Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 386795, 17 pages
doi:10.1155/2009/386795

Research Article

Subgraphs Matching-Based Side Information Generation for Distributed Multiview Video Coding

Hongkai Xiong,[1,2] Hui Lv,[1] Yongsheng Zhang,[1] Li Song,[1] Zhihai He,[3] and Tsuhan Chen[2]

[1] *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[2] *Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[3] *Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211, USA*

Correspondence should be addressed to Hongkai Xiong, xionghongkai@sjtu.edu.cn

Received 23 April 2009; Revised 25 September 2009; Accepted 9 December 2009

Recommended by Mei-Juan Chen

We adopt constrained relaxation for distributed multiview video coding (DMVC). The novel framework integrates the graph-based segmentation and matching to generate interview correlated side information without knowing the camera parameters, inspired by subgraph semantics and sparse decomposition of high-dimensional scale invariant feature data. The sparse data as a good hypothesis space aim for a best matching optimization of interview side information with compact syndromes, from inferred relaxed coset. The plausible filling-in from a priori feature constraints between neighboring views could reinforce a promising compensation to interview side-information generation for joint multiview decoding. The graph-based representations of multiview images are adopted as constrained relaxation, which assists the interview correlation matching for subgraph semantics of the original Wyner-Ziv image by the graph-based image segmentation and the associated scale invariant feature detector MSER (maximally stable extremal regions) and descriptor SIFT (scale-invariant feature transform). In order to find a distinctive feature matching with a more stable approximation, linear (PCA-SIFT) and nonlinear projections (Locally linear embedding) are adopted to reduce the dimension SIFT descriptors, and TPS (thin plate spline) warping model is to catch a more accurate interview motion model. The experimental results validate the high-estimation precision and the rate-distortion improvements.

1. Introduction

Multiview video coding (MVC) has played a new paradigm of a wide variety of interactive multimedia applications. In many MVC systems, the fundamental efforts have been dedicated to investigating the adjacent views in addition to the traditional temporal and spatial correlations within a single view. The availability of multiple views benefits many image processing tasks such as enhancement, segmentation, or object recognition. However, the existing interview prediction assumes that the video frames from different views can be freely exchanged or simultaneously available at the encoder [1]. We should be aware that the communication between cameras with tremendous data volume is impractical. Inspired from lossless Slepian-Wolf and lossy Wyner-Ziv source coding theory [2, 3] where separate encoding of correlated sources can approach the rate of joint entropy, provided joint decoding is executed with known correlation. Distributed Multiview Video Coding (DMVC) is normally emerging to attain benefits inherent to distributed video coding (DVC) [4].

Suppose X and Y are correlated sources termed as source data and side information [2]. Traditional source coding assumes that Y should be available at both encoder and decoder, and then the rate-distortion (R-D) function for X given Y is $R_{X|Y}(D)$. Conversely, distributed source coding (Wyner-Ziv) theorem assumes that Y is only available at decoder, and encoder could only access to the correlation between X and Y, and corresponding rate-distortion function is denoted as $R_{X|Y}^{WZ}(D)$. Surprisingly, a rate loss $R_{X|Y}^{WZ}(D) - R_{X|Y}(D) = 0$ is proved feasible for Gaussian memoryless source and mean square error (MSE) distortion metric [3]. Pradhan et al. also have proved that there is no rate loss for arbitrary side information Y and independent Gaussian noise in theory [5]. For general distribution and arbitrary distortion metric, Zamir [6] has proved that the rate loss is less than 0.5 bit/sample. Several practical Slepian-Wolf and Wyner-Ziv video coding approaches have been proposed [7–13], where temporal prediction for the side information of the estimated frame is fulfilled at the decoder side other than the encoder side. Pradhan and Ramchandran [8] contribute a DVC framework based on syndrome for cosets, which encodes the residue of Wyner-Ziv frame with traditional block-based prediction coding scheme on the scale of motion and computation. Because the operational block length is small, PRISM might adopt relative short BCH block codes. Aaron et al. [9] develop a transform-domain DVC scheme with intraframe encoding and interframe decoding, which uses Turbo coder for each subband. Impressively, both side information Y and correlated channel between coded source and side information would impose the essential constraints on DVC coding performance.

Due to the extremely large amount of data associated with Multiview video, efficient compression techniques are essential for 3D scene communication by exploiting the inherent similarities of the Multiview imagery: interview and temporal similarity. As we know, temporal similarities have been motivated as a variety of motion compensated-prediction (MPC) methods in hybrid video compression standards, for example, MPEG-4, H. 264, and WM9. According to the level of geometric redundancy for Multiview imagery, various Multiview video compression algorithms could be categorized into three classes: 3D model-based algorithms, disparity/depth-based algorithms, and distributed compression.

In 3D model-based/model-aided algorithms, the geometry of the objects of the scene is recovered using camera parameters, which are obtained by camera calibration of shape-from silhouettes techniques. Scene geometry is explicitly used to convert images to view-dependent texture maps prior to compression [14–16]. However, there is a high degree of freedom between multiple views, and 3D scene geometry is impractical to be available or accurately estimated for intermap correlation.

In disparity/depth-based algorithms, scene geometry is implicitly used by performing disparity prediction and compensation across the different views or combing depth information of each view. It is noted that disparity is the displacement of corresponding points from different shooting positions of the cameras. A typical example is scalable hybrid predictive coding (SHPC) algorithm [17, 18], where one view is compressed as a based layer by normal single-view compression and other views as enhancement layer(s) in combination with multiple depth information. Herein, disparity-compensated prediction (DCP) is used to reduce the interview redundancy. Joint Video Team (JVT) has also been developing Joint Multiview Video Model (JMVM) based on H.264/AVC-based trajectory [19]. However, disparity estimation (DE) to obtain the dense map of the corresponding points from different view is still an open challenge for computer vision paradigm.

Distributed compression algorithms compress each video stream individually without geometric priors. For DMVC, flexible prediction fusion methods between temporal and view correlations have been seriously considered to generate the side information at the decoder. Previously, Zhu et al. simply absorbed Wyner-Ziv coding to compress data acquitted by large light field system [20]. And then Aritgas et al. use View Synthesis Prediction to compensate interview correlated side information [21]. However, VSP needs depth information for each frame and is not realistic due to complex appearance of real scenes [22]. In [23], a mix prediction method is applied through wavelet transform. However, the coding performance is limited without explicit inference of correlation. Revisiting the transformation from signal to bases, we have recognized that it generally achieves two desirable properties: variable decoupling and dimension reduction. It is shown in harmonic analysis that the Fourier, wavelet, and ridgelet bases are independent components for various ensembles of mathematical functions. Unfortunately, the ensemble of natural images is obviously different from those functional classes so that it degrades the correlation estimation and rate-distortion performance in DMVC. Therefore, image components must be adapted to natural images, and it leads to sparse coding with overcomplete basis or dictionary. Going beyond the image bases, the texton-like representation consisting of a number of image bases at various geometric, photometric, and dynamic configurations is taken into account. The basic idea has been presented in our previous work [24], where a feature-based Wyner-Ziv coding framework (FWZC) for DMVC is explored to preserve the constrained relaxation with multiple side information implication and high-level features matching at the decoder.

In this paper, we present a novel graph matching-based FWZC scheme. It integrates graph-based segmentation and matching to generate interview correlated side information with a significant rate-distortion performance and without knowing the camera parameters. It is inspired by subgraph semantics and sparse decomposition of high-dimensional scale invariant feature data. The sparse feature data as a good hypothesis space are employed to enable best matching optimization of interview side information with compact syndromes, from inferred relaxed coset. Obviously, a priori knowledge extracted from multiple image descriptions of neighboring views should reinforce a plausible compensation and approximation to the original information in a converged sense. The graph-based representations of Multiview images are adopted as constrained relaxation, which assists the interview correlation matching for subgraph semantics of the original Wyner-Ziv image by the graph-based image segmentation and the associated scale invariant feature detector MSER (maximally stable extremal regions) and descriptor SIFT (scale-invariant feature transform). In order to find a distinctive feature matching with a more stable approximation, linear (PCA-SIFT) and nonlinear projections (Locally linear embedding, LLE) are adopted to reduce the dimension of high-dimensional SIFT descriptors, and TPS (thin plate spline) warping model is to catch a more accurate interview motion model in 3D angle of view.

This paper is organized in the following manner. Section 2 presents the DMVC architecture and highlights the formulation of subgraph-based DMVC scheme with constrained relaxation. In Section 3, a detailed implementation

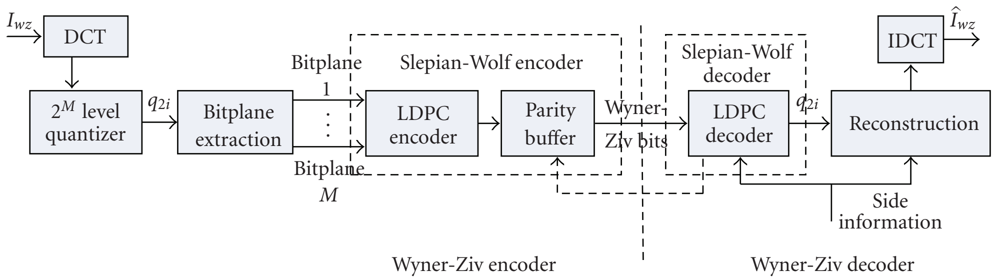


Figure 1: A practical Wyner-Ziv video codec architecture.

of the subgraph-based DMVC scheme is illustrated, involving with graph-based image segmentation, feature extraction and sparse description, and graph-based matching with warps. Section 4 presents the experiment results. Section 5 concludes the paper and discusses future directions.

2. Problem Statement

2.1. Distributed MVC. In video coding standardized by MPEG or the ITU-T H.26x recommendations, the encoder and decoder jointly exploit the statistics of the source signal. Separate encoding of correlated sources can approach the rate of joint entropy, provided joint decoding is executed with known correlation. Figure 1 shows a practical Wyner-Ziv video codec architecture.

