# Conditional Random Field based Side-information Fusion for Distributed Multi-view Video Coding

Yongsheng Zhang[1], Hongkai Xiong[1], Hao Wang[1], Chang Wen Chen[2]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Computer Science and Engineering, State University of New York at Buffalo, NY 14260-2000, USA

***Abstract*— This paper presents a new temporal and inter-view side-information fusion algorithm for distributed multi-view video coding (DMVC). Unlike existing fusion algorithms in DMVC schemes that produce the fusion mask by finding the motion vector outliers, it introduces conditional random fields (CRF) to exploit the intrinsic geometric regularity and temporal consistency constraint in multi-view video sequences. Specifically, Wyner-Ziv (WZ) frames are modeled by CRF with the temporal and the inter-view side-information as two observations. The observation distribution models the local accuracy of the temporal and the inter-view side-information. The transition distribution of the CRF model represents the local geometric regularity, e.g., the edge directions and the local smoothness of the WZ frame. Its parameters are trained from previously decoded WZ frames, and the inference is made on trained weights to generate fused side-information. The accurate modeling is validated to show a significant performance gain over the existing fusion algorithms by experiments.**

## I. INTRODUCTION

Recently, multiview video systems have become more and more popular, *e.g.*, 3-D television, free viewpoint television, and wireless sensor networks. In view that multiview video requires much more bandwidth for transmission than single-view video, how to efficiently compress the multiview video has become a popular research topic. Since the multiview video consists of video sequences captured by multiple cameras towards the same scenario but from different angles and locations, significant correlation exists among views. To improve the compression efficiency by exploiting the inter-view correlation together with the temporal inter-frame correlation, Joint Video Team (JVT) has been developing the Joint Multiview Video Model (JMVM) [1] based on H.264/AVC which assumes that the video frames from different views can be freely exchanged or simultaneously available at the encoder. We should be aware that the communication between cameras with tremendous data volume is impractical, and its high encoding complexity becomes a big burden for multiview video capturing. To alleviate the encoding complexity while maintaining the coding efficiency, distributed multiview video coding (DMVC) scheme [2] has been concerned to attain benefits inherent to the Wyner-Ziv (WZ) theorem.

Wyner and Ziv have proved that even if correlated sources are separately encoded without getting information from each other, the coding performance can be as good as joint encoding if the compressed signals can be jointly decoded [3]. The most attractive advantage of WZ video coding, namely, distributed video coding (DVC), is that it can shift the computation-intensive motion estimation process from the encoder side to the decoder side, and thus significantly reduce the encoding complexity.

In DVC applications, a widely accepted approach to improve the rate-distortion (RD) performance is to produce side-information with higher quality [4]–[6]. In a typical DMVC framework, there are two kinds of side-information for a WZ frame: temporal side-information and inter-view side-information. The former is typically generated by motion compensated interpolation (MCI) [4], and the latter is usually generated by affine transform [2], [5]. Guo *et. al.* proposed to exploit the inter-view correlation by a six-parameter global affine transform model [2]. To model the depth variation between neighboring views, Xiong *et. al.* proposed a sub-graph matching-based inter-view side-information generation algorithm using SIFT (scale-invariant feature transform) descriptor and sub-graph segmentation [5]. Given these derived temporal and inter-view side-information, a fusion algorithm is desired to select better side-information at different regions to produce the final side-information for WZ decoding. Xiong *et. al.* adopted a fusion algorithm by analyzing the motion field consistency [5]. The temporal side-information is replaced by the inter-view side-information in regions with intensive motion. Artigas *et. al.* have ever fused side information by analyzing the motion compensation error or neighboring frame error to predict the reliability of temporal side-information [6]. However, those fusion methods only exploit the relative motion between frames without taking into account the spatial consistency property within the WZ frame.

In this paper, we propose a new side-information fusion method for DMVC applications based on Conditional Random Field (CRF) modeling [7], where the spatial consistency constraint along with the temporal coherent property of video sequences is exploited to improve fusion algorithm performance. Specifically, the DMVC side-information fusion problem is formulated as a CRF model to reflect interactions among neighboring sites on a 2-D lattice. The CRF model consists of two components: a local decision term and an interaction term. The local decision term decides the label of a given site based only on observations and ignoring labels of neighboring sites. The interaction term can be seen as a data-dependent smoothing function, which penalizes the label deviation between neighboring sites. Parameters of the CRF model are trained from previously decoded WZ frames.

