Incorporating Feature Point-based Motion Hypotheses in Distributed Video Coding

Ralph Hänsel, Henryk Richter, Erika Müller
Institute of Communications Engineering
University of Rostock
Rostock, Germany
{*ralph.haensel, henryk.richter, erika.mueller*}*@uni-rostock.de*

***Abstract*—Emerging mobile video applications introduce demands (e.g. low encoding complexity) which do not fit well to conventional video coding (e.g. H.264/AVC, MPEG-4). Upcoming coding schemes, following the distributed video coding (DVC) paradigm, are well suited for mobile video encoding due to their low encoder complexity.**

The performance of a DVC scheme is mainly influenced by the Slepian-Wolf coding performance and the side information quality. The side information is commonly obtained by temporal interpolation, where fast inhomogeneous motion is still challenging. In this paper, fast motion is negotiated by applying a multiple motion hypotheses pixel-based temporal interpolation method. Global and local motion estimation are incorporated. Simulation results show side information quality improvements by up to $0.6\,\mathrm{dB}$ for the sequence Stefan. Furthermore, state-of-the-art side information generation approaches are outperformed by $1.2\,\mathrm{dB}$ in terms of overall RD performance.

I. Introduction

Distributed Video Coding (DVC) is an interesting topic, which raises a strong research interest since the beginning of the last decade. Despite extensive research activities a lot of challenges regarding to DVC are still unsolved (e.g. feedback channel, source modeling, fast motion).

Conventional video coding schemes (e.g. H.264/AVC) are very well suited for broadcasting scenarios due to its high encoding and low decoding complexity. On the other hand, DVC is well suited for emerging application scenarios requiring low encoding complexity (e.g. mobile video encoding). A low complexity video encoder is achieved by shifting the complex motion estimation part to the decoder. Furthermore, DVC is also capable to assure robust video transmission and efficient multiview video coding. An extensive overview of further application scenarios for DVC is given in [7], [10].

Distributed Video Coding (DVC) is based on the theories of Slepian and Wolf [12] as well as Wyner and Ziv [15]. They prove that the rate distortion performance of a coding scheme does not depend on the availability of side information (SI) at the encoder. The SI only needs to be available at the decoder. In terms of conventional video coding the SI is the motion compensated reference frame. In DVC, side information is commonly obtained by temporal interpolation of adjacent key frames at the decoder.

One of the unsolved challenges in DVC is high rate distortion (RD) performance equivalent to conventional inter frame video coding (e.g. H.264/AVC). The RD performance on the one hand depends on the Slepian-Wolf (SW) coding performance. On the other hand, a high side information Y quality is essential for high reconstruction quality as well as low data rate. The side information is commonly estimated by temporal interpolation (sec. III). Fast motion is very challenging due to low temporal correlation between adjacent frames. This problem cannot be solved by expanding the search range (max motion vector length $|\mathbf{mv}|$). Block based motion estimation (ME) tends to mismatching in case of a large search window. Thus, the necessary true motion is not obtained.

In this paper, a multiple motion hypotheses temporal interpolation method is proposed (sec. IV). It incorporates multiple initial values for motion estimation and a rather small search range. Therefore, the probability of successful matching and obtaining the true motion is increased. Three types of hypotheses are applied. The global motion hypothesis reflects the camera motion, whereas the local motion hypothesis corresponds to the motion of a segment. Finally, a zero motion hypothesis is also taken as initial value. Simulation results show an increased side information (SI) quality as well as an improved overall RD performance of $1.6\,\mathrm{dB}$ and $1.2\,\mathrm{dB}$, respectively (sec. V). Conclusions and some remarks on further work are given in section VI.

II. Pixel domain DVC scheme

A pixel domain DVC (distributed video coding) scheme (fig. 1) is proposed in this paper. It is chosen due to its lower encoding complexity compared to DCT (discrete cosine transform)-based DVC coding schemes. Furthermore, pixel domain DVC is essential for a flexible decoding process.

