JOINT DISPARITY AND MOTION ESTIMATION USING OPTICAL FLOW FOR MULTIVIEW DISTRIBUTED VIDEO CODING

Matteo Salmistraro*, Lars Lau Rakêt[†], Catarina Brites[‡], João Ascenso[‡], Søren Forchhammer*

*DTU Fotonik, Technical University of Denmark, {matsl, sofo}@fotonik.dtu.dk
[†]DIKU, University of Copenhagen, Denmark, larslau@diku.dk
[‡]Instituto Superior Técnico, Portugal, {catarina.brites, joao.ascenso}@lx.it.pt

ABSTRACT

Distributed Video Coding (DVC) is a video coding paradigm where the source statistics are exploited at the decoder based on the availability of Side Information (SI). In a monoview video codec, the SI is generated by exploiting the temporal redundancy of the video, through motion estimation and compensation techniques. In a multiview scenario, the correlation between views can also be exploited to further enhance the overall Rate-Distortion (RD) performance. Thus, to generate SI in a multiview distributed coding scenario, a joint disparity and motion estimation technique is proposed, based on optical flow. The proposed SI generation algorithm allows for RD improvements up to 10% (Bjøntegaard) in bit-rate savings, when compared with block-based SI generation algorithms leveraging temporal and inter-view redundancies.

Index Terms— Distributed Video Coding, Multiview Video, Disparity Estimation, Motion Estimation, Optical Flow.

1. INTRODUCTION

In recent years, the Distributed Video Coding (DVC) [1, 2] paradigm has been considered as a promising approach for multiview scenarios [3]. DVC empowers an emerging set of applications, such as visual sensor networks, where each sensing node has limited computational resources, thus requiring low-complexity encoding but also efficient video compression. DVC is based on two information theoretic results from the 1970s, the Slepian-Wolf [4] and Wyner-Ziv (WZ) [5] theorems. In particular, the WZ theorem considers the setup where a source is independently lossy encoded but jointly decoded with a correlated signal, commonly referred to as Side Information (SI). Compared to predictive video coding, DVC exploits the source redundancy partially or totally at the decoder. This enables to leverage inter-camera redundancy without inter-camera communication. In Multiview DVC (M-DVC), the SI creation and fusion techniques play a critical role in the overall compression performance. Inter-view SI is generated by exploiting the inter-

view correlation between cameras, and the intra-view SI is generated exploiting the temporal correlation. Once the two estimations are generated, fusion techniques are typically applied to obtain the final SI, i.e. the two estimations are combined according to their reliability [6, 7]. The better the quality of the fused SI frame, the smaller the number of 'errors' the DVC decoder has to correct and, thus, less redundancy bits are transmitted. However, an alternative SI creation approach has been proposed in [8], the MultiView Motion Estimation (MVME) technique. First, MVME estimates the disparity between temporally aligned frames in the central and lateral (left or right) views and then, motion is estimated for each matched block in the lateral view. The motion vectors obtained for the lateral view are then applied to the central WZ frame to generate the SI. To estimate the motion and disparity of each block, MVME uses a block matching algorithm. However, Optical Flow (OF) for motion estimation can lead to higher SI quality when compared with classical block-based SI generation methods [9], such as Overlapped Block Motion Compensation (OBMC) which is an efficient intra-view SI generation algorithm [10], relying on the use of weighted average of multiple candidate blocks. Motion estimation based on OF produces a dense motion field, where the displacement of each pixel is influenced by the displacement of all other pixels through total variation regularization, allowing for higher flexibility in the motion estimation compared with e.g. OBMC [10]. Thus, the optical flow framework is exploited into a novel SI creation solution, called Time-Disparity OF (TDOF) with the following contributions: 1) the use of OF for estimating the motion of the current view given the lateral views in a DVC setup; and 2) the handling of occlusion through filtering and joint interpolation of scattered sets. To allow for better inter-view matching quality, a pre-alignment step is introduced, to handle areas lying outside the field of view of one camera but available in the other view. TDOF shares with MVME the general concept of using the motion of lateral views to generate SI. Finally, the robustness of the proposed TDOF method is analysed using an on-line correlation noise modelling, as opposed to many M-DVC works still relying on off-line modelling [6, 7].

The rest of the paper is organized as follows: in Section 2 the adopted DVC architecture is presented, in Section 3 the proposed SI generation method is described and Section 4 assesses its performance.