The rest of this paper is organized as follows. The DMVC codec structure is introduced in Section II. The proposed CRF-based fusion method is presented in Section III and evaluated with experiments in Section IV. Section V concludes this paper.

## II. DISTRIBUTED MULTI-VIEW VIDEO CODING FRAMEWORK

In a multi-view video coding (MVC) scheme, multiple cameras are used to capture the same scene from different directions. Since the angle between camera views is small, high inter-view redundancy exists between video sequences captured from neighboring cameras. To eliminate inter-view communication and reduce computational complexity while preserving high coding efficiency, distributed multiview video coding (DMVC) scheme has been proposed [2], [8].

In DMVC, video frames are divided into two categories: key frames and WZ frames. Key frames are encoded with conventional hybrid predictive video coding schemes, *e.g.*, H.264/AVC Intra or Inter coding, and used as reference to produce side-information for WZ decoding, while WZ frames are intra encoded with WZ encoder and jointly decoded at the decoder side. The temporal and inter-view

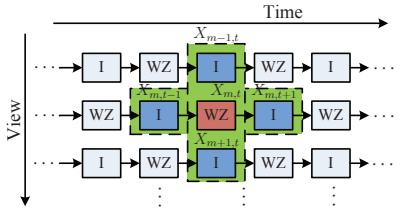



Fig. 1: A typical DMVC frame structure.

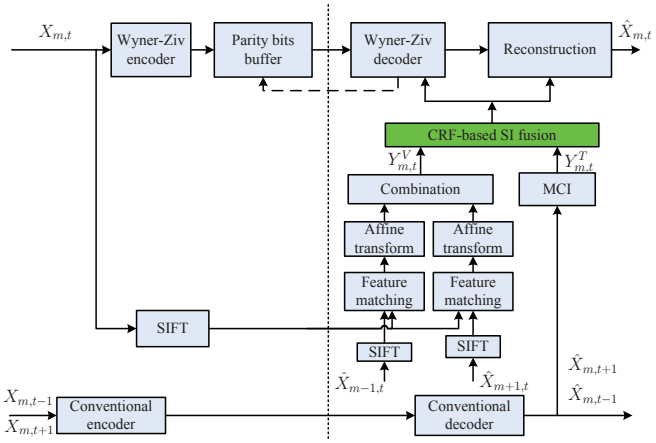



Fig. 2: Codec architecture of WZ-DMVC schemes.

redundancy are only exploited at the decoder side. Figure 1 shows the frame structure of a typical DMVC scheme. As usually assumed in the literature [2], [5], the conventional frames are encoded with H.264/AVC Intra mode.

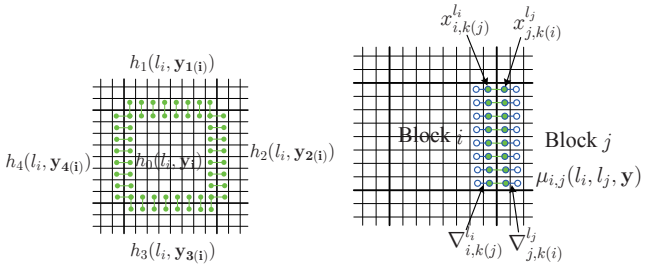
Figure 2 presents the codec architecture of the DMVC scheme adopted in this paper. For a WZ frame $X_{m,t}$, its temporal side-information $Y^T_{m,t}$ is first generated from temporal neighboring frames $X_{m,t-1}$ and $X_{m,t+1}$ by motion compensated interpolation with spatial motion smoothing [4]. The inter-view side-information $Y^V_{m,t}$ is generated based on frames $X_{m-1,t}$ and $X_{m+1,t}$ from neighboring cameras through sub-graph matching-based inter-view side-information generation algorithm using SIFT descriptor and sub-graph segmentation [5]. With side-information $Y^T_{m,t}$ and $Y^V_{m,t}$, the proposed fusion algorithm is used to produce the final side-information $Y_{m,t}$. $Y_{m,t}$ together with the received WZ bit-stream will be fed into the WZ decoder to produce the decoded WZ frame $\hat{X}_{m,t}$.

## III. SIDE-INFORMATION FUSION

In the literature [5], [6], fusion algorithms are typically realized by analyzing the consistency of motion fields. For regions with intensive motion, the temporal side-information is replaced by the inter-view side-information to produce the final side-information for WZ decoding [5]. However, this approach does not exploit the interaction constraints between neighboring blocks and the spatial smoothness property of natural images. In this paper, we propose to formulate the side-information fusion problem as a CRF model to represent not only the local label association of individual blocks but also interactions between neighboring blocks.