A blockwise DCT is firstly applied to a Wyner-Ziv frame I_{wz}. For each DCT transform coefficient band of a Wyner-Ziv frame I_{wz} (even frames), the Wyner-Ziv codec makes use of a quantizer, a bit-plane extraction, and a Slepian-Wolf codec (Turbo or LDPC) to generate layered parity bits. These parity bits are punctured and transmitted upon request by the decoder through a feedback channel. At the decoder side, the odd frames are conventionally decoded to generate the side information. The side information can be seen as a noisy version of the Wyner-Ziv frames, and the decoder employs a Laplacian noise model for error correction of received codes. The Laplacian parameter is estimated by observing the statistics from decoded frames. More parity bits from the encoder buffer through feedback are requested once the decoder cannot reliably decode the original symbols.

The decoder and the reconstruction modules assume a Laplacian residual distribution between Wyner-Ziv frame I_{wz} and side information Y. Let d be the difference between corresponding coefficients in I_{wz} and Y, then the distribution of d can be approximated as $f(d) = (\alpha/2)e^{-\alpha|d|}$ for each subbands. Let c_j^i denote the ith bit of a coefficient c_j and let $\hat{c}_j^i$ denote the estimated reconstruction value for c_j^i. The probability can be computed using the residual distribution model as follows:

$$P = \frac{\alpha}{2}e^{-\alpha|d_{c_j^i}|}, \quad d_{c_j^i} = \left(m_i I\left(\hat{c}_j^i\right) + \text{offset}\right)i - I\left(y_j\right), \tag{1}$$

where m_i represents the magnitude of ith bit-plane, $I(\hat{c}_j^i)$ indicates the possible value of c_j^i (1 or 0), y_j indicates the coefficient of side information corresponding to c_j, and offset is an estimated value used to compensate the lower part of c_j. Because the lower bit-plane of c_j is still not decoded, the value of offset is decided in terms of the distribution parameter and the quantization step size.

DMVC is normally emerging to attain benefits inherent to distributed video coding for the Multiview camera setup in Figure 2. It arranges Intraframes and Wyner-Ziv frames, noted by I and WZ, respectively, in an interlaced way. Two directions are defined: Temporal Directions, from which the intraview side information is generated by the temporal interpolation and View Direction, from which constrained relaxation matching is applied to infer the interview correlated side information. The whole DMVC system consists of independent encoder and joint decoder. Thus, low encoding complexity and high coding performance can be achieved.

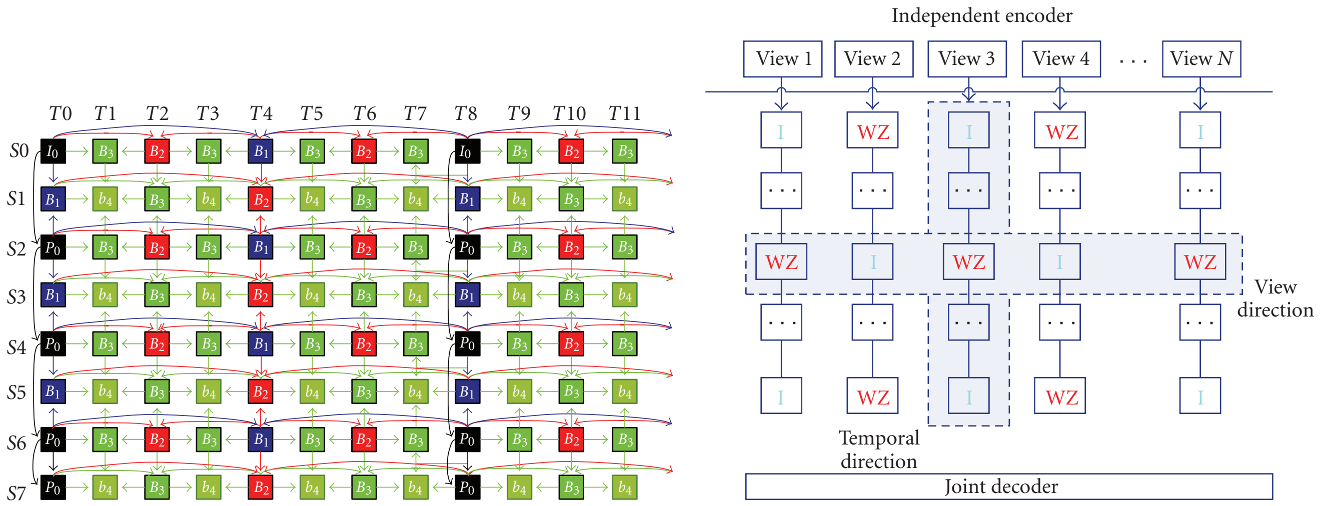
2.2. Feature-Based Wyner-Ziv Coding with Constrained Relaxation. The basic idea of Slepian-Wolf coding theory is to partition the space of all possible source outcomes into disjoint bins (sets). Usually, these bins (sets) are used as the cosets of some linear channel code for the specific correlation model. FWZC extended this idea by using high-level features, $F(I_{wz})$, as constraints. I_{wz} is WZ frame, and the group of features $F(I_{wz})$ constructs a relaxed frame coset, $\mathfrak{U}$, which consists of a set of frames that are inferred as all the possible representations of I_{wz} under the available constraints and syndromes.

$$\mathfrak{U}\left\{\hat{I}_{wz}(g) = Y \oplus Z(g)\right\}, \tag{2}$$

where $\oplus$ is the operation that uses $z(g)$ (parity bits gradually received) to correct Y, which is the approximation (side information) to I_{wz}. Thus, the fundamental envision is to use $F(I_{wz})$ to decode the original video frame by finding the best match in $\mathfrak{U}\{I_{wz}(g)\}$. This can be formulated as

$$\hat{I}_{wz} = \arg\min_{g \in \mathfrak{U}} \left\| I_{wz} - \hat{I}_{wz}(g) \right\|. \tag{3}$$

Figure 3 depicts the coding structure of the FWZC system. Compared with the traditional DCT domain Wyner-Ziv coding in Figure 1, this coding procedure imposes two new modules; one is the feature extraction/matching module and the other is the side information fusion module. It extracts scale invariant local features as the high-level constraints, which are transmitted to the decoder to notify the decoder how the source frame looks like (Y), the more



(a) JMVM

(b) DMVC

Figure 2: The Multiview camera setup and coding structure: (a) spatial-temporal prediction structure based on H.264/MPEG-AVC hierarchical B pictures; (b) joint spatial-temporal side information generation process for DMVC.

distinctive the information ($F(I_{wz})$) is, the easier the target I_{wz} can be identified, so that fewer parity bits $Z(g)$ are required to decode I_{wz}. It can be equivalent to a learning-based optimization problem from sparse data. In essence, prior information can be used for choosing an efficient input representation, or for choosing a good hypothesis space that leads to enhanced performance of the learning machine. In view of the given set of data provided by the random sampling with a noisy function, such a problem is ill-posed as there exists an infinity of functions that pass through the data. The common way with regard to regularization theory is by means of a stabilizer assuming that the function presents some intrinsic properties, for example, smoothness. It induces the underlying problem in (3) of finding the function that minimizes the functional combination of the empirical convex loss and prior information, associating with different approximation on the balance between fitness and prior constraints.

The incurred attempt is dedicated to finding distinctive frame information $F(I_{wz})$ to generate more likely approximation Y to I_{wz}.

Specifically, the "Side information generation" module in Figure 3 firstly generates a temporal side information. In terms of the obtained MVs, the "Arbitrator" module selects desired regions where spatial cue is required and requests local features of these regions from the encoder. With the received local features, interview side information is generated in the "Side information generation" module. Finally, the "Arbitrator" fuses the interview side information and temporal side information to generate the final side information for Multiview Wyner-Ziv decoding. This setup deduces the computational complexity of the encoder by only extracting local features from partial samples of the frame.

Noted, once there exists bad matching within either wide view images or occlusion, we might (1) use the RANSAC algorithm [25] to cope with a large proportion of outliers in the candidate point data. It uses the smallest point set possible beyond conventional sampling techniques; (2) we can use the image extrapolation or inpainting approaches [26] to synthesis the bad matching region from both surrounding areas and interview reference images via a partial difference equation (PDE). Typically, it could be interpreted as an iterative optimization algorithm to approximate the minimum of the energy using belief propagation.

2.3. Subgraphs-Based Matching for FWZC. In this paper, we explore this feature-based idea with multiple representations of graphs to break those bottlenecks previously mentioned. Usually, the vertices-based features (point) are good for texture (high entropy), while edge-based features (lines, curves, axes, sketches) are good for cartoon (low entropy). As natural images are decomposed as texture and cartoon, the mixed graph-based representations are attained through the graph-based image segmentation and the associated scale invariant feature detector MSER and descriptor SIFT implications. To find distinctive feature matching throughout the over-complete space in a more stable approximation, PCA-SIFT and TPS warping models are adopted to reduce the dimension of SIFT descriptors and catch a more accurate interview motion model in 3D angle of view.

Graphs here can be attained through the effective image segmentation; meanwhile the points are produced by dimension reduction which has a low dimension of feature descriptor. Through such representations, the high-level feature aggregation $F(I_{wz})$ is supplemented so as to make a more distinctive constraint for FWZC. Since the subgraph-based matching method is to exploit the interview correlations, we focus on the generation of interview side information Y_v.

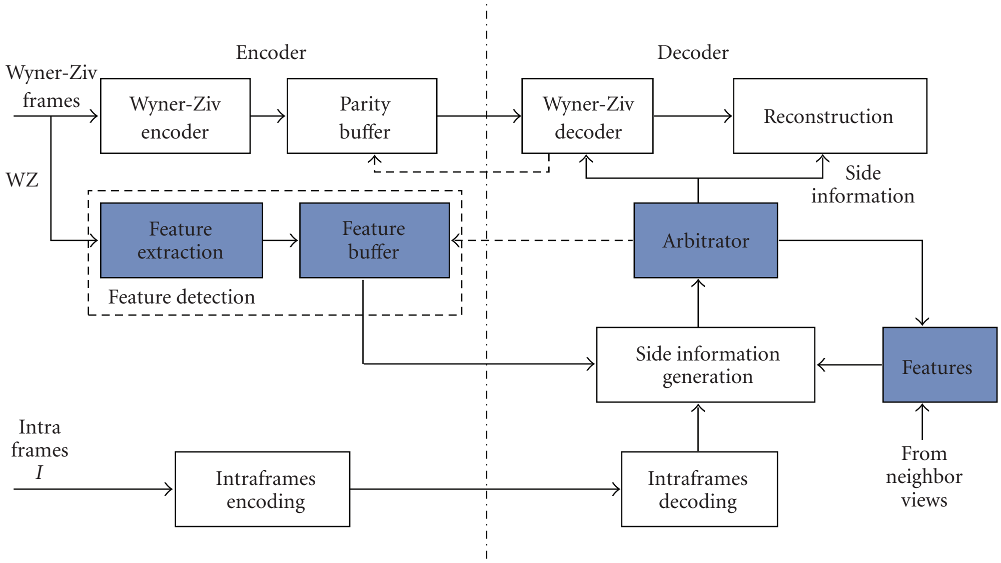


Figure 3: The block diagram of the codec of Feature-based WZC scheme.

