# AN EFFECTIVE EPIPOLAR GEOMETRY ASSISTED MOTION ESTIMATION TECHNIQUE FOR MULTI-VIEW IMAGE AND VIDEO CODING

*Jiangbo Lu* *, *Hua Cai* †, *Jian-Guang Lou* †, *and Jiang Li* †

* Department of ESAT, University of Leuven and Multimedia Group, IMEC, Leuven, Belgium
† Media Communication Group, Microsoft Research Asia, Beijing, China

## ABSTRACT

To efficiently encode data-intensive multi-view imaging content, conventional hybrid predictive coding methodologies choose to address the compression by exploiting temporal and inter-viewpoint redundancy. However, their key yet time-consuming component, motion estimation (ME), is usually not efficient in inter-viewpoint prediction because inter-viewpoint motion is quite different from temporal motion. In essence, inter-viewpoint correlation is subject to epipolar geometry, which provides constraints for multi-view image sequences. A fast inter-viewpoint ME technique is hence proposed in this paper to accelerate the encoding by employing epipolar geometry. Theoretical analysis and experimental results prove that the proposed ME algorithm can greatly reduce search region and effectively track large and irregular motion that is typical for convergent multi-view camera setups. As a result, compared with fast full search at large search size adopted in H.264, our proposed ME algorithm can obtain a similar coding efficiency while achieving a speedup ratio of 2.9.

***Index Terms***— Multi-view video, multi-view image, source coding, motion estimation, epipolar geometry, H.264/AVC

## 1. INTRODUCTION

Multi-view video or free viewpoint video is an exciting application, because it enables users to interactively watch a static or dynamic scene from different viewing angles. As a brand new application, it has received increasing recent attention. Generally, to provide a smooth multi-perspective viewing experience, content producers need to capture a distinct scene with ideal quality from multiple camera positions, such as the convergent multi-view camera setup in Fig. 1, where the cameras are posed inward to capture the scene from different angles. Usually, the simultaneous multiple video streams from multi-view cameras are referred to as multi-view video. A multi-view 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 the still images taken by multiple cameras at one certain time instance, so it is essentially a multi-view image sequence or a *frozen moment* sequence as named in [1].

Though multi-view 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 multi-view image/video encoder becomes an indispensable necessity. For the same reason, multi-view video coding is identified by MPEG 3D audio and video (3DAV) ad hoc group as one of the most challenging issues associated with such new applications as free viewpoint video [2], [3]. In this paper, we center our study on effectively coding multi-view captured content. More specifically, we are especially interested in coding the multi-view image and video captured by convergent multi-view 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 games, and event broadcasting. Meanwhile, the multi-view content captured from convergent viewpoints is more difficult for a multi-view image or video encoder than that captured from a parallel multi-view camera setup, which can be regarded as a simplified form of the convergent setup once the angular difference between adjacent camera's viewing directions is decreased to zero.

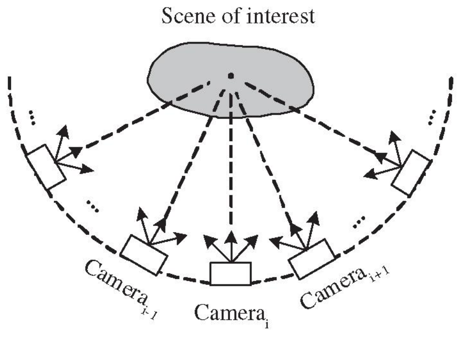

**Fig. 1**. A convergent multi-view camera setup.

While multi-view 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 [4] have already been proposed. The investigation in [5] provides clear evidence that there are technologies that significantly outperform today's available reference method (AVC simulcast). In addition to exploiting temporal redundancy to achieve coding gains, inter-viewpoint redundancy is also exploited in the schemes [5] by performing inter-viewpoint prediction across different views.