In the first step, the input video sequences is split up into key- and Wyner-Ziv (WZ) frames. On the one hand, the key frames K are en- and decoded by an H.264 intra en-/decoder. The key frame quality is adjusted by the quantization parameter QP. On the other hand, each pixel of a WZ frame X_i is quantized in the Wyner-Ziv encoder. Therefore, a 2^M-step linear quantization is applied. Subsequently, each bit plane $q_i^{(b)}$ of a quantization symbol q_i is Slepian-Wolf (SW) encoded. A binary turbo encoder [11]

978-963-8111-76-0

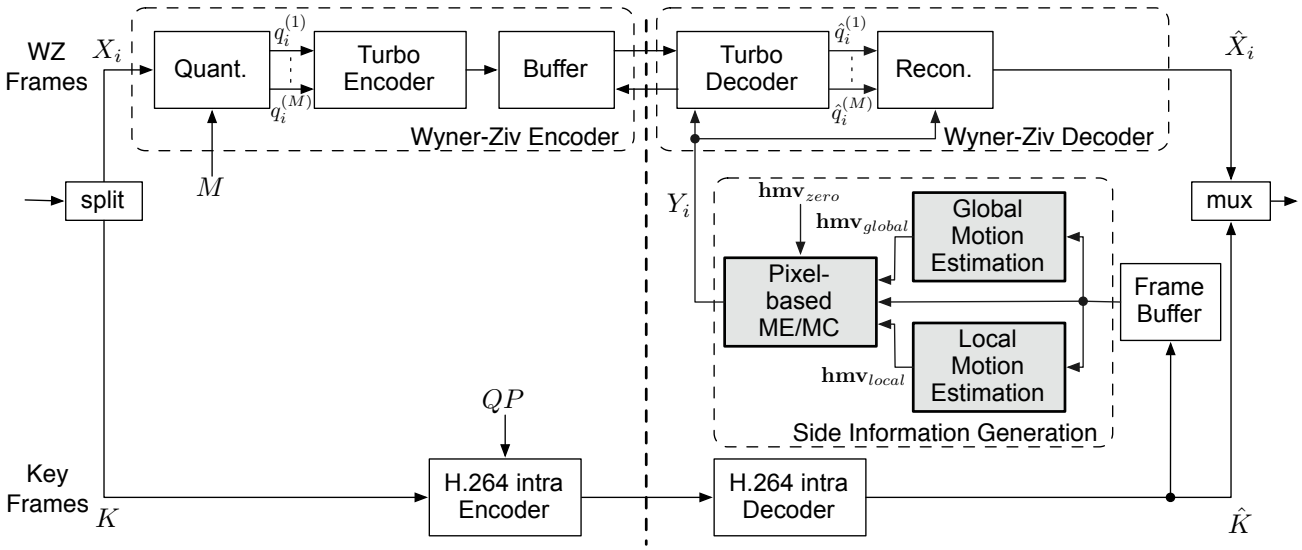


Figure 1. Proposed pixel domain DVC scheme and *Multiple Motion Hypothesis Pixel-based Temporal Interpolation* (MPBTI)

is applied to implement SW coding. The generated parity symbols are stored in a buffer and send to the decoder on request. At the receiver, the initial side information Y is generated by temporal interpolation (sec. III). Finally, the side information Y is corrected in the SW decoding and subsequent reconstruction process to obtain the decoded Wyner-Ziv frame.

III. Side information estimation in DVC

A. Temporal interpolation

Side information generation in low complexity DVC, commonly exploits the temporal correlation. Fast motion is challenging for the block-based temporal interpolation (TI) applied in *Bidirectional motion estimation with spatial smoothing* (BiMESS [2]). This approach was extended by a hierarchical motion estimation applying different block sizes to cope with fast motion (*Motion compensated temporal interpolation* – MCTI [1], [3]). Finally, block artifacts are a drawback of the mentioned algorithms.

The later issue is solved by *Pixel-Based Temporal Interpolation* (PBTI, [13]). The fast motion challenge was negotiated in [6] by estimating a global motion model at the encoder, which increases the encoding complexity. Finally, mesh-based TI [8] performs well in deformation areas.

B. Temporal extrapolation

Fast motion is also very challenging for temporal extrapolation (TX) approaches [4], [5]. Temporal extrapolation do not require small key frame distances, due to incorporating decoded WZ frames. Motion estimation is slightly weaker compared to TI. Finally, the TX approach shows good results in occluded areas.

