CODING OF MULTI-VIEW IMAGE SEQUENCES WITH VIDEO SENSORS

Markus Flierl[*] *and Bernd Girod*

Max Planck Center for Visual Computing and Communication
Stanford University, California
mflierl@stanford.edu

ABSTRACT

We investigate coding of multiple image sequences with video sensors. The video sensors are arranged in an array to monitor the same scene from different view points. Furthermore, the sensors are connected to a central decoder via a network. Note that the video sensors process highly view-correlated images. This correlation can be exploited if the video sensors operate in a collaborative fashion. On the contrary, temporal correlation among the images of each sequence can be exploited locally at each sensor. Collaborative coding of the multi-view videos can be achieved by distributed processing of the multi-view imagery. If the video sensor network utilizes a central decoder, the view-correlation can be exploited by centralized disparity compensation at the decoder. But before the decoder is able to apply disparity compensation efficiently, accurate disparity values have to be estimated at the central decoder. This paper discusses the impact of disparity fields at the central decoder and uses these estimates for centralized disparity compensation at the decoder to improve the efficiency of the video sensor network.

Index Terms— Multi-view video coding, video sensors

1. INTRODUCTION

Video camera arrays can be used to sample dynamic scenes by generating multi-view image sequences. Such arrays may be part of three-dimensional TV systems which enable users to view a distant 3D world freely [1]. Unfortunately, the high processing and bandwidth requirements of end-to-end 3D TV exceed the capabilities of most systems [2]. A critical component of such systems is the coding engine that compresses the multi-view video data into a rate-distortion efficient representation. The most straightforward approach to the multi-view coding problem is to temporally encode the individual video streams independent of one another [2]. But efficient coding can be achieved by exploiting the correlation in temporal direction as well as the correlation among the views.

The correlation among the views can be exploited with two different approaches: In one possible compression scenario, encoders of the sensor signals are connected and compress the signals jointly. The disadvantage of this scenario is that the encoder have to share, i.e., exchange, their information and that coding decisions of each encoder are directly affected by several neighboring sensors. In an alternative compression scenario, each encoder operates independently but relies on a joint decoding unit that receives all coded sensor signals. This is also known as distributed source coding. The advantage of this scenario is that the encoder do not have to share their information directly. A special case of this scenario is source coding with side information. Wyner and Ziv showed that for certain cases the encoder does not need the side information to which the decoder has access to achieve the rate-distortion bound [3].

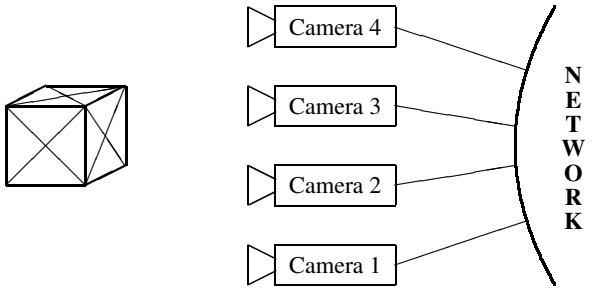

Fig. 1. Capturing multi-view sequences with video sensors.

Fig. 1 depicts the distributed source coding scenario where the video sensors are connected to a central decoder via a network. The central decoder is able to perform efficiently disparity estimation on previously decoded images of the multi-view image sequences. With this information available, highly efficient disparity compensated side information can be generated to decode the current frames of the multi-view image sequences more efficiently.

The paper is organized as follows: Section 2 outlines our coding scheme with video sensors. We discuss the encoder of a video sensor, the central decoder and the way it exploits the side information, as well as the disparity compensation of the side information. Section 3 studies the impact of the disparity compensation on the rate efficiency of the distributed scheme. Experimental results with the coding scheme are presented in Section 4.

2. CODING SCHEME WITH VIDEO SENSORS

Our coding scheme uses video sensors to capture multi-view image sequences of a dynamic scene. The video sensors operate in a distributed source coding scenario where each video sensor is directly connected to the central decoder. The temporal correlation within each image sequence is exploited locally at the video sensor. The inter-view correlation may be exploited at the central decoder to further improve the efficiency of the coding scheme. The central decoder is able to perform efficiently disparity estimation on decoded images of the multi-view image sequences. With this information available, highly efficient disparity compensated side information can be generated to decode the frames of the multi-view image sequences more efficiently.

