X be a binary image of size m-by-n, in which pixels $X(i,j)$ form an i.i.d. equiprobable Bernoulli random process. Define L to be a random integer representing the disparity shift and constrain $|L| \leq l \ll n$. Define also random indices $M_1 \leq M_2$ and $N_1 \leq N_2$ to be the vertical and horizontal boundaries of the disparity region, respectively. Thus, D is the

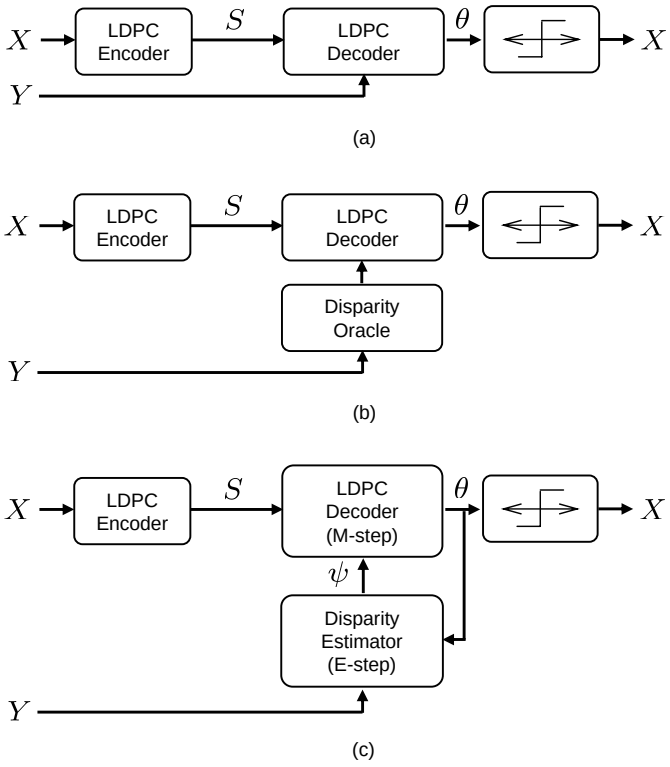

Fig. 3. (a) Distributed compression assuming no disparity (b) Distributed compression with a disparity oracle (c) Distributed compression with unsupervised learning of disparity D via EM

tuple (L, M_1, M_2, N_1, N_2). Let R and Z be (M_2-M_1+1)-by-(N_2-N_1+1) and m-by-n binary images, respectively, where $R(i,j)$ and $Z(i,j)$ form i.i.d. Bernoulli random processes with $P\{R(i,j) = 1\} = 0.5$ and $P\{Z(i,j) = 1\} = \epsilon \leq 0.5$. Generate the image Y as follows using R to fill in the newly revealed area and Z as noise. Notice that the pixels $Y(i,j)$ form an i.i.d. equiprobable Bernoulli random process.

$$\begin{aligned} Y &:= X \\ Y(M_1 : M_2, N_1 : N_2) &:= R \\ Y(M_1 : M_2, N_1 + L : N_2 + L) &:= X(M_1 : M_2, N_1 : N_2) \\ Y &:= Y \oplus Z \end{aligned}$$

We denote the *a posteriori* probability distribution of D as $P_{app}\{D\}$ and that of X as

$$\begin{aligned} P_{app}\{X\} &\approx \prod_{i,j} P\{X(i,j)\} \\ &= \prod_{i,j} \theta(i,j)^{X(i,j)}(1-\theta(i,j))^{1-X(i,j)}, \end{aligned}$$

where the $\theta(i,j) = P_{app}\{X(i,j) = 1\}$ are parameters. Thus, θ can be interpreted as a soft estimate of X.