### A. Conditional Random Field

Conditional random field (CRF) is a type of discriminative undirected probabilistic graphical model. It is usually used for passing




(a) Association feature (b) Interaction feature

Fig. 3: Features exploited in the CRF model.

or labeling sequential data, such as natural language processing or biological sequences [7]. Recently, it is extended to the 2D lattice graphical models in computer vision applications, *e.g.*, image segmentation and object recognition [9], to capture the label associations at individual sites as well as the interactions between neighboring sites on a 2D grid lattice.

In CRF, each vertex of the graph represents a random variable whose distribution is to be inferred, and edges in the graph present interactions between a pair of random variables. The random variables $l_i$ obey the Markov property with respect to the graph, *i.e.*, $p(l_i|\mathbf{y}, \mathbf{l}_{S/i}) = p(l_i|\mathbf{y}, \mathbf{l}_{\mathcal{N}_i})$, where $S/i$ represents the set of all sites in the graph except site $i$, $\mathcal{N}_i$ is the set of neighbors of site $i$ in the graph, and $\mathbf{l}_\Omega$ represents the set of labels for sites in set $\Omega$.

CRF model is globally conditioned on all observations in $\mathbf{y}$, which is different from the Markov random field (MRF). The condition of positivity requirement has been assumed implicitly, *i.e.*, $p(\mathbf{l}|\mathbf{y}) > 0$ for $\forall \mathbf{l}$. According to the Hammersley-Clifford theorem [10], and assuming only up to pairwise clique potentials to be nonzero, distribution $p(\mathbf{l}|\mathbf{y})$ can be written as

$$p(\mathbf{l}|\mathbf{y}) = \frac{1}{Z} \exp\left( \sum_{i \in S} A_i(l_i, \mathbf{y}) + \sum_{i \in S} \sum_{j \in \mathcal{N}_i} I_{i,j}(l_i, l_j, \mathbf{y}) \right), \quad (1)$$

where $Z$ is a normalization factor. The association potential $A_i(l_i, \mathbf{y})$ measures how well the label of a specific site matches the overall observation $\mathbf{y}$ ignoring labels of neighboring sites. The interaction potential $I_{i,j}(l_i, l_j, \mathbf{y})$ is a data dependent smoothing function, which models the interactions between labels of site $i$ and site $j$ given observation $\mathbf{y}$. In what follows, we will explain the association potentials $A_i(l_i, \mathbf{y})$ and the interaction potentials $I_{i,j}(l_i, l_j, \mathbf{y})$ in details.

### B. Association potential

In a CRF model, association potential $A_i(l_i, \mathbf{y})$ models the cost of assigning label $l_i$ to site $i$ given observation $\mathbf{y}$. It is defined as

$$A(l_i|\mathbf{y}) = \mathbf{w}_i^T \mathbf{f}_i(\mathbf{y}), \quad (2)$$

where $\mathbf{w}_i$ are model parameters, function $\mathbf{f}_i(\mathbf{y})$ maps the observation $\mathbf{y}$ to a feature vector for each site $i$, such that $\mathbf{f}_i : \mathbf{y} \to R^p$. It is worth to mention that the feature function $\mathbf{f}_i(\mathbf{y})$ is a function of the whole set of observation $\mathbf{y}$, on the contrary, MRF model only uses data from the individual site to define the association potential.

For the side-information fusion problem, a WZ frame is divided into non-overlapping blocks with fixed size. Each block is considered as a site of the CRF model, and feature function $\mathbf{f}_i^l(\mathbf{y})$ represents the cost of assigning label $l_i$ to site $i$ given the whole observation $\mathbf{y}$. For simplicity, only observations from site $i$ and its neighboring

sites $\mathcal{N}(i)$ are considered in the association potential. Since the MV thresholding fusion algorithm [5] exploits the smoothness property of the motion field and produces a coarse reliability map of the temporal side-information, we adopt the MV thresholding fusion result as an initial setup for the CRF model.