Fig. 2 shows the distributed coding scheme with disparity compensation at the central decoder. N multi-view image sequences are

[*]This work has been supported by the Max Planck Center for Visual Computing and Communication.

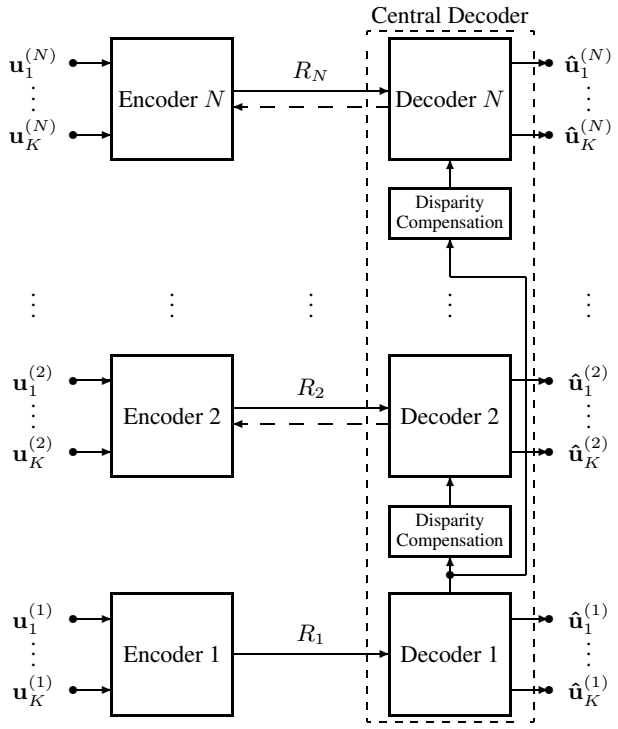

Fig. 2. Distributed coding scheme with disparity compensation at the central decoder.

represented by $\mathbf{u}_k^{(n)}$ with $k = 1, 2, \ldots, K$ temporally successive frames of $n = 1, 2, \ldots, N$ views. The coding scheme comprises N encoder that operate independently as well as one central decoder. The latter is made up of $N-1$ "Wyner-Ziv" decoder $n = 2, \ldots, N$ that are dependent on *Decoder 1*. The side information for *Decoder* n with $n = 2, \ldots, N$ can be improved by performing disparity compensation. As the video signals are not stationary, *Decoder* n with $n = 2, \ldots, N$ is decoding with feed-back.

2.1. Encoder of a Video Sensor

As mentioned before, the video sensor is able to exploit the temporal correlation. In our scheme, K temporally successive images of a sequence are encoded with a motion-compensated lifted wavelet transform [4]. Each temporal subband is decorrelated by a spatial transform. The coefficients of this transform are encoded with nested lattice codes. Fig.3 provides an overview of the encoder of a video sensor.

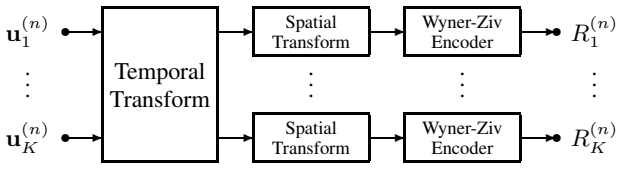

Fig. 3. Encoder of a video sensor exploiting temporal correlation.

The temporal transform performs a dyadic decomposition with a motion-compensated lifted Haar wavelet. It provides K output pictures that are decomposed by a spatial 8×8 DCT. The motion information that is required for the motion-compensated wavelet transform is estimated in each decomposition level depending on the results of the lower level. The correlation of motion information between two image sequences is not exploited. For the motion-compensated lifted Haar wavelet, the even frames of the image sequence are used to predict the odd frames with the estimated motion vectors. The prediction step is followed by an update step which uses the negative motion vectors as an approximation. We use a block-size of 16×16 in the prediction step and select the motion vectors such that they minimize a Lagrangian cost function based on the squared error in the high-band. Additional scaling factors in low- and high-band are used to normalize the transform.