Although inter-viewpoint prediction can greatly improve the coding performance of a multi-view image or video encoder, it also significantly increases computational costs. The reason is that inter-viewpoint redundancy has to be exploited by conducting spatial ME across different views, while ME is usually the most time-consuming component in a conventional video encoder, especially when variable block-size ME is performed. For example, it has been found that variable block-size ME consumes heavy computation time of a H.264 encoder. More specifically, multi-prediction modes, multi-reference frames and higher motion vector resolution adopted in ME

* The work presented in this paper was carried out at Microsoft Research Asia, Beijing, China.

1-4244-0481-9/06/$20.00 ©2006 IEEE

of H.264 can consume 60% (1 reference frame) to 80% (5 reference frames) of total encoding time of the H.264 codec [6].

Obviously, an effective inter-viewpoint ME technique is highly desirable for most recently developed multi-view image or video encoders with hybrid temporal-viewpoint prediction structures [5]. Considering that epipolar geometry is a powerful and obtainable geometry constraint for multi-view images, we therefore propose an effective ME scheme to accelerate the inter-viewpoint prediction and coding based on knowledge of epipolar geometry. Concentrated on speeding integer pixel motion search, the proposed ME technique comprises a new starting search point prediction method and a motion refinement region reduction scheme. Independent of complicated vision algorithms and fully compatible to H.264/AVC syntax, the proposed technique promises real application. To the best of our knowledge, this is the first attempt at explicitly employing solid geometry constraints in advanced video coding standard.

Section 2 reviews the basic principles of epipolar geometry, and provides our proposal's foundation. Section 3 presents the proposed epipolar geometry assisted ME technique from a theoretical perspective, while the optimal parameter setting is derived from experiments on high-quality multi-view image sequences. More simulation results are reported in Section 4. We conclude our work and give future directions in Section 5.

## 2. BRIEF REVIEW OF EPIPOLAR GEOMETRY

Epipolar geometry, as a specific example of multi-view geometry, is the only available geometry constraint between a stereo pair of images of a single scene [7]. It has been extensively studied in computer vision. Because epipolar geometry plays a fundamental role in our proposed ME technique, we briefly review its basis before presenting our proposed methods.

Let us consider a stereo imaging setup as shown in Fig. 2, where $C_1$ and $C_2$ are the optical centers of the first and the second cameras, and the plane $I_1$ and $I_2$ are the first and the second image planes. Given a point $P$ in a 3D space, let us denote its projection on the second image plane as $p_2$. According to epipolar geometry, 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

$$\widetilde{p}_1^T F \widetilde{p}_2 = \widetilde{p}_1^T l_1 = 0, \tag{1}$$

where $\widetilde{p}_1$ and $\widetilde{p}_2$ are the homogeneous coordinates of $p_1$ and $p_2$, and $F$ is called the *fundamental matrix* (FM). It is a $3 \times 3$ matrix, determined by the intrinsic matrix and the relative position of the two cameras. Therefore, from Eqn. (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 [8]. If camera geometry is calibrated, the general case when both intrinsic and extrinsic camera parameters are known, FM can be easily calculated from camera projective matrices. In this paper, we concentrate singularly on multi-view image/video compression by assuming that reliable FM is already computed for current multi-view camera setup during a preprocessing phase.