The disparity estimator in Fig. 3(c) performs the E-step of

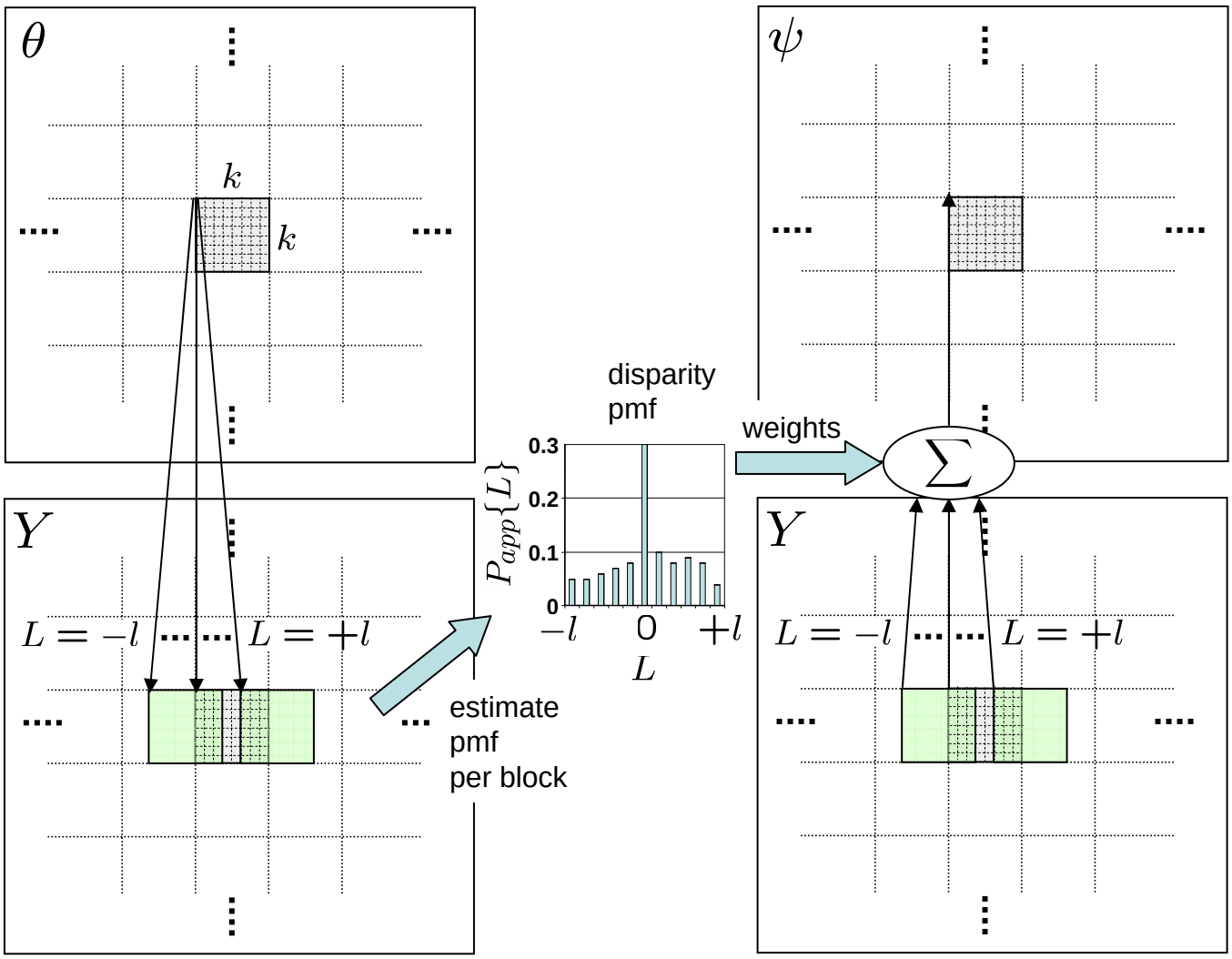

Fig. 4. The E-step: disparity estimation and side information blending

EM; namely, the following refinement of $P_{app}\{D\}$.

$$\begin{aligned} P_{app}\{D\} &:= P\{D|Y,S;\theta\} \\ &\propto P\{D\}P\{Y,S|D;\theta\} \\ &\approx P_{app}\{D\}P\{Y|D;\theta\} \end{aligned}$$