2. MULTIVIEW DVC CODING ARCHITECTURE

The proposed M-DVC codec is based on the monoview DVC codec presented in [10] and is depicted in Fig. 1. The three-camera multiview setup depicted in Fig. 2 is considered here where central view frames can be WZ or Intra coded according to a fixed GOP structure. The left and right views, $\widetilde{I_r}$ and $\widetilde{I_l}$, are independently encoded, and although they can be coded with any available video coding solution, the H.264/AVC Intra coding scheme has been adopted here, as it is typically done in literature [8]. The encoder of the central view divides the frames into Key Frames (KFs) and WZ frames, X. The KFs are coded independently, using H.264/AVC Intra coding and the WZ frames are DCT transformed, quantized and organized in bitplanes. Each bitplane is fed into a Low-Density Parity Check Accumulate (LDPCA) encoder [11] that generates the syndromes, which are stored in a buffer and sent to the decoder upon request. The M-DVC decoder uses already decoded frames from central and lateral views to generate the SI frame Y and a residual frame R, which corresponds to the estimation of $X - Y$. The soft probabilities of each bitplane are then calculated with a Laplacian correlation noise model, derived from R: all the residuals used in this work are estimated without using X. A feedback channel allows the decoder to request new syndromes (as in the Stanford DVC codec [1]) if the received syndromes are not enough to successfully decode the source (bitplane). To improve the reliability of the decoded bitplane, an additional 8-bit CRC is used to check for any remaining decoding errors. Once all the bitplanes of a given DCT band are decoded the corresponding coefficients are reconstructed [12].
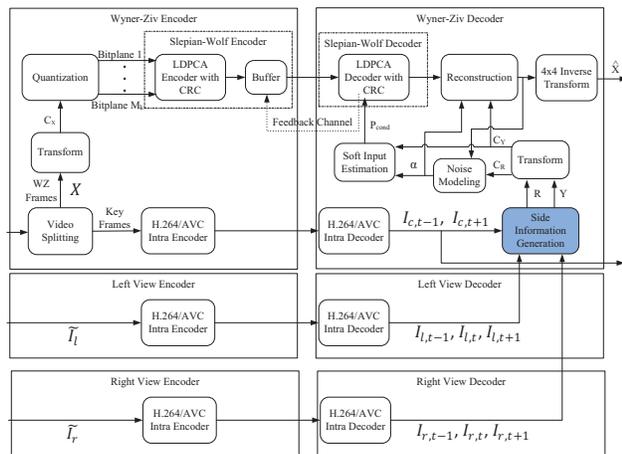

Fig. 1. Proposed M-DVC coding architecture.

3. TDOF SIDE INFORMATION GENERATION

Consider X as the WZ frame to be coded. The TDOF (Time-Disparity OF) approach makes use of three frames in the right view, three frames in the left view and $I_{c,t-1}$ and $I_{c,t+1}$ in the central view. The disparity field between $I_{c,t-1}$ and $I_{l,t-1}$ is first calculated. Then, for each point (which may be a non-integer position) hit by a disparity vector in $I_{l,t-1}$, the motion vector between this point and its corresponding point in $I_{l,t}$ is calculated, as depicted in Fig. 2. The motion vector is then applied to the corresponding pixel, $\mathbf{x}$, in $I_{c,t-1}$, obtaining a scattered set of points $S_{l,t-1}$ for the SI frame Y. The set of frames $I_{c,t-1}$, $I_{l,t-1}$ and $I_{l,t}$ constitute a "path". Applying this procedure to the other three paths, three new sets of scattered points can be obtained: $S_{l,t+1}$ (path $I_{c,t+1}$, $I_{l,t+1}$ and $I_{l,t}$), $S_{r,t-1}$ (path $I_{c,t-1}$, $I_{r,t-1}$ and $I_{r,t}$) and $S_{r,t+1}$ (path $I_{c,t+1}$, $I_{r,t+1}$ and $I_{r,t}$). The described solution differs from [8] because it is proposed here to calculate the disparity and motion with an OF technique followed by filtering and joint interpolation, i.e. the fusion of the scattered sets. The TDOF SI generation algorithm can be divided into three steps, corresponding to the three introduced novelties: 1) pre-alignment of the time aligned frames to remove unmatched areas, 2) OF calculation, 3) scattered sets filtering and joint interpolation, to obtain the final SI.
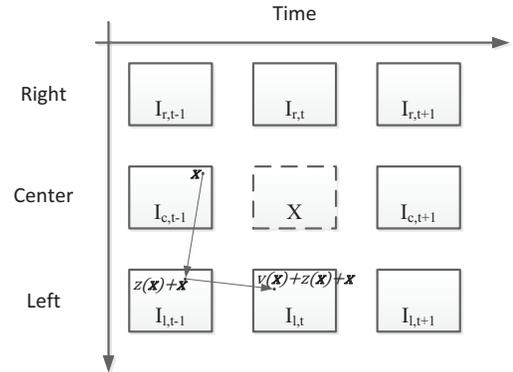

Fig. 2. Three-camera setup, depicting the path for the set $S_{l,t-1}$.