Encoder 1 in Fig. 2 encodes the side information for all *Decoder* n, $n = 2, \ldots, N$, and does not employ distributed source coding principles. A scalar quantizer is used to represent the DCT coefficients of all temporal bands. The quantized coefficients are simply run-level encoded. On the other hand, each *Encoder* n, $n = 2, \ldots, N$ is designed for distributed source coding and uses nested lattice codes to represent the DCT coefficients of all temporal bands. In particular, a 1-dimensional nested lattice code [5] is used where the cosets are constructed in a memoryless fashion [6].

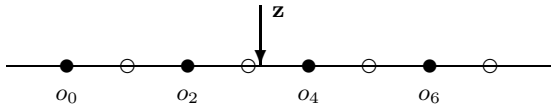

Fig. 4. Coset-coding of transform coefficients where *Encoder 2* transmits at a rate R_{TX} of 1 bit per transform coefficient.

Fig. 4 explains the coset-coding principle. Assume that *Encoder* N transmits at a rate R_{TX} of 1 bit per transform coefficient and utilizes two cosets $\mathcal{C}_{1,0} = \{o_0, o_2, o_4, o_6\}$ and $\mathcal{C}_{1,1} = \{o_1, o_3, o_5, o_7\}$ for encoding. Now, the transform coefficient o_4 shall be encoded and the encoder sends one bit to signal coset $\mathcal{C}_{1,0}$. With the help of the side information coefficient $\mathbf{z}$, the decoder is able to decode o_4 correctly. If *Encoder* N does not send any bit, the decoder will decode o_3 and we observe a decoding error.

Consider the 64 transform coefficients $\mathbf{c}_i$ of the 8×8 DCT at *Encoder* N. The correlation between the i-th transform coefficient $\mathbf{c}_i$ at *Encoder* N and the i-th transform coefficient of the side information $\mathbf{z}_i$ depends strongly on the coefficient index i. In general, the correlation between corresponding DC coefficients ($i = 0$) is very high, whereas the correlation between corresponding high-frequency coefficients decreases rapidly. To encounter the problem of varying correlation, we adapt the transmission rate R_{TX} to each transform coefficient. For weakly correlated coefficients, a higher transmission rate has to be chosen.

Adapting the transmission rate to the actual correlation is accomplished with nested lattice codes [5]. The idea of nested lattices is, roughly, to generate diluted versions of the original coset code. As we use uniform scalar quantization, we consider the 1-dimensional lattice. The fine code shall have a minimum Euclidean distance Q. The nested codes are coarser and the union of their cosets gives the fine code.

The binary representation of the quantized transform coefficients determines its coset representation in the nested lattice. If the transmission rate for a coefficient is $R_{TX} = \mu$, then the μ least significant bits of the binary representation determine the ν-th coset $\mathcal{C}_{\mu,\nu}$. For highly correlated coefficients, the number of required cosets and, hence, the transmission rate is small. To achieve efficient entropy coding of the binary representation of all 64 transform coefficients, we define bit-planes. Each bit-plane is run-length encoded and transmitted to *Decoder* N upon request.

2.2. Decoder using Side Information

At *Encoder* N, the quantized transform coefficients are represented with 10 bit-planes, where 9 are used for encoding the absolute value, and one is used for the sign. *Encoder* N is able to provides the full bit-planes, independent of any side information at the *Decoder* N. *Encoder* N is also able to receive a bit-plane mask to weight the current bit-plane. The masked bit-plane is run-length encoded and transmitted to *Decoder* N.

Given the side information at *Decoder* N, masked bit-planes are requested from *Encoder* N. For that, *Decoder* N sets the bit-plane mask to indicate the bits that are required from *Encoder* N. Dependent on the received bit-plane mask, *Encoder* N transmits the weighted bit-plane utilizing run-length encoding. *Decoder* N attempts to decode the already received bit-planes with the given side information. In case of decoding error, *Decoder* N generates a new bit-plane mask and requests an additional weighted bit-plane.