Figure 4 illustrates the generic interview side information process at the decoder. Given the colocated left (right) view of the current WZ frame, I_v, the attribute graph is generally given by 3-tuple $G = (V, E, D)(\in F(I_{wz}))$, with V being a set of vertices consisting of distinctive feature points, E being groups of edges belonging to various subregions, and D, being descriptors for each $v_i \in V$ allowing for significant levels of local shape distortion and change in illumination. This colocated view feature information (V,E,D) can be determined at the decoder due to the availability of I_v. For WZ frame, I_{wz}, its attribute graph is given by $G' = (V', E', D')(\in F(I_{wz}))$. The vertices-based feature (V', D') is extracted at the encoder and will be transferred to the decoder; the edge-based feature (E') is determined at the decoder.

The goal of segmentation at the decoder in *Step 1* is to split each image into $n + 1$ regions that are likely to contain similar disparities that make a promising compensation for separated regions. A graph partition of G is denoted by

$$G = \{g_0, g_1, \dots, g_n\}. \quad (4)$$

Each subgraph has an amibute graph $g_i = (V_i, E_i, D_i)$. We can denote the $n + 1$ subgraphs matching functions by

$$\Psi_i : V_i \longrightarrow V_i' \cup \{\phi\}, \quad \text{given } \{D_i, D_i'\}, \quad i = 0, 1, \dots, n. \quad (5)$$

For each $v_j \in V_i$, $\Psi_i(v_j) \in V_i'$ or $\Psi_i(v_j) \in \phi$ indicating no match. Features are efficiently matched in *Step 2* by identifying the nearest neighbor keypoint that has the minimum Euclidean distance for the invariant descriptor vector based on dimension reduction. In this step, adjacent views correlations are exploited through graph segmentation and point features matching. Since we do this at the decoder where I_{wz} is not available, the feature information (V', D') should be extracted at the encoder in advance and transferred to the decoder. As the result of matching, the graph G' is also partitioned into $n + 1$ subgraphs $G' = \{g_0', g_1', \dots, g_n'\}$.

Now we have pairs of matched attribute graphs

$$(g_i, g_i') \quad i = 0, 1, \dots, n. \quad (6)$$

In *Step 3*, interview side information, Y_v^l and Y_v^r from the left and right views of the WZ frame are obtained by the geometric transform, TPS warping $F_i(x, y)$.

$$\begin{gathered} Y_{v_i}^l = F_i(g_i), \qquad Y_{v_i}^r = F_i(g_i), \\ Y_v^l = Y_{v_0}^l \cup Y_{v_1}^l \cup \cdots \cup Y_{v_n}^l, \\ Y_v^r = Y_{v_0}^r \cup Y_{v_1}^r \cup \cdots \cup Y_{v_n}^r. \end{gathered} \quad (7)$$

Finally, a view fusion method is applied to generate the interview side information Y_v in *Step 4*

$$Y_v = C\left(Y_v^l, Y_v^r\right). \quad (8)$$

The subgraph matching-based side information generation algorithm is summarized in Table 1.

3. Implementation Issues

3.1. Graph-Based Segmentation. Firstly, a rough segmentation of the correlated left (right) view image I_v is performed using the graph-based segmentation method [27]. Through blurring with a Gaussian filter, a segmentation consisting of a small number of large regions is obtained. All feature nodes $v_i \in V$ are divided into a small unknown number of $n + 1$ subgraphs for the graph matching in *Step 2* (refer to (4)). We assume that $n+1$ should be small and the subregions are large enough so that sufficient feature points for each $v_j \in V_i$ are contained to make the accurate matching. The first layer g_0 is always made of the background and small subregions whose

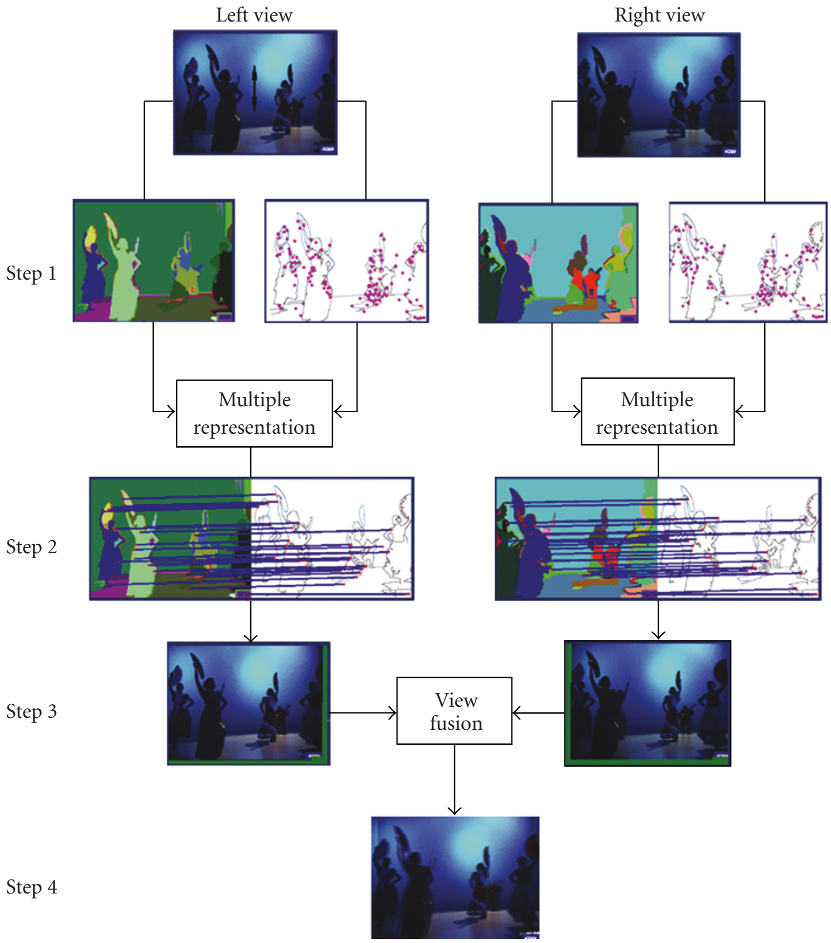


Figure 4: The subgraphs matching-based side information generation procedure: (1) *Step 1* applies segmentation and PCA-SIFT algorithms to obtain the multiple representations; (2) *Step 2* uses minimum Euclidean distance to find pairs of matching subgraphs; (3) left and right interview side information is generated by TPS warping; (4) view fusion methods are used to make the final accurate interview side information.

feature points are not sufficient. The segmentation results with different scale of subgraph matching-based semantic parameters are shown in Figure 5, where k sets a scale of observation when a larger value causes a preference for larger components, σ is the smoothing factor when 0 labels nonsmoothing.

3.2. Affine Invariant Features Extraction of Subgraphs. Many existing algorithms of image matching are difficult to handle the viewpoint change. Recently, local invariant features are shown to be robust for occlusion, background clutter, and content changes [28]. The definition is implicated by the observation that even though the regions themselves are covariant, the normalized image pattern they cover and the feature descriptors derived from them are typically invariant. Among all the popular scale invariant feature detectors and affine invariant feature detectors, maximally stable extremal regions (MSER) algorithm is evaluated to obtain the best results as shown in Figure 6. Also, scale-invariant feature transform (SIFT) algorithm is identified as the best descriptor which is most resistant to common image deformations [29]. Based on MSER and SIFT, a robust affine invariant features extraction is put forward to benefit the subsequent subgraph matching.

The SIFT algorithm was recently identified as the most resistant to image deformations and affine distortion between different views. In given graph representation $G = (V, E, D), V$ is a set of vertices consisting of such distinctive feature points, localized at local peaks in a scale-space search and stable over transformations; and D as descriptor represents the local image gradients in the feature point's neighborhood. At the decoder, features of the colocated left and right views of WZ frame can be extracted. At the encoder, WZ frames are extracted the associated features.

Principal component analysis (PCA) has been widely used in data analysis, and PCA-based SIFT introduces a more compact, distinctive, and accurate local descriptors [30]. It reduces the dimension of the descriptor through the following transform:

$$y = A_k(x - u_x), \quad (9)$$

Table 1: The subgraph matching-based side information generation procedure.

	Subgraphs Matching-based Algorithm
Step 1.	Use graph-based segmentation, subgraph-based scale invariant feature detector and descriptor, and dimension reduction algorithms to obtain the graph-based sparse data space;
Step 2.	Use minimum Euclidean distance to find pairs of matching subgraphs;
Step 3.	Generate left and right interview side information by TPS warping model;
Step 4.	Use View fusion methods to generate the final interview side information.