The approximation in the last step is reasonable since θ is iteratively reconciled with S during the M-step. But the E-step, as written above, is expensive due to the vast number of possible values of D. To reduce computation, the disparity estimator first learns only the shift L block-by-block as shown in the left hand side of Fig. 4. For a specified blocksize k, every k-by-k block of θ is compared to the colocated block of Y as well as all those shifted between $-l$ and l pixels horizontally. For a block $\theta_{u,v}$ with top left pixel located at (u,v), the distribution on the shift $L_{u,v}$ is updated by

$$P_{app}\{L_{u,v}\} :\approx P_{app}\{L_{u,v}\}P\{Y_{u,v+L_{u,v}}|L_{u,v};\theta_{u,v}\},$$

where $Y_{u,v+L_{u,v}}$ is the k-by-k block of Y with top left pixel at $(u, v + L_{u,v})$.

Finally, the disparity estimator creates estimates $\psi_{u,v}$ of the block $X_{u,v}$ by blending estimates from each of the blocks $Y_{u,v+L_{u,v}}$ according to the distribution $P_{app}\{L_{u,v}\}$, as shown in the right hand side of Fig. 4. More generally, this can be described as

$$\psi(i,j) = \sum_d P_{app}\{D=d\}P\{X(i,j)|D=d,Y\}.$$

The LDPC decoder performs the M-step; namely, the following maximization of the likelihood of Y and the syndrome S.

$$\begin{aligned} \theta &:= \arg\max_\theta P\{Y,S;\theta\} \\ &= \arg\max_\theta \sum_d P\{D=d\}P\{Y,S|D=d;\theta\} \\ &\approx \arg\max_\theta \sum_d P_{app}\{D=d\}P\{Y,S|D=d;\theta\} \end{aligned}$$

True maximization is intractable, so we approximate it with an iteration of LDPC decoding.

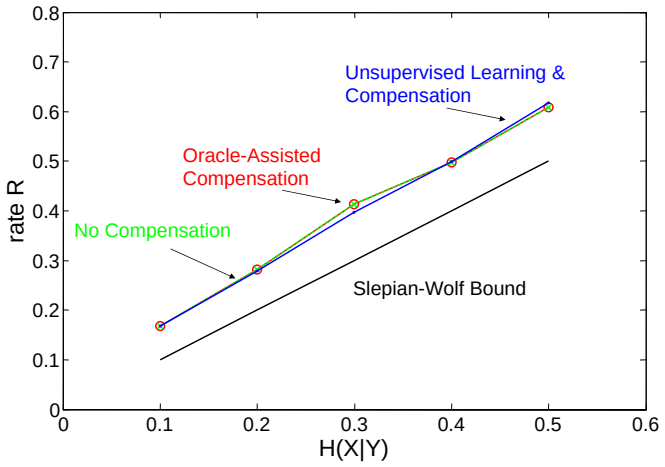

Fig. 5. Rate (in bit/pixel) required to communicate X for the different systems shown in Fig. 3, when there is no disparity. The codes used are regular of degree 3.

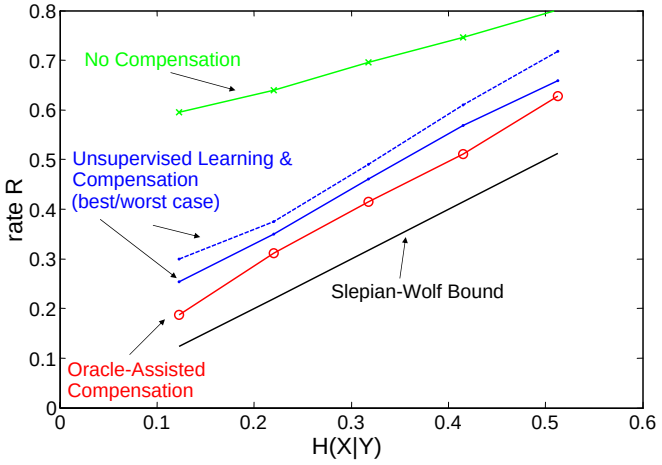

Fig. 6. Rate (in bit/pixel) required to communicate X for the different systems shown in Fig. 3, when there is disparity of size 32-by-32 and shift 5. The codes used are regular of degree 3.