3.1 Pre-alignment

To allow for higher matching quality and higher robustness during the disparity estimation, the two temporally co-located frames, (e.g. $I_{l,t-1}$ and $I_{c,t-1}$ for the set $S_{l,t-1}$), of each path are pre-aligned. The pre-alignment phase removes unmatched lateral (two) bands, one in the lateral frame and one in the central frame to assure that no significant occlusions can occur. This allows for a higher quality match, because wrong estimations in those bands would influence the quality of the whole SI frame, given the particular formulation of the OF problem. Consider the path leading to the calculation of $S_{l,t-1}$ and the frames $I_{l,t-1}$ and $I_{c,t-1}$ with dimension $m \times n$ pixels (with m being the number of col-

umns). The average global disparity d_G between these two frames is calculated by minimizing:

$$d_G = \underset{q}{\text{argmin}} \sum_{i=0+q\chi(q)}^{m-1+q(\chi(-q))} \sum_{j=0}^{n-1} \frac{|I_{l,t-1}(i,j) - I_{c,t-1}(i-q,j)|}{(m-|q|)n} , \quad (1)$$

where the indicator function χ is defined as: $\chi(q) = 1$ if $q \geq 0$, and $\chi(q) = 0$ otherwise, and q is the search range employed in (1). In this way, the left (right) lateral band of a frame which has no correspondence in the right (left) band of the other frame is removed, generating the corresponding aligned set of frames $I_{l,t-1}^a$ and $I_{c,t-1}^a$. The relative distance between cameras does not change and it is the same between center and left and between center and right, therefore the same disparity may be used for pre-aligning the other left path, leading to generate set $S_{l,t+1}$, while the opposite value can be used for the two remaining sets.

3.2 Optical Flow Calculation

With the aligned frames $I_{l,t-1}^a$ and $I_{c,t-1}^a$, the disparity field z can be estimated by minimizing the data fidelity term:

$$C_D(x,z) = |I_{l,t-1}^a(x + z(x)) - I_{c,t-1}^a(x)|. \quad (2)$$

$z(x)$ is, in general, a vector-valued function, therefore there are two unknowns at every point in (2). To solve this issue, the TV-L^1 formulation [13] has been adopted. TV-L^1 relies on the L^1-norm of the OF constraint (2), and a Total Variation (TV) regularization term: the 1-Jacobian of the field $(J_1 z(x))$ [14] is adopted here. For the TV-L^1 problem, a computationally efficient solution exists to minimize:

$$E_D(z) = \int \lambda_D C_D(x,z) + |J_1 z(x)| dx, \quad (3)$$

where x is a 2D point in $I_{c,t-1}$. OF based disparity estimation produces a dense field, since a disparity vector is calculated for each pixel. The calculation of the disparity vector for point x is influenced by the quality of all the other matches (i.e. magnitude of the constraint) and the smoothness of the disparity field. It is here proposed to directly generate a motion field having source in $I_{c,t-1}^a$: the constraint for the motion estimation is defined taking into account the motion field v and the disparity z.

$$C_T(x,z,v) = |I_{l,t-1}^a(x + z(x)) - I_{l,t}^a(x + z(x) + v(x))|, \quad (4)$$

The minimization of the energy, leading to the estimation of v is performed jointly with the already calculated z:

$$E_T(v,z) = \int \lambda_T C_T(x,z,v) + |J_1 v(x)| dx. \quad (5)$$

With this approach, each point x in $I_{c,t-1}^a$ is directly coupled with its corresponding motion vector $v(x)$, which can be used to project x into a new location, obtaining the scattered set $S_{l,t-1}$, composed by the elements p_x.