## 3. PROPOSED EPIPOLAR GEOMETRY ASSISTED MOTION ESTIMATION TECHNIQUE

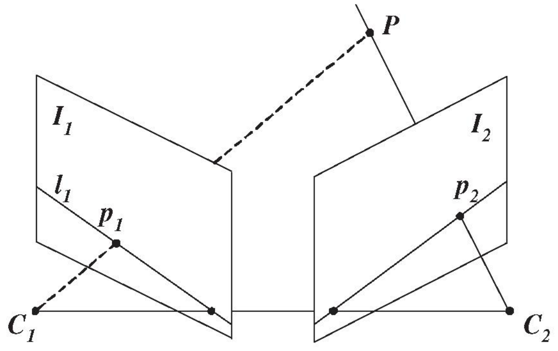

**Fig. 2**. Epipolar geometry.

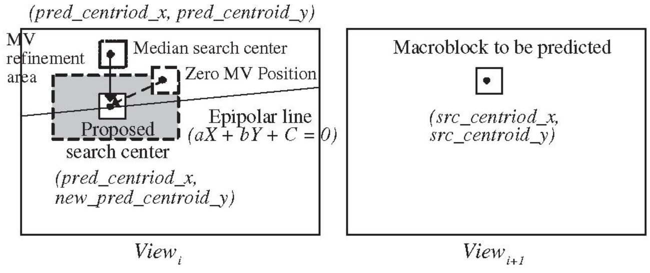
(pred_centriod_x, pred_centroid_y)
MV refinement area
Median search center
Zero MV Position
Epipolar line (aX + bY + C = 0)
Proposed search center
(pred_centriod_x, new_pred_centroid_y)
Macroblock to be predicted
(src_centriod_x, src_centroid_y)
View i
View i+1

**Fig. 3**. The proposed search center in $View_i$.

Different from traditional ME algorithms, which are used to temporally perform prediction, our epipolar geometry assisted ME technique is proposed to effectively deal with spatial prediction or inter-viewpoint prediction. Such a difference in the application scenarios actually accounts for quite different ME design principles from the traditional temporal ones. In general, temporal motion cannot be characterized in an adequate way, especially when there is sudden motion or a scene change, because the object motion or camera movement involved is not absolutely predictable. On the contrary, for a multi-view image sequence, inter-viewpoint correlation is mainly determined by the geometry relationship between the scene and multiple camera setups, so the spatial motion or essentially the disparity vector relating two adjacent views is subject to the epipolar constraint. This makes inter-viewpoint motion search effective in a predictable and reduced search space, because inter-viewpoint motion by its nature is more structured than temporal one. In fact, to track the large and irregular (depth-dependent) motion typical for convergent multi-view camera setup, traditional full-search ME and most fast-ME algorithms, such as four-step search (FSS) [9] and predictive algorithm (PA) [10], have to greatly amplify the motion refinement grid to prevent the search points from dropping into a local minimum in the earlier search stages. Our experiments show that in some situations, a refinement grid size of $\pm 16 \times \pm 8$ is still insufficient to approach a good quality and complexity tradeoff.

Exploiting the special property of epipolar geometry associated with multi-view captured images, we propose an effective ME algorithm that can track the real motion even with a largely reduced motion refinement area, so the coding efficiency is guaranteed while the ME complexity can be reduced. Our proposed ME technique consists of a new starting search point prediction method and a motion refinement region reduction scheme.

The proposed motion starting search point is based on the hypothesis that widely-used median motion vector (MV) predictor is vertically away from the epipolar line. This results in an implicit assumption that horizontal component of median predicted MV (*pred_mv_x*) is of high prediction confidence, so we only need to rectify the vertical component of median predicted MV (*pred_mv_y*) to make the proposed search center lie on the epipolar line, as shown in Fig. 3. In fact, it makes sense to assume that *pred_mv_x* is reliable enough since horizontal motion is much more intensive than vertical motion for most typical multi-view camera setups, where epipolar lines primarily have a slope angle much less than 45 degree, and horizontal motion can thus be more reliably predicted with little influence from the unstable factors. It is shown in our experiments that the correlation coefficient between horizontal components of the best MVs obtained from our proposed ME and that from fast full search (FFS) adopted in H.264 reference software [11], can still be larger than 0.70, even when the motion search region shrinks along its vertical dimension significantly in our proposed ME schemes.

In addition, since it is well-known in correspondence problem that a larger matching window is desirable to achieve reliable matching, we thus apply epipolar constraint only to motion search at a block-size of $16 \times 16$. Namely, for mode 1 in H.264, the prediction outliers from small matching windows can be kept away from destroying the smooth motion field, and also the computation increase is kept marginal (ME overhead lower than $1\%$), for epipolar constraint is solely performed at macroblock (MB) level.