Decoder N has the following options for each mask bit: If a bit in the bit-plane is not needed, the mask value is 0. The mask value is 1 if the bit is required for error-free decoding. If the information at the decoder is not sufficient for this decision, the mask is set to 2 and the encoded transform coefficient that is used as side information is transmitted to *Encoder* N. With this side information $\mathbf{z}_i$ for the i-th transform coefficient $\mathbf{c}_i$, *Encoder* N is able to determine its best transmission rate $\mu = R_{TX}[i]$. This information is incorporated into the current bit-plane and transmitted to *Decoder* N: Bits that are not needed for error-free decoding are marked with 0. Further, 1 indicates that the bit is needed and its value is 0, and 2 indicates that the bit is needed with value 1.

Decoder N aims to estimate the i-th transform coefficient $\hat{\mathbf{c}}_i$ based on the current transmission rate $\mu = R_{TX}[i]$, the partially received coset $\mathcal{C}_{\mu,\nu}$, and the side information $\mathbf{z}_i$.

$$\hat{\mathbf{c}}_i = \underset{\mathbf{c}_i \in \mathcal{C}_{\mu,\nu}}{\operatorname{argmin}} \, [\mathbf{c}_i - \mathbf{z}_i]^2 \quad \text{given} \quad \mu = R_{TX}[i] \tag{1}$$

With increasing number of received bit-planes, i.e. increasing transmission rate $R_{TX}[i]$, this estimate gets more accurate and stays definitely constant for rates beyond the critical transmission rate $R^*_{TX}[i]$. Therefore, a simple decoding algorithm is as follows: An additional bit is required if the estimated coefficient changes its value when the transmission rate increases by 1. An unchanged value for an estimated coefficient is just a necessary condition for having achieved the critical transmission rate. This condition is not sufficient for error-free decoding and, in this case, *Encoder* N has to determine the critical transmission rate to resolve any ambiguity.

Note that *Decoder* N receives the coded information in bit-plane units, starting with the plane of least significant bits. With each new bit-plane, *Decoder* N utilizes a coarser lattice where the number of cosets as well as the minimum Euclidean distance increases exponentially.

2.3. Disparity-Compensated Side Information

To improve the efficiency of any *Decoder* n, $n = 2, \ldots, N$, the side information from *Decoder 1* is disparity compensated in the image domain. Disparity estimation can be performed on previously decoded images of the multi-view image sequences as they are available at the central decoder. With this information, disparity compensated side information can be generated to decode the current frames of the multi-view image sequences $n = 2, \ldots, N$ more efficiently. As long as temporally previous frames are not available, the side information for *Decoder* n, $n = 2, \ldots, N$, is less correlated and *Encoder* n has to transmit at a higher bit-rate.

3. IMPACT OF DISPARITY COMPENSATION

We consider briefly the general problem of coding a pair of multi-view images. Two coding scenarios are compared - the centralized and the distributed scenario. In the centralized scenario, both multi-view images are known to the encoder. The encoder is able to estimate the disparity between the two images arbitrarily accurate. That is, the disparity between a pair of images is known at the encoder for encoding both images. In the distributed scenario, the true disparity is unknown to each encoder as only one image is available at each encoder. The encoder / decoder may assume an uncertainty interval for the expected disparity. Due to this uncertain disparity information at the distributed encoder / decoder, we will observe a rate loss for distributed coding.

We study this problem with the assumption that the images are scalar Gaussian random fields over a two-dimensional orthogonal grid [4]. Let $\mathbf{u}$ and $\mathbf{w}$ be two such model images. We characterize their relationship by considering the conditional entropy rate $H(\mathbf{u}|\mathbf{w})$. We consider their relationship if the disparity between the images is known, i.e., if the disparity is a deterministic value. We compare to the case where the disparity is uncertain due to a distributed coding scenario.

We model the second image as a shifted version of $\mathbf{u}$, corrupted by additive white Gaussian noise. First, we assume that the disparity is known and, hence, the shift is deterministic. In this case, the mutual information rate between the images $\mathbf{u}$ and $\mathbf{w}$ is independent of the value of the shift, given a constant variance of the additive noise. That is, a deterministic shift does not change the mutual information rate. Second, we assume uncertain disparity information and assign a probabilistic distribution to the shift. Let the resulting image be $\mathbf{s}$. In this case, the mutual information rate between the images $\mathbf{u}$ and $\mathbf{s}$ depends strongly on the variance of this distribution.

