An Epipolar Geometry-Based Fast Disparity Estimation Algorithm for Multiview Image and Video Coding

Jiangbo Lu, *Student Member, IEEE*, Hua Cai, *Member, IEEE*, Jian-Guang Lou, *Member, IEEE*, and Jiang Li, *Senior Member, IEEE*

***Abstract*—Effectively coding multiview visual content is an indispensable research topic because multiview image and video that provide greatly enhanced viewing experiences often contain huge amounts of data. Generally, conventional hybrid predictive-coding methodologies are adopted to address the compression by exploiting the temporal and interviewpoint redundancy existing in a multiview image or video sequences. However, their key yet time-consuming component, motion estimation (ME), is usually not efficient in interviewpoint prediction or disparity estimation (DE), because interviewpoint disparity is completely different from temporal motion existing in the conventional video. Targeting a generic fast DE framework for interviewpoint prediction, we propose a novel DE technique in this paper to accelerate the disparity search by employing epipolar geometry. Theoretical analysis, optimal disparity vector distribution histograms, and experimental results show that the proposed epipolar geometry-based DE can greatly reduce search region and effectively track large and irregular disparity, which is typical in convergent multiview camera setups. Compared with the existing state-of-the-art fast ME approaches, our proposed DE can obtain a similar coding efficiency while achieving a significant speedup for interviewpoint prediction and coding. Moreover, a robustness study shows that the proposed DE algorithm is insensitive to the epipolar geometry estimation noise. Hence, its wide application for multiview image and video coding is promising.**

***Index Terms*—Disparity estimation (DE), epipolar geometry, fast motion estimation (ME), H.264/AVC, multiview image, multiview image compression, multiview video, multiview video compression, video coding.**

I. Introduction

MULTIVIEW video or free-viewpoint video is an exciting application, because it enables users to watch a static or dynamic scene from different viewing angles. As a brand new application, multiview video has recently received increasing attention. Generally, to provide a smooth multiperspective viewing experience, content producers need to capture a distinct scene with ideal quality from multiple camera positions, such as the convergent multiview camera setup shown in Fig. 1, where the cameras are positioned inwards to capture the scene from different angles. Usually, the simultaneous multiple video streams from multiview cameras are referred to as multiview video. A multiview video sequence can be naturally regarded as a temporal sequence of special-visual-effect snapshots, captured from different viewpoints at multiple times. Such a special snapshot is comprised of all of the still images taken by multiple cameras at one certain time instance, so it is essentially a multiview image set or a *frozen moment* sequence [1]. On the other hand, a multiview image is a degenerated form of multiview video in which the temporal axis has completely shrunk to zero, but its production does not require multiple cameras all the time. For instance, a multiview image set can be generated by steadily moving a single video camera around a scene of interest along a predefined capture trajectory. Fig. 2 depicts the concepts of multiview video, multiview image set, and their relationship. As a typical example, Fig. 3 shows the snapshots at the same time instance from two high-quality multiview video sequences, *Breakdancer* and *Ballet*, which are captured by arranging eight cameras along a one-dimensional (1-D) arc spanning about 30° from one end to the other [2].

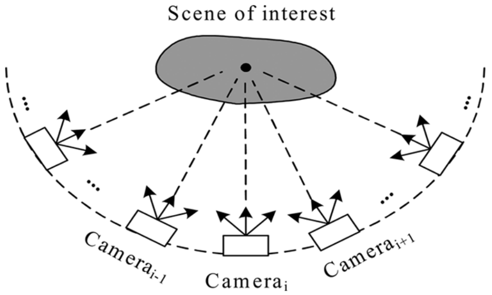


Fig. 1. Convergent multiview camera setup.

Though multiview image/video is capable of providing the exciting viewing experience, it is achieved at the expense of large storage and transmission bandwidth. To overcome these problems, a specially designed multiview image/video encoder becomes an indispensable necessity. For the same reason, multiview video coding is identified by the MPEG 3-D audio and video (3DAV) ad hoc group as one of the most challenging problems associated with such new applications as free-viewpoint video [3], [4]. In this paper, we center our study on effectively coding multiview-captured content. More specifically, we are especially interested in coding the multiview

Manuscript received April 12, 2006; revised December 4, 2006. The work presented in this paper was carried out at Microsoft Research Asia, Beijing, China.

J. Lu is with the Department of Electrical Engineering, University of Leuven and Multimedia Group, IMEC, Leuven 3001, Belgium (e-mail: jiangbo@imec.be).

H. Cai, J.-G. Lou, and J. Li are with the Media Communication Group, Microsoft Research Asia, Beijing 100080, China (e-mail: huacai@microsoft.com; jlou@microsoft.com; jiangli@microsoft.com).

Digital Object Identifier 10.1109/TCSVT.2007.896659

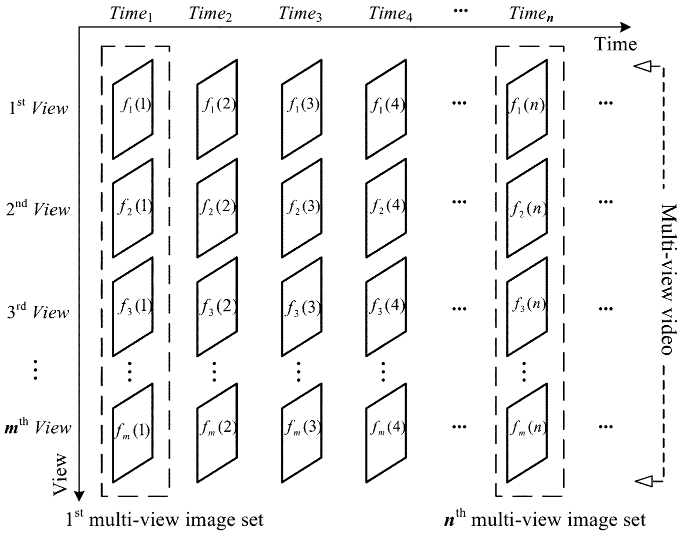


Fig. 2. Multiview video and multiview image set. A multiview video sequence consists of m conventional video streams captured from m viewpoints ($f_i(j)$ denotes the jth frame of the ith view), while a typical multiview image set can be constructed by all of the still images from m viewpoints at any particular time instant, e.g., the first multiview image set and the nth one.

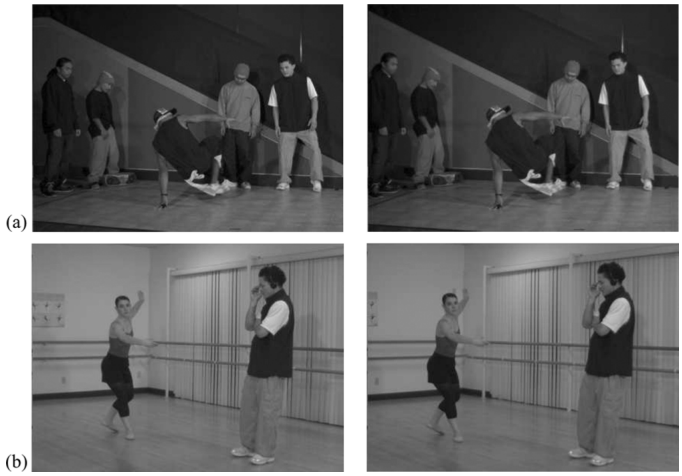


Fig. 3. Two adjacent multiview images captured by arranging cameras on an arc. (a) *Breakdancer*. (b) *Ballet*.

image and video captured by convergent multiview camera setups as shown in Fig. 1, because such a setup usually finds a wide application in movies, advertising, educational video (such as surgical instructions), sports events, and event general broadcasting. Meanwhile, the multiview content captured from convergent viewpoints is more difficult for a multiview image or video encoder than that captured from a parallel multiview camera setup, which can be regarded as a simplified form of the convergent setup once the angular difference between two adjacent cameras' viewing directions is decreased to zero. The reason is that the disparity intensity and disparity directions for the foreground objects and background scenes are often much more intensive and heterogeneous across the multiview frames captured using a convergent camera setup.

Although multiview video coding is still an active ongoing activity in the MPEG 3DAV ad hoc group, several competitive and promising coding techniques based on the H.264/AVC framework [5] have already been proposed. The investigation [6] clearly shows that there are technologies that significantly outperform today's available reference method (AVC simulcast). In addition to exploiting temporal redundancy to achieve coding gains, interviewpoint redundancy is also exploited in these schemes [6] by performing interviewpoint prediction across different views. The interviewpoint prediction is also an indispensable coding tool for multiview image coding, in which the multiview image set consists of the still images captured from different viewing directions.

Though the interviewpoint prediction can greatly improve the coding performance of a multiview image or video encoders, it also significantly increases computational costs. This is because interviewpoint redundancy is typically exploited by conducting interviewpoint disparity estimation (DE) across different views based on motion estimation (ME) approaches, while ME is usually the most time-consuming component in a conventional video encoder, especially when the variable block-size ME is performed. For example, it has been found that the variable block-size ME consumes heavy computation time of a H.264 encoder. More specifically, multiprediction modes, multireference frames, and higher motion vector resolution adopted in ME of H.264 can consume 60% (1 reference frame) to 80% (5 reference frames) of the total encoding time of the H.264 codec [7].

Obviously, a fast interviewpoint DE technique is desirable for most of the recently developed multiview image or video encoders with hybrid temporal-viewpoint prediction structures [6]. Actually, in the past few years, numerous fast ME algorithms have been proposed for alleviating the heavy computational load of ME while maintaining its prediction performances. For instance, four-step search (FSS) [8], diamond search (DS) [9], predictive algorithm (PA) [10], adaptive rood pattern search [11], hierarchical estimation with SAD reusing [12], and hybrid unsymmetrical-cross multihexagon-grid search (UMHexagonS) [7] adopted in the H.264 JM 10.1 model [13]. However, these fast ME algorithms, which were essentially proposed to accelerate temporal prediction, may render inefficient the direct application to interviewpoint prediction [14], [15], because such differences in the application scenarios actually dictate quite different ME and DE design principles and the associated prediction performance. In fact, to track the large and irregular (depth-dependent) disparity typical for convergent multiview camera setups, traditional full-search ME and most fast-ME algorithms have to greatly amplify the motion refinement grid to prevent the search points from dropping into a local minimum in the earlier search stages. Otherwise, the resulting coding efficiency will significantly drop.