The association features measure the deviation from the initial estimate and the discontinuity strength on boundaries between neighboring blocks. For the side-information fusion problem in this paper, the association feature vector $\mathbf{f}_i^l(\mathbf{y})$ is composed of five elements, as shown in Figure 3a: feature $h_0(l_i, \mathbf{y}_i)$ measures the deviation cost from initial estimate and feature $h_1(l_i, \mathbf{y}_{1(i)})$ to $h_4(l_i, \mathbf{y}_{4(i)})$ measures the discontinuity cost on the four block boundaries, *i.e.*,

$$\begin{aligned}\mathbf{f}_i^l(\mathbf{y}) = \quad & \{h_0(l_i, \mathbf{y}_i), h_1(l_i, \mathbf{y}_{1(i)}), h_2(l_i, \mathbf{y}_{2(i)}), \\ & h_3(l_i, \mathbf{y}_{3(i)}), h_4(l_i, \mathbf{y}_{4(i)})\}. \end{aligned} \quad (3)$$

where the on-site deviation cost is

$$h_0(l_i, \mathbf{y}_i) = \frac{1}{B} \sum_b (x_{i,b}^{l_i} - y_{i,b})^2, \quad (4)$$

$b \in [1, B]$ is the index of pixels in block $i$, $B$ is the block size, $x_{i,b}^{l_i}$ is the $b$-th pixel in block $i$ of the temporal or inter-view side-information according to the label $l_i$. The boundary discontinuity cost is defined as

$$h_t(l_i, \mathbf{y}_{i,t}) = \frac{1}{K} \sum_{k \in S_{i,t}} (x_{i,k(j)}^{l_i} - y_{j,k(i)})^2. \quad (5)$$

$t \in [1, 4]$ is the index of four boundaries, $k$ is the index of boundary pixels, $S_{i,t}$ is the index set of boundary pixel for boundary $t$, $y_{j,k(i)}$ is the $k$-th pixel on the boundary of block $j$ close to block $i$.

### C. Interaction potential

The interaction potential is defined as a function of observation $\mathbf{y}$ and labels of neighboring sites. The label assignment should minimize the discontinuity strength on boundaries between neighboring blocks. For neighboring block $i$ and block $j$, their interaction potential is defined as

$$I(l_i, l_j, \mathbf{y}) = v_{i,j} \mu_{i,j}(l_i, l_j, \mathbf{y}), \quad (6)$$

where feature function $\mu_{i,j}(l_i, l_j, \mathbf{y})$ is defined as

$$\mu_{i,j}(l_i, l_j, \mathbf{y}) = \frac{1}{K} \sum_{k \in S_{i,t}} \left( x_{i,k(j)}^{l_i} - x_{j,k(i)}^{l_j} - \nabla_{i,j,k}^{l_i,l_j} \right)^2, \quad (7)$$

where $k$ is the index of boundary pixels, $x_{i,k(j)}^{l_i}$ is the $k$-th pixel on the boundary of block $i$ close to block $j$ given label assignment $l_i$, $\nabla_{i,j,k}^{l_i,l_j}$ is the target gradient between the $k$-th pixel pair on the boundary between block $i$ and block $j$ given label assignment $l_i$ and $l_j$, respectively.

$$\nabla_{i,j,k}^{l_i,l_j} = \frac{1}{2} \left( \nabla_{i,k(j)}^{l_i} + \nabla_{j,k(i)}^{l_j} \right), \quad (8)$$

where $\nabla_{i,k(j)}^{l_i}$ is the gradient for the $k$-th pixel on boundary of block $i$ to block $j$, and $\nabla_{j,k(i)}^{l_j}$ is the gradient for the $k$-th pixel on boundary of block $j$ to block $i$, as shown in Figure 3b.

It is worth to mention that the interaction feature $\mu_{i,j}(i, j, \mathbf{y})$ measures the boudnary discontinuity strength of assigning label $l_i$ and label $l_j$ to neighboring block $i$ and $j$, respectively. While the association feature $h_t(l_i, \mathbf{y}_{i,t})$ in Eq. (5) measures the discontinuity strength between current label assignment $l_i$ and the observations in neighboring blocks.

### D. Training and Inference

Let $\theta = \{\mathbf{w}, \mathbf{v}\}$ be the set of CRF parameters. These parameters are trained with standard maximum-likelihood approach using previously decoded WZ frames involving the evaluation of the normalization factor $Z$. In general, the evaluation of $Z$ is a NP-hard problem. Its parameters can be estimated with either sampling techniques or some approximations. In this paper, we adopt the pseudo-likelihood formulation due to its simplicity and consistency of the estimate for the large lattice limit.