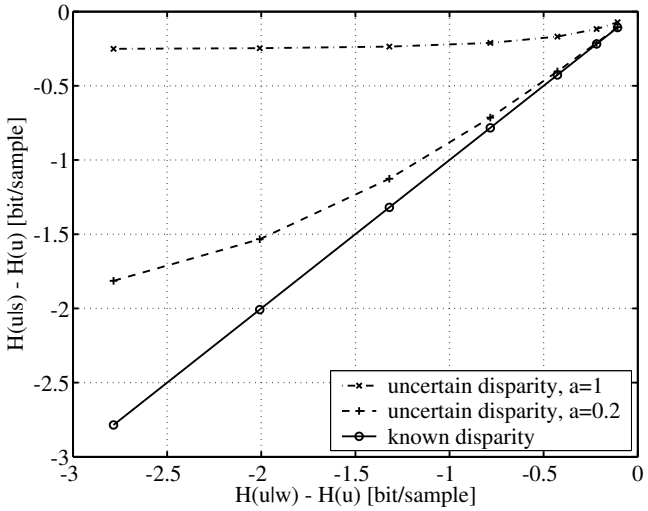

Fig. 5. Conditional entropy rate difference $H(\mathbf{u}|\mathbf{s}) - H(\mathbf{u})$ that is gained when knowing the disparity up to the uncertainty a over the conditional entropy rate difference $H(\mathbf{u}|\mathbf{w}) - H(\mathbf{u})$ when knowing the disparity exactly. For a given uncertainty a, the components of the disparity vector are uniformly distributed in the interval $[-a, a]$.

Fig. 5 depicts the conditional entropy rate difference $H(\mathbf{u}|\mathbf{s}) - H(\mathbf{u})$ that is gained when knowing the disparity up to the uncertainty a over the conditional entropy rate difference $H(\mathbf{u}|\mathbf{w}) - H(\mathbf{u})$ when knowing the disparity exactly. Only if the disparity is exactly known

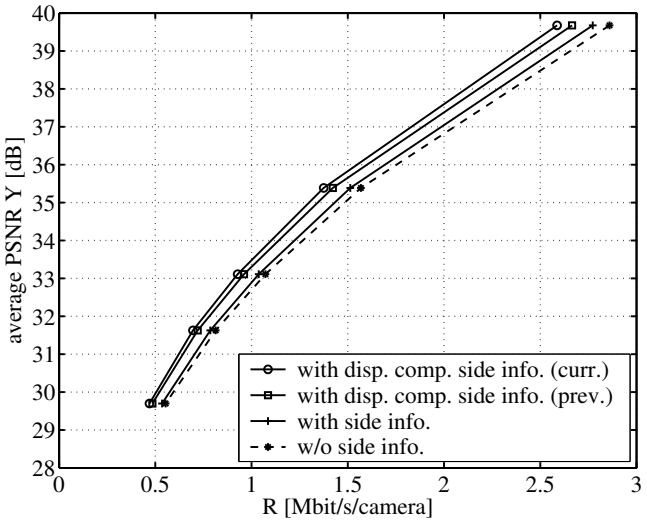

Fig. 6. Average luminance PSNR vs. bit-rate per camera at the Wyner-Ziv encoder / decoder for the multi-view sequence *Jungle* with $N = 8$ views. Compared is decoding with disparity-compensated side information, decoding with coefficient side information only, and decoding without side information. $K = 8$ is used.