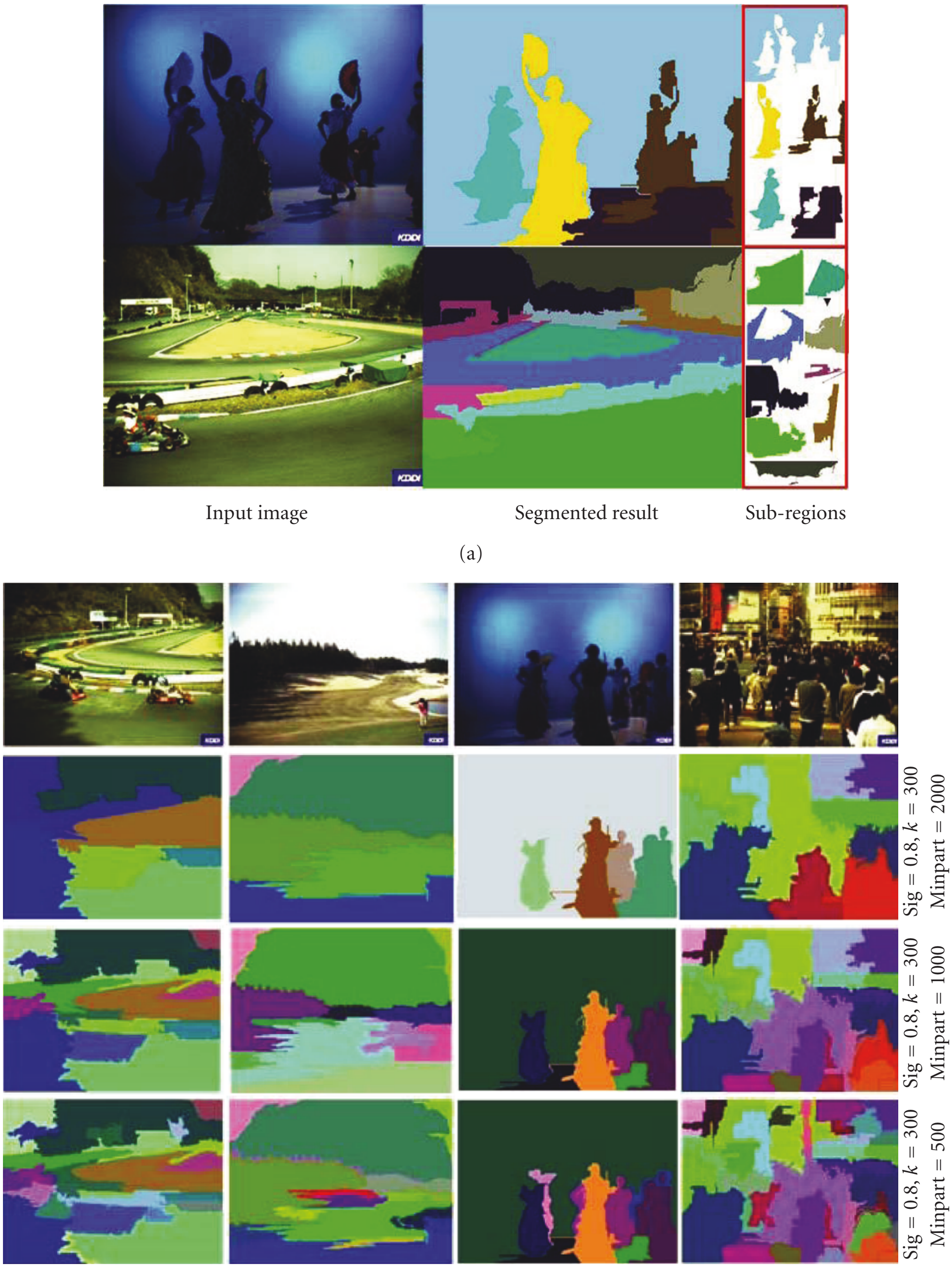
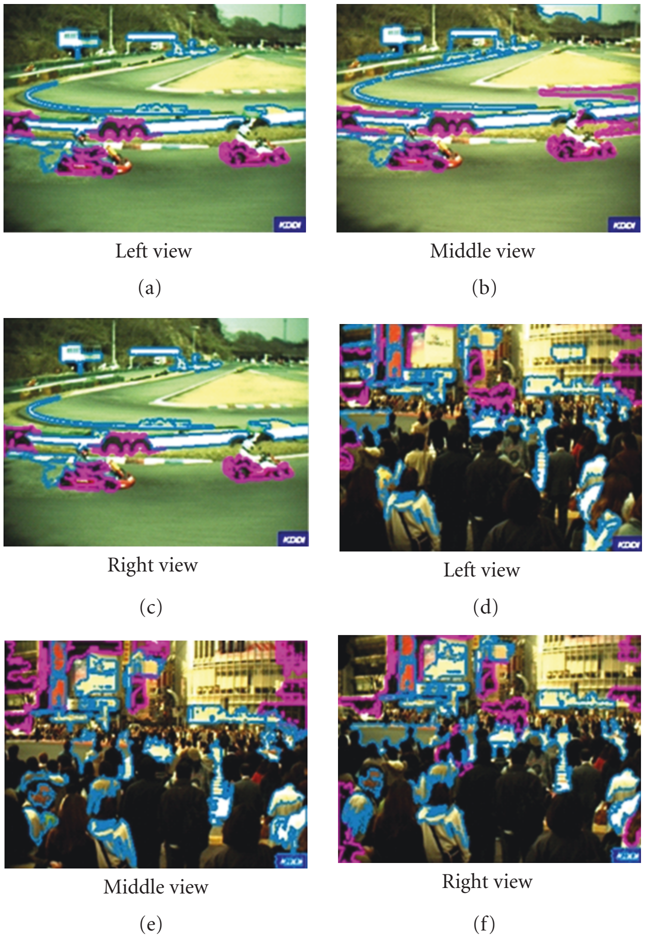


Figure 5: Subgraph-based image segmentation results with different scale of semantic parameters.


Left view

(a)

Middle view

(b)

Right view

(c)

Left view

(d)

Middle view

(e)

Right view

(f)

Figure 6: MSERs for the left, middle, and right views.

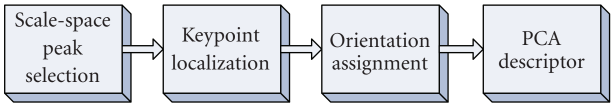


Figure 7: Four stages in PCA-SIFT.

Table 2: SIFT versus PCA-SIFT ($n = 20$).

Sequence	*RACE1*	*FLAMENCO*	*GOLF*
SIFT (dB)	30.51	25.08	26.32
PCA-SIFT (dB)	30.63	25.23	26.35

where x is normalized image gradient vector, the projection matrix, A_k presents the offline computed eigenspace, and y is the k dimension vector of PCA-SIFT descriptor. The local image patches surrounding each interest point are normalized so that their dominant orientation is in the same direction, which creates the redundancy that makes PCA effective. This normalized local gradient image patch is transformed into a 41 by 41 vector whose dot product is computed with the 20 prelearned PCA basis vectors. The dot product produces a signed 20-element integer vector which is the descriptor vector for that interest point.

Through four stages of PCA-SIFT in Figure 7, features are matched by identifying the nearest neighbor in the database which stores the candidate features extracted from the left and right views. The nearest neighbor is the keypoint which has the minimum Euclidean distance for the invariant descriptor vector. The solution of the feature matching problem is

$$\arg\min_{v'_k}\left\|v_j - v'_k\right\| \quad (k = 1, 2, \ldots, l). \tag{10}$$

For each feature point $v_j \in V_i$, $v_k \in V'_i$, and l is the number of feature points in V'_i. Having found pairs of matched feature points $\{v_j, \hat{v}'_j\}$: $\hat{v}'_j \in V'_i$ *or* $\hat{v}'_j \in \phi$, (5) can be determined for each subgraph.

Table 2 chooses three Multiview video sequences “Flamenco1”, “Race1”, and “Golf”, to compare the prediction results between SIFT and PCA-SIFT. In this paper, the dimension of PCA-SIFT feature space n is set to 20. According to the analysis in [30], this setup achieves a good trade-off between matching accuracy and feature space dimension. In the following of this paper, all the experiments are performed with $n = 20$. Both SIFT and PCA-SIFT use

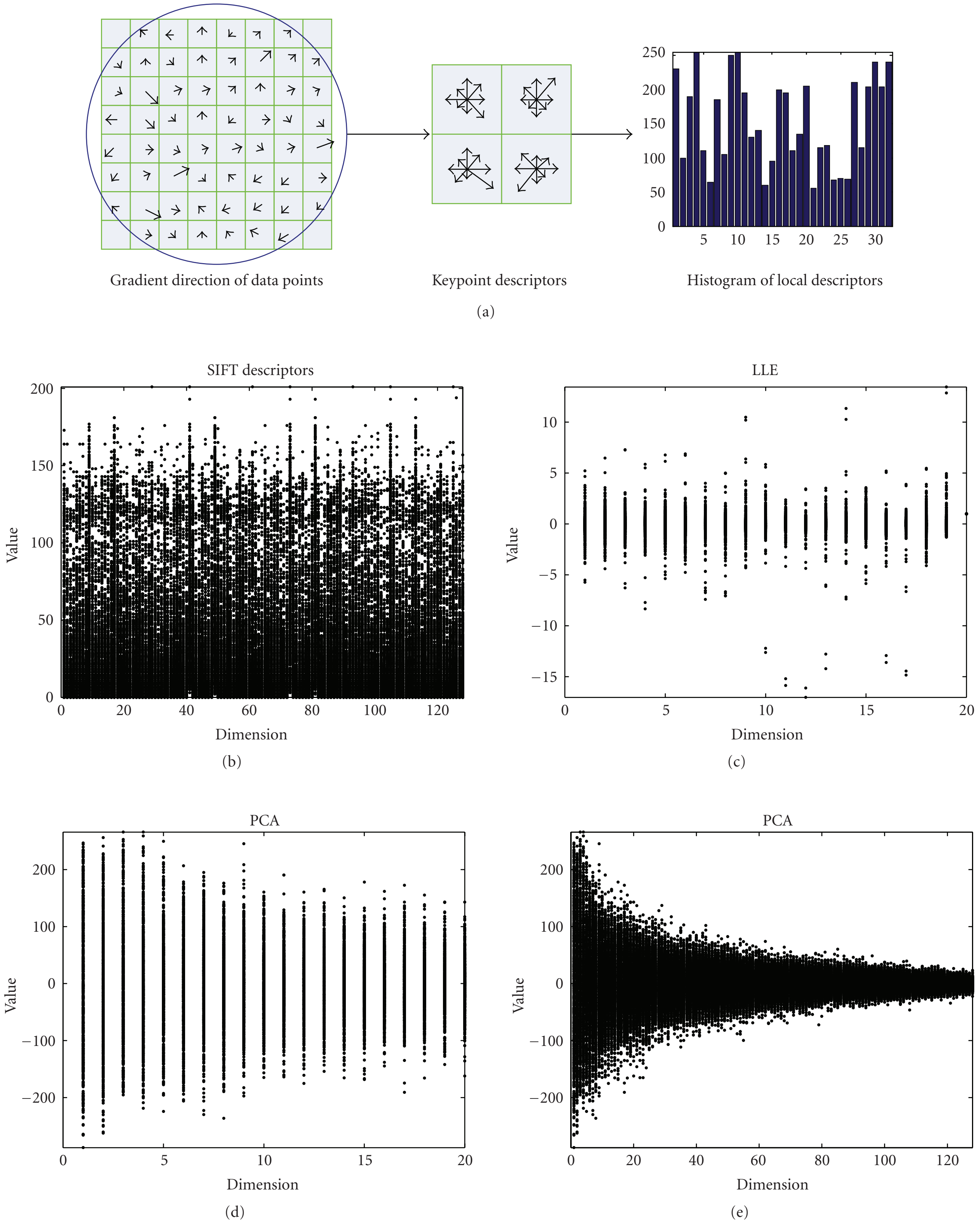
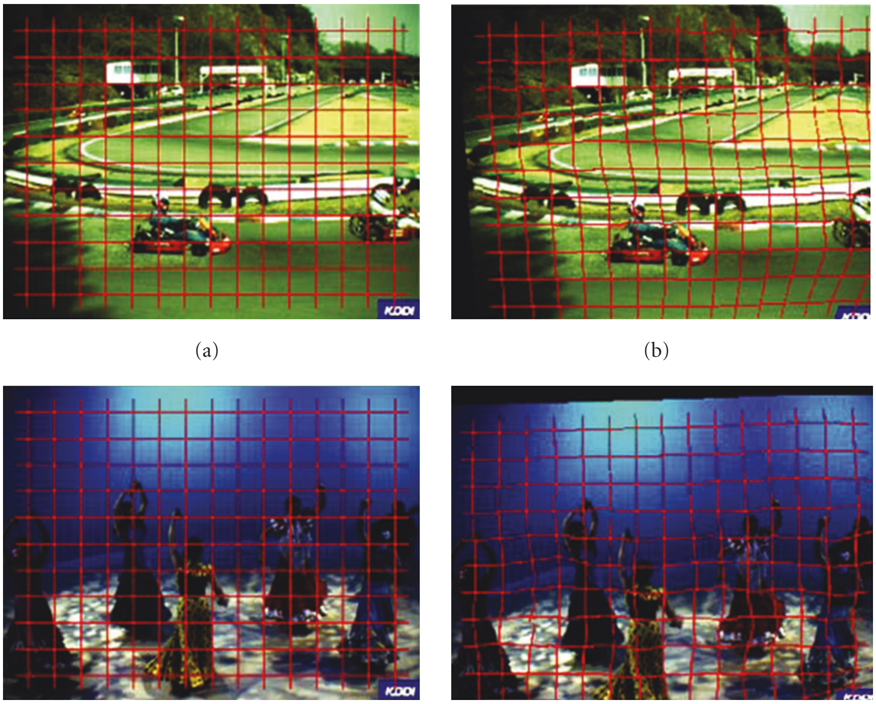