$$\begin{aligned}\hat{\theta}^{ML} = \quad & \arg\max_\theta \sum_{m=1}^{M} \Big( \sum_{i \in S} A_i(l_i, \mathbf{y}, \theta) + \\ & \sum_{i \in S} \sum_{j \in \mathcal{N}(i)} I_{i,j}(l_i, l_j, \mathbf{y}, \theta) - \log Z - \frac{1}{2\tau^2} \sum_k \theta_i^2 \Big), \end{aligned} \quad (9)$$

where $M$ is the number of decoded WZ frames used for training, and it is set to 1 in experiments.

$$Z = \sum_{l_i = \{-1, 1\}} \exp \left( \sum_{i \in S} A_i(l_i, \mathbf{y}, \theta) + \sum_{i \in S} \sum_{j \in \mathcal{N}(i)} I_{i,j}(l_i, l_j, \mathbf{y}, \theta) \right). \quad (10)$$

If $\tau$ is given, the penalized log pseudo-likelihood in Eq. (9) is convex with respect to the parameter $\theta$, and can be easily maximized using gradient descent algorithm.

For a WZ frame, the goal of CRF fusion is to find the optimal label assignment $\mathbf{l}$ over the frame with respect to the defined cost function. In this paper, the belief propagation algorithm is adopted for the approximate inference process [11].

## IV. EXPERIMENTAL RESULTS

In this section, we conduct experiments to evaluate the performance of the proposed CRF-based side-information fusion algorithm for DMVC applications. In experiments, two different fusion algorithms are evaluated: the proposed CRF-based algorithm and the MV thresholding fusion algorithm [5]. Two SIF ($320 \times 240$) video sequences (*Race* and *Flamenco2*) are tested. The frame rate of WZ frame is $15FPS$ and only WZ frames are evaluated for RD performance in experiments. The motion vector smoothing bidirectional motion compensation algorithm with block size $16 \times 16$ and search range 16 is used to generate temporal side-information [4], and sub-graph matching based affine transform is used to generate the inter-view side-information [5]. The LDPCA approach with block length 6336 is adopted as the Slepian-Wolf codec [12]. One decoded WZ frame is used to train CRF parameters, *i.e.*, $M$ is set to 1 in equation (9).

Figure 4 presents the PSNR of the fused side-information with different fusion algorithms. The temporal and inter-view side-information are also shown for comparison. The results show that the proposed CRF-based fusion algorithm achieves significant performance gain. Since parameters of the proposed algorithm are trained from previously decoded WZ frames, we also evaluate the influence of the previously decoded WZ frame quality to the fusion results. Results in Figure 4 show that higher quality WZ frame is helpful in the training process and produces more accurate side-information, where the fused side-information with WZ frame QP of 22 is better in some frames than that with WZ frame QP of 28. However, the proposed fusion algorithm achieves superior performance than the MV thresholding fusion algorithm in both cases.

Table I presents the average PSNR of fused side-information with different fusion algorithms, and these fusion results are also compared to the temporal and inter-view side-information. Results in Table I show that the PSNR of fused side-information of the "Race" sequence

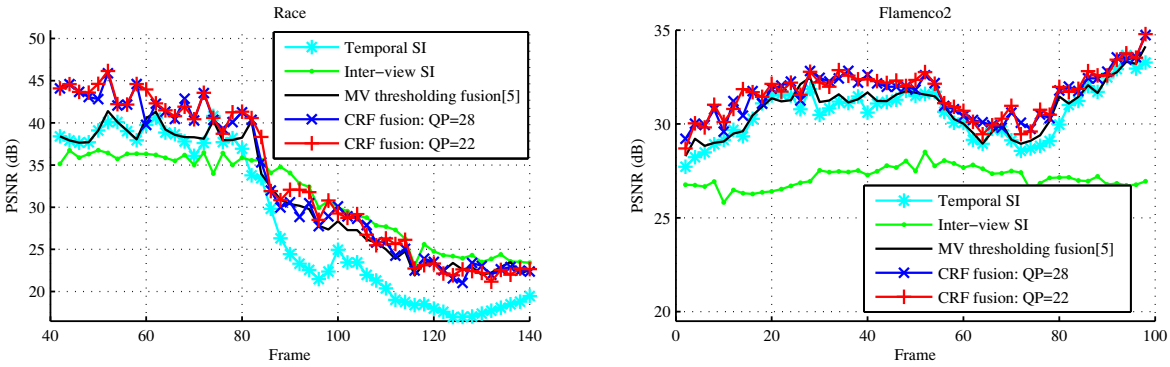



Fig. 4: The fused side-information PSNR of each WZ frame with different fusion methods.

