HUMAN VISUAL SYSTEM AWARE DECODING STRATEGIES FOR DISTRIBUTED VIDEO CODING

Sören Sofke, Ralph Hänsel, Erika Müller

University of Rostock, Institute of Communications Engineering, Germany

{soeren.sofke,ralph.haensel,erika.mueller}@uni-rostock.de

ABSTRACT

Distributed Video Coding (DVC) is a recent paradigm that offers emerging capabilities in contradiction to the established conventional video coding techniques. Based on the Slepian-Wolf (SW) and Wyner-Ziv (WZ) theorems, a DVC system has an outstanding low encoding complexity, by shifting the computational complex process of correlation exploration to the decoder. The fundamental DVC architecture is organized to reach a competitive Rate-Distortion (RD) performance in terms of PSNR, despite its low correlation with the human visual system (HSV). In contrast, this paper addresses the problem by proposing three image processing tools for exploiting spatio-temporal correlations to reduce the *perceptual* distortion of WZ frames. The proposed WZ pixel domain framework offers a comparable RD-performance referring to H.264 AVC intra coding.

Index Terms— Distributed Video Coding, Perceptual Wyner-Ziv Coding, Human Visual System

1. INTRODUCTION

State-of-the-art video codecs, e.g. MPEG-2, H.264/AVC or VC-1, gain their high RD-performance by exploiting spatio-temporal redundancy at the encoder. Since the encoding is much more complex then the decoding process, this class of codecs fits perfectly in a broadcasting scenario (single transmitter, multiple receivers). However, there is a growing number of applications requiring for low complexity encoding solutions. DVC is known to be a coding technique that allows exploiting redundancy at the decoder and not at the encoder, anymore. This approach allows to design new ubiquitous video capturing devices with low power and lightweight calculating capacities for ultra mobile computing. Girod et. al. [1] introduced one of the first practical Wyner-Ziv (WZ) video codec architecture that has become the most popular WZ reference.

A pixel domain WZ video codec, adopted from [1], is used in this paper, as illustrated in Figure 1. Based on some previously transmitted intra coded key frames (H.264/AVC), the WZ decoder generates the so-called Side Information (SI) by temporal interpolation - an estimation for the WZ frame.

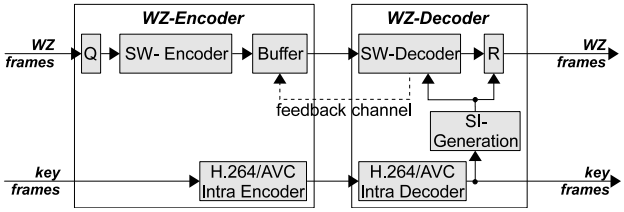

Fig. 1. Pixel domain WZ video codec.

Moreover, pixel values of WZ frames are quantized (Q) before the SW-Encoder performs bit plane extraction over the resultant quantization symbols. Each bit plane is turbo encoded separately and its parity bits are stored into a buffer. Its systematic partition is discarded, since it is already available at the decoder as SI. Only a subset of parity bits are transmitted to correct the SI estimation errors. The difference between the SI and the WZ frame is modeled as additive noise, assuming to have a Laplacian probability density distribution. A feedback channel is used to request for more parity bits until the current plane is decoded successfully. Finally, Minimum Mean Square Error (MMSE) reconstruction [2] (R) corrects the SI estimation errors.

In this paper, perceptual aware (plausible) WZ coding is established. Insight into the codec architecture leads to some major requirements for perceptual adjusted WZ coding, that is quite different from conventional WZ decoding techniques. The decoder has to aspire for very plausible spatial and temporal visual content during the decoding process, in order to increase the perceptual quality of the final decoded frames. In Section 2 principles of image Quality Assessment (QA) are introduced to reveal what the HVS is sensitive to and to derive plausible coding strategies. In Section 3, a set of three advanced image processing tools for perceptual adjusted correlation exploitation are proposed, to better model the correlation between information already available at the decoder. Notably, a fine granular SI interpolation algorithm, a SI refinement algorithm and an algorithm for decoder side spatial correlation exploration are derivated. The performance of the developed tools is summarized in Section 4, highlighting their gains of up to 4-9 dB. Conclusions are given in Section 5.

2. QUALITY ASSESMENT

For perceptual coding, it is important to understand what the HVS is sensitive to. Therefore, different Full Reference (FR) QA metrics are studied. This objective metrics attempt to quantify the visibility of errors of distorted images by modeling the human visual perception, instead of traditional error summation (e.g. PSNR, MSE, SAD).

Weighted PSNR (wPSNR) - Considering the spatial Contrast Sensitivity (CS) threshold of vision, it is suggested in [3] to filter the residual between a reference image and a distorted image using the CS function frequency response matrix to model the perceptual difference between images.