$$p_x = [I_{c,t-1}^a(x), x + v(x), C_T(x,z,v)]. \quad (6)$$

3.3 Scattered Sets Filtering and Joint Interpolation

Once the four sets are available, it is possible to perform interpolation and obtain four SI frame estimations: $Y_{l,t-1}$, $Y_{l,t+1}$, $Y_{r,t-1}$ and $Y_{r,t+1}$ which averaged could lead to the final SI Y, mimicking the procedure used for MVME [8]. However, the use of the OF based technique presented in the previous Section allows higher granularity. Thus, the following solution is proposed; first, the scattered sets are fused:

$$S_l = S_{l,t-1} \cup S_{l,t+1} \text{ and } S_r = S_{r,t-1} \cup S_{r,t+1}. \quad (7)$$

This fusion allows the handling of holes in SI, due to occlusions and disocclusions. Moreover, when some points are wrongly matched with other points (that occurs when their true match is occluded), the density of points in some areas increases. Therefore, it is proposed to process each fused set (S_l and S_r) to remove points from too dense areas: for each p_x having $C_T(x,z,v) \geq \Psi$, where Ψ is a threshold, the Φ closest neighbors are selected, including p_x. Among them the neighbor having the highest value of $C_T(x,z,v)$ is removed. Once the two sets have been filtered, they are interpolated to obtain the values for the pixel locations, using linear triangular interpolation. The interpolation is divided in two phases: first, using the scattered points, a piecewise triangular surface is generated. Then, for each point having integer coordinates, a bivariate linear interpolation is applied inside the triangle it belongs to [15]. This leads to the generation of the two joint estimations Y_l and Y_r. The final SI Y and its corresponding residual estimate R are calculated as:

$$Y = \frac{1}{2}(Y_l + Y_r) \text{ and } R = Y_l - Y_r. \quad (8)$$

These calculations are carried out only for points which do not belong to the occluded regions identified in the pre-alignment phase. For the pixels belonging to these regions, Y_l or Y_r is used as the final SI, depending on which one is available. For what concerns the correlation noise (or residual) in those regions, if Y_l is available the residual is calculated as $Y_{l,t-1} - Y_{l,t+1}$, otherwise $Y_{r,t-1} - Y_{r,t+1}$ is used.

4. EXPERIMENTAL RESULTS

In Table 1, a detailed description of the test conditions is presented. To ensure a representative set of scenarios, video sequences [16] with still cameras (*Outdoor* and *Book Arrival*) and moving cameras (*Kendo* and *Balloons*), with different depth structures have been selected. All sequences were downsampled to CIF resolution. For the first two sequences, the distance between two consecutive cameras is

6.5cm, while for the latter two the distance is 5cm. To analyze the robustness of the TDOF method, consecutive cameras for the first two sequences are not used, leading to higher disparity. The central view has been coded using a GOP 2 structure; all experiments are conducted only for the luminance component, as usual in DVC. The RD performance of the proposed solution is assessed using four RD points, obtained using four quantization tables (Q_i) of the DISCOVER project [17] and varying the Quantization Parameter (QP) of the KFs accordingly, as shown in Table 2. The KFs are H.264/AVC Intra coded (Main profile). The QPs are chosen to minimize the PSNR variation in the central view between KFs and WZ frames. The left and the right views are H.264/AVC Intra coded (Main profile), with the same QPs used for KFs.

Sequence	Frame Rate	Coded Frames	Views
Outdoor	15 fps	100	6,8,10
Book Arrival	15 fps	100	6,8,10
Kendo	30 fps	300	3,4,5
Balloons	30 fps	300	3,4,5

Table 1. Test Conditions

For the OF calculation, the energies E_D and E_T are minimized, through an iterative procedure, in a coarse-to-fine pyramid, following the general implementation described in [14, 18]: 70 pyramid levels are used, and linear interpolation is used to upscale the flows from a coarser level to a finer one. Since the OF formulation treats a frame like a continuous function, bicubic interpolation is used. After extensive experiments, it has been determined that $\lambda_T = 115$ and $\lambda_D = 25$ are appropriate for good RD performance, and they are the same for all RD points. The λ_D and λ_T values are different since the disparity field is usually much smoother than the motion field, therefore a high value of λ_D is not required to ensure disparity matching, because usually $|J_1 v(x)| > |J_1 z(x)|$. The proposed TDOF method is compared with alternative SI generation solutions to justify some of the algorithm steps, namely: 1) Y_{avg}: The SI corresponds to the average of the OF-generated estimations $Y_{l,t-1}$, $Y_{l,t+1}$, $Y_{r,t-1}$ and $Y_{r,t+1}$; and 2) Y_l (resp. Y_r): The SI is generated from the left (resp. right) view through OF, scattered set filtering and joint interpolation (see Section 3.3). Table 3 shows the average SI quality for all frames for these three optical flow based solutions.