FIGURE 8: SIFT descriptors with 128 dimensions; and dimension reduction with LLE versus PCA.

Table 3: PSNR gain of Proposed Graph-based versus other three schemes.

Proposed Graph-based versus (dB)	RACE1	FLAMENCO1	FLAMENCO2	Golf
Temporal ME	1.5∼2.5	1.0	1.0∼1.2	0.5∼0.8
Feature-based	0.3∼0.5	0.4∼0.5	0.5∼0.6	0.3
264 I frames	6.5∼9	2.5∼4.5	2.5∼4	4.8∼6



(a)

(b)

(c)

(d)

Figure 9: Frame warping using TPS.

the same 6-parameter affine transform prediction model, and equal number of matching keypoints (9 pairs of features are selected for each WZ frame) to generate interview side information. From the average PSNR of the estimated frames, we can see that PCA-SIFT's matching accuracy at the keypoint level translates into good performance.

As a nonlinear dimensionality reduction, locally linear embedding (LLE) succeeds in identifying the underlying structure of the manifold and does not involve local minima. Its procedure can be described as follows:

(1) compute the neighbors of each data point;

(2) compute the weights that best reconstruct each data point from its neighbors, minimizing the cost by constrained linear fits: $\varepsilon(W) = \sum_i |\overline{X}_i - \sum_j W_{i,j}\overline{X}_j|^2$;

(3) compute the vectors best reconstructed by the weights minimizing the quadratic form by its bottom nonzero eigenvectors: $\phi(R) = \sum_i |\overline{R}_i - \sum_j W_{i,j}\overline{R}_j|^2$,

where $\overline{X}$ is the 128-dimension SIFT descriptors, and $\overline{R}$ represents the reduction data.

Figure 8 shows an illustrative comparison of a sampled frame by linear (PCA) and nonlinear LLE.

3.3. Thin Plate Spline (TPS). Thin-Plate Spline warps have been shown to be very effective as a parameterized model of the optic flow field between images of various deforming surfaces. The close-form minimizer of TPS is parameterized by a global affine matrix **d** and a local warping coefficient matrix **c**. Giving K pairs of matched feature points $\langle v_j, \hat{v}'_j \rangle$ for each subregion extracted from (10) as control points, the spatial interpolation function can be written for each subregion:

$$F_i(z, \mathbf{d}, \mathbf{c}) = z \cdot \mathbf{d} + \sum_{j=1}^{K} \phi\left(\left\| z \quad - v_j \right\|\right) \cdot c_j, \qquad (11)$$

where $v_j \in V_i$, **d** is a 3×3 matrix as affine transform, and **c** is a $K \times 3$ matrix as the nonaffine deformation. The kernel function $\phi(\|z - v_j\|)$ is a $1 \times K$ vector for each point z, where each entry $\phi_i(z) = (\|z - v_i\|^2 \log \|z - v_i\|)$ for 2D coordinates. The solution of optimum $\{\mathbf{d}, \mathbf{c}\}$ could be

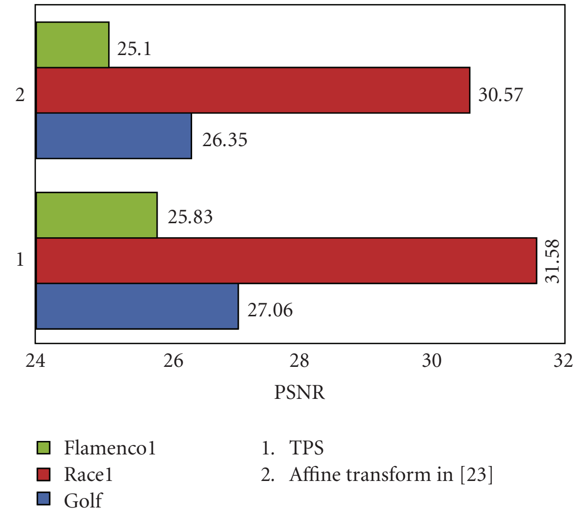


Figure 10: The PSNR of the estimated frames in the view direction; "1" denotes the proposed scheme involving TPS in combination with subgraph matching; "2" represents the scheme using global affine transform for interview side information generation [23].

attained by Tikhonov regularization minimizing the energy function. Therefore, the interview side information can be retrieved by (8). Figure 9 displays this warping transform by using TPS.

To evaluate TPS's effectiveness in the proposed approach, WZ frames of three Multiview video sequences are estimated from the view direction by 6-parameters global affine transform and TPS warping of the graph-based matching. The number of features from the encoder to assist the affine transform and TPS warping is 9 and 50, respectively, so that the overheads for the features are nearly equal due to the low-dimension descriptor of PCA-SIFT. Figure 10 shows the average PSNR of the estimated frames (luminance). It can be observed that TPS warping with graph matching works better than the global affine transform especially for the sequence with high motion, for example, RACE1. Figure 11 illustrates the individual frame's PSNR of affine transform and TPS warping for the second view of the "RACE1" and "GOLF" Multiview video sequences. It is shown that each subgraph extracted from the original Wyner-Ziv target image is more accurately estimated with the proposed scheme.

3.4. Side Information Fusion

3.4.1. Temporal Side Information. Temporal side information, Y_t, is predicted from the temporal direction according to the algorithm in [31]. As displayed in Figures 12(a) and 12(b), forward motion estimation is applied to get the candidate motion vectors for each nonoverlapped block in the interpolation frame I_{wz}. From the available candidate vectors, the motion vector that intercepts the interpolated frame closer to the center of block is under consideration. Now that each block in the interpolated image has a motion vector, bidirectional motion compensation is performed to obtain the interpolated frame.

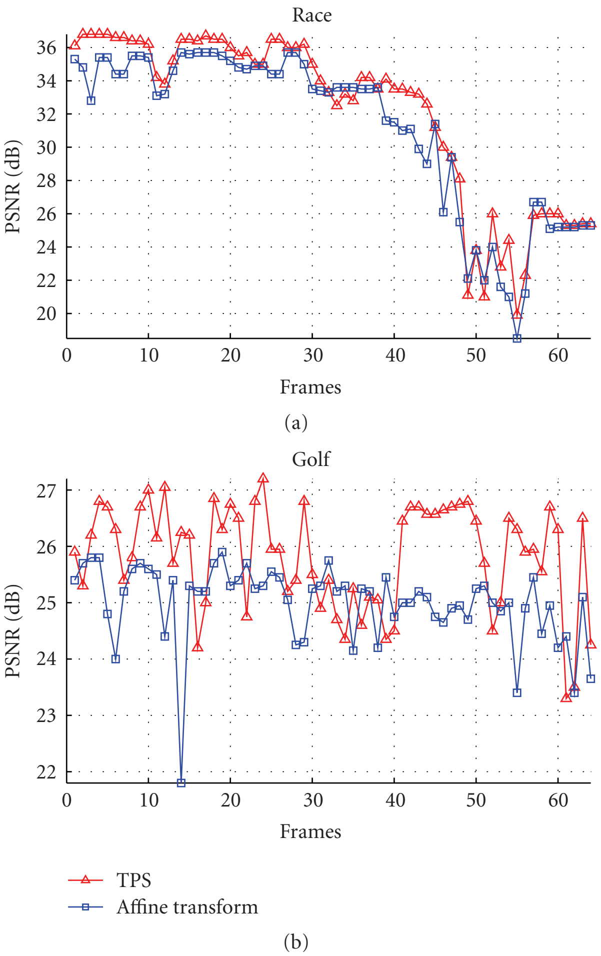


Figure 11: The PSNR (luminance) comparison of TPS and affine transform for (a) Race1 and (b) Golf Multiview sequences.