Mean Structural Similarity (mSSIM) - The perceptual quality of a distorted image is assessed by modeling its structural similarity referring to its reference, as proposed in [4]. Luminescence, contrast and structural distortion are taken into account.

Visual Information Fidelity (VIF) - An information fidelity approach is used in [5] to quantify the relation of Shannon information loss between a reference image and its distorted version. Above, natural scene statistics in conjunction with HVS modeling are exploited.

To compare the performance of the introduced QA metrics, their correlation to representative subjective image distortion ratings, provided by [6], is analyzed. Validation criteria are Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and linear Correlation Coefficient (CC). The studied QA metrics are nonlinearly mapped to predict the subjective Distortion (D) using a logistic function:

$$D(metric) = \frac{\beta_1 - \beta_2}{1 + \exp\frac{metric-\beta_3}{\beta_4}} + \beta_2 \qquad (1)$$

whereby β_i is obtained by the Nelder-Mead method. Discussing the results presented in Table 1, all introduced QA metrics outperform the widely used PSNR predicting the true perceptual distortion. VIF gives superior accuracy. Hence, in this paper it is used as the reference metric to validate the final WZ coding performance. Also wPSNR and mSSIM are useful to derive new concepts that come to more perceptual adjusted results in WZ video coding.

Table 1. Correlation between subjective D and predicted D(metric)

Metric	RMSE	MAE	CC
PSNR	9.1235	7.3248	0.8240
wPSNR	7.3025	5.6470	0.8913
mSSIM	8.1261	6.2751	0.8634
log_{10}VIF	**5.0266**	**3.8667**	**0.9500**

3. PERCEPTUAL CODING TECHNIQUES

This section intends to introduce three novel image processing tools for WZ coding to purposefully improve the perceptual quality of WZ coded video frames.

3.1. Pixel-Based Temporal Interpolation (PBTI)

Algorithms for SI generation presented in the literature are based almost exclusively on block-based motion estimation and linear compensation techniques [7, 8]. These approaches are adopted from conventional residual-based video coding schemes, where the encoder reduces the amount of information by expressing the motion of many pixels by fewer motion vectors. In DVC, this kind of information reduction is of no account, since motion compensation is performed at the decoder. Empirical studies demonstrate that block-based motion compensation leads to blurry and blocky artifact-oriented SI of low objective and subjective quality.

PBTI is an advanced approach for perceptual aware SI generation and low distortion regarding the WZ frame. As exhibited in Section 2, the structural information of an image is essential for the quality of visual perception. While using block-based motion estimation technologies, the structural information becomes distorted, especially for sequences with morphologic or inhomogeneous motion characteristics. Hence, it is proposed to use fine granular motion estimation at pixel level. The design of PBTI is described in the following walkthrough:

1. Unidirectional motion estimation (forward and backward) is performed between key frame $t-1$ and $t+1$ to predict the motion of each pixel (x_0, y_0). To conserve the spatial reference, a modified block matching algorithm is adopted. The actual pixel is located in the center of the Matching Window (MW) surrounded by its spatial neighborhood. A Mean weighted Absolute Difference (MwAD) is used as block matching criteria, that has to be minimized regarding the motion vector (Δx,Δy). By using a Gaussian Window (GW) as weight, differences in the center of the matching block are stronger penalized than distortions near the search window border (see Equation 2).

$$(\Delta x, \Delta y) = \min_{\Delta x, \Delta y} \sum_{x,y} |MW_{t-1,x_0,y_0}(x,y) - MW_{t+1,x_0+\Delta x,y_0+\Delta y}(x,y)| \times GW(x,y) \qquad (2)$$

2. The initial Side Information (iSI) is obtained by linear pixel interpolation regarding the two temporal adjacent key frames and the corresponding motion vector fields.
3. Post processing is performed by bidirectional PBTI for all unallocated areas of the iSI.

The final SI distinguishes itself with very plausible image content, clear and sharp edges, without blurry and blocky dis-

continuation as known from block-based algorithms. In comparison to the often used BiMESS algorithm proposed in [8], it gains up to 2-4 dB in PSNR.