Sequence	Q_1	Q_4	Q_7	Q_8
Outdoor	38	32	28	23
Book Arrival	39	36	29	25
Kendo	39	36	29	22
Balloons	33	30	24	20

Table 2. QPs for Right and Left Views and KFs

As shown, the scattered set filtering and joint interpolation technique is able to outperform, in SI quality, the simple

average Y_{avg} of the OF-generated estimations and a single view (left or right) joint estimation (Y_l or Y_r).

The RD performance of the proposed M-DVC solution (with the TDOF method) is compared with the RD performance of three M-DVC codecs integrating the following benchmark SI generation solutions: 1) OBMC [10]; 2) DCVP (Disparity Compensated View Prediction) [19] applying OBMC between $I_{r,t}$ and $I_{l,t}$; and 3) MVME [8]. The OBMC and DCVP use the correlation noise (or residual) estimation of [10], MVME and TDOF use the residual estimation in (8).

Sequence	Y_{avg}	Y_l	Y_r	TDOF
Outdoor	34.19	36.02	35.63	36.65
Book Arrival	37.51	37.56	37.51	38.48
Kendo	36.38	37.57	37.79	38.93
Balloons	39.69	40.82	40.82	41.47

Table 3. SI PSNR [dB] for alternative OF-based SI Generation Methods

Sequence	MVME		OBMC		DCVP	
	PSNR [dB]	Rate [%]	PSNR [dB]	Rate [%]	PSNR [dB]	Rate [%]
Outdoor	0.30	-4.13	0.43	-5.90	0.44	-6.14
Book Arrival	0.50	-7.46	0.29	-4.45	2.55	-34.21
Kendo	0.71	-10.40	0.58	-8.63	0.61	-9.03
Balloons	0.59	-8.27	0.21	-2.92	1.48	-19.51

Table 4. BD Gains of the Proposed TDOF Solution Regarding alternative SI Generation Methods

In the filtering of the scattered set, proposed for the TDOF method, $\Phi=3$, and Ψ is chosen such that the number of points having $C_T(x, z, v) \geq \Psi$ is less than or equal to the 1% of the total number of points in the fused scattered sets S_l or S_r. Table 4 shows the Bjøntegaard (BD) [20] PSNR gains and bitrate savings between the TDOF SI generation method and the alternatives MVME, OBMC and DCVP. Rate and PSNR are calculated on all the frames of the central view. The OF-based solution outperforms all the other proposed solutions, for all the sequences; the highest gain when compared with MVME and OBMC is obtained for the *Kendo* sequence (0.71dB and 0.58dB respectively), which has medium complex depth structure and a complex object motion. For what concerns DCVP, the highest gains are obtained for the sequences *Book Arrival* and *Balloons* (2.55dB and 1.49dB respectively) with relatively low motion activity and rather complex depth structure, making the temporal interpolation task much simpler than disparity compensation. In Fig. 3, the RD performance results obtained for *Outdoor* and *Kendo* are presented; again, only the central view KFs and WZ frames rate and PSNR is considered. As shown in Fig. 3, the SI PSNR gains are reflected in

M-DVC codec RD performance improvements. A similar trend is also followed by the other two sequences; RD results are not shown here for these two sequences due to paper length constrains.