3.4.2. Side Information Fusion. Interview side information Y_v is obtained by the combination of TPS warping results from the left and right views by (8) (simply average the two results). Figure 13 shows the side information frames generated from the temporal direction and the view direction. Temporal method works well for the prediction of objects with small motion, while the view prediction with proposed scheme, subgraphs-based matching, has advantages for those with high motion and can significantly reduce the effect of ghost compared to [9] (the small icon "KDDI" is erased in advance and later added when the interview fusion is finished; the blank area is filled with the average of adjacent two frames in the temporal direction). Furthermore, an inherent data fusion algorithm to reconstruct more accurate side information from temporal and view side should be applied. The final side information Y is generated by

$$Y = C(Y_t, Y_v). \tag{12}$$

We generate a fusion mask for Y, where 1 indicates that the pixel is taken from the interview side information, and 0 the pixel taken from the temporal side information. In this work, we adopt the intensity of MVs as criteria to measure

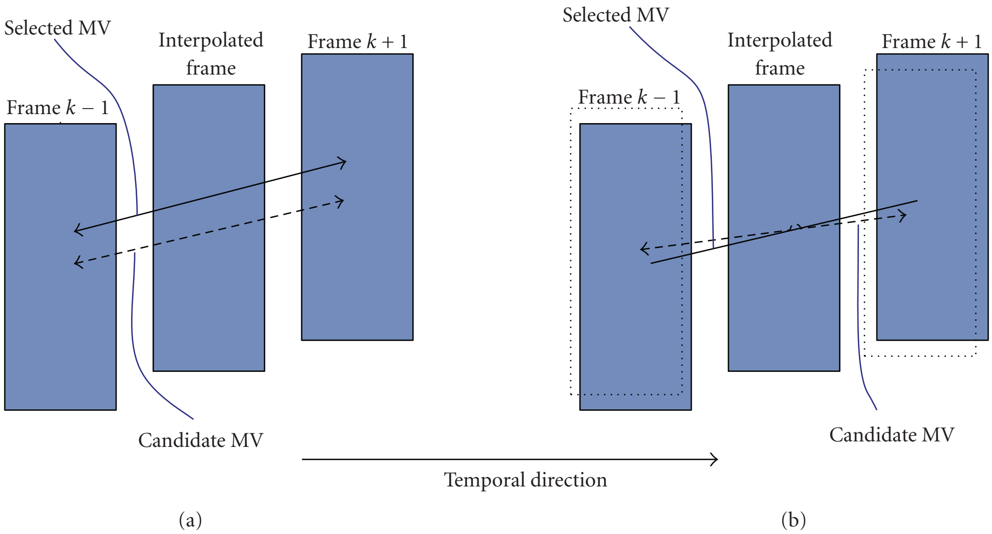
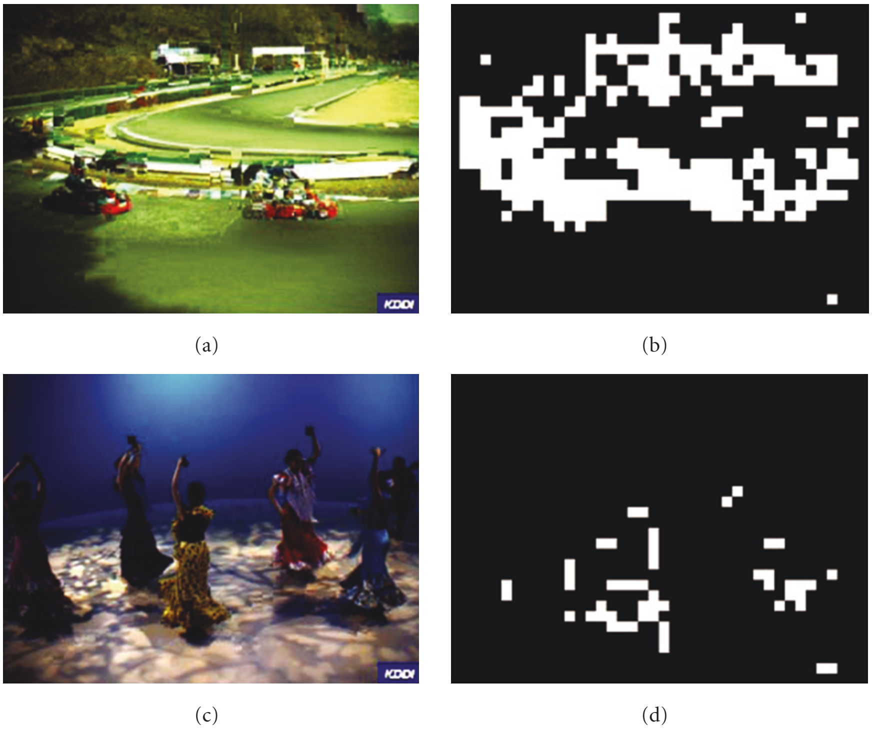


Figure 12: (a) Selection of the motion vector; (b) Bidirectional motion estimation.


(a)

(b)

(c)

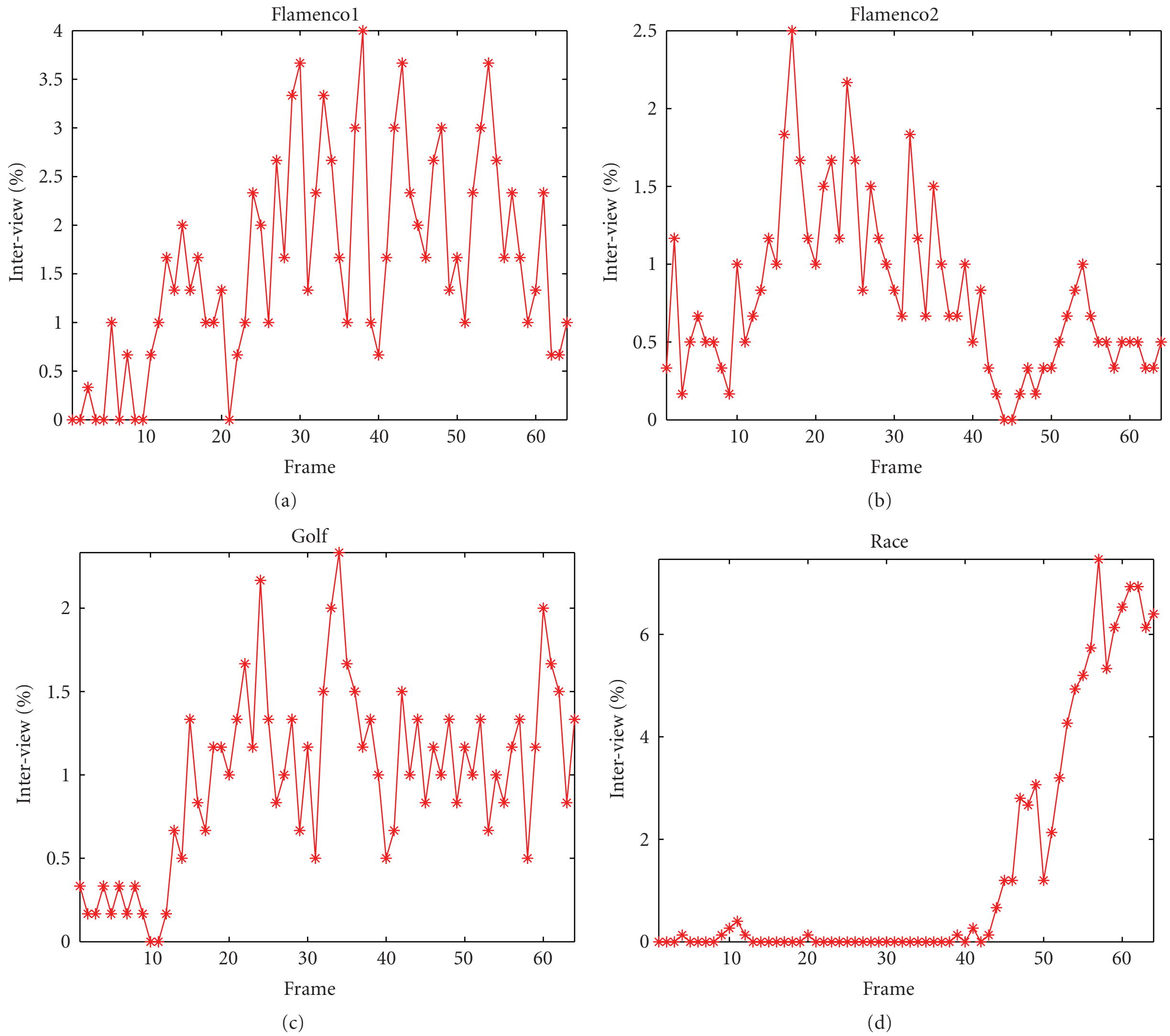
(d)

Figure 13: The fusion masks, white areas indicate that temporal side information is unreliable and interview side information is used.

the reliability of interview side information and temporal side information [32]. Since temporal motion estimation performs poorly in regions where motion is high, it is obvious that the motion vectors from temporal estimation can be used as criteria for fusion. The abrupt changes of the direction of the motion vectors, which have low spatial coherence, are considered to be incorrect motion vectors when compared to the true motion field. They can be detected by the weighted vector median filters, extensively used for noise removal. Through these abrupt motion vectors, we set the corresponding block to ones in fusion masks. Figure 13 shows an example of the fusion mask, where white areas indicate temporal estimation unreliable and thus enable interview estimation.

From Figure 13, it can be seen that the temporal side information shown on the left side has a bad estimation in areas with high motion so that these areas should be determined by the interview side information.

Figure 14 shows the percentage of MBs from interview side information in each frame of some sequences.



(a)

(b)

(c)

(d)

Figure 14: The percentage of interview MBs in each frame.

It demonstrates that interview side information contributes very few in frames with low motion, for example, the first 40 frames of "Race" sequence. Obviously, the interview side information is more helpful to improve the quality of fused side information for frames with intensive motion.