In general, temporal motion cannot be characterized in an adequate way, especially when there is sudden motion or a scene change. This is because the object motion or camera movement involved is not absolutely predictable. On the contrary, for multiview visual contents, interviewpoint correlation is mainly determined by the geometry relationship between the scene and multiple camera setups. As a result, the disparity search across neighboring views can become more effective in a predictable and reduced search space, because the interviewpoint motion is highly dependent on the depths of the objects and the scene, which by its nature is much more structured than the temporal motion.

It has been discovered that the interviewpoint motion or essentially the disparity vector relating two adjacent views is subject to the epipolar constraint, which is the only available geometry constraint between a pair of stereo images of a single scene [16]. However, so far, little effort has been made to integrate this theoretical principle from stereo vision with the emerging practical multiview image/video encoders. Among these very limited previous works, Hata and Etoh [17] proposed to search the best block match only along the one-dimensional (1-D) epipolar line, and the resultant displacement along the epipolar line is encoded to replace the conventional two-dimensional (2-D) motion vectors. Though epipolar geometry provides a geometrical constraint for the correspondences, limiting the search range to the epipolar line is not an optimal way for interviewpoint coding, because video coding does not target at finding true correspondences, but rather reference blocks that minimize the coding cost. As a consequence, there is only considerably minor improvement in the multiview image coding efficiency based on this approach [17], and this does not warrant the extra increase in the decoding complexity and the changes of bit-stream syntax. In addition to this work, a view synthesis prediction technique [18], [19] has recently been proposed to improve the interviewpoint coding efficiency by exploiting the known camera geometrical parameters. Based on the specific techniques of depth plane traversing and 3-D projection, it demonstrates an improved interviewpoint coding efficiency for the rectified multiview video frames, but the encoding complexity is increased as well because of the additional view synthesis prediction and coding mode decision [19]. On the other hand, traversing through different hypothesized depths in the 3-D space for the current block to be encoded is exactly equivalent to searching along the projected epipolar line in the reference viewpoint image. Aiming at effective utilization of epipolar geometry for fast block-based DE, we show in our recent work [20] that the disparity search space can be largely reduced to accelerate the entire DE process, while the interviewpoint coding efficiency does not noticeably degrade. However, this is still a suboptimal approach exploiting epipolar constraints for fast DE, and it also lacks of a rigorous justification of the proposed disparity search center and search space. Furthermore, a complete fast DE algorithm is not provided to show clear advantages over state-of-the-art fast ME approaches when adopted for DE.

Considering that epipolar geometry is a powerful and obtainable geometry constraint for multiview visual contents, we propose a generic fast DE framework to accelerate the interviewpoint prediction based on the knowledge of epipolar geometry, which remains a very valuable but little explored topic so far. More precisely, our research objective is to propose a generic fast DE approach that effectively exploits the epipolar geometry to strike an optimal tradeoff between the interviewpoint coding efficiency and the associated DE computational complexity, under the conventional predictive video coding framework. To assure its general application, the proposed DE algorithm is also desired to be compatible with the existing video coding standards. The contributions of this paper can be summarized as follows.

1) Based on a novel DE technique exploiting the property of epipolar geometry, a few rarely reported optimal disparity vector distribution properties are revealed.
2) Motivated by these unique observations, we propose a basic framework of the epipolar geometry-based fast DE technique, which is generally applicable to the full-search and conventional fast-search algorithms.
3) An adaptive and fast DE scheme is proposed out of the above basic framework, and it is specifically implemented in a H.264 video encoder, with the explicit consideration of the variable block-size motion search structure adopted in H.264/AVC.
4) Two additional local improvements are proposed to further accelerate the proposed algorithm.
5) The sensitivity of the proposed fast DE technique to the epipolar geometry estimation noise is also examined for the first time, where the Gaussian random noise with different intensities is imposed onto the reference epipolar geometry constraint acquired in an ideal situation.

To the best of our knowledge, this is the first attempt at explicitly employing solid geometry constraints in an advanced video coding standard. Independent of complicated vision algorithms and fully compatible with the H.264/AVC standard, the proposed technique assures its real application.

The remainder of this paper is organized as follows. In Section II, the property of epipolar geometry, as the theoretical foundation of this paper, is first briefly reviewed. We then derive the research motivations from some rarely reported observations. Section III presents the basic framework of the proposed epipolar geometry-based fast disparity search. We then propose an adaptive and fast implementation out of the basic framework in Section IV, which is specifically implemented in a H.264 video encoder. Section V describes two additional methods that can be used to further accelerate the proposed algorithm. The experimental results are shown in Section VI, along with the investigation of the sensitivity of the proposed DE algorithm to the epipolar geometry estimation noise. We conclude our work and discuss future research options in Section VII.

II. Brief Review of Epipolar Geometry and Research Motivations From Observations

A. Brief Review of Epipolar Geometry

Epipolar geometry, as a specific example of multiview geometry, is the only available geometry constraint between a stereo pair of images of a single scene [16]. It has been extensively studied in computer vision.

Let us consider a stereo imaging setup as shown in Fig. 4. Let C_1 and C_2 be the optical centers of the first and second cameras and let the plane I_1 and I_2 be the first and second image planes. According to epipolar geometry, for a given image point p_2 in the second image, its corresponding point p_1 in the first image is constrained to lie on line l_1. This line is called the epipolar line of p_2. The epipolar constraint can be formulated as follows:

$$l_1 = F \cdot \tilde{p}_2 \tag{1}$$

$$\tilde{p}_1^T \cdot F \cdot \tilde{p}_2 = \tilde{p}_1^T \cdot l_1 = 0 \tag{2}$$

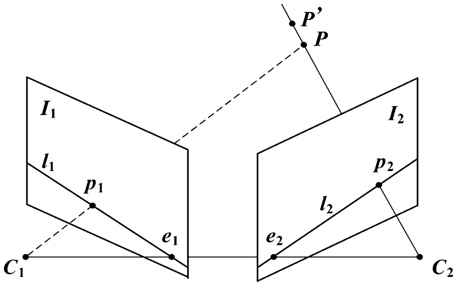


Fig. 4. Epipolar geometry.

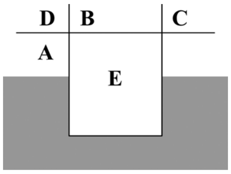


Fig. 5. Neighboring blocks for the median MV prediction (E is the current block).

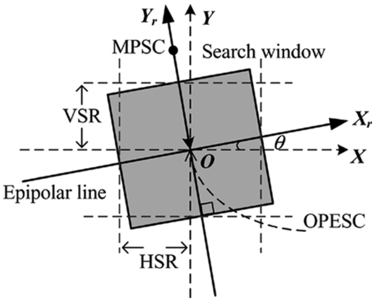


Fig. 6. Representation of the optimal DV in the rotated coordinate system aligned to the epipolar line.

where $\tilde{p}_1$ and $\tilde{p}_2$ are the homogeneous coordinates of p_1 and p_2, and F is called the *fundamental matrix* (FM). It is a 3 × 3 matrix, determined by the intrinsic matrix and the relative position of the two cameras. Therefore, from (1), it is clear that, once F is available, the equation of epipolar line l_1 can be computed to significantly reduce the search space of correspondence by following the obtained epipolar constraint. Actually, there are many ways to determine FM. A good review of existing techniques for estimating FM is presented in [21]. If camera geometry is calibrated, in which case both intrinsic and extrinsic camera parameters are known, FM can be easily and precisely calculated from them through a few matrix multiplications [21] as

$$F = A_1^{-T}[t]_\times R A_2^{-1} \tag{3}$$

where A_1 and A_2 are the intrinsic matrices of the first and second cameras, respectively, and (R, t) is the rigid transformation (rotation and translation) which brings points expressed in the second camera coordinate system to the first one.[1] In reality, it is often a common practice for multiview video system designers to first calibrate camera geometry before real capture sessions [1], [2], and camera parameters are also identified as the important side information to be encoded and transmitted in MPEG's requirements on multiview video coding [22]. In this paper, we hence concentrate on the multiview image/video compression by assuming that reliable FM has already been computed for the current multiview camera setup during a preprocessing phase.

B. Observations and Research Motivations

Though epipolar geometry provides geometrical constraints to correspondences, it does not suggest on its own where to find the correct corresponding point along the epipolar line. The problem becomes even more complicated when putting epipolar geometry under the video coding context, because achieving the optimal video coding efficiency is the key design concern other than finding geometrically correct corresponding points. To motivate effective utilization of epipolar constraints for fast DE on the rigorously justified foundation, two fundamental questions should be appropriately addressed first.

1) How do we decide on an optimal disparity starting search point and a special search window shape in conjunction with epipolar constraints, so that there is a very high probability to find the minimum video coding cost positions around this search center?
2) How do we further reduce the disparity search space to accelerate the epipolar geometry based DE, without noticeably compromising the interviewpoint video coding efficiency?

Generally speaking, the prediction accuracy of the starting motion search point is of critical importance to the search pattern and size adopted in a fast ME algorithm, which jointly determine the search speed and the resulting performance. As a consequence, most of the recent fast ME algorithms make use of the median predicted starting point to lead the subsequent motion search to the promising area, which ideally is around the global minimum. More clearly, the starting search point is obtained from the median motion vector (MV) predictor, where the median value of the left, top, and top-right (or top-left) neighboring blocks' MVs are used for the prediction (see Fig. 5). The effectiveness of such a median predicted start point is validated not only by its good application in an early termination algorithm [23], but also by the high concentration of optimal MVs centered at this point in a rood shape [24].