5. CONCLUSION

In this paper, a new OF-based method for joint disparity and motion estimation for SI generation in M-DVC, called TDOF, is proposed; the TDOF method includes techniques to filter erroneous interpolations and to jointly interpolate sets of scattered points. The TDOF SI generation method leads to bitrate savings up to 10%, 8.6% and 34% when compared with MVME, OBMC and DCVP, respectively.
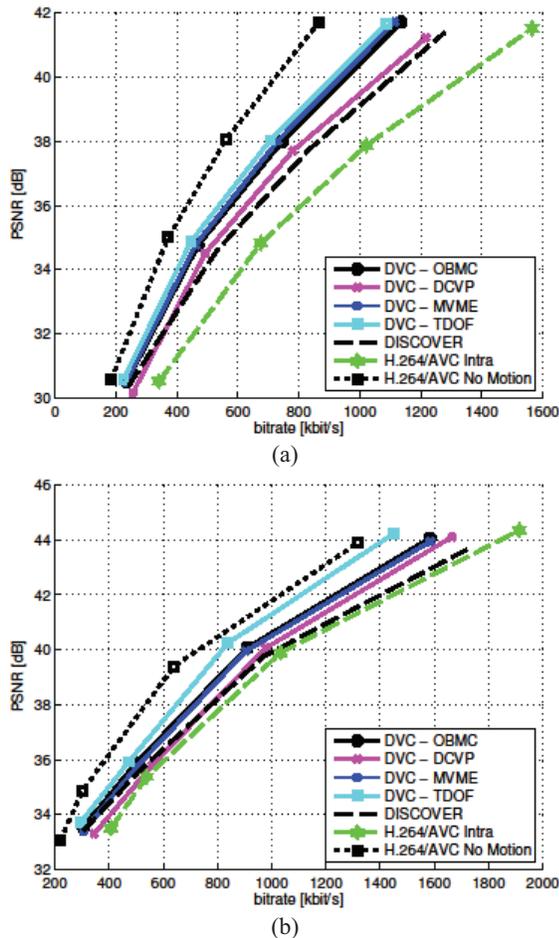

Fig. 3. RD performance for *Outdoor* (a) and *Kendo* (b) sequences.

REFERENCES

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

[2] 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.

[3] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, and J. Ostermann, "Distributed Monoview and Multiview Video Coding," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 67–76, 2007.

[4] D. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[5] A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[6] T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Fusion Schemes for Multiview Distributed Video Coding," in *Proc. EUSIPCO 2009*, Scotland, 2009, vol. 1, pp. 559–563.

[7] F. Dufaux, "Support Vector Machine Based Fusion for Multi-View Distributed Video Coding," in *Proc. DSP*, 2011, pp. 1–7.

[8] X. Artigas, F. Tarrés, and L. Torres, "Comparison of Different Side Information Generation Methods for Multiview Distributed Video Coding," in *Proc. SIGMAP 2007*, Barcelona, Spain, 2007.

[9] Huynh Van Luong, L. L. Raket, Xin Huang, and S. Forchhammer, "Side Information and Noise Learning for Distributed Video Coding Using Optical Flow and Clustering," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4782–4796, Dec. 2012.

[10] X. Huang and S. Forchhammer, "Cross-Band Noise Model Refinement for Transform domain Wyner–Ziv Video Coding," *Signal Process. Image Commun.*, vol. 27, no. 1, pp. 16–30, Jan. 2012.

[11] D. Varodayan, A. Aaron, and B. Girod, "Rate-Adaptive Codes for Distributed Source Coding," *EURASIP Signal Process. J.*, vol. 86, no. 11, pp. 3123–3130, Nov. 2006.

[12] D. Kubasov, J. Nayak, and C. Guillemot, "Optimal Reconstruction in Wyner-Ziv Video Coding with Multiple Side Information," in *Proc. IEEE MMSP, October, 2007*, 2007, pp. 183–186.

[13] C. Zach, T. Pock, and H. Bischof, "A Duality based Approach for Realtime TV-L^1 Optical Flow," in *Proc. Proceedings of the 29th DAGM conference on Pattern recognition*, Heidelberg, Germany, 2007, pp. 214–223.

[14] L. L. Rakêt, L. Roholm, A. Bruhn, and J. Weickert, "Motion Compensated Frame Interpolation with a Symmetric Optical Flow Constraint," in *Advances in Visual Computing*, vol. 7431 , 2012, pp. 447–457.

[15] I. Amidror, "Scattered Data Interpolation Methods for Electronic Imaging Systems: a Survey," *J. Electron. Imaging*, vol. 11, no. 2, p. 157, Apr. 2002.

[16] A. Smolic, G. Tech, and H. Brust, "Report on Generation of Stereo Video Database," Technical report D2.1, Jul. 2010.

[17] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER Codec: Architecture, Techniques And Evaluation," in *Proc. PCS*, Lisboa, Portugal, 2007.

[18] L. L. Rakêt, "Optical Flow C++ Implementation." [Online]. Available: http://image.diku.dk/larslau/.

[19] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Multiview Distributed Video Coding with Encoder Driven Fusion," in *Proc. EUSIPCO 2007,* Poznan, Poland, 2007.

[20] G. Bjøntegaard, "Calculation of Average PSNR Differences between RD Curves," in *ITU-T Q6/SG16, Doc. VCEG-M33, in: 13th Meeting*, Austin, USA, 2001.