C. Spatial interpolation

The spatial correlation is exploited in [14] for side information generation. At first, a subset of pixel is decoded by incorporating temporal side information. Subsequently, a joint spatial and temporal side information method and SW decoding are applied to reconstruct the second subset of pixel. The spatial side information is estimated based on spatial interpolation. Thus, fast motion is less challenging, because spatial interpolation does not depend on high temporal correlation (slow linear motion).

IV. Multiple Motion Hypotheses Pixel-based Temporal Interpolation

The aim of the proposed side information generation method (fig. 1) is to cope with fast motion and thus the mismatching problem. Therefore, the proposed *Multiple Motion Hypotheses Pixel-based Temporal Interpolation* (MPBTI) method incorporates multiple robust motion hypotheses and a rather small search range. These hypotheses include the global motion $\mathbf{hmv}_{global}$, the local motion $\mathbf{hmv}_{local}$ and no motion $\mathbf{hmv}_{zero}$. They are incorporated by the pixel-based motion estimation/compensation step, which reduces the mismatching probability and block artifacts significantly. Finally, the encoder should stay at very low complexity, by applying the proposed algorithms only at the decoder side.

The proposed *Multiple Motion Hypotheses Pixel-based Temporal Interpolation* (MPBTI) approach incorporates both adjacent key frames K_{i-1}, K_{i+1} and follows these steps:

1) Global motion estimation $\rightarrow \mathbf{hmv}_{global}$
2) Local motion estimation $\rightarrow \mathbf{hmv}_{local}$
3) Forward pixel-based motion estimation
4) Bidirectional pixel-based motion estimation
5) Motion compensation

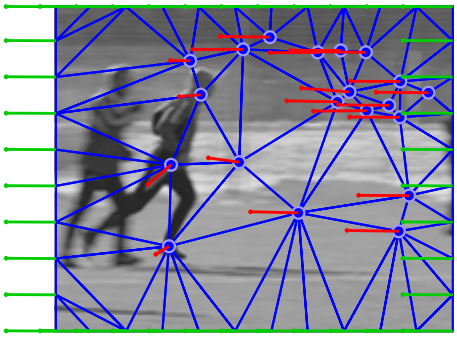
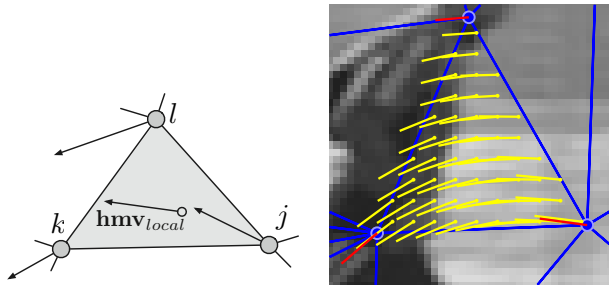


Figure 2. Motion of feature points



(a) adjacent feature points $tri_{j,k,l}$ (b) $\mathbf{hmv}_{local}$ for each pixel

Figure 3. Local motion vector interpolation based on triangle mesh

A. Global motion estimation

The global motion hypothesis $\mathbf{hmv}_{global}$ is obtained by global motion estimation. Therefore, feature point matching is applied to obtain robust motion vectors for characteristic areas (fig. 2). The SIFT [9] algorithm is used for feature point fp_i detection and matching. Finally, a vector median filter (eq. 1) is applied on the set of feature point motion vectors $\{\mathbf{mv}_i\}$, obtaining a robust global motion estimation by eliminating outliers.

$$\sum_j ||\mathbf{hmv}_{global} - \mathbf{mv}_j||_2 \quad \leq \quad \sum_j ||\mathbf{mv}_i - \mathbf{mv}_j||_2 \quad (1)$$

B. Local motion estimation

The local motion hypothesis $\mathbf{hmv}_{local}$ represents a more detailed hypothesis compared to the global motion estimation. Local motion is estimated based on the feature points fp_i and its motion vectors $\mathbf{mv}_i$ estimated in the previous step. Delaunay triangulation is performed to divide the frame into triangles $tri_{j,k,l}$ (fig. 2), where additionally feature points are placed at the frame border having global motion. The local motion hypothesis $\mathbf{hmv}_{local}$ for each pixel is calculated by linear interpolation of the motion vectors of the adjacent feature points (fp_j, fp_k, fp_l, fig. 3(a)). Motion vectors are interpolated by disjoint linear interpolation of its x- and y-components. Finally, a dense initial motion vector field is obtained (fig. 3(b)).