Although the median predicted search center (MPSC) generally performs well in the temporal motion search, motivated by epipolar geometry principle, we may speculate that the best matching point or physically correct correspondence for two adjacent multiview images subject to the epipolar constraint, should predominantly stay on or near the corresponding epipolar line rather than surrounding the MPSC. In line with this speculation, we project the MPSC orthogonally onto the epipolar line, and its projection point is chosen as our proposed starting search center (see Fig. 6). To compare the disparity prediction accuracy of such an orthogonal projection epipolar search center (OPESC) with that of widely used MPSC, we

[1] Notations in (3): $[t]_\times$ is the antisymmetric matrix defined by t such that $[t]_\times x = t \times x$ for all 3-D vector x, and $A^{-T} = (A^{-1})^T = (A^T)^{-1}$ is a concise notation for any invertible square matrix A.

TABLE I
CONVERGENT MULTIVIEW IMAGE SET OR VIDEO SEQUENCES

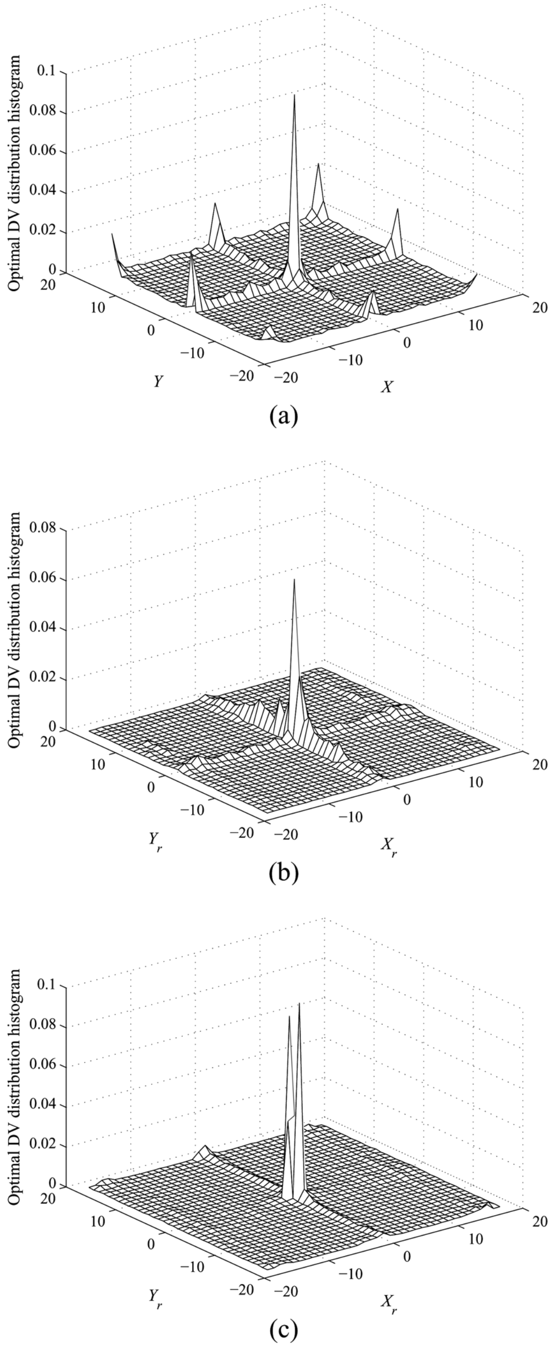
	Ballet	*Breakdancer*	*Dinosaur*
Resolution	1024×768	1024×768	720×576
Angular intervals	∼4°	∼4°	10°
Viewpoint number	8	8	36

devise an adequate search window for the epipolar geometry-based full search, which is then fairly compared with the fast full search (FFS) supported in the H.264 JM model.

The specific experimental settings are as follows. The search window used in FFS is $[-16, 16] \times [-16, 16]$, and the search window for the epipolar geometry-based full search is a square measured in a rotated coordinate system (defined by O, X_r, and Y_r) aligned to the epipolar line, i.e., the shaded region in Fig. 6. The principal set of integer pixel search points for epipolar geometry-based full search is$\{(x,y)|x = Round(x_0 + \Delta x), y = Round(y_0 + \Delta x \cdot tg\theta)$, where $x_0 = -y_0 \cdot tg\theta, y_0 \in [-16, 16]$, and $\Delta x \in [-16, 16]\}$, represented in the original coordinate system (defined by O, X, and Y). θ is the slope angle of the current epipolar line, and $Round(\cdot)$ is the operator to round the value to the nearest integer. Hence, the number of the integer pixel global search candidates for epipolar geometry-based full search is 1089 (33 × 33), which is the same as that of FFS in this experiment.

By applying FFS and epipolar geometry-based full search to the ME module of the H.264 JM 10.1 model in turn, we encode three convergent multiview captured real sequences, i.e., the second multiview image set of *Ballet* [25], the first multiview image set of *Breakdancer* [25], and all of the 36 multiview images of *Dinosaur* [26]. Among them, *Breakdancer* is a test sequence adopted in MPEG 3DAV's call for proposal on multiview video coding [27], and *Dinosaur*, from the University of Hannover, is a popular multiview image set adopted in vision research. *Ballet* is captured with a similar multiview camera setup to that was used for *Breakdancer*, but the former sequence has a larger color contrast between the foreground dancer and background walls, compared with the contrast seen in the latter one. Table I describes the basic configuration information of these multiview image set or video sequences. After executing different DE methods with the rate-distortion (R-D) optimization enabled, the resulting optimal disparity vectors (DV) in terms of H.264 R-D optimized costs are represented by the relative distance and direction to MPSC or OPESC. Since we are targeting a generic fast DE paradigm that can be easily adopted and extended in different video coding standards, we specifically focus on presenting and analyzing the experimental results for the most common DE procedure in this section, i.e., DE performed for the macroblock size of 16 × 16. Also, it is worth to noting that, when epipolar geometry-based full search is adopted, the optimal DV distributions are considered in the rotated coordinate system.

Fig. 7 shows the optimal integer pixel DV distribution histograms for *Ballet*, using different disparity search schemes and reference points for building the distribution histograms. More specifically, we consider the following three combinations of the DE algorithms and the distribution reference points, corresponding to Fig. 7(a)–(c).



(a)

(b)

(c)

Fig. 7. Optimal DV distribution histograms for *Ballet* (a) using FFS and MPSC as the reference point, (b) using epipolar geometry-based full search and MPSC as the reference point, and (c) using epipolar geometry-based full search and OPESC as the reference point.

(a) FFS is used to obtain optimal DVs, and the reference point for building optimal DV distribution histogram is MPSC.
(b) Epipolar geometry-based full search is used to obtain optimal DVs, and the reference point for building optimal DV distribution histogram is MPSC.
(c) Epipolar geometry-based full search is used to obtain optimal DVs, and the reference point for building optimal DV distribution histogram is OPESC.

As can be observed from Fig. 7(a) and (b), there are obvious distribution leakages from the central area around MPSC using

TABLE II
OPTIMAL DV DISTRIBUTION PERCENTAGES FOR *Ballet*

Ballet: optimal DV distribution percentage	Integral area $[-1,1]\times[-1,1]$	Integral area $[-16,16]\times[-4,4]$
Using FFS and MPSC as the reference point	15.80%	43.68%
Using epipolar geometry-based full search and MPSC as the reference point	16.09%	39.62%
Using epipolar geometry-based full search and OPESC as the reference point	31.49%	47.96%

TABLE III
OPTIMAL DV DISTRIBUTION PERCENTAGES FOR *Breakdancer*

Breakdancer: optimal DV distribution percentage	Integral area $[-1,1]\times[-1,1]$	Integral area $[-16,16]\times[-4,4]$
Using FFS and MPSC as the reference point	32.12%	65.51%
Using epipolar geometry-based full search and MPSC as the reference point	30.42%	56.32%
Using epipolar geometry-based full search and OPESC as the reference point	53.20%	69.25%

TABLE IV
OPTIMAL DV DISTRIBUTION PERCENTAGES FOR *Dinosaur*

Dinosaur: optimal DV distribution percentage	Integral area $[-1,1]\times[-1,1]$	Integral area $[-16,16]\times[-4,4]$
Using FFS and MPSC as the reference point	70.96%	86.11%
Using epipolar geometry-based full search and MPSC as the reference point	18.98%	33.43%
Using epipolar geometry-based full search and OPESC as the reference point	80.13%	87.97%

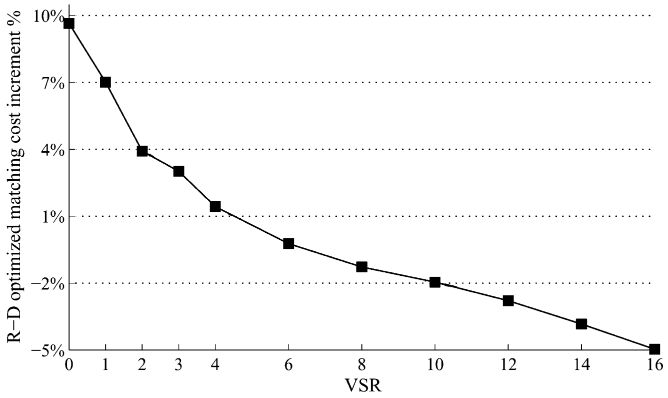


Fig. 8. Average R-D optimized matching cost increment rate versus VSR, for epipolar geometry-based full search in comparison with FFS (HSR $= 16$, VSR $= 16$).

either FFS or epipolar geometry-based full search. However, the optimal DV distribution centered at OPESC based on epipolar geometry-based full-search algorithm demonstrates a much better congregate property than that of previous approaches, as shown in Fig. 7(c).

Table II compares the corresponding optimal DV distribution percentages in one center square area ($[-1,1]\times[-1,1]$), and one center band area ($[-16,16]\times[-4,4]$) of different schemes for *Ballet*. The comparison results reveal the same finding from Fig. 7, i.e., the OPESC adopted in epipolar geometry-based full search has a much higher probability to be chosen as the minimum R-D cost point than the widely used MPSC. The optimal DV distribution probabilities for *Breakdancer* in Table III and *Dinosaur* in Table IV also show this similar property.