Iterating between the E-step and the M-step in this way provides a coarse profile of the disparity, limited by the granularity of k-by-k blocks. We refine the estimate of D by estimating the disparity region boundary variables $\{M_1, M_2, N_1, N_2\}$, once several contiguous blocks agree upon a value for L. For simplicity, instead of maintaining probability distributions on the values of $\{M_1, M_2, N_1, N_2\}$, we estimate a single value for each boundary variable and refine it during every iteration. This improves the fineness of compensation of the side information beyond the granularity of k-by-k blocks.

The decoding algorithm terminates successfully when the thresholded estimate $\hat{X} = \mathbf{1}\{\theta > 0.5\}$ yields syndrome equal to S.

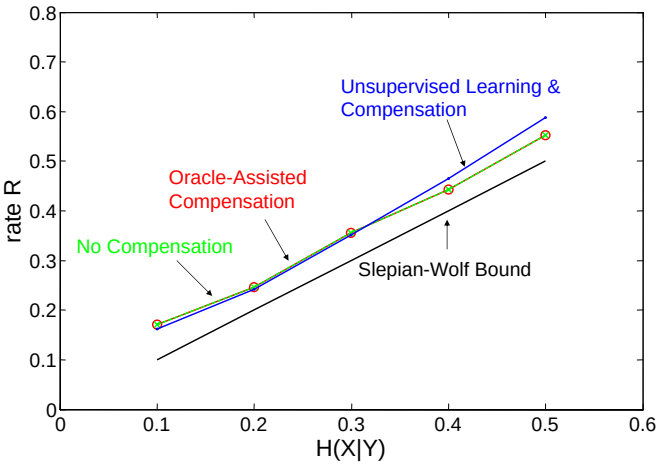

Fig. 7. Rate (in bit/pixel) required to communicate X for the different systems shown in Fig. 3, when there is no disparity. The codes used are irregular of degree ranging from 2 to 21.

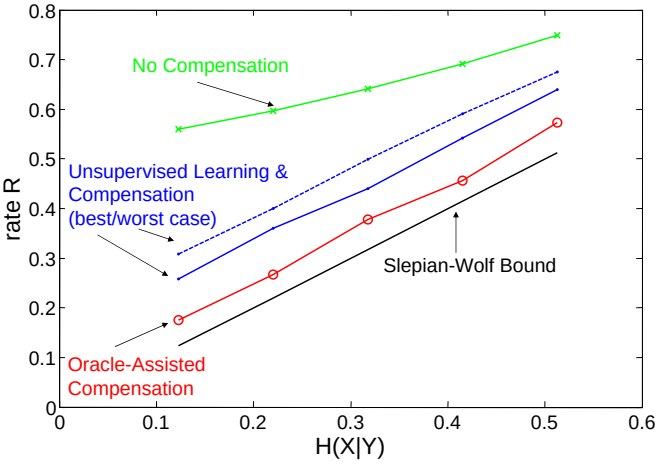

Fig. 8. Rate (in bit/pixel) required to communicate X for the different systems shown in Fig. 3, when there is disparity of size 32-by-32 and shift 5. The codes used are irregular of degree ranging from 2 to 21.

IV. Simulation Results

For our simulations, we select the following constants: image height $m = 72$, image width $n = 88$, maximum horizontal shift $l = 5$, blocksize $k = 8$. The camera noise parameter $\epsilon = P\{Z(i,j) = 1\}$ ranges between 0.01 and 0.11. The distributions of $L_{u,v}$ are initialized to

$$P_{app}\{L_{u,v}\} := \begin{cases} 0.75, & \text{if } L_{u,v} = 0; \\ 0.025, & \text{if } L_{u,v} \neq 0. \end{cases}$$

Rate control is implemented as follows. After 150 decoding iterations, if $\hat{X}$ still does not satisfy the syndrome condition, the decoder requests additional transmission rate from the encoder. We employ rate-adaptive codes as described in [8].