3.2. Guided Side Information Refinement (gSIR)

After the SI is estimated, the SW decoder starts to correct the remaining SI estimation errors. Even if the SI is plausible itself, there might be a spatial object misregistration between SI and WZ objects, resulting from nonlinear or inhomogeneous real object motion. The SW-decoder starts to displace already spatial plausible estimated areas. This property is desirable, since it forces the SI to be also temporally plausible. Admittedly, quantization is applied to the WZ frames at the encoder and only the most significant bit planes are decoded. The dynamic range above the quantization level preserves the spatial mismatch and leads to a loss of sharpness and comes up with ghosting artifacts, notably in areas of clear shapes.
To overcome this problem, the gSIR technique is proposed. A partially decoded SI is refined by sequential motion compensation. In comparison to refinement techniques known from the literature, e.g. [7], the gSIR is performed at pixel level. It works as follows:

1. A partially decoded SI is used for gSIR by performing the above introduced unidirectional PBTI. Here, the already decoded bit planes are exploited as sampling points to create an improved motion field for nonlinear temporal interpolation.
2. The guided SI and the partial decoded SI are merged using the multiple SI reconstruction approach from [2]. This procedure results in a refined SI, that is spatial and temporal more plausible, especially in the undecoded bit planes.
3. Step (1) and (2) are repeated iteratively for all bit planes to decode.

gSIR is an important tool to ensure the reliability of decoded WZ frames. It increases the PSNR by up to 3 dB, notably for high motion sequences, where the initial motion estimation is less accurate.

3.3. Exploitation of Spatial Correlation by Inpainting

Usually, transform domain WZ coding is employed to take an advantage of spatial redundancy as trade-off between additional encoder complexity and overall coding performance [8]. This techniques lead to some very annoying coding artifacts, as known from conventional DCT or Wavelet coding. Furthermore, spatial decorrelation at the encoder increases its complexity and thereby contradicts the WZ principle by exploiting redundancy at the encoder.
To exploit spatial correlation at the decoder, it is intended to utilize digital inpainting [9]. Inpainting is descended from art reconstruction and competent to fill undefined image areas.

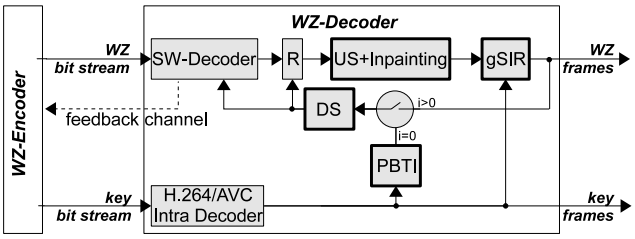

Fig. 2. Proposed WZ decoder.

The proposed method downsamples the WZ frames at the encoder and only uses a subset of its pixels for SW-encoding. At the decoder, an inpainting algorithm is applied to fill up all the missing elements and presents the human observer a plausible image without the need to exactly resample the original. The proposed inpainting provides perceptual aware image content with respect to the contrast sensitivity threshold of the HSV.
Here, a very simple boundary value (BV) solver is used. Based on the position of already reconstructed pixels, an elliptic Partial Differential Equation (PDE) is formulated to be applied on the domain of the unknown pixels. The neighborhood of the unknown pixels supplies BV for the PDE. The benefit of using a downsampled image for WZ coding is to reduce the data rate. Additionaly, it halves the SW-encoder complexity.

3.4. Overall Architecture for Plausible WZ Video Coding

The decoder architecture for perceptual WZ coding combining the above introduced image processing methods is shown in Figure 2.

1. The WZ frame is downsampled (DS) regarding a chess-like pattern (2:1) at the encoder. WZ encoding is performed on the downsampled frame.
2. The decoder generates the SI using the proposed PBTI algorithm. The SI is downsampled.
3. Plane by plane, the downsampled SI is decoded and reconstructed. After each iteration, the partial decoded SI is up sampled (US) and inpainting is performed to fill the unknown pixels. Then, the inpainted SI is used for gSIR taking into account the adjacent key frames.

4. SIMULATION RESULTS

The proposed algorithms have been sequentielly integrated into a basic Wyner-Ziv codec (Section 1). Two test sequences with medium and high amount of motion were analysed (`foreman` and `soccer`, qcif@30Hz, GOP=2). The RD-performance is shown in Figure 3 and Figure 4. A H.264 AVC inta codec is used for key frame coding, as known to be the most efficient intra codec in the literature. Hence, its RD-Performance is also presented as reference.

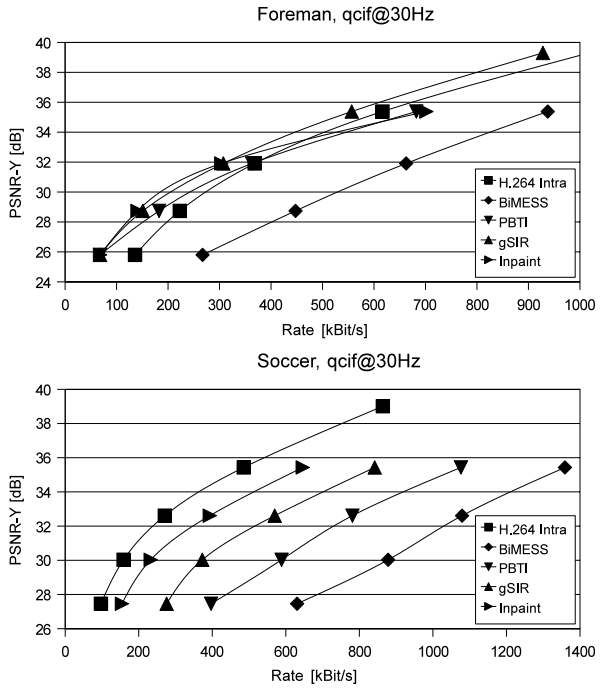

Fig. 3. RD-Performance in PSNR.