Based on the observations from the above experiments, it can be concluded that the proposed OPESC integrated with an epipolar line-aligned search window can provide a far more accurate disparity prediction than MPSC can. This rarely observed fact justifies the validity and advantages of our proposed fast DE algorithm, which exploits the high concentration of optimal DVs that distribute around OPESC. Furthermore, Fig. 7(b) and (c) clearly shows that, even if exactly the same set of optimal DVs are considered, OPESC is still a much better optimal DV distribution center than MPSC is. Hence, the necessity of projecting MPSC orthogonally to acquire OPESC is confirmed for the macroblock-level disparity search.

Based on these results, it is also evident that the search candidates closer to the OPESC have a higher possibility of finally being chosen as the optimal DVs. Therefore, a good tradeoff point should exist between prediction accuracy (or coding efficiency) and computational load, when the horizontal search range (HSR) or vertical search range (VSR) of the search window is varied (see Fig. 6). Considering that, for most typical multiview camera setups, horizontal disparity is the principal component of the whole movement and thus much more important than vertical disparity, we may speculate that the VSR can be greatly reduced to achieve a coding speedup without causing noticeable degradation in the optimal disparity search results. Fig. 8 shows a typical percentage increase of the R-D optimized costs in H.264 versus different VSR values for epipolar geometry-based full search in comparison with FFS (with a search window of 33 × 33), and the best tradeoff between the computational saving and the performance loss can be identified. For example, VSR $= 4$ can be chosen in epipolar geometry based full search to reduce the complexity while achieving a decent coding performance. Fig. 8 also indicates that the better optimal DV congregate property of epipolar geometry-based full search (compared with FFS) demonstrated in Fig. 7 will not result in an increase in the matching cost, and hence no compromise in the coding efficiency is to be incurred. As a comparative experiment, we also examine the impact of VSR reduction on the disparity prediction performance of FFS in dealing with multiview image sets. The average matching cost increment rate of VSR-varying FFS with a rectangle search shape is demonstrated in Fig. 9, and apparently it is much larger

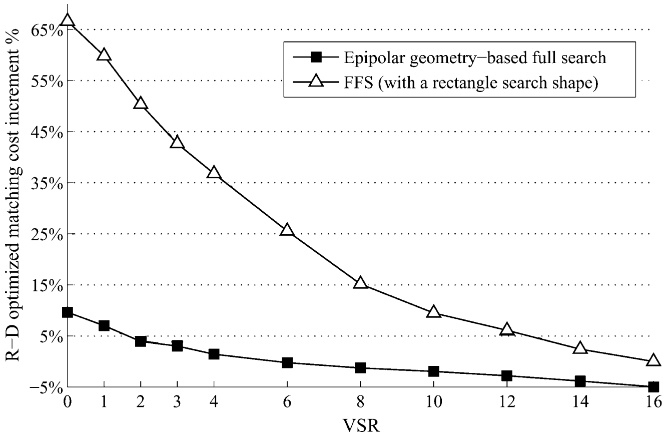


Fig. 9. Average R-D optimized matching cost increment rate of VSR-varying epipolar geometry-based full search, and VSR-varying FFS with a rectangle search shape, in comparison to FFS (HSR $= 16$, VSR $= 16$).

than that of epipolar geometry-based full search, when the same VSR value is considered.

From Fig. 9, it can also be observed that when VSR increases from 1 to 16, the average matching cost increment rate of VSR-varying epipolar geometry-based full search converges far more rapidly to its minimum value, than that of VSR-varying FFS. This observation clearly indicates that the search window of our proposed epipolar geometry-based fast DE algorithm can significantly shrink along its vertical dimension, to achieve a good tradeoff between the complexity and disparity prediction performance.

Motivated by the foregoing seldom-reported observations, we propose a novel epipolar geometry-based fast-disparity-search algorithm for multiview image and video coding. The proposed algorithm can effectively track the real disparity even with a largely reduced disparity refinement area. Thus, the coding efficiency is guaranteed while the DE complexity is reduced. Basically, our proposed DE technique consists of a starting search-point prediction method, a physically correct rotated disparity search region, and a disparity refinement region reduction scheme, as presented in Sections III and IV.

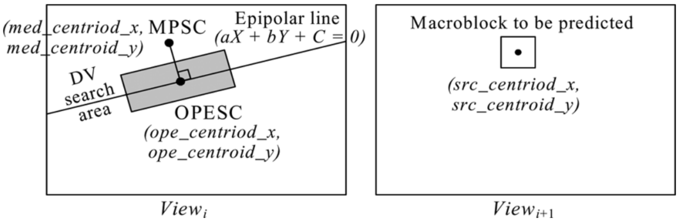


Fig. 10. Proposed search center (OPESC) and the epipolar line-aligned search space (the shaded area).

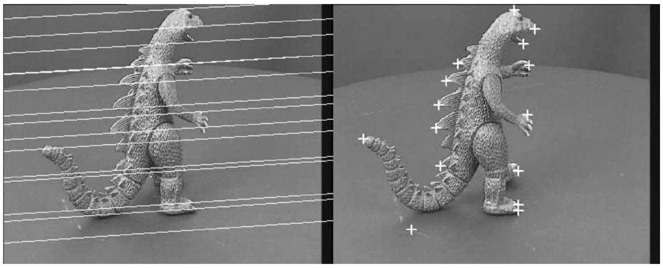


Fig. 11. Epipolar constraints for two adjacent *Dinosaur* images. The epipolar lines displayed in the left image are computed from the corresponding cross centers in the right image by imposing epipolar geometry constraints.

III. Basic Framework of the Proposed Fast DE Technique

As shown in Fig. 10, the key ideas of our proposed epipolar geometry-based fast disparity estimation (EGB) are twofold. The first one is to transform the commonly adopted MPSC to obtain its orthogonal projection point (i.e., OPESC) on the corresponding epipolar line. Then, the disparity search is performed in a largely reduced epipolar line-aligned search space centered at OPESC.

More clearly, as noted in Fig. 10, we consider two neighboring views $View_i$ and $View_{i+1}$ and denote the input FM relating them be F. Given the centroid coordinates of the current macroblock to be predicted in $View_{i+1}$, the corresponding epipolar line equation $(aX + bY + c = 0)$ in $View_i$ can be computed by multiplying F by the homogeneous centroid coordinates using (1). We propose this centroid-based epipolar line calculation in that the resulting epipolar line computed from the centroid, unlike the top-left corner of the macroblock [17], can more precisely reflect the average epipolar constraint for a group of pixels. After calculating the epipolar line equation, the proposed starting search point OPESC $(ope\_centroid\_x(y))$ and its corresponding initial DV $(ope\_dv\_x(y))$ for the current macroblock under search $(src\_centroid\_x(y))$, can be derived from the MPSC point $(med\_centroid\_x(y))$ given by the median predicted DV $(med\_dv\_x(y))$ as follows:

$$\begin{cases} med\_centroid\_x(y) = src\_centroid\_x(y) + med\_dv\_x(y) \\ ope\_centroid\_x(y) \\ \quad = ORTHO\_PROJ(med\_centroid\_x(y), a, b, c) \\ ope\_dv\_x(y) = ope\_centroid\_x(y) - src\_centroid\_x(y) \end{cases} \tag{4}$$

In addition, since it is well known in the correspondence problem that a larger matching window is desirable to achieve the reliable matching, we thus propose applying (4) only to the disparity search at the macroblock level, i.e., for a block-size of 16×16. Although the fine-grained block disparity search for submacroblocks are supported in a few recent video coding standards, e.g., MPEG-4 [28] and H.264/AVC [5], we only transform MPSC to obtain OPESC for the macroblock level disparity search, so that the prediction outliers from small matching windows or submacroblocks can be kept away from destroying the smooth disparity field. Additionally, the computation increase can be kept marginal (the DE overhead is lower than 1.2% for our proposed fast DE implemented in H.264 in Section V), since the computation of epipolar constraint is solely performed at the macroblock level. As a typical example, Fig. 11 provides a visual inspection of the epipolar constraints relating two adjacent *Dinosaur* multiview images.

Thus far, we have presented the key techniques of the proposed fast DE that exploit the special property of epipolar geometry. In fact, as an effective and generally applicable DE technique, the proposed scheme can work well with the full-search and fast ME algorithms for multiview image and video coding to achieve a good tradeoff between the quality and complexity.

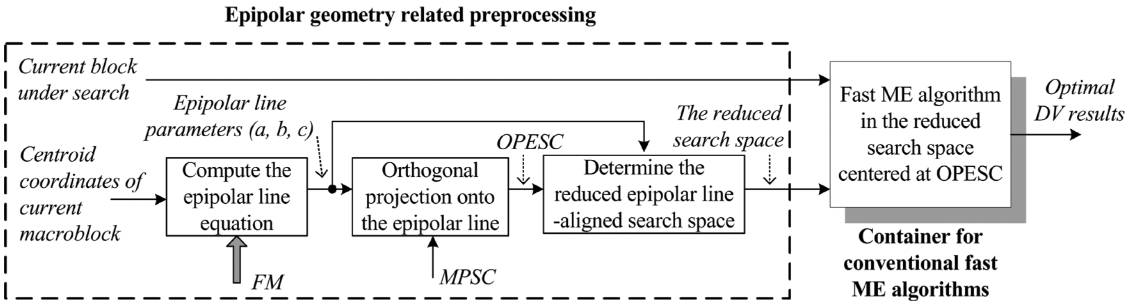

Epipolar geometry related preprocessing
Current block under search
Epipolar line parameters (a, b, c)
Centroid coordinates of current macroblock
Compute the epipolar line equation
FM
OPESC
Orthogonal projection onto the epipolar line
MPSC
The reduced search space
Determine the reduced epipolar line -aligned search space
Fast ME algorithm in the reduced search space centered at OPESC
Container for conventional fast ME algorithms
Optimal DV results

Fig. 12. Basic framework of the proposed fast DE technique.

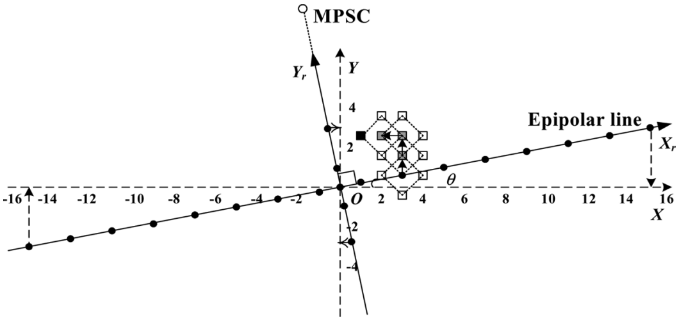


Fig. 13. Major search process and search patterns of the proposed EGB.