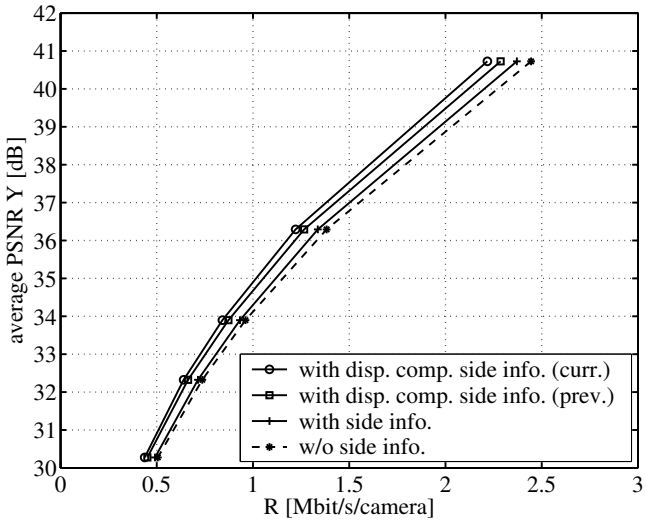

Fig. 7. Average luminance PSNR vs. bit-rate per camera at the Wyner-Ziv encoder / decoder for the multi-view sequence *Uli* with $N = 8$ views. Compared is decoding with disparity-compensated side information, decoding with coefficient side information only, and decoding without side information. $K = 8$ is used.

at the distributed decoder, then there is no rate loss when compared to the centralized coding scenario. If the side information at the distributed decoder is altered due to uncertain disparity information, the conditional entropy rate difference $H(\mathbf{u}|\mathbf{s}) - H(\mathbf{u})$ increases up to zero, where the side information is not helpful at all. This happens when increasing the disparity uncertainty interval $[-a, a]$.

4. EXPERIMENTAL RESULTS

For the experiments, we choose the MPEG-3DAV multi-view image sequences *Jungle* and *Uli*, each with $N = 8$ views, at a spatial resolution of 256×192 [7]. We divide each view with 240 frames at 30 fps into groups of $K = 8$ pictures. The GOPs of view 4 are encoded with *Encoder I* at high quality by setting the quantization parameter $QP = 2$. This coded version of view 4 is used for disparity compensation. The compensated frames provide the side information for *Decoder* n, $n = 2, \ldots, N$, to decode the remaining views.

Figs. 6 and 7 show the average luminance PSNR over the bit-rate per camera of the distributed codec *Encoder* n, $n = 2, \ldots, N$, for the multi-view sequences *Jungle* and *Uli*, respectively. The rate-distortion points are obtained by varying the quantization parameter for the nested lattice in *Encoder* n, $n = 2, \ldots, N$, where the minimum lattice distance is $Q = 2QP$. Disparity estimates derived from previously decoded images (*prev.*) are less accurate and, hence, less efficient than estimates derived from current image pairs (*curr.*). When compared to decoding without side information, decoding with coefficient side information reduces the bit-rate of both multi-view sequences by up to 3%. Decoding with disparity-compensated side information reduces the bit-rate of both by up to 10%.

5. CONCLUSIONS

This paper investigates coding of multi-view image sequences with video sensors that are connected to a central decoder. The video sensors process highly view-correlated images. To take advantage of this correlation, the video sensors operate in a collaborative fashion. The coding scheme utilizes a central decoder which exploits the view-correlation by centralized disparity compensation at the decoder. We have found that distributed coding results in an uncertainty of the disparity information at the Wyner-Ziv decoder. This degrades the coding efficiency when compared to centralized encoding. In the experiments, disparity-compensated side information reduces the bit-rate by up to 10% over decoding without side information. Without disparity compensation, this gain decreases to 3%.

6. REFERENCES

[1] M. Tanimoto, "Free viewpoint television - FTV," in *Proceedings of the Picture Coding Symposium*, San Francisco, CA, Dec. 2004.

[2] A. Vetro, W. Matusik, H. Pfister, and J. Xin, "Coding approaches for end-to-end 3D TV systems," in *Proceedings of the Picture Coding Symposium*, San Francisco, CA, Dec. 2004.

[3] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, pp. 1–10, Jan. 1976.

[4] M. Flierl and B. Girod, "Video coding with motion-compensated lifted wavelet transforms," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 561–575, Aug. 2004.

[5] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1250–1276, June 2002.

[6] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[7] 3DTV Network of Excellence, "Public software and data repository," Nov. 2005. [Online]. Available: https:// www. 3dtv-research. org/ publicSwLibrary. php