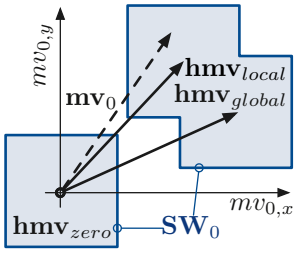


Figure 4. Final search window $\mathbf{SW}_0$ construction

C. Pixel-based motion estimation

Pixel-based motion estimation (PBME) in conjunction with DVC was proposed in [13] to reduce block artifacts. In this paper, the method is extended by having a more adaptive non-rectangular search window $\mathbf{SW}_0$ (fig. 4).

The search window $\mathbf{SW}_0$ is obtained by the union of the hypotheses search windows (eq. 2, fig. 4). The zero motion search window $\mathbf{SW}_{zero}$ for example is given in equation 3, where a small search range of $SR = 6$ is applied. The global motion and local motion search windows $\mathbf{SW}_{global}$ and $\mathbf{SW}_{local}$, respectively, are in addition shifted by the corresponding motion hypothesis $\mathbf{hmv}_{global}$ or $\mathbf{hmv}_{local}$.

$$\mathbf{SW}_0 = \mathbf{SW}_{zero} \cup \mathbf{SW}_{local} \cup \mathbf{SW}_{global} \quad (2)$$

$$\mathbf{SW}_{zero} = \{\mathbf{mv} | -SR \leq mv_x \leq SR \wedge -SR \leq mv_y \leq SR\} \quad (3)$$

An equivalent rectangular search window covers a larger area. Thus, the complexity and probability of mismatching is reduced by applying a non-rectangular search window.

Subsequently, a dense motion vector field is estimated by PBME. Therefore, a weighted block matching approach is applied to obtain the motion vector $\mathbf{mv}_0 = (mv_{0,x}, mv_{0,y})$ (eq. 4, $\mathbf{mv}_0 \in \mathbf{SW}_0$). The neighborhood of each pixel (matching window, MW) in the previous key frame K_{i-1} is matched to the subsequent key frame K_{i+1}. The neighborhood is weighted by a Gaussian window (GW, $\sigma_{GW} = 5$) and the motion estimation is performed at full pixel accuracy.

$$\mathbf{mv}_0 = \arg\min_{\mathbf{mv}_0} \sum_{(x,y)\in MW} GW(x,y) \times |K_{i-1}(x,y) - K_{i+1}(x + mv_{0,x}, y + mv_{0,y})| \quad (4)$$

D. Bidirectional pixel-based motion estimation

Subsequently, the bidirectional pixel-based motion estimation (BiPBME, eq. 5) is applied for motion field refinement. Therefore, a search window $\mathbf{SW}_1$ (eq. 6) aligned with the motion vector $\mathbf{mv}_0$ is employed. Furthermore, a small search range $SR = 1$ is used to reduced the mismatching probability and computational complexity. The bidirectional motion estimation is performed at half pixel accuracy.

(a) BiMESS [2] 19.8 dB (b) PBTI [13] 20.1 dB (c) proposed **MPBTI 23.8 dB** (d) Original

Figure 5. Soccer side information example, WZ-frame 61, QCIF, 30 fps

$$\mathbf{mv}_1 = \arg\min_{\mathbf{mv}_1} \sum_{(x,y)\in MW} GW(x,y) \times |K_{i-1}(x - mv_{1,x}, y - mv_{1,y}) - K_{i+1}(x + mv_{1,x}, y + mv_{1,y})| \quad (5)$$

where $\mathbf{mv}_1 \in \mathbf{SW}_1$.

$$\mathbf{SW}_1 = \{\mathbf{mv} | -SR + mv_{0,x}/2 \leq mv_x \leq SR + mv_{0,x}/2 \wedge -SR + mv_{0,y}/2 \leq mv_y \leq SR + mv_{0,y}/2\} \quad (6)$$

E. Pixel-based motion compensation

Finally, pixel-based motion compensation (PBMC) at half pixel accuracy is applied, incorporating the neighboring key frames. The final side information Y_i is obtained by averaging both motion compensated adjacent key frames K_{i-1}, K_{i+1}.

