

# FAST MOTION ESTIMATION WITH DUAL SEARCH WINDOW FOR STEREO 3D VIDEO ENCODING

<sup>1</sup>Michal Joachimiak, <sup>2</sup>Kemal Ugur

<sup>2</sup>Jani Lainema, <sup>1</sup>Moncef Gabbouj

<sup>1</sup>Dept. of Signal Processing,  
Tampere University of Technology,  
Tampere, Finland

<sup>2</sup>Nokia Research Center,  
Nokia Corp.,  
Tampere, Finland

## ABSTRACT

Stereoscopic 3D video is becoming a reality in many application areas, ranging from high quality entertainment to mobile video services. Due to the need to process two views, the complexity of 3D video applications is significantly higher than traditional 2D counterparts. In order to enable real-time 3D video services in mobile devices, this paper proposes a novel algorithm which reduces the complexity of stereo video encoding with improvement of coding efficiency. A novel search window center prediction method is proposed that exploits the correlation between two views. Experimental results show that the average encoding time of the second view can be decreased by 80% with an increase in coding efficiency of up to 2%. The state-of-art fast motion estimation methods for stereoscopic 3D video encoding show coding efficiency decrease, whereas proposed method achieves the speed-up with increase in coding efficiency, making it suitable for high quality 3D video applications.

**Index Terms**— MVC, search center prediction, fast motion estimation, interview correlation

## 1. INTRODUCTION

With advances of capture, encoding and display technologies, the 3D video is reaching the consumer domain. The acquisition setup for the 3D video encoding comprises of cameras observing the scene at slightly different angles. The cameras have partially overlapping fields of view. The video sequences, corresponding to different camera viewpoints, named shortly views, are bundled into a single 3D video sequence. In Multiview Video Coding (MVC) the base view is the view for which the only possible prediction is the one using the reference frames in the same view. Views which can refer to other views are called auxiliary. In the stereoscopic 3D video coding two views are constructed in a way that camera optical axes are parallel. Stereoscopic camera records the same scene using two aligned sensors. Since the left and right view images are destined for the reception at the same time instance it is necessary to perform an inter-camera synchronization during capturing. Also color and illumination compensa-

tion has to be implemented. These requirements, impose a high spatial correlation between views at the same temporal location. The correlation is inversely proportional to the distance between cameras.

The 3D video encoding standard called MVC was recently published by the JVT of ITU-T and MPEG as an extension of H.264/AVC [1]. It utilizes the same inter-frame prediction based on motion estimation representation as H.264/AVC, which means it supports multiple reference frames and variable macroblock sizes with tree-structured sub-macroblock partitions. Additionally, disparity estimation is possible. The encoding of MVC results in great computational complexity increase due to processing of auxiliary views. For this reason, similarly to the H.264/AVC video encoding, one of the major computationally complex parts of the stereoscopic 3D video encoding is the motion estimation for optimal macroblock mode selection done by means of cost calculation for rate-distortion optimization (RDO). This exhaustive operation can take up to 50-80% [2] of the total encoding time.

The interview correlation can be advantageous for fast motion estimation and fast disparity estimation of the auxiliary views. A range of strategies, that utilize the interview correlation, for reducing the amount of operations for auxiliary views can be found in the literature. They can be divided into methods which select only few motion vectors to search for the match, the methods which select only a subset of modes to search, and the methods which perform a search range decrease to limit the number of possible matches in the search window. They are analyzed in detail in section 2. For all methods, the size of the set of potential macroblock matches searched in the auxiliary view is substantially decreased, comparing to the full macroblock search. Thus, the majority of fast motion estimation methods share a common tradeoff which is the loss in coding efficiency.

The proposed fast motion estimation method can be assigned to the third type of methods. It performs motion vector search over two smaller windows. The coordinates of the center of the first window are calculated as a median from the coordinates of the vectors of neighboring macroblocks. The

center of the second window is predicted using the motion information from the base view. Thanks to this prediction method, in opposite to other fast motion estimation methods for multiview video coding, together with the speedup, the coding efficiency can be increased. The proposed method can be used with fast disparity estimation methods to improve encoding speed gain and coding efficiency at the same time. The method [2] is implemented and tested for comparison.

The paper is organized as follows. In section 2 the complexity of disparity and motion estimation, together with existing approaches to reduce it, is analyzed. In section 3, a novel search center prediction method together with dual search window algorithm is presented. Section 4 describes the results and section 5 concludes the paper.