Therefore, we illustrate a general framework of the proposed fast DE technique in Fig. 12, which clearly shows the integration interface for plugging in the existing fast ME algorithms or motion search heuristics to leverage the advantages of the proposed fast DE framework. The delimited part on the left in Fig. 12 includes the proposed normative modules preparing epipolar geometry-based information, and FM is the only extra information to be provided for this processing. The block with a shadow following the normative procedure on the right represents the container, where various fast ME algorithms can be included and conducted in a reduced search region flexibly centered at OPESC.

IV. Adaptive and Fast Implementation of the Proposed EGB Algorithm in H.264

As a specific instance out of the proposed epipolar geometry-based fast DE framework, a practical fast-disparity-search technique is proposed here. It mainly comprises two major search steps. The first major search process is a rotated unsymmetrical rood-pattern search with a sampling grid size of 2, marked as the dots in Fig. 13. More specifically, here we set the user-configurable parameter HSR to 16 and VSR to 4 as the size of the adopted unsymmetrical rood pattern, in order to track the principal component of true disparity along the epipolar line with a largely reduced number of initial search candidates. The second major step is a recursive local diamond search (for up to four recursive times) to bring the disparity search to the minimum matching cost point, marked as the squares in Fig. 13. Because the sampling grid size of the unsymmetrical rood pattern is 2, a recursive loop of four is sufficient for the refinement task.

As a state-of-the-art video coding standard, H.264 is chosen as the multiview encoder for implementing and integrating our proposed EGB algorithm, and the key search procedure is illustrated in Fig. 13. Since the variable block-size ME is supported in H.264, we explicitly take into account of this feature in the design of our proposed EGB, and an adaptive and fast implementation of the EGB in a H.264 encoder is presented below.

Being the proposed starting search point, OPESC is first computed and checked at the start of the disparity search to obtain the initial R-D optimized matching cost, which is defined in H.264 as the following rate-constrained Lagrangian cost function:

$$J(dv) = \mathrm{SAD}(dv) + \lambda \cdot R(dv - med\_pred\_dv) \qquad (5)$$

where SAD (sum of absolute difference) is a block-wise distortion measurement, R gives the bit rate of encoding the difference between the current DV and the median predicted DV, and λ is the Lagrangian multiplier. In fact, when the OPESC can already give a fairly good estimation of the optimal matching point, the unsymmetrical rood-pattern search or even the entire search process for current block disparity search can be early terminated with only a negligible loss in the prediction performance. To adaptively decide the case when the unsymmetrical rood-pattern search can be skipped and when the entire search process can be terminated early, we adopt the similar basic thresholds (i.e., $\{TH_1 = 1000, TH_2 = 800, TH_3 = 7000\}$) defined in the simplified UMHexagonS, as a fast ME algorithm in H.264 JM 10.1 model [13]. Because H.264 features variable block-size ME ($\mathrm{MODE} = 1$ to 7, corresponding to seven inter prediction modes 16 × 16, 16 × 8, 8 × 16, 8 × 8, 8 × 4, 4 × 8, and 4 × 4, respectively), we also explicitly deal with this multiple modes DE in our adaptive fast DE implementation, where the basic threshold set $\{TH_1 = 1000, TH_2 = 800, TH_3 = 7000\}$ for MODE 1 is adaptively modified to generate appropriate MODE-dependent thresholds for the submacroblock disparity search. Specifically, MODE-dependent thresholds are obtained by linearly scaling the basic threshold set, and the scaling coefficient is determined in proportion to the area ratio between the current block and macroblock (MODE 1).

Including such a series of adaptive thresholds, the overall EGB algorithm can be depicted as in Fig. 14, with the detailed description as follows.

Step 1) Compute the corresponding epipolar line equation $(aX + bY + c = 0)$ based on the centroid coordinates of current macroblock to be predicted, and this step is only executed at the start of disparity search for a new macroblock.

Step 2) Perform the orthogonal projection of MPSC onto the epipolar line to get OPESC position. This step is performed only for MODE 1. For submodes defined in H.264, go to Step 3) and approximate MPSC as OPESC. Because the median DV predictor for submacroblocks can actually benefit from the accurate disparity prediction conducted at the macroblock level, MPSC predicted for each submacroblock can be approximated as OPESC for the current block, in order to save the computational complexity. Our experiments prove that there is little loss in coding efficiency due to this simplification.

Step 3) Calculate the R-D optimized matching cost at OPESC. If the cost is small enough (compared with MODE-dependent TH_1), do a local diamond search, then terminate the searching process for the current block and go to Step 10).

Step 4) Conduct a local diamond search centered at OPESC to locate the position with the minimum R-D optimized matching cost, which is to serve as the center for the unsymmetrical rood-pattern search later. This step is involved to refine OPESC, since the calculation of OPESC is based on the fundamental matrix, which may be affected by digitization effects and computation precision.

Step 5) Adaptively determine whether to bypass or conduct the unsymmetrical rood pattern search, based on the comparison result between the current minimum matching cost and different thresholds for DE under different MODE (i.e., MODE-dependent TH_2 and MODE-dependent TH_3). If skipped, go to Step 7). Otherwise continue with Step 6).

Step 6) Conduct the unsymmetrical rood-pattern search in a center-biased search order.

Step 7) For submodes defined in H.264, calculate the R-D optimized matching cost for the predicted point based on the up-layer DV prediction (similar to UMHexagonS [7]). If it has a smaller matching cost than the smallest one found in the previous steps, store the current cost and position.

Step 8) Check whether the search process can be terminated now by comparing the minimum cost with the same thresholds (MODE-dependent TH_1) used in Step 3). If the search can be terminated, go to Step 10). Otherwise, continue with Step 9).

Step 9) Conduct a recursive local diamond search (for a maximum looping times of four) to refine the disparity search.

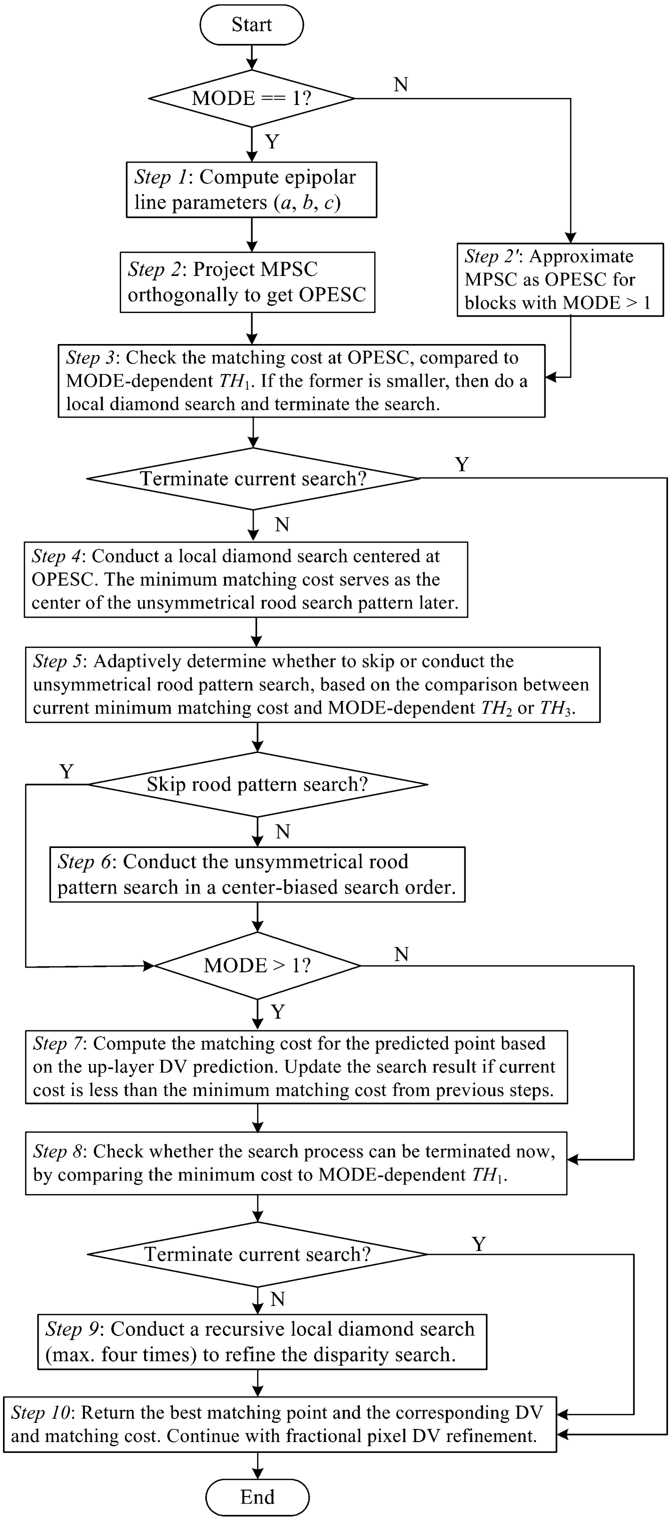

Start
MODE == 1?
N
Y
Step 1: Compute epipolar line parameters (a, b, c)
Step 2: Project MPSC orthogonally to get OPESC
Step 2′: Approximate MPSC as OPESC for blocks with MODE > 1
Step 3: Check the matching cost at OPESC, compared to MODE-dependent TH1. If the former is smaller, then do a local diamond search and terminate the search.
Terminate current search?
Y
N
Step 4: Conduct a local diamond search centered at OPESC. The minimum matching cost serves as the center of the unsymmetrical rood search pattern later.
Step 5: Adaptively determine whether to skip or conduct the unsymmetrical rood pattern search, based on the comparison between current minimum matching cost and MODE-dependent TH2 or TH3.
Y
Skip rood pattern search?
N
Step 6: Conduct the unsymmetrical rood pattern search in a center-biased search order.
MODE > 1?
N
Y
Step 7: Compute the matching cost for the predicted point based on the up-layer DV prediction. Update the search result if current cost is less than the minimum matching cost from previous steps.
Step 8: Check whether the search process can be terminated now, by comparing the minimum cost to MODE-dependent TH1.
Terminate current search?
Y
N
Step 9: Conduct a recursive local diamond search (max. four times) to refine the disparity search.
Step 10: Return the best matching point and the corresponding DV and matching cost. Continue with fractional pixel DV refinement.
End

Fig. 14. Flowchart of the proposed adaptive and fast EGB algorithm implemented in H.264.