We describe the starting search point computation as noted in Fig. 3. We consider two neighboring views, $View_i$ and $View_{i+1}$, and denote the input FM relating them be $F$. Given the coordinates of the centroid of current MB (*src_centroid_x*(*y*)) 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 Eqn. (1). We propose this centroid-based epipolar line calculation in that the resulting epipolar line computed from the centroid rather than the top-left corner of the MB, can more precisely reflect the average epipolar constraint for a group of pixels. After calculating the epipolar line equation, the proposed starting search point and the rectified vertical component of starting MV (*new_pred_mv_y*) for current MB can be derived from median predicted search point (*pred_centroid_x*(*y*)) as follows,

$$\begin{aligned} &pred\_centroid\_x = src\_centroid\_x + pred\_mv\_x \\ &new\_pred\_centroid\_y = (a \times pred\_centroid\_x + c)/-b \qquad (2) \\ &new\_pred\_mv\_y = new\_pred\_centroid\_y - src\_centroid\_y. \end{aligned}$$

Considering that in usual situations most epipolar lines have fairly small slope angles, we devise two reduced motion search spaces centered on the epipolar geometry rectified start search point (illustrated in Fig. 4). The first proposed search space is a rectangle area with the horizontal search range (HSR) larger than the vertical search range (VSR), and the other one is a parallelogram aligned to epipolar line with the same parameters. The motion search is performed in a center-biased order, e.g., from 0 to $e$ when VSR is 2, as shown in Fig. 4. To guarantee this center-biased MV selection, when a few candidates give the same rate-distortion (R-D) cost, we also measure their Euclidean distances to the center and then select and record the one with the shortest distance. When traversing the parallelogram search space, an incremental computation of $y$ coordinate is adopted instead of deriving current $y$ directly from Eqn. (2), i.e., the immediate previous $y$ coordinate is recorded, so we only need one floating-point addition of the constant $-a/b$ to get current $y$ coordinate.

Fig. 5 shows the typical percentage increase of the R-D optimized cost in H.264 and the SAD (sum of absolute differences) versus the VSR value for our proposed ME algorithm in comparison to FFS (with a search range $\pm 16 \times \pm 16$), and the best trade-off between computational saving and performance loss can be identified. For example, VSR = 4 can be chosen to reduce the complexity while achieving a decent coding performance.

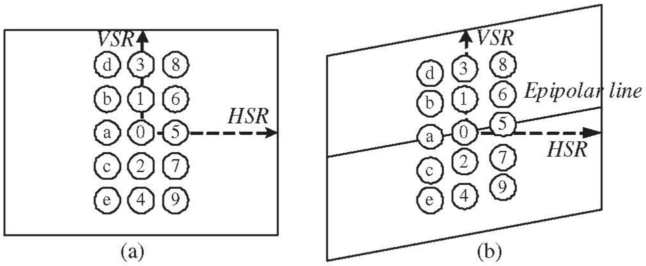

**Fig. 4**. Two proposed search spaces with search order.

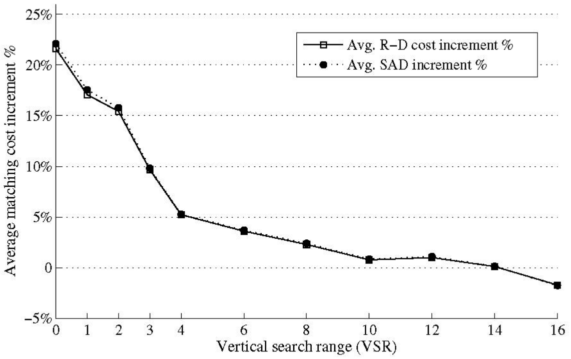

**Fig. 5**. Average matching cost increment % versus VSR.

As an effective and generally applicable ME technique, the proposed scheme can work well with full search and fast ME algorithms for multi-view image/video coding to achieve a good tradeoff between quality and complexity. As a specific example, the proposed technique is applied to FFS to produce two fast variants with different search spaces (Fig. 4). We choose to abbreviate the first variant in rectangle search shape to RFFS, and the parallelogram one to PFFS.