## 2. COMPLEXITY OF DISPARITY AND MOTION ESTIMATION

The motion estimation (ME) is an inter-frame macroblock prediction tool. In the single-view video coding, where only temporal motion is used, the motion information depends mainly on moving objects in the observed scene or on the camera movement. In multiview video coding disparity estimation is used additionally. It is different from motion estimation as it depends on the depth of the object in the scene and the physical camera adjustment. For objects close to the camera setup the disparity is high, for objects more distant it is low. In the situation when the motion of the macroblock in temporal direction is close to zero, the disparity of the same macroblock does not have to exhibit the same property. For both motion estimation types the rate-distortion optimization (RDO) algorithm is used to select the best macroblock mode and the best reference frame. The amount of macroblocks to check in the search area, the number of reference frames, as well as multiple views are the factors causing the computational complexity increase.

In the literature three main strategies for motion estimation complexity reduction in the auxiliary views of the 3D video, based on interview correlation, can be found. The first is to calculate or explicitly choose the motion vector from the set of the previously found motion vectors. The second is based on decreasing the number of macroblock modes tested during fast search. The last one is to decrease the size of the search window, so the number of potential macroblock matches is substantially lower.

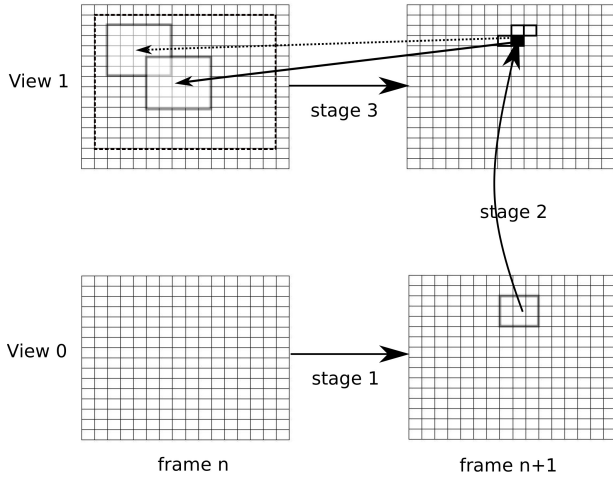
An example of the first method can be found in [3]. The set of potential motion vectors is created out of the motion vectors of the neighboring macroblocks in the same view and motion vectors of the macroblocks in the neighborhood of the corresponding macroblock in the base view. The set of five initial guesses is created. From these initial predictions one is selected using RDO. The refining search over a very small search window is possible. The same principle can be used also for disparity estimation. A method which reuses dis-

parity vectors from the same frame, from the corresponding frame in the neighboring view and from the previous frame in the same view can be found in [4]. Similarly to [3] the refining search is performed.

The second approach is to perform macroblock mode candidates pre-selection. When early decision can be made, the motion estimation is either not done, which is the case for early SKIP or DIRECT modes, or is performed for a limited set of macroblock modes. In certain conditions ([5]) it is possible to reduce the set of modes according to the direction of prediction, either temporal or interview, found during full macroblock search for mode 16x16. In [6] the mode pre-decision is based on modes selected for spatially neighboring macroblocks in the same view and corresponding macroblock in the neighboring view. For encoding speed increase, the algorithm can decide to select SKIP mode and do not check other modes. When the early SKIP decision can not be made, the vectors in auxiliary views are not searched but reused from the base view. In [7] for every macroblock in auxiliary view, the mode types of the corresponding macroblock in the base view and its neighboring macroblocks are used to infer the set of modes to check.

Third solution is to decrease the size of the search window in the auxiliary view. In this case, the most problematic is an accurate prediction of the initial motion vector, which is set as the center of the search window. It is possible to infer the initial motion vector from the motion information for the base view. When accurate prediction can be found, it is possible to minimize the negative effect of the search window size decrease, which is the drop in coding efficiency. The resulting search range is typically a fraction of the range used for the base view. The division factor can be estimated using the motion information stored during the full search in the base view. Motion field can be used to assign a macroblock to the motion type. Based on that, the division factor can be decided [7]. Adaptive, horizontal and vertical search range reduction for disparity estimation can be found in [8]. An example of the search range decrease based on physical camera setup is shown in [9]. The disadvantage in this case is the need to know the camera geometry in advance. An adaptive search range adjustment for disparity estimation, based on disparity mode is done in [2]. In this approach the amount of motion between views has influence on the fraction by which the search range is decreased. Also the search window center is shifted to the direction estimated by the global disparity.