Figs. 5 and 6 show the performance of the systems in Fig. 3 using regular codes of degree 3, when there is no disparity and when there is disparity, respectively. In Fig. 6,

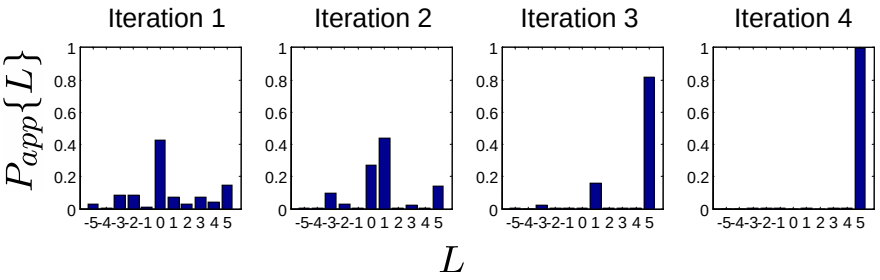

Fig. 9. Evolution of a disparity probability distribution for a sample 8-by-8 block

the disparity region is 32-by-32 pixels in size and the shift is $L = 5$. For the proposed scheme of Fig. 3(c), we show results when the disparity region is aligned with the 8-by-8 block grid (best case) and when it is offset from the grid by 4 pixels horizontally and vertically (worst case). Figs. 7 and 8 show the corresponding performance using irregular codes with degree distribution ranging from 2 to 21, when there is no disparity and when there is disparity, respectively.

Figs. 5 and 7 indicate that with no disparity, the compression performance is identical for all three systems shown in Fig. 3. This is because they all generate the same side information for the LDPC decoder. When disparity exists, Figs. 6 and 8 demonstrate that only the oracle-assisted system performs as close to the Slepian-Wolf bound as before. The system that assumes no disparity performs up to 3 times worse than before because the side information is very unreliable in the disparity region. The proposed unsupervised learning system does significantly better than this, and comes close to the performance of the impractical oracle-assisted scheme.

To illustrate the progress of the unsupervised learning decoding algorithm, we show how the disparity probability distribution evolves for a sample 8-by-8 block in Fig. 9.

V. Conclusions

For the simplified problem of distributed compression of random dot stereograms, unsupervised learning of disparity is superior to ignoring disparity and is also practical. To our knowledge, there is no literature on applying unsupervised learning of disparity to either realistic distributed stereoscopic image compression or realistic low complexity video compression. This suggests an interesting research direction.

VI. Acknowledgments

We thank Prof. Bernd Girod and Markus Flierl for useful discussions.

References

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

[2] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Proc. Asilomar Conf. on Signals, Syst., Comput.*, Pacific Grove, CA, 2002.

[3] R. Puri and K. Ramchandran, "PRISM: a new robust video coding architecture based on distributed compression principles," in *Proc. Allerton Conf. Commun., Contr. and Comput.*, Allerton, IL, 2002.

[4] A. Aaron, S. Rane, and B. Girod, "Wyner-Ziv video coding with hash-based motion compensation at the receiver," in *Proc. IEEE International Conf. on Image Processing*, Singapore, 2004.

[5] B. Julesz, "Binocular depth perception of computer generated patterns," *Bell Sys. Tech. J*, vol. 38, pp. 1001–1020, 1960.

[6] A. Liveris, Z. Xiong, and C. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes," *IEEE Commun. Lett.*, vol. 6, no. 10, pp. 440–442, Oct. 2002.

[7] R. G. Gallager, "Low-density parity-check codes," *Cambridge MA: MIT Press*, 1963.

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