V. Simulation Results

A comprehensive set of video sequences (Coastguard, Foreman, Soccer, Stefan) is used to evaluate the performance of the proposed coding scheme. The well known BiMESS, PBTI and the proposed MPBTI methods are compared in terms of side information quality, RD-performance and computational complexity. Furthermore, QCIF resolution sequences with 15 fps and 30 fps are used. The rate and distortion values only rely on the luminance component.

A. Side information quality

The visual side information quality for the sequence Soccer at 30 fps is shown in figure 5. The result of the BiMESS [2] side information generation suffer from block artifacts and motion estimation mismatching (fig. 5(a)). The blocking artifacts are significantly reduced by applying the PBTI [13] algorithm (fig. 5(b)). Finally, the visual and objective quality is significantly improved applying the proposed MPBTI approach (fig. 5(c)) due to reduced mismatching probability. Especially the interpolation quality of fast moving objects is enhanced (background and right border). However, the highly inhomogeneous motion at the left frame border is still challenging. Fast motion at the frame borders leads to an inherent problem of temporal interpolation. If an object is only visible in one frame, temporal interpolation will not work, because the visibility of the object in both adjacent frames is mandatory.

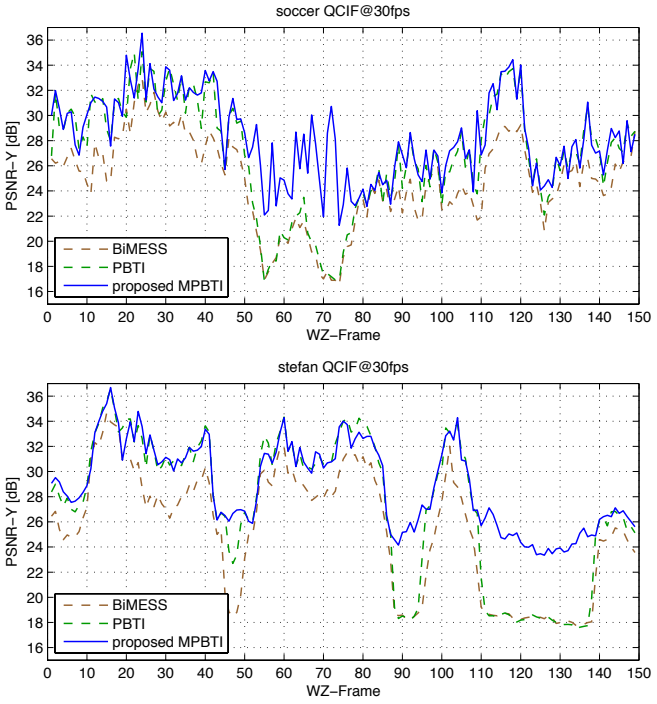


Figure 6. Side information quality, Soccer, Stefan, QCIF, 30 fps

Figure 6 shows the side information PSNR (peak signal to noise ratio) for each WZ frame of the sequences Soccer and Stefan at 30 fps. The proposed MPBTI outperforms BiMESS and PBTI especially for frames with very fast motion (Soccer 50 . . . 80, Stefan 110 . . . 140). However, the multiple motion hypotheses approach might be misleading for a few low motion frames, where PBTI outperforms MPBTI by a very small margin (e.g. Stefan frame 80).

Regarding to the average side information PSNR (tab. I), MPBTI outperforms state-of-the-art side information generation methods (PBTI, MCTI) by up to 0.6 dB. Very low motion sequences at high frame rates (e.g. Coastguard, Foreman) are slightly challenging, because the hypotheses are sometimes misleading.