All methods for fast motion estimation and majority of methods for fast disparity estimation, based on the search range reduction, exhibit coding efficiency drop. The exception is fast disparity estimation by Zhu et al in [8]. The proposed method is able to speedup the ME process during coding the auxiliary view of the stereoscopic 3D, with coding efficiency increase at the same time.

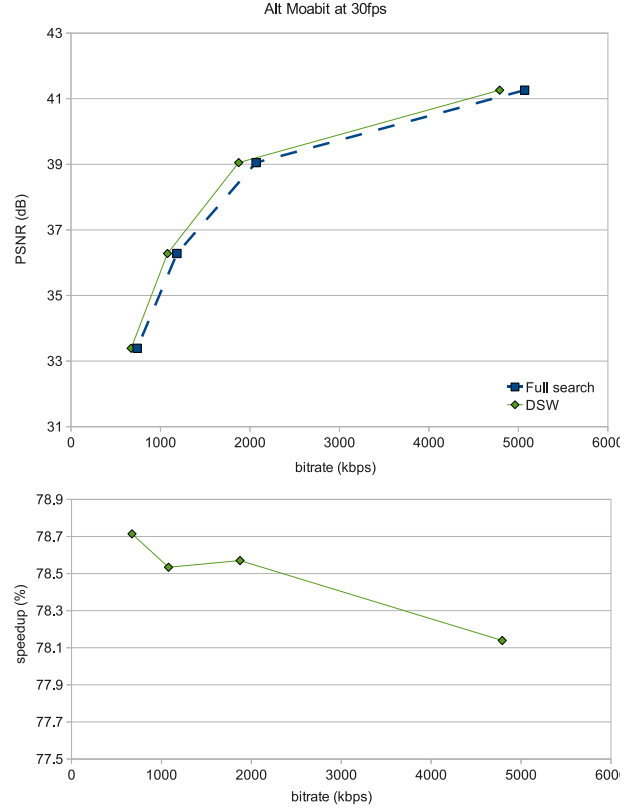


**Fig. 1.** Dual search window algorithm. Squares represent blocks in video frames. Instead of bigger search window (drawn using dashed line) two small search windows are used. Center for the second search window (pointed by the dashed line) is predicted using motion information from the base view.

### 3. PROPOSED DUAL SEARCH WINDOW ALGORITHM

Since video sequences provided to the stereoscopic encoder come from two neighboring views, the temporal motion in the left view is similar to the temporal motion in the right view. During full macroblock search, the RDO procedure selects (by means of Lagrangian optimization) one macroblock mode with corresponding motion vector pointing to the best matched macroblock in the reference frame. Exploiting the temporal motion similarity between views, it is possible to reuse the motion field calculated in the base view, during ME in the secondary view.

For every macroblock in the second view, the following steps are performed. Motion vectors for the set of selected modes (16x16, 16x8, 8x16 and 8x8), in the base view, are stored (stage 1 on Fig. 1). For each candidate motion partition, we compute two search window centers. First one is computed by taking the median value of the motion vectors of neighboring macroblocks (upper, upper-right, left) in the same view, and it is called the spatial motion center. The second search window center is computed by taking the median value of the motion vectors in a 3x3 neighboring window of the corresponding macroblock in the base view, and it is called the view motion center (stage 2 on Fig. 1). The motion estimation in the auxiliary view (stage 3 on Fig. 1) is performed over two search windows. To increase the encoding speed, the size of the windows is decreased. For the computational complexity simulations the division factor of 4 was tested for the search range reduction.



**Fig. 2.** Speedup and corresponding rate-distortion plot for Alt Moabit sequence.

To prevent double searching the same area, in case when two search windows are overlapping, a window overlap detection mechanism was implemented.

### 4. EXPERIMENTAL RESULTS

The proposed algorithms were implemented on the publicly available stereoscopic MVC encoder [10]. In addition to the dual search window algorithm (DSW), the ISRP algorithm [2] was implemented in the encoder to have a comparison to the different fast estimation method. The experiments are conducted by following the common test conditions for MVC defined by JVT [11]. The search range for the anchor was set to 64. Two reference frames are used per view. The sequences are encoded in low delay mode, using one I-frame per view followed by P frames. Tree-structured hierarchical motion partitions were enabled as defined in the H.264/AVC standard. The same simulation conditions are used for DSW and ISRP. The only difference is the range adjustment decided by these algorithms. Video sequences selected for tests are “Akko & Kayo” and “Rena” from Nagoya University, “Book Arrival”, “Alt Moabit”, “Car” and