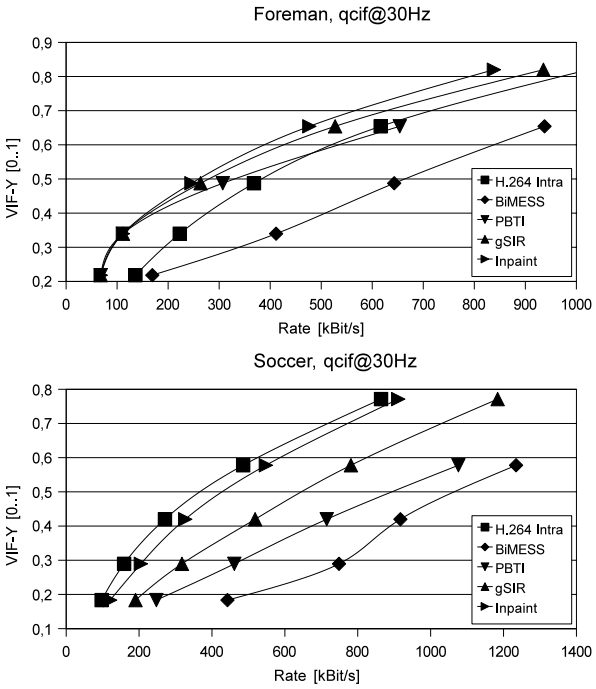

Fig. 4. RD-Performance in VIF.

The proposed PBTI achieves a quality gain of up to 3 dB for `soccer` and up to 2.5 dB for `foreman` compared to BiMESS. Additionaly, the gSIR improves the RD-performance for the high motion sequence `soccer` by more than 2 dB. Furthermore, inpaiting improves the PSNR for the `soccer` sequence by additionally 2 dB. Whereas, the SW-encoding complexity is reduced by a factor of two, because only a subset of pixels is encoded. The overall performance is comparable to state-of-the-art H.264 AVC intra coding.

As shown in Figure 4, the proposed framework also achieves coding gain in term of VIF, proofed to be more correlated to the HSV. Furthermore, inpainting attains the overall best performance, also for the `foreman` sequence. Consequently, inpainting is a suitable tool for spatial correlation exploitation in the proper sense of distributed video coding.

5. CONCLUSION AND FURTHER WORK

Perceptual aware metrics are analysed to design specifically improved image processing tools for plausible WZ coding. The proposed tools increase the RD-performance in terms of PSNR by 4-9 dB as well as in terms of the more meaningfull VIF metric and enables the pixel domain WZ codec to reach the performance of an advanced intra codec. Furthermore, the encoder complexity for pixel domain WZ coding is reduced by downsampling and inpainting. Ongoing research is focused on improving inpainting strategies for WZ-coding as counterpart of transform domain coding to further increase the RD-performance considering low encoder complexity.

6. REFERENCES

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

[2] D. Kubasov, J. Nayak, and C. Guillemot, "Optimal reconstruction in wyner-ziv video coding with multiple side information," in *Proc. MMSP*, 1–3 Oct. 2007, pp. 183–186.

[3] M. Miyahara, K. Kotani, and V.R. Algazi, "Objective picture quality scale (pqs) for image coding," *IEEE Trans. on Communications*, vol. 46, no. 9, pp. 1215–1226, Sept. 1998.

[4] Z. Wang, L. L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, pp. 121–132, 2004.

[5] Hamid R. Sheikh and Alan C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, pp. 430–444, 2006.

[6] H. Sheikh, L. Wang, Z. Cormack, and A. Bovik, *LIVE Image Quality Assessment Database Release 2*, http://live.ece.utexas.edu/research/quality.

[7] A.B.B. Adikari, W.A.C. Fernando, H.K. Arachchi, and W.A.R.J. Weerakkody, "Sequential motion estimation using luminance and chrominance information for distributed video coding of wyner-ziv frames," *Electronics Letters*, vol. 42, no. 7, pp. 398–399, 30 March 2006.

[8] F. Pereira, J. Ascenso, and C. Brites, "Studying the gop size impact on the performance of a feedback channel-based wyner-ziv video codec," in *Proc. of PSIVT*, December, 2007, pp. 801–815.

[9] M. Bertalmo, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," *Proc. SIGGRAPH*, pp. 417–424, 2000.