Step 10) Record the obtained best integer pixel DV value and the associated minimum matching cost, and continue with the fractional pixel DV refinement.

Since we focus on the fast disparity search at integer pixel precision, we choose to adopt the conventional techniques of fast fractional pixel motion search in our implementation. For example, in this paper, we choose the Center Biased Fractional Pel Search (CBFPS) strategy with a diamond search pattern used in UMHexagonS [7].

V. Accelerating the Proposed EGB Algorithm From Two Specific Aspects

The adaptive and fast implementation discussed in Section IV demonstrates the major advantages of the proposed EGB algorithm in computational load, but we can nevertheless further accelerate the EGB algorithm from two specific aspects. First, we propose an incremental computation method for deriving epipolar line equation and rood search pattern, which can largely reduce the number of floating-point operations. Then, motivated by the positive results of the optimal DVs' directional distribution probabilities, we revise the original equal-arm rood-pattern search by proposing a MPSC-favored unequal-arm rood-pattern search, which can decrease the candidate pattern points for every block under disparity search. This results in a negligible loss in the coding performance.

A. *Incremental Computation Method for Epipolar Line Equation and Rood-Pattern Search*

With double-precision floating-point arithmetic intensively used to derive the epipolar line equation and the positions of an unsymmetrical rood pattern, we propose an incremental computation method to largely decrease the overhead of floating-point operations. In addition, gaining a significant speedup for this part of processing, the proposed method is also suitable for the platforms with weak floating-point processing capability, e.g., PDA and Pocket PC.

We reuse the notations defined in Section II. For each macroblock to be predicted in I_2, the most straightforward approach to obtain the corresponding epipolar line l_1 in I_1 is to left multiply the macroblock centroid coordinates p_2 by F using (1). However, this approach results in nine floating-point multiplications and six floating-point additions for each macroblock. Considering the regular macroblock encoding order and the fixed macroblock size, we propose an efficient incremental epipolar line equation computation method as follows.

1) Traverse along the same scanline (same y)

$$\begin{aligned} l_{\text{curr}} &= F \cdot \tilde{p}_{\text{curr}} \\ &= F \cdot \begin{bmatrix} x_{\text{prev}} + 16 \\ y_{\text{curr}} \\ 1 \end{bmatrix} \\ &= F \cdot \begin{bmatrix} x_{\text{prev}} \\ y_{\text{curr}} \\ 1 \end{bmatrix} + F \cdot \begin{bmatrix} 16 \\ 0 \\ 0 \end{bmatrix} \\ &= l_{\text{prev}} + \Delta l_x. \end{aligned} \tag{6}$$

TABLE V
Speedup of Floating-Point (FP) Operations on an Intel Pentium 4 Processor

	Breakdancer	*Ballet*	*Dinosaur*
FP operations of the straightforward method	9 FP multiplications and 6 FP additions per macroblock		
FP operations of the proposed method	3 FP additions, 1 FP store, and 1 FP load per macroblock		
Speedup ratio	4.04	3.68	3.35

2) Traverse to the start of a new scanline (start x at 8)

$$\begin{aligned} l_{\text{curr}} &= F \cdot \tilde{p}_{\text{curr}} \\ &= F \cdot \begin{bmatrix} 8 \\ y_{\text{prev}} + 16 \\ 1 \end{bmatrix} \\ &= F \cdot \begin{bmatrix} 8 \\ y_{\text{prev}} \\ 1 \end{bmatrix} + F \cdot \begin{bmatrix} 0 \\ 16 \\ 0 \end{bmatrix} \\ &= l_{\text{prev}} + \Delta l_y. \end{aligned} \tag{7}$$

Therefore, to get the corresponding epipolar line equation for any macroblock in a frame, we only need to compute the epipolar line parameters l_0 (a_0, b_0, and c_0) for the first macroblock of this frame and the horizontal increment vector Δl_x and vertical increment vector Δl_y. Increasing the immediate previously visited macroblock's epipolar line parameters l_{prev} by Δl_x or Δl_y, the current epipolar line equation (i.e., l_{curr}) can be easily obtained with only three floating-point additions. To compare the real run-time speed of the proposed incremental computation method and the original method, we encode different multiview image sets by applying these two approaches alternately to derive the epipolar line equations. Table V shows the speedup ratio of the proposed floating-point operation reduction scheme in comparison to the straightforward computation that frequently requires matrix multiplication. Without any comprise of the coding efficiency, the proposed incremental computation scheme can bring a speedup factor of more than 3.3 times of the original one, specifically for this portion of processing on an Intel Pentium 4 processor. Moreover, the removal of expensive floating-point multiplications and the reduction of floating-point additions actually guarantee the applications of the proposed EGB on the portable platforms, where the floating-point processing is costly computation-wise.

Following the same approach, we can simplify the floating-point computation for the coordinates of unsymmetrical rood pattern. Only the OPESC needs to be first calculated, while all of the other positions can be incrementally obtained by adding the epipolar line gradient $-a/b$. To guarantee a center-biased rood search order, two temporary variables are used to store the previous positions leading to positive and negative directions. This improvement achieves a speedup factor of 1.53 for the calculation of unsymmetrical rood pattern on an Intel Pentium 4 processor, in comparison with directly applying the epipolar line equation to derive the rood search pattern.

TABLE VI
OPTIMAL DV DISTRIBUTION PERCENTAGES WITH REGARD TO THE RELATIVE LOCATION OF THE OPTIMAL MATCHING POSITIONS TO MPSC

	Breakdancer	*Ballet*	*Dinosaur*
Percentages in the MPSC-located closed half plane	82.19%	73.98%	96.01%
Percentages in the enlarged MPSC-located half plane	90.00%	81.71%	98.48%

The resulting complexity overhead of these two modules consumes less than 1.2% of the total DE computational cost. In fact, this portion of calculation can be further accelerated for online multiview video coding by looking up the prestored tables, if the initial multiview camera setup stays unchanged in the capture process. Because the epipolar constraint only concerns the relative positions and poses of the cameras, the complexity of the scenes or the content of the frames to be encoded does not affect the epipolar line equations and rood search patterns.

B. A MPSC-Favored Unequal-Arm Rood Search Pattern

Considering that the block-matching criterion used in H.264 is the R-D optimized cost in which the difference between the candidate DV and the median predicted DV is also considered, we expect that the optimal DV in terms of the R-D optimized cost will most likely appear in the same closed half plane where MPSC stays. This is because of the impact from the DV difference term in the matching criterion [refer to (5)]. Here, the closed half plane is defined as a planar region consisting of all points on one side of the epipolar line and the points on the epipolar line. This allows the X_r-axis of the rotated coordinate system (O, X_r, and Y_r) illustrated in Fig. 6.

To validate this hypothesis, we use epipolar geometry-based full search (described in Section II) to derive the optimal DVs for different multiview sequences, and then the optimal DVs' distribution probabilities are collected with regard to the relative location of the corresponding optimal matching positions to MPSC. For instance, if and only if the Y_r coordinates of both the optimal matching position and MPSC have the same sign, the occurrence times of the optimal DV in the MPSC-located closed half plane gets increased by one. Table VI proves that there is indeed a dominant percentage of the optimal DV distribution in the MPSC-located closed half plane and an even higher percentage in the enlarged MPSC-located half plane. The latter is constructed by extending the MPSC-located closed half plane by a distance of $1/cos\theta$ in the MPSC-opposite direction along the Y_r-axis, which is shown as the shaded region in Fig. 15.

Based on these positive statistical results, we propose an unequal-arm rood search pattern to favor the MPSC direction, i.e., we only keep one search candidate in the MPSC-opposite half plane. In addition, the MPSC is also included as a search candidate at the search start to check whether the rood pattern search can be bypassed. Demonstrating the similar PSNR values, only the bit-rate increase ratio of this unequal-arm rood pattern is reported in comparison to its counterpart with an equal arm length in Table VII, which is small.

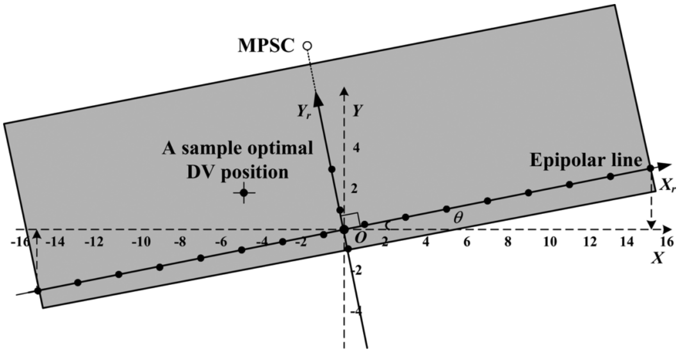


Fig. 15. MPSC-favored unequal-arm rood search pattern.

TABLE VII
BIT-RATE INCREASE RATIO OF THE UNEQUAL-ARM ROOD PATTERN SEARCH COMPARED WITH THE EQUAL-ARM ROOD PATTERN SEARCH

Bit-rate increase ratio	VSR = 4	VSR = 8	VSR = 16
Breakdancer	0.10%	-0.45%	0.64%
Ballet	0.96%	-0.46%	0.87%
Dinosaur	-0.07%	-0.12%	-0.19%
Reduced rood search points per block	1	3	7

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

Our experiments are based on the JM version 10.1 of H.264 reference software [13]. The baseline profile is used to configure the encoder. We set the number of reference frames to 1, and all frames, except for the first one, are encoded as P-frames. R-D optimization and CAVLC entropy encoding are enabled. For our proposed EGB, the HSR is set to 16, while VSR is reduced to 4. All of the techniques studied in Sections IV and V have been integrated into our proposed EGB, and they are activated throughout the following experiments.