**Table 1.** Comparison of the average speedup for the secondary view. The results in columns correspond to proposed DSW algorithm, ISRP implementation [2], both methods run at the same time, DSW run with interview prediction off.

sequence	proposed DSW[%]	ISRP [2][%]	DSW+ISRP[%]	DSW + interview OFF[%]
Akko & Kayo	-40.94	-43.14	-60.26	-79.98
Book Arrival	-44.67	-44.39	-61.08	-78.76
Car	-47.24	-40.68	-62.30	-80.09
Hands	-46.23	-42.80	-66.58	-82.96
Alt Moabit	-44.40	-45.50	-61.22	-78.49
Rena	-48.03	-42.35	-58.72	-79.50
<b>Average</b>	-45.25	-43.15	-61.69	-79.96

**Table 2.** Bjontegaard delta rate results. The anchor uses 64x64 search range.

sequence	proposed DSW[%]	ISRP [2][%]	DSW+ISRP[%]	DSW + interview OFF[%]
Akko & Kayo	-0.94	0.82	0.60	-0.39
Book Arrival	-0.16	0.73	0.58	-0.59
Car	-1.54	-0.15	-0.24	-1.36
Hands	-0.40	0.14	0.06	-0.29
Alt Moabit	-1.31	-0.19	-0.66	0.92
Rena	-1.01	0.01	-0.25	-0.54
<b>Average</b>	-0.89	0.23	0.02	-0.38

“Hands” provided by Heinrich Hertz Institute. Two views out of every sequence are chosen for the experiment. The set is composed out of sequences characterized by high global motion (“Akko&Kayo”, “Car”, “Alt Moabit”, “Rena”), the sequence with small local motion (“Book Arrival”), and the sequence with high local irregularities (“Hands”). The Table 1 shows the encoding speedup for the second view corresponding to the sequences. The speedups are averaged over four QPs used (22,27,32,37). However, as can be seen from the exemplary speedup plot on Fig. 2, the standard deviation over QPs is not high. The columns in the Table 1 correspond to the speedup for different configuration scenarios. Correspondingly: only DSW is used, only ISRP is used, both DSW and ISRP are used, and DSW with interview prediction turned off (disparity estimation not used). The simulations show that the speed performance of both algorithms is approximately the same. However, they can be combined to implement more efficient coding with higher speed. Additionally, when the disparity estimation is not used, the speedup for DSW algorithm can be doubled. The anchor in these tests was the full macroblock search for base and secondary view with search range 64 for both views. It can be seen that the proposed method can improve coding efficiency together with encoding speed improvement. For scenario when the interview prediction is used the speed gain is approximately the same to the ISRP, however the Bjontegaard delta bitrate results show that the difference of coding efficiency between the proposed DSW algorithm and the ISRP is more than 1% favoring DSW (see the first and second column in the Table 2). The results from simulations, where both DSW and ISRP were used, are

**Table 3.** Bjontegaard delta rate results for the the DSW where search range for the anchor was decreased to 16x16.

sequence	DSW[%]	DSW+interview OFF[%]
Akko & Kayo	-0.89	-0.48
Book Arrival	-0.18	-0.80
Car	-1.54	-1.40
Hands	-0.28	-0.19
Alt Moabit	-1.41	-8.71
Rena	-1.06	-0.59
<b>Average</b>	-0.89	-2.03

shown. The average computational time saving is more than 60% and the average Bjontegaard delta bitrate is close to zero.

The Table 2 presents the Bjontegaard delta bitrate results for four scenarios. first column corresponds to the DSW algorithm, second to the ISRP, third to the scenario where both were used. The last one displays the results for DSW algorithm used with interview prediction disabled. It can be seen from the Table 2 that in the case of DSW the most coding efficiency is gained for sequences with high global motion. However for other sequences there is still benefit. For all implementations for which results are shown in the tables 1 and 2 the initial motion search range was set to 64. The same range was used for the anchor. In the case of DSW the range reduction to 16, was applied for auxiliary view. For the ISRP its native range reduction method was used.

The second simulation scenario tested was with interview prediction disabled. The DSW algorithm was used for two

reference frames in the temporal direction. The speedup for the secondary view is shown in the last column of the Table 1. The anchor was also the full macroblock search for both views with range 64. Additionally, interview prediction was turned off. The average speedup is approximately 80%.