In our future work, we plan to derive thresholds to adaptively switch among different VSR values, which will depend on the precision of FM, image resolution, camera setup, noise intensity and other factors.

## 4. EXPERIMENTAL RESULTS

Our experiments are based on the latest JM version 10.1 of H.264 reference software [11]. Baseline profile is used to configure the encoder. We set the number of reference frames to 1. All frames except for the first one are encoded as P-frames. R-D optimization and CAVLC entropy encoding are enabled. For our proposed fast ME algorithms, the HSR is set to 16, while VSR is reduced to 4. We adopt the recent *Breakdancer* and *Ballet* multi-view video sequences from Microsoft Research (MSR) in our experiments, which are captured in high-quality ($1024 \times 768$) by arranging eight cameras on an arc that were adjusted with precise camera parameters [12]. All of the experiments were carried out on a computer with a Intel Pentium 4 processor.

We encode the first multi-view images of *Breakdancer* and the second ones of *Ballet* with different QP values to compare the R-D performance and average integer pixel ME speed of FFS (with a search range ±16×±16), hybrid unsymmetrical-cross multihexagon-grid search (UMHexagonS) [6], as a fast ME algorithm adopted in H.264, and the proposed RFSS (VSR=4) and PFSS (VSR=4). Table 1 and 2 summarize the results for *Breakdancer* and *Ballet*, where for the convenience of comparison, ΔPSNR, ΔBitrate, and ΔSpeed represent the average PSNR gains (dB), the average bitrate increase rate (%), and the average integer pixel ME speed-up ratio, when they are compared respectively with the absolute performance of FFS in the third column. For both sequences, the proposed RFSS (VSR=4) and PFSS (VSR=4), only cause a negligible PSNR degradation of 0.02 dB on average, but can achieve a speedup factor of about 2.9 in comparison to FFS in the entire range of bitrate.

We have also modified UMHexagonS by integrating our proposed epipolar-geometry rectified search center and reducing VSR to 4. Compared with FFS again, the modified UMHexagonS can achieve an average integer pixel ME speedup by a factor of 3.79 at a PSNR loss of 0.04 dB for *Breakdancer*, and a factor of 3.48 at a PSNR loss of 0.02 dB for *Ballet*.

**Table 1**. R-D performance and average integer pixel ME speed comparison for *Breakdancer*.

| QP | Performance | FFS | UMHexagonS | RFSS (VSR=4) | PFSS (VSR=4) |
|---|---|---|---|---|---|
| 24 | ΔPSNR | 39.59 | 0.00 | 0.01 | 0.01 |
| | ΔBitrate | 6,942.96 | 0.70% | 2.26% | 2.24% |
| | ΔSpeed | 1.00 | 2.23 | 2.75 | 2.82 |
| 28 | ΔPSNR | 38.24 | -0.01 | -0.01 | -0.01 |
| | ΔBitrate | 2,915.14 | -0.47% | 2.21% | 1.86% |
| | ΔSpeed | 1.00 | 2.58 | 2.79 | 3.00 |
| 32 | ΔPSNR | 37.10 | -0.03 | -0.03 | -0.03 |
| | ΔBitrate | 1,459.03 | 1.88% | 2.90% | 3.97% |
| | ΔSpeed | 1.00 | 3.02 | 2.84 | 3.12 |
| 36 | ΔPSNR | 35.84 | -0.06 | -0.05 | -0.05 |
| | ΔBitrate | 887.21 | -0.40% | 1.67% | 1.30% |
| | ΔSpeed | 1.00 | 3.58 | 3.04 | 2.76 |
| Avg. | ΔPSNR | 0.00 | -0.03 | -0.02 | -0.02 |
| | ΔBitrate | 0.00% | 0.43% | 2.26% | 2.34% |
| | ΔSpeed | 1.00 | 2.85 | 2.86 | 2.93 |