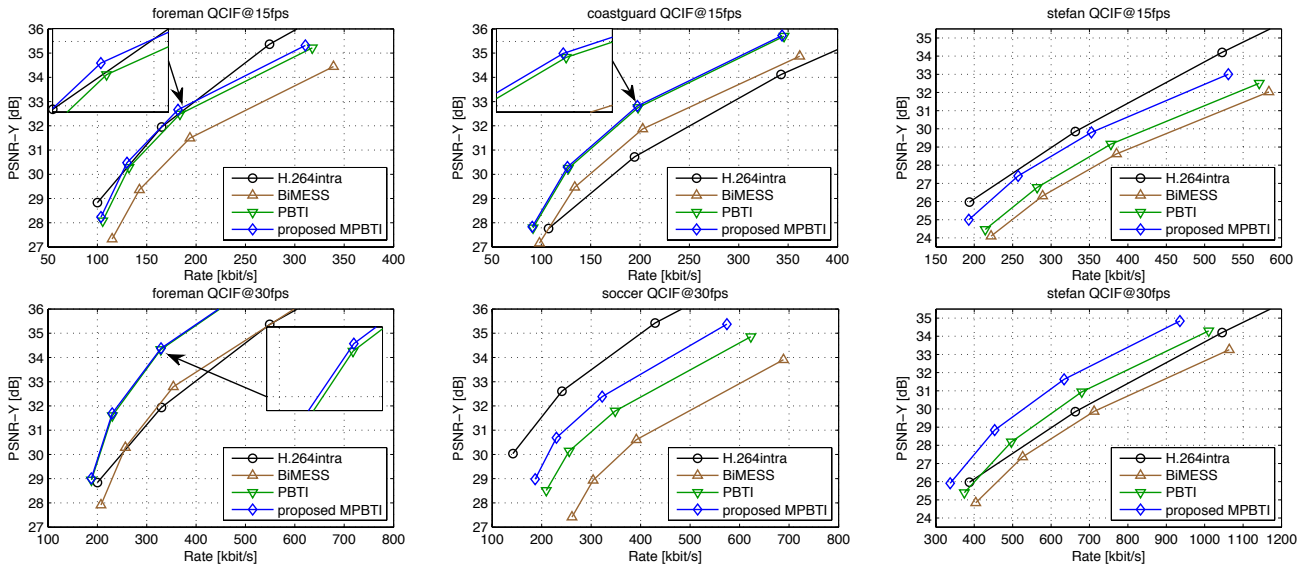


Figure 7. RD performance, quantization parameter set: $(QP, M) = \{(40, 1), (35, 1), (30, 1), (25, 2)\}$

Table I
AVERAGE SIDE INFORMATION QUALITY, PSNR [DB]

QCIF 15 fps	BiMESS [2]	PBTI [13]	MCTI [3]	**MPBTI** proposed
Coastguard	30.3	32.4	31.6	**32.8**
Foreman	28.0	29.8	29.1	**30.2**
Soccer	21.0	21.4	22.3	**22.9**
Stefan	21.7	22.7	–	**24.2**

QCIF 30 fps	BiMESS [2]	PBTI [13]	MI-2 [4]	**MPBTI** proposed
Coastguard	36.3	**39.1**	37.5	**38.9**
Foreman	33.6	**36.8**	34.9	**36.3**
Soccer	24.8	26.9	–	**28.3**
Stefan	25.3	27.3	–	**28.9**

values for MCTI and MI-2 are take for the corresponding literature

B. Overall RD performance

The overall RD (rate-distortion) performance of the DVC coding scheme (fig. 1) incorporating the proposed MPBTI is shown in figure 7. The overall RD performance considers the H.264 intra encoded key frames as well as the WZ frames reconstruction quality and data rate. A fixed set of key frame quantization parameter (QP) and WZ frame quantization parameter (M) was used.

The proposed MPBTI method outperforms all mentioned side information generation algorithm in terms of RD performance. For sequences with low motion (e.g. Coastguard, Foreman) MPBTI shows a slightly increased performance compared to PBTI. In most cases it shows better reconstruction quality compared to the conventional H.264 intra coding scheme. The proposed MPBTI outperforms PBTI for fast motion sequences (e.g. Soccer, Stefan) by up $1.2\,\mathrm{dB}$. Furthermore, the proposed DVC coding scheme shows an RD performance close to H.264 intra for Stefan at $15\,\mathrm{fps}$, which is remarkable for a DVC coding scheme. The sequence Soccer is still challenging for DVC, because of its very fast and inhomogeneous motion.

C. Computational complexity

The encoding complexity is very critical in a DVC coding application. Therefore, the proposed MPBTI algorithm only affects the decoding complexity. The encoding complexity is retained at its low value.