The last test suite was implemented to show how the coding efficiency is influenced by the new prediction method. The simulations were run against the anchor for which the search range in the secondary view was reduced by the same amount as it is done in DSW algorithm (results in the Table 3). For the first column in the Table 3 two reference frames were used. The first frame corresponding to the temporal direction is processed by DSW. For the second one corresponding to the interview direction the disparity estimation is done in the standard way. Because the search range reduction is made only for the first reference frame, the same reduction is performed on the anchor. The results in the second column are obtained for the DSW algorithm performed on two reference frames in temporal direction. The search range was reduced for both reference frames for the anchor as well. To show the maximum possible encoding speed increase, the interview prediction was turned off, as the disparity estimation is not changed by the DSW. The anchor measurements were done with disparity estimation turned off as well. From the second column in the Table 3 it can be seen that applying DSW on more reference frames improves the coding efficiency even further. The high coding efficiency increase in the case of "Alt Moabit" is caused by very high motion present in this sequence.

The measurements of the performance of the DSW algorithm show that the overhead of data collection and median calculation costs about 5% of the speedup in case of two reference frames processed by the DSW (i.e. without DSW the speedup would be 85%). For the DSW applied for a single reference frame it is about 2.5% of the speedup.

## 5. CONCLUSIONS

In order to enable high quality real-time stereo video services on resource constrained devices, a novel method for the search range center prediction in motion estimation for the secondary view is presented. The proposed method assumes that the temporal motion of both views are highly correlated as they describe the same scene but from slightly different viewpoints. This assumption is exploited by reusing the motion search results of the base view in the motion search of the auxiliary view. The motion search is performed on two smaller windows with reduced size, enabling encoding speedup. In addition to the standard search window center prediction method used for the first window in the auxiliary view, the center of the second motion search window is chosen adaptively based on the encoding results of the first view. The proposed method can be used with other search range reduction methods (e.g. [2]) to improve the overall coding efficiency. Experimental results show that average encoding time

of the second view can be decreased by 80% while increasing the coding efficiency by 2%.

## 6. REFERENCES

- [1] *Advanced Video Coding for General Audio-Visual Services* ITU-T Rec. H.264 - ISO/IEC 14496-10, 2009.
- [2] X. Xu, Y. He, "Fast Disparity Motion Estimation in MVC Based on Range Prediction," in *15th IEEE International Conference on Image Processing*, San Diego, USA, 2008, pp. 2000-2003.
- [3] L. Ding, P. Tsung, W. Chen, S. Chien and L. Chen, "Fast Motion Estimation with Inter-view Motion Vector Prediction for Stereo and Multiview Video Coding," in *Acoustics, Speech and Signal Processing*, Las Vegas, 2008, pp. 1373-1376.
- [4] P. He, M. Yu, Z. Peng, G. Jiang, "Fast Mode Selection and Disparity Estimation for Multiview Video Coding," in *Third Int. Symposium on Intelligent Information Technology Application Workshops*, IITAW '09, Nanchang, 2009, p. 209.
- [5] W. Zhu, W. Jiang, and Y. Chen, "A Fast Inter Mode Decision for Multiview Video Coding," in *International Conference on Information Engineering and Computer Science*, ICIECS 2009, Wuchan, Dec. 2009, p. 1.
- [6] L. Ding, P. Tsung, et al., "Computation-Free Motion Estimation with Inter-View Mode Decision for multiview Video Coding," in *3DTV Conference*, Kos Island, 2007, p.1.
- [7] L. She, Z. Liu, T. Yan, Z. Zhang, and P. An, "View-Adaptive Motion Estimation and Disparity Estimation for Low Complexity Multiview Video Coding," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 20, p. 925, Jun. 2010.
- [8] W. Zhu, X. Tian, F. Zhou and Y. Chen, "Fast Disparity Estimation Using Spatio-Temporal Correlation of Disparity Field for Multiview Video Coding," *IEEE Trans. Consum. Electron.*, vol. 56, pp. 957-964, May 2010.
- [9] Y. Kim, J. Kim, and K. Sohn, "Fast Disparity and Motion Estimation for Multi-View Video Coding," *IEEE Trans. Consum. Electron.*, vol. 53, p.712, May 2007.
- [10] "Nokia's MVC Software," Internet: <http://research.nokia.com/page/4988>, Nov. 2010 [Feb. 18,2011].
- [11] Y. Su, A. Vetro, A. Smolic, *Common Test Conditions For Multiview Video Coding*, ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/SG16, Doc. JVT-U211, Oct. 2006.