TABLE I: The average side-information PSNR with different fusion algorithms.

| | | Test sequences | |
|---|---|---|---|
| | | Race | Flamenco2 |
| Temporal SI (dB) | | 28.306 | 30.608 |
| Inter-view SI (dB) | | 31.145 | 27.140 |
| MV threshold [5] (dB) | | 31.305 | 30.848 |
| CRF Fusion (dB) | QP=28 | 32.782 | 31.544 |
| | QP=22 | 33.161 | 31.615 |

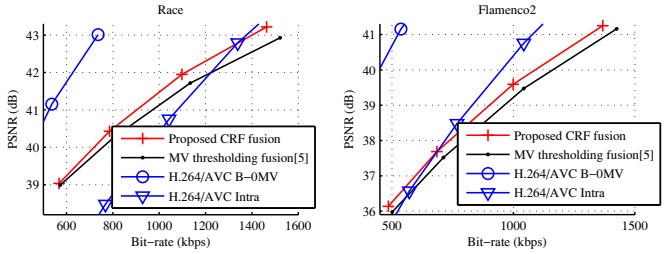



Fig. 5: Rate-distortion performance of DMVC scheme with the proposed side-information fusion algorithm.

with WZ frame QP of 22 is about $0.379dB$ higher than that with WZ frame QP of 28, while the difference is only $0.071dB$ for the "Flamenco2" sequence. These results show that the WZ frame quality used for training has slightly influence to the training accuracy. The results also show that even with WZ frame QP of 28, the proposed algorithm achieves up to $1.477dB$ and $0.696dB$ performance gain compared with the MV thresholding fusion algorithm for the "Race" and "Flamenco2" respectively.

Figure 5 presents the rate-distortion performance of WZ frames with different side-information fusion algorithm. Results in Figure 5 show that the proposed fusion algorithm achieves notable performance gain compared with the fusion algorithm based on MV thresholding [5].

## V. CONCLUSIONS

In this paper, we proposed a CRF-based temporal and inter-view side-information fusion algorithm for DMVC applications, where the association potential measures the cost of a label assignment for a given site only based on observations and ignoring labels of other sides, while the interaction potential imposes spatial consistency constraint for label pairs of adjacent sites. Experimental results show a considerable performance gain compared with existing fusion algorithms.

## REFERENCES


[1] A. Vetro, Y. Su, H. Kimata, and A. Smolic, "Joint multiview video model JMVM 2.0," ITU-T and ISO/IEC Joint Video Team, Doc. JVT-U207, 2006

[2] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Wyner-Ziv based multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 6, pp.713-724, Jun. 2008.

[3] A. Wyner, "Recent Results in the Shannon Theory," *IEEE Transactions on Information Theory*, Vol. 20, no. 1, pp. 2- 10, Jan. 1974.

[4] Ascenso, J. and Pereira, F., "Advanced Side Information Creation Techniques and Framework for Wyner-Ziv Video Coding," *Journal of Visual Communication and Image Representation, Special Issue Resource-aware adaptive video streaming*, vol. 19, no. 8, pp. 600-613, Dec. 2008.

[5] H. Xiong, H. Lv, Y. Zhang, L. Song, Z. He, and T. Chen, "Subgraphs matching-based side information generation for distributed multiview video coding," *EURASIP Journal on Advances in Signal Processing*, Article ID 386795, 2009.

[6] X. Artigas, E. A., and L. Torres, "Side Information Generation for Multiview Distributed Video Coding Using a Fusion Approach," *in Proc. Nordic Signal Processing Symposium (NORSIG 2006)*, Jun. 2006, pp. 250-253.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". *in Proc. International Conference on Machine Learning*, 2001, pp. 282-289.

[8] T. Maugey, W. Miled, M. Cagnazzo and B. Pesquet-Popescu, "Fusion schemes for multiview distributed video coding," *in Proc. 17th European Signal Processing Conference*, Glasgow, Scotland, Aug, 2009, pp. 559-563.

[9] S. Kumar and M. Hebert, "Discriminative Random Fields," *International Journal of Computer Vision*, vol. 68, pp. 179-201, Apr. 2006.

[10] J. M. Hammersley and P. Clifford, "Markov random fields and Gibbs random fields", *Israel Journal of Mathematics*, vol. 14, no. 1, 1973, pp., 92-103.

[11] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, no. pp. 689-695, Dec. 2000.

[12] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive codes for distributed source coding," *Signal Processing*, vol. 86, no. 11, pp. 3123-3130, Nov. 2006.