We focus the experiments on two multiview video sequences (i.e., *Breakdancer* and *Ballet*) and one multiview image set, *Dinosaur*. In fact, the first two multiview video sequences can be regarded as two temporal sets of multiview image sets, which consist of multiple images simultaneously captured from all the viewpoints at one certain time instance. Hence, we encode the first multiview image set of *Breakdancer*, the second one of *Ballet*, and all of the thirty-six multiview images of *Dinosaur* with different QP (quantization parameter) values to compare the R-D performance and integer pixel DE speed of FFS (with a search window size of 33 × 33), UMHexagonS as a fast ME algorithm adopted in H.264 JM model [7], UMHexagonS-2 as a simplified and accelerated variant of UMHexagonS, and our proposed EGB. The FM is the only side information to be fed into the H.264 JM model, when the proposed EGB is chosen to accelerate the interviewpoint prediction.

TABLE VIII
AVERAGE (AND WORST CASE) NUMBER OF SAD OPERATIONS FOR DIFFERENT DE ALGORITHMS ($\mathrm{QP} = 28, \mathrm{MODE} = 1-7$)

	FFS	UMHex-agonS	UMHex-agonS-2	EGB
Breakdancer	1089 (1089)	659.82 (2008)	140.08 (1917)	68.27 (209)
Ballet	1089 (1089)	731.01 (2008)	186.47 (1917)	83.87 (209)
Dinosaur	1089 (1089)	386.86 (2008)	83.55 (1917)	47.89 (209)

TABLE IX
R-D PERFORMANCE AND THE INTEGER PIXEL DE SPEED COMPARISON FOR *Breakdancer*

QP	Performance	FFS	UMHex-agonS	UMHex-agonS-2	EGB
24	ΔPSNR	39.59	0.00	0.00	0.00
	ΔBitrate	6,942.96	0.70%	3.69%	2.52%
	ΔExe. Time	9.77	3.90	1.39	1.00
28	ΔPSNR	38.24	-0.01	-0.03	-0.03
	ΔBitrate	2,915.14	-0.47%	1.26%	1.69%
	ΔExe. Time	10.08	3.72	1.44	1.00
32	ΔPSNR	37.10	-0.03	-0.04	-0.07
	ΔBitrate	1,459.03	1.88%	0.89%	-0.27%
	ΔExe. Time	10.31	2.99	1.38	1.00
36	ΔPSNR	35.84	-0.06	-0.05	-0.08
	ΔBitrate	887.21	-0.40%	-0.93%	-0.73%
	ΔExe. Time	13.08	3.45	1.77	1.00
Avg.	ΔPSNR	0.00	-0.03	-0.03	-0.04
	ΔBitrate	0.00%	0.43%	1.23%	0.80%
	ΔExe. Time	10.81	3.52	1.50	1.00

A. Comparison of Coding Performance and Execution Speed of Different DE Algorithms

Because SAD operation is the most time-consuming component in DE, we first report the average number of SAD operations for different DE algorithms in Table VIII, and the worst case number of SAD operations is also provided in the brackets. Note that here the *normalized computation cost* [24] is used to convert the submode SAD operations to the relative cost, according to the *unit cost* that is defined as the SAD operation for a 16 × 16 block. For instance, an 8 × 8 block SAD operation is 1/4 of one *unit cost*.

Tables IX–XI summarize the detailed experimental results for *Breakdancer*, *Ballet*, and *Dinosaur*, respectively. For the convenience of comparison, $\Delta\mathrm{PSNR}$, $\Delta\mathrm{Bitrate}$, and $\Delta\mathrm{Exe.Time}$ represent the average PSNR gain (dB), the average bit-rate increase rate (%), and the average relative execution time cost of the integer pixel DE, respectively. All three fast DE schemes are compared with the absolute coding performance of FFS in the third column, with the exception that EGB is used as the reference when the relative time cost is measured. When the average integer pixel DE speed of the proposed EGB is measured, all of the epipolar geometry related complexity overhead is also included. The proposed EGB only cause a negligible PSNR degradation of 0.08 dB at maximum, but can achieve a maximum speedup factor of 12.21, 3.52, and 1.59 in comparison with FFS, UMHexagonS, and UMHexagonS-2 on average.

TABLE X
R-D PERFORMANCE AND THE INTEGER PIXEL DE SPEED COMPARISON FOR *Ballet*

QP	Performance	FFS	UMHex-agonS	UMHex-agonS-2	EGB
24	ΔPSNR	41.17	0.00	-0.01	0.00
	ΔBitrate	5,287.37	0.76%	3.42%	2.56%
	ΔExe. Time	8.55	3.50	1.61	1.00
28	ΔPSNR	39.94	-0.02	-0.03	-0.02
	ΔBitrate	3,236.30	-0.88%	1.17%	0.09%
	ΔExe. Time	9.36	3.63	1.65	1.00
32	ΔPSNR	38.38	-0.04	-0.05	-0.05
	ΔBitrate	2,002.94	-0.11%	0.74%	1.53%
	ΔExe. Time	9.84	3.26	1.47	1.00
36	ΔPSNR	36.61	-0.03	-0.06	-0.07
	ΔBitrate	1,315.54	1.26%	0.71%	1.03%
	ΔExe. Time	10.42	3.22	1.62	1.00
Avg.	ΔPSNR	0.00	-0.02	-0.03	-0.04
	ΔBitrate	0.00%	0.26%	1.51%	1.30%
	ΔExe. Time	9.54	3.40	1.59	1.00

TABLE XI
R-D PERFORMANCE AND THE INTEGER PIXEL DE SPEED COMPARISON FOR *Dinosaur*

QP	Performance	FFS	UMHex-agonS	UMHex-agonS-2	EGB
24	ΔPSNR	40.61	-0.01	-0.03	-0.03
	ΔBitrate	4,317.64	0.29%	1.60%	0.92%
	ΔExe. Time	11.42	3.27	1.13	1.00
28	ΔPSNR	38.89	-0.03	-0.05	-0.04
	ΔBitrate	2,835.72	0.47%	1.38%	0.60%
	ΔExe. Time	11.74	2.70	1.16	1.00
32	ΔPSNR	36.61	-0.04	-0.05	-0.07
	ΔBitrate	1,771.00	0.85%	1.68%	0.52%
	ΔExe. Time	12.34	2.17	1.27	1.00
36	ΔPSNR	34.03	-0.02	-0.04	-0.04
	ΔBitrate	1,020.82	0.95%	1.97%	0.69%
	ΔExe. Time	13.34	2.23	1.23	1.00
Avg.	ΔPSNR	0.00	-0.03	-0.04	-0.05
	ΔBitrate	0.00%	0.64%	1.66%	0.68%
	ΔExe. Time	12.21	2.59	1.20	1.00

Since the multiview image and video sequences under experiments have a relatively large portion of low-textured background or ground, especially for *Dinosaur* where there is a large percentage of texture-less blue background and low-textured turn table regions, we believe that the proposed EGB algorithm can demonstrate an even better coding performance and an improved speedup ratio for more difficult multiview sequences. This is especially true when the usual temporal prediction oriented ME methods fail to track the large and irregular interviewpoint disparity and predict the best matching blocks with rich textures in these situations. On the other hand, as a specific example out of the basic framework for epipolar geometry-based fast DE techniques, the proposed EGB may have a good potential to be further improved, e.g., the HSR and VSR can be adaptively decided according to the multiview camera setup and the complexity of the scene. Based on this first trial in exploiting geometry constraints for effective multiview image and video coding, more research interests may be aimed at reshaping and integrating the conventional fast ME algorithms

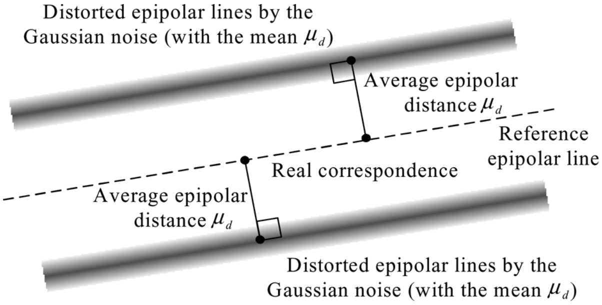

Distorted epipolar lines by the Gaussian noise (with the mean μ_d)
Average epipolar distance μ_d
Reference epipolar line
Real correspondence
Average epipolar distance μ_d
Distorted epipolar lines by the Gaussian noise (with the mean μ_d)

Fig. 16. Average epipolar distance μ_d and distorted epipolar lines by the Gaussian noise.

with this epipolar geometry-based framework. Therefore, more fine-tuned solutions and extensions beyond what is discussed in this paper can be envisioned in the near future.

B. Evaluation of the Sensitivity of the Proposed EGB to the FM Estimation Noise

Although the techniques to estimate the FM are not within the scope of this paper, it is of importance to examine the robustness of the proposed EGB to the FM estimation noise, because a noise-insensitive property is highly desired to guarantee the advantages of the proposed algorithm, in case that camera geometry is not known and the FM has to be estimated.

Since the calibrated camera projection matrices are available for all the sequences adopted in this paper, we can very precisely derive the FM from them, and the epipolar lines calculated from these FM can be regarded as the solid references. Considering that average epipolar distance (the average distance of a point to its corresponding epipolar line, denoted as μ_d hereinafter) is one of common metrics adopted to calibrate the precision of the estimated FM [21], [29], [30], we thus distort the reference epipolar lines by imposing the Gaussian noise, with the mean μ_d as large as four or six pixels and the standard deviation being 0.5 pixel, as shown in Fig. 16.