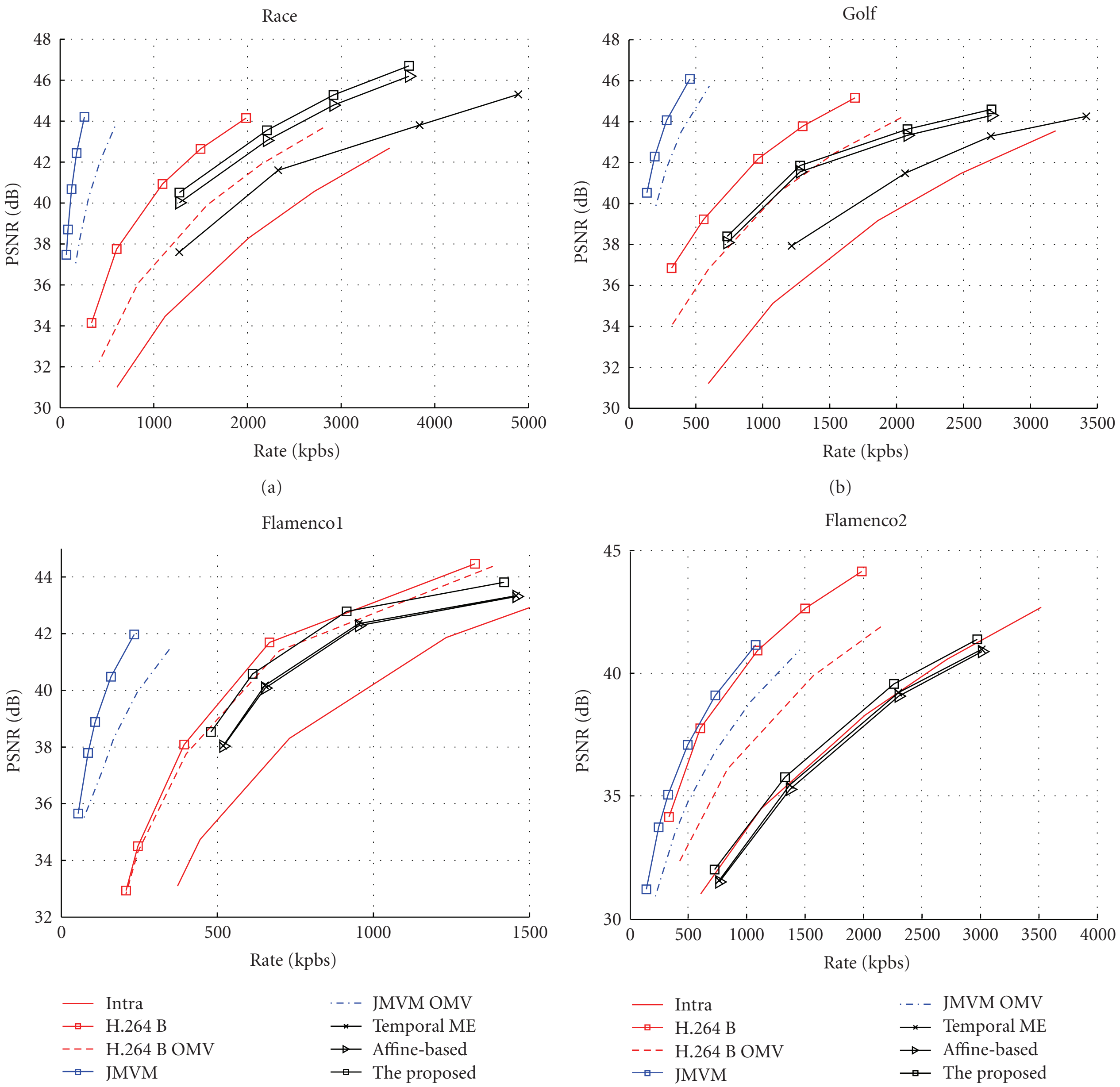
4. Experimental Results

We have tested the proposed scheme on Multiview sequences from KDDI Lab, where three views with 128 frames (320 × 240) of each sequence are chosen. DMVC's structure is IWIW in an interlaced way. The Wyner-Ziv frame rate is 15 f/s. It is assumed that the I frames are available at the decoder perfectly reconstructed. Each Wyner-Ziv frame is predicted from both the temporal direction by interpolation solution presented in Section 3.4.1 and view direction with constrained relaxation of subgraphs matching described in Section 3.1–3.3. The LDPC adopted in our DVC scheme has block length with $L = 6336$ bits and its source rate is 2/66,3/66,4/66,...,66/66 [33]. The source node degree distribution is irregular with

$$\lambda(x) = 0.316x^1 + 0.415x^2 + 0.128x^6 + 0.069x^7 + 0.020x^{18} + 0.052x^{20}. \tag{13}$$

The parameters of graph-based image segmentation are manually set in this paper, where the observation scale k is set to 300, smoothing factor σ is 0.8, and the minimum subregion size is 1000. The PCA method is adopted in experiments to reduce the dimension of deduced SIFT descriptors. The number of PCA-SIFT based features transmitted upon request is around 50 with the dimensionality of the feature space $n = 20$. The projection matrix used in the PCA-SIFT is precomputed once and stored.

Figure 15 gives the R-D curves of four Multiview video sequences from eight MVC coding methods. These methods are grouped into three categories: the JMVM coding method, the H.264/AVC coding method, and the DMVC method.



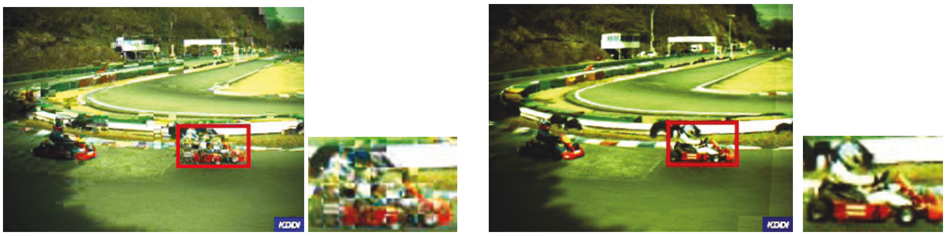
(a)

(b)

(c)

(d)

Figure 15: R-D curves of "Race1", "Flamenco1", "Golf", and "Flamenco2" Multiview sequences.



(a)

(b)

Figure 16: Side information generation from (a) the temporal direction, (b) from the view direction with the graph-based matching.

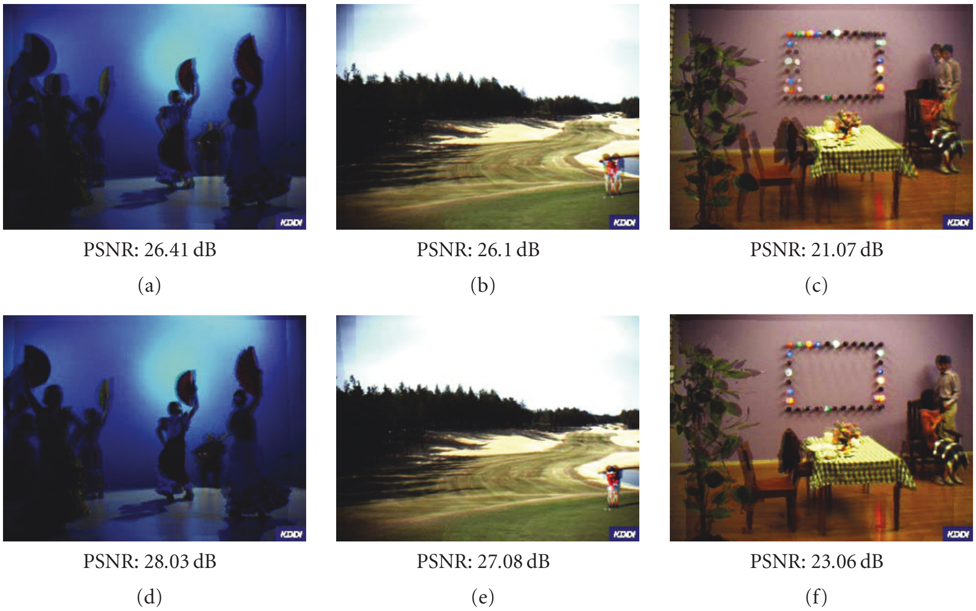
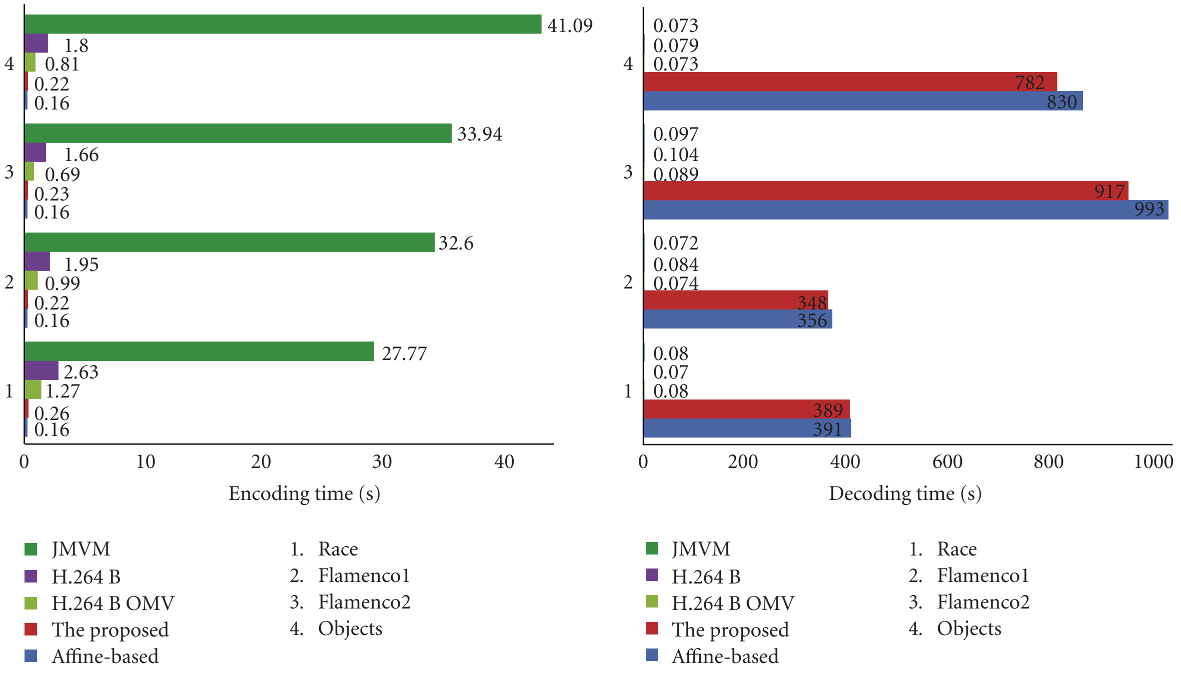


Figure 17: (a, b, and c) interview side information generated by 6-parameters global affine transform. (d, e, and f) interview side information generated by TPS warping of the graph-based matching.