The DVC decoding complexity is dominated by the side information generation and SW decoding complexity. PBTI and MPBTI are more complex than the block based BiMESS method. The forward motion estimation dominates the complexity of PBTI and MPBTI. Applying a search range of $SR = 10$ in PBTI $21 \times 21 = 441$ matching window MW comparisons per pixel are required. The search range is reduced to $SR = 6$ in MPBTI, due to incorporating multiple motion hypotheses. Therefore, a maximum number of $13 \times 13 \times 3 = 507$ matching window comparisons per pixel are required. Thus, the decoding complexity is only slightly increased by MPBTI in the worst case (non overlapping search windows) compared to PBTI.

VI. CONCLUSIONS AND FURTHER WORK

A *Multiple Motion Hypotheses Pixel-based Temporal Interpolation* (MPBTI) approach is proposed in this paper. It negotiate the challenge of fast inhomogeneous motion for side information generation in distributed video coding (DVC). Therefore, local, global and zero motion hypotheses are incorporated based on SIFT feature point extraction and

matching. The proposed MPBTI outperforms state-of-the-art side information generation in terms of side information quality by up to $0.6\,\mathrm{dB}$ and in terms of RD performance by up to $1.2\,\mathrm{dB}$. Very inhomogeneous motion and occlusions are still challenging.

Further work should include the complexity reduction of MPBTI. Furthermore, the temporal interpolation weakness at the frame border and in occluded areas should be solved by motion field extrapolation.

REFERENCES

[1] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret. The discover codec: Architecture, techniques and evaluation. In *Proc. Picture Coding Symposium (PCS)*, 2007.

[2] J. Ascenso, C. Brites, and F. Pereira. Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding. In *Proc. Speech and Image Processing, Multimedia Communications and Services, Conference on (EC-SIP-M, EURASIP)*, 2005.

[3] J. Ascenso and F. Pereira. Hierarchical motion estimation for side information creation in wyner-ziv video coding. In *Proc. of the 2nd international conference on Ubiquitous information management and communication (ICUIMC)*, 2008.

[4] S. Borchert, R. Westerlaken, R. Gunnewiek, and R. Lagendijk. On extrapolating side information in distributed video coding. In *Proc. Picture Coding Symposium (PCS)*, 2007.

[5] S. Borchert, R. Westerlaken, R. Gunnewiek, and R. Lagendijk. Motion compensated prediction in transform domain distributed video coding. In *Multimedia Signal Processing (MMSP), IEEE Workshop on*, 2008.

[6] F. Dufaux and T. Ebrahimi. Encoder and decoder side global and local motion estimation for distributed video coding. In *Proc. Multimedia and Signal Processing (MMSP), IEEE Workshop. on*, 2010.

[7] F. Dufaux, W. Gao, S. Tubaro, and A. Vetro. Distributed Video Coding: Trends and Perspectives. *Image and Video Processing, EURASIP Journal on*, 2009:13, 2009.

[8] D. Kubasov and C. Guillemot. Mesh-based motion-compensated interpolation for side information extraction in distributed video coding. In *Proc. Image Processing (ICIP), IEEE Int. Conference on*, 2006.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[10] F. Pereira, L. Torres, C. Guillemot, T. Ebrahimi, R. Leonardi, and S. Klomp. Distributed video coding: Selecting the most promising application scenarios. *Signal Processing: Image Communication*, 23:339–352, 2008.

[11] D. Rowitch and L. Milstein. On the performance of hybrid fec/arq systems using rate compatible punctured turbo (rcpt) codes. *Communications, IEEE Transactions on*, 48(6):948–959, June 2000.

[12] D. Slepian and J. Wolf. Noiseless Coding of Correlated Information Sources. *Information Theory, IEEE Transactions on*, 19(4):471–480, July 1973.

[13] S. Sofke, R. Hänsel, and E. Müller. Human Visual System aware Decoding Strategies for Distributed Video Coding. In *Proc. Picture Coding Symposium (PCS)*, Chicago, May 2009.

[14] M. Tagliasacchi, A. Trapanese, S. Tubaro, J. Ascenso, C. Brites, and F. Pereira. Exploiting spatial redundancy in pixel domain wyner-ziv video coding. In *Image Processing (ICIP), IEEE Int. Conference on*, 2006.

[15] A. D. Wyner and J. Ziv. The Rate-Distortion Function for Source Coding with Side Information at the Decoder. *Information Theory, IEEE Transactions on*, 22(1):1–10, Jan 1976.