**Table 2**. R-D performance and average integer pixel ME speed comparison for *Ballet*.

| QP | Performance | FFS | UMHexagonS | RFSS (VSR=4) | PFSS (VSR=4) |
|---|---|---|---|---|---|
| 24 | ΔPSNR | 41.17 | 0.00 | 0.00 | 0.00 |
| | ΔBitrate | 5,287.37 | 0.76% | 2.14% | 1.91% |
| | ΔSpeed | 1.00 | 2.15 | 3.02 | 2.66 |
| 28 | ΔPSNR | 39.94 | -0.02 | -0.01 | -0.01 |
| | ΔBitrate | 3,236.30 | -0.88% | 0.82% | -0.07% |
| | ΔSpeed | 1.00 | 2.62 | 3.06 | 2.83 |
| 32 | ΔPSNR | 38.38 | -0.04 | -0.03 | -0.03 |
| | ΔBitrate | 2,002.94 | -0.11% | 3.17% | 3.60% |
| | ΔSpeed | 1.00 | 2.73 | 2.65 | 2.90 |
| 36 | ΔPSNR | 36.61 | -0.03 | -0.03 | -0.03 |
| | ΔBitrate | 1,315.54 | 1.26% | 2.66% | 1.83% |
| | ΔSpeed | 1.00 | 3.22 | 2.86 | 3.12 |
| Avg. | ΔPSNR | 0.00 | -0.02 | -0.02 | -0.02 |
| | ΔBitrate | 0.00% | 0.26% | 2.20% | 1.82% |
| | ΔSpeed | 1.00 | 2.68 | 2.89 | 2.88 |

## 5. CONCLUSIONS AND FUTURE WORKS

An effective ME technique based on epipolar constraint is proposed for speedy multi-view image and video coding. By employing epipolar constraints, we can greatly reduce the search range to accelerate the encoding process. This technique can work with almost any ME algorithms, and the encoded bitstream is fully compatible with H.264 syntax.

Future directions may include adaptively choosing VSR to achieve a smart complexity-scalable encoder and investigating the impact of FM's precision on MV prediction. Implementing floating-point based epipolar line calculation with integer arithmetic may further improve run-time performances.

## 6. REFERENCES

[1] Jian-Guang Lou, Hua Cai, and Jiang Li, "A real-time interactive multi-view video system," in *The 13th ACM International Conference on Multimedia (ACMMM 2005)*, pp. 161-170, Singapore, Nov. 2005.

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

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

[4] "Advanced video coding for generic audio-visual services," Int. Telecommun. Union-Telecommun. (ITU-T) and Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC), Recommendation H.264 and ISO/IEC 14996-10 AVC, 2003.

[5] ISO/IEC JTC1/SC29/WG11, "Survey of Algorithms Used for Multi-view Video Coding (MVC)," Doc. N6909, Hong Kong, China, January 2005.

[6] Z.B. Chen, P.Zhou, and Y.He, "Fast Integer Pel and Fractional Pel Motion Estimation for JVT," JVT-F017, 6th Meeting, Awaji, Japan, Dec. 5-13, 2002.

[7] R.Hartley and A.Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, UK, 2000.

[8] Z. Zhang, "Determine the Epipolar Geometry and Its Uncertainty: A Review," *International Journal of Computer Vision*, Kluwer Publ., vol. 27, no. 2, pp. 161-195, Mar. 1998.

[9] L.M. Po and W.C. Ma, "A Novel Four-Step Search Algorithm for Fast Block Motion Estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.6, no.3, pp.313-317, June 1996.

[10] A. Chimienti, C. Ferraris, and D. Pau, "A Complexity-Bounded Motion Estimation Algorithm," *IEEE Transactions on Image Processing*, vol.11, no.4, pp.387-392, April 2002.

[11] JM Reference Software Version 10.1. [Online]. Available: http://iphome.hhi.de/suehring/tml/download/jm10.1.zip.

[12] MSR 3D Video Sequences. [On-line]. Available: http://www.research.microsoft.com/vision/ImageBasedRealities/3DVideoDownload/.