It can be found from Table XII that the degradation of the coding efficiency due to the noisy FM estimation $(\mu_d = 4)$ is very limited (bit-rate increase ratio below 5%), in comparison with the reference coding cases when cameras are calibrated precisely. It is also evident that adaptively increasing the VSR for the proposed DE method is an effective means to counter the FM noise with an increased intensity $(\mu_d = 6)$. In a good review of FM estimation techniques [21], we can find that even a suboptimal FM estimation algorithm can achieve a high estimation precision or a small average epipolar distance $(\mu_d \leq 1)$. This fact and the above results indicate that our proposed DE algorithm can actually tolerate a fairly large fundamental matrix estimation noise.

Because the proposed EGB algorithm is insensitive to the fundamental matrix estimation noise, it can be generally applied to accelerate multiview image and video coding, even if camera projection matrices are unavailable. This study is also meaningful for investigating the possible methods to approximate our current floating-point based epipolar geometry processing in the future.

TABLE XII
BIT-RATE INCREASE RATIO OF EGB WITH THE NOISY FM ESTIMATION ($\mu_d = 4, 6$) COMPARED WITH THE REFERENCE EGB ($\mu_d = 0$)

Bit-rate increase ratio	$\mu_d = 6$ VSR = 4	$\mu_d = 6$ VSR = 4	$\mu_d = 6$ VSR = 6
Breakdancer	1.42%	5.12%	4.04%
Ballet	1.83%	2.76%	2.28%
Dinosaur	1.84%	3.12%	2.20%

VII. CONCLUSION

A novel and effective DE technique based on epipolar geometry is proposed for speedy multiview image and video coding. The principle of epipolar geometry and the optimal DV distribution histograms presented at the beginning place a solid foundation for this paper. As a result, by employing the epipolar constraint, the proposed fast DE algorithm can largely reduce the computational cost while maintaining the high coding efficiency. Because the basic framework of the proposed epipolar geometry-based fast DE technique is a generic model, various conventional fast ME algorithms can actually be integrated to leverage its strength in interviewpoint disparity prediction. Furthermore, the proposed DE method can tolerate a relatively large FM estimation noise, and the encoded bitstream is fully compatible with the H.264/AVC syntax, both of which assure its real applications. To the best of our knowledge, this is the first attempt at explicitly employing solid geometry constraints in an advanced video coding standard.

Future directions may include adaptively choosing HSR and VSR to achieve a smart complexity-scalable encoder. Implementing floating-point-based epipolar line calculation with integer arithmetic may further improve run-time performances.

ACKNOWLEDGMENT

The authors would like to thank D. Daniels for his editing of this paper.

REFERENCES

[1] J.-G. Lou, H. Cai, and J. Li, "A real-time interactive multi-view video system," in *Proc. ACM Int. Conf. Multimedia*, Singapore, Nov. 2005, pp. 161–170.

[2] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video interpolation using a layered representation," in *Proc. ACM Conf. Comput. Graphics*, 2004, pp. 600–608.

[3] A. Smolić and D. McCutchen, "3DAV exploration of video-based rendering technology in MPEG," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 3, pp. 348–356, Mar. 2004.

[4] A. Smolić and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proc. IEEE*, vol. 93, Special Issue on Advances in Video Coding and Delivery, no. 1, pp. 98–110, Jan. 2005.

[5] *Advanced Video Coding for Generic Audio-Visual Services*, ISO/IEC 14996-10 AVC, Int. Telecommun. Union-Telecommun. (ITU-T) and Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC), 2003, Recommendation H.264 and.

[6] *Survey of Algorithms Used for Multi-View Video Coding (MVC)*, ISO/IEC JTC1/SC29/WG11, Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC), Jan. 2005, Doc. N6909.

[7] Z. B. Chen, P. Zhou, and Y. He, "Fast integer pel and fractional pel motion estimation for JVT," presented at the JVT-F017, 6th Meeting, Awaji, Japan, Dec. 2002, unpublished.

[8] L. M. Po and W. C. Ma, "A novel four-step search algorithm for fast block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 313–317, Jun. 1996.

[9] S. Zhu and K.-K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 287–290, Feb. 2000.

[10] A. Chimienti, C. Ferraris, and D. Pau, "A complexity-bounded motion estimation algorithm," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 387–392, Apr. 2002.

[11] Y. Nie and K.-K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 11, no. 12, pp. 1442–1449, Dec. 2002.

[12] H. F. Ates and Y. Altunbasak, "SAD reuse in hierarchical motion estimation for the H.264 encoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, Mar. 2005, pp. 905–908.

[13] H.264/AVC Reference Software ver. JM version 10.1e [Online]. Available: http://www.iphome.hhi.de/suehring/tml/download/old_jm/jm10.1.zip

[14] H. Aydinoglu and M. H. Hayes, "Compression of multi-view images," in *Proc. IEEE Int. Conf. Image Process.*, Austin, TX, Nov. 1994, pp. 385–389.

[15] J. López, J. H. Kim, A. Ortega, and G. Chen, "Block-based illumination compensation and search techniques for multiview video coding," presented at the Picture Coding Symp., San Francisco, CA, 2004, unpublished.

[16] R. Hartley and A. Zisserman*, Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[17] K. Hata and M. Etoh, "Epipolar geometry estimation and its application to image coding," in *Proc. IEEE Int. Conf. Image Process.*, Kobe, Japan, Oct. 1999, pp. 472–476.

[18] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View synthesis for multiview video compression," presented at the Picture Coding Symp., Beijing, China, 2006.

[19] S. Yea, J. Oh, S. Ince, E. Martinian, and A. Vetro, "Report on core experiment CE3 of multiview coding," presented at the JVT-T123, 20th Meeting, Klagenfurt, Austria, Jul. 2006.

[20] J. Lu, H. Cai, J.-G. Lou, and J. Li, "An effective epipolar geometry assisted motion-estimation technique for multi-view image coding," in *Proc. IEEE Int. Conf. Image Process.*, Atlanta, GA, Oct. 2006.

[21] Z. Zhang, "Determine the epipolar geometry and its uncertainty: A review," *Int. J. Comput. Vis.*, vol. 27, pp. 161–195, Mar. 1998.

[22] *Requirements on Multi-View Video Coding v.6,*, ISO/IEC JTC1/SC29/WG11, Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC), Apr. 2006, Doc. N8064.

[23] L. Yang, K. Yu, J. Li, and S. Li, "An effective variable block-size early termination algorithm for H.264 video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 784–788, Jun. 2005.

[24] K.-K. Ma and G. Qiu, "Unequal-arm adaptive rood pattern search for fast block-matching motion estimation in the JVT/H.26L," in *Proc. IEEE Int. Conf. Image Process.*, Barcelona, Spain, Sep. 2003, pp. 901–904.

[25] MSR 3-D Video Sequences [Online]. Available: http://www.research.microsoft.com/vision/InteractiveVisualMediaGroup/3DVideoDownload/

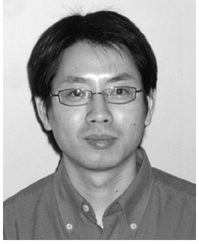
[26] Dinosaur Sequence From University of Hannover [Online]. Available: http://www.robots.ox.ac.uk/vgg/data1.html

[27] *Call for Proposals on Multi-View Video Coding*, ISO/IEC JTC1/SC29/WG11, Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC), Jul. 2005, Doc. N7327.

[28] *Information Technology-Coding of Audio-Visual Objects, Part 1: Systems, Part 2: Visual, Part 3: Audio*, ISO/IEC JTC1/SC29/WG11, Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC), 1998, FCD 14496.

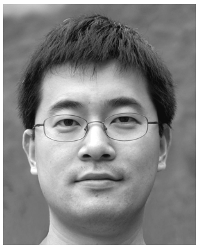
[29] R. I. Hartley, "In the defense of eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 6, pp. 580–593, Jun. 1997.

[30] Z. Zhang and C. Loop, "Estimating the fundamental matrix by transforming image points in projective space," *Comput. Vis. Image Understanding*, vol. 82, pp. 174–180, May 2001.



Jiangbo Lu (S'06) received the B.S. and M.S. degrees from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively, both in electrical engineering. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering at University of Leuven, Leuven, Belgium.

Since October 2004, he has been a Ph.D. Researcher with the Multimedia Group, IMEC, Leuven, Belgium. From April 2003 to August 2004, he was a GPU Architecture Design Engineer with VIA-S3 Graphics, Shanghai, China. In 2002 and 2005, he was a visiting student with Microsoft Research Asia, Beijing, China. His research interests include video coding, motion analysis, multiview imaging systems, stereo vision, video-based rendering, and general-purpose GPU computing.



Hua Cai (M'04) received the B.S. degree from the Shanghai Jiaotong University, Shanghai, China, in 1999, and the Ph.D. degree from the Hong Kong University of Science and Technology (HKUST) in 2003, all in electrical and electronic engineering.

He joined Microsoft Research Asia, Beijing, China, in December 2003 and is currently a Researcher with the Media Communication Group. His research interests include digital image/video signal processing, image/video coding and transmission, multiview video systems, multiview video coding and transmission, and mobile media computing.

Jian-Guang Lou (M'04) received the B.Sc. degree and the M.Sc. degree in automation from Zhejiang University, Hangzhou, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Institution of Automation, Chinese Academy of Science, Beijing, China, in 2003.

He joined Microsoft Research Asia, Beijing, China, in 2003. His main research interests include computer vision, image processing, multiview video, and multimedia systems.

Jiang Li (SM'04) received B.S. degrees in applied physics and applied mathematics from Tsinghua University, China, in 1989, the M.S. degree in optics from the Physics Department, Zhejiang University, Hangzhou, China, in 1992, and the Ph.D. degree in applied mathematics from the State Key Laboratory of Computer Aided Design and Computer Graphics, Zhejiang University, in 1998.

He joined the Visual Computing Group, Microsoft Research Asia, Beijing, as Researcher in January 1999. He became Lead Researcher of the Internet Media Group in 2002 and Research Manager of the Media Communication Group in 2004. He invented Bi-level video, Portrait video, and Watercolor video, which are suitable to very low-bandwidth networks. He released Microsoft Portrait, the first video communication software prototype on Pocket PC and Smartphone. He is now leading the Media Communication Group in the research of mobile video communication, multiparty conferencing, multiview video, and peer-to-peer streaming. Before joining Microsoft, he was an Associate Professor with the Physics Department, Zhejiang University.