(a) Encoding

(b) Decoding

Figure 18: The average coding time comparison for different coding paradigms.

The “Intra” in Figure 15 is the result of Intracoding with H.264/AVC [34]. And “H.264 B” stands for result of H.264/AVC with motion search open. The frame structure is “I-B-I-B-I-· · ·”, the search range is 32 and reference frame number is set to 2. The configuration of “H.264 B 0MV” is similar with that of “H.264 B” except that the bidirectional motion search is closed. The “JMVM” stands for the result of JMVM with motion search on, where the GOP is set to 15, the search range is set to 96, the max iterations for bi-directional search is 4, and the search range for iteration is set to 8. The setup of “JMVM 0MV” is similar to that of “JMVM” just with motion search turned off. The rest three methods are DMVC method based on Wyner-Ziv coding. There, frame structure is "I-B-I-B-I-· · · " as shown in Figure 2(b). The difference between these three DMVC approaches is their different side information generation method. "Temporal ME" generates side information bidirectional motion compensation where the search range is 16. The "Affine-based" method [23] generates interview side information with affine transform, and fuses that with the results of "Temporal ME" to produce the final side information for Wyner-Ziv decoding. And "The proposed" stands for the result of the proposed subgraph-based method.

The results in Figure 15 show that the proposed graph-based DMVC approach outperforms H.264/AVC-based intra coding up to 4-5 dB, DMVC with temporal prediction about 0.5–1.5 dB, and the "Affine-based" DMVC scheme about 0.3–0.5 dB. In terms of motion classification in four video sequences, it is supposed that the proposed DMVC scheme has quite high precision for objects with high motion so as to bring a significant improvement in a rate-distortion sense.

More results of various texture images selected from related sequences are presented in Figure 16. Figure 17 shows the side information frames generated from the temporal direction and the view direction. It can be seen that each subgraph extracted from the original Wyner-Ziv target image is more accurately estimated with the proposed approach. In the contours of separate regions of same images, it has significantly reduced the ghost effect and attained better PSNR values.

To analyze the additional computations for feature extraction, we performed simulations of DMVC schemes with and without local feature extraction process, and compared their encoding complexity with existing typical coding schemes, for example, "JMVM", "H.264/AVC B", and "H.264/AVC B 0MV" without bidirectional motion search. It is worth mentioning that the feature extraction processing in the proposed DMVC scheme would give a hint for interview side information generation, which could be regarded as motion-compensated prediction in H.264 or joint Multiview video coding. It is performed on Windows XP SP3 system with Intel Core 2 CPU 1.86 GHz and 2.00 GB memory.

Figure 18 presents the average encoding/decoding time of a WZ frame, a B-frame of H.264/AVC, or a B frame of JMVM coding paradigm. The experimental results in Figure 18(a) demonstrate that although noticeable additional computations are introduced by the feature extraction process, the total coding complexity of the proposed scheme is still significantly lower than conventional prediction schemes. In fact, the computational complexity has been mainly transferred to the decoder in distributed video coding sense. Generally, the total decoding of Wyner-Ziv coding might typically take hundreds to thousands of seconds for the WZ frames, as shown in Figure 18(b), which is far beyond the additional computation burden of feature matching at the decoder side.

The communication overhead, induced by the local feature descriptors, has been taken into account of the overall bit-rate in Figure 15. It has demonstrated a superior rate-distortion performance of the proposed scheme compared with a variety of existing schemes. In fact, the communication overhead is correlated with a source. For a sequence with smooth motion, the temporal side information is relatively high and there would be few frames to transmit features to the decoder, such as "Flamenco 2". For example, the communication overhead for "Race" and "Flamenco 2" is about 28 kbps and 11 kbps on the average.

5. Conclusion

This paper proposes a novel graph matching-based FWZC scheme for DMVC. It devotes graph-based representations of Multiview images to generate interview correlated side information without knowing the camera parameters. The sparse feature set as a good hypothesis space aims for a best matching optimization of interview side information with compact syndromes, from inferred relaxed coset. The plausible filling-in from a priori feature constraints between neighboring views could reinforce a promising compensation to interview side information generation for joint Multiview decoding. The graph-based representations of Multiview images are adopted as constrained relaxation, which assists the interview correlation matching for subgraph semantics of the original Wyner-Ziv image by the graph-based image segmentation and the associated scale invariant feature detector MSER and descriptor SIFT. In order to find distinctive feature matching with a more stable approximation, linear and nonlinear projections are adopted to reduce the dimension of high-dimensional SIFT descriptors, and TPS warping model is to catch a more accurate interview motion model in 3D angle of view.

Acknowledgments

The authors are grateful to anonymous reviewers for their valuable comments. Their thorough technical and editorial comments significantly improved the quality of this paper. The work was supported in part under NSFC Grants no. 60632040, no. 60772099, no. 60928003, no. 60736043, and the National High Technology Research and Development Program of China (863 Program) (no. 2006AA01Z322).

References

[1] Y. Ho, S. Yoon, and S.-Y. Kim, "A framework for multi-view video coding using layered depth images," ISO/IEC JTC1/SC29/WG11 M11582, Hong Kong, January 2005.

[2] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[3] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, 1976.

[4] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, and J. Ostermann, "Distributed monoview and multiview video coding: basics, problems and recent advances," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 67–76, 2007.

[5] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1181–1203, 2003.

[6] R. Zamir, “The rate loss in the Wyner-Ziv problem,” *IEEE Transactions on Information Theory*, vol. 42, no. 6, part 2, pp. 2073–2084, 1996.

[7] R. Puri and K. Ramchandran, “PRISM: a new robust video coding architecture based on distributed compression principles,” in *Proceedings of the 40th Annual Allerton Conference on Communication, Control, and Computing*, Allerton, Ill, USA, October 2002.

[8] 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, 2003.

[9] A. Aaron, S. Rane, E. Setton, and B. Girod, “Transform-domain Wyner-Ziv codec for video,” in *Visual Communications and Image Processing 2004*, vol. 5308 of *Proceedings of SPIE*, pp. 520–528, San Jose, Calif, USA, January 2004.

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

[11] C. Brites, J. Ascenso, J. Q. Pedro, and F. Pereira, “Evaluating a feedback channel based transform domain Wyner-Ziv video codec,” *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 269–297, 2008.

[12] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, “The Discover codec: architecture, techniques and evaluation,” in *Proceedings of the Picture Coding Symposium (PCS ’07)*, Lisbon, Portugal, November 2007.

[13] R. Martins, C. Brites, J. Ascenso, and F. Pereira, “Refining side information for improved transform domain Wyner-Ziv video coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 9, pp. 1327–1341, 2009.

[14] G. Ziegler, H. P. A. Lensch, N. Ahmed, M. Magnor, and H.-P. Seidel, “Multi-video compression in texture space,” in *Proceedings of the International Conference on Image Processing (ICIP ’04)*, vol. 4, pp. 2467–2470, Singapore, October 2004.

[15] M. Magnor and B. Girod, “Model-based coding of multiviewpoint imagery,” in *Visual Communication and Image Processing*, vol. 1 of *Proceedings of SPIE*, pp. 14–22, Perth, Australia, June 2000.

[16] M. Droese, C. Clemens, and T. Sikora, “Extending single-view scalable video coding to multi-view based on H.264/MPEG-4 AVC,” in *Proceedings of IEEE International Conference on Image Processing (ICIP ’06)*, pp. 2977–2980, Atlanta, Ga, USA, October 2006.

[17] K. Müller, A. Smolic, M. Droese, P. Voigt, and T. Wiegand, “Multi-texture modelling of 3D traffic scenes,” in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME ’03)*, Baltimore, Md, USA, July 2003.

[18] N. Ozbek and A. M. Tekalp, “Scalable multi-view video coding for interactive 3DTV,” in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME ’06)*, pp. 213–216, Toronto, Canada, July 2006.

[19] M. Tanimoto, T. Fujii, K. Yamamoto, et al., “Response to call for proposals on multi-view video coding: overview and coding tool,” ISO/IEC JTC1/SC29/WG11 MPEG2006/m12945, Bangkok, Thailand, January 2006.

[20] X. Zhu, A. Aeron, and B. Girod, “Distributed compression for large camera arrays,” in *Proceedings of IEEE Workshop on Statistical Signal Processing (SSP ’03)*, St Louis, Mo, USA, September 2003.

[21] X. Artigas, E. Angeli, and L. Torres, “Side information generation for multiview distributed video coding using a fusion approach,” in *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG ’06)*, pp. 250–253, Reykjavik, Iceland, June 2006.

[22] R.-S. Wang and Y. Wang, “Multiview video sequence analysis, compression, and virtual viewpoint synthesis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 397–410, 2000.

[23] 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, 2008.

[24] H. Lv, H. Xiong, Y. Zhang, and Z. He, “Side information generation with constrained relaxation for distributed multiview video coding,” in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS ’08)*, pp. 3450–3453, Seattle, Wash, USA, May 2008.

[25] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[26] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proceedings of the ACM Conference on Computer Graphics (SIGGRAPH ’00)*, pp. 417–424, New Orleans, La, USA, July 2000.

[27] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[28] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[29] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] Y. Ke and R. Sukthankar, “PCA-SIFT: a more distinctive representation for local image descriptors,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’04)*, vol. 2, pp. 506–513, July 2004.

[31] J. Ascenso, C. Brites, and F. Pereira, “Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding,” in *Proceedings of the 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, Slovakia, July 2005.

[32] M. Ouaret, F. Dufaux, and T. Ebrahimi, “Fusion-based multiview distributed video coding,” in *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 139–144, Santa Barbara, Calif, USA, October 2006.

[33] D. Varodayan, A. Aaron, and B. Girod, “Rate-adaptive distributed source coding using low-density parity-check codes,” in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, pp. 1203–1207, Pacific Grove, Calif, USA, November 2005.

[34] JVT, “Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14 496-10 AVC)—Joint Committee Draft,” Joint Video Team (JVT) of ISO/IEC MPEG and ITU-TVCEG, document JVT-G050r1.doc, 2003.