Predictive fast motion/disparity search for multiview video coding

PoLin Lai and Antonio Ortega

Signal and Image Processing Institute, Dept. of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564

ABSTRACT

We consider the problem of complexity reduction in motion/disparity estimation for multiview video coding. We propose predictive fast search algorithms that, after either the motion field or the disparity field has been estimated, obtain with low complexity a good set of candidate vectors for the other field. The proposed scheme performs predictive motion search from view to view and predictive disparity search from one time instant to another time. We also propose an efficient search pattern that starts with the candidate vectors from the proposed algorithms. Simulation results show a very significant reduction in encoding complexity with slight coding efficiency degradation as compared to the full search in both motion and disparity estimations.

Keywords: Multiview video coding, predictive search, motion/disparity estimation, vector candidates

1. INTRODUCTION

In multiview video systems, scenes are captured simultaneously from multiple cameras. These cameras are set to shoot the scenes from different locations and/or different angles. They provide digital video data that could be useful in several applications, such as surveillance systems, on-demand telecommunications, entertainment, and immersive virtual reality. Multiview video contains very large amounts of data as compared with monoscopic video, with the amount of added data increasing with the number of views. Multiview video coding (MVC) has recently become an active research area focused on compression for efficient storage and transmission of multiview video data.[1] Simulcast is a straightforward coding scheme for multiview video in which each view sequence is encoded independently (see Figure 1). This allows temporal redundancy to be exploited, using standard block-based motion compensation techniques.

In a multiview video scenario, because different views are capturing the same scene, there exists an additional source of redundancy, namely, cross-view redundancy. Similar to motion estimation (ME), we can use the block matching procedure to find block correspondence from view to view, through disparity estimation (DE). This cross-view redundancy is not exploited in the straightforward simulcast scheme. A MVC technique that exploits both temporal and spatial redundancy, can be constructed as follows: A given frame in view v at time t, can use past frames within view v as temporal reference for motion estimation, while using reconstructed frames from neighboring views as spatial reference for disparity compensation. We will denote it as the "dual search scheme". Previous research[2–5] has shown that combining motion compensation and disparity compensation will achieve higher coding efficiency as compared to simulcast. Some of this prior work[2,3] focuses in stereo video coding, which is a special case of multiview when the number of views is equal to two. The method by Li and He[4] divides multiview video into stereo pairs and within each pair the disparity is only estimated at certain times (i.e., not for every frame). This coding scheme only achieved slight gain as compared to simulcast, because the cross-view redundancy was not fully exploited. Other MVC techniques perform the disparity estimation at every time instant,[5] which leads to a consistent coding gain, but requires a much higher coding complexity as compared to simulcast, because both motion and disparity estimations have to be performed. These proposed schemes illustrate the tradeoff between coding efficiency and complexity. If the frames can have both temporal and spatial references for block matching, significant coding efficiency improvements over simulcast will be achieved. But

Further author information:
PoLin Lai: E-mail: polinlai@usc.edu, Antonio Ortega: E-mail: ortega@sipi.usc.edu

if only one of the two fields is estimated for some frames/views in order to reduce encoder complexity, this will result in a loss in coding efficiency.

In this paper, we propose fast predictive search algorithms that can be used when either the motion or the disparity field is available and we wish to estimate the other field efficiently. The main novelty is that we show that, for a current frame to be encoded, if, say, its motion field is available, and disparity information is available for the previous frames, then it is possible to obtain good disparity vector candidates for this frame with low complexity. Likewise, it is also possible to obtain good motion candidates when the disparity field is available for the current frame, and the motion field is available for frames in other views. We also propose a new search strategy that can better exploit the characteristics of the vector candidates obtained with our method. The rest of the paper is organized as follows. In Section 2 we first explain the dual search MVC scheme that exploits both temporal and cross-view redundancy. Then we introduce the proposed predictive search algorithms. The new search pattern adopted in this paper is also included in Section 2. Two sets of simulation results, one on the original images and one using a coding approach based on H.264/AVC,[6] are provided in Section 3. In this section we also discuss the bit allocation issue[7] and we demonstrate improved overall performance is achievable for both the dual full search scheme and the proposed algorithms when simple bit allocation rules are used. Conclusions are presented in Section 4.

2. DUAL SEARCH MULTIVIEW VIDEO CODING SCHEME AND THE PROPOSED PREDICTIVE SEARCH ALGORITHMS

2.1. The dual multiview video coding scheme

To fully exploit temporal and spatial redundancy in MVC, both motion and disparity estimations should be performed. Figure 2 shows the basic structure of such a dual search coding scheme.

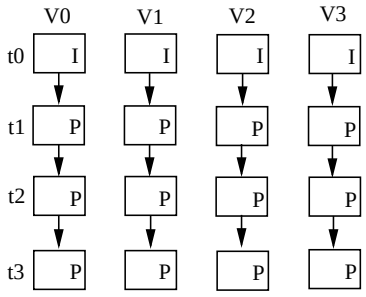

Figure 1. Simulcast for multiview video coding

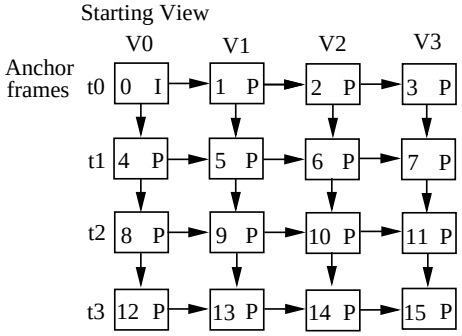

Figure 2. Dual search and compensation scheme for multiview video coding

Figure 2 illustrates an MVC coding structure with 16 frames forming one group of pictures (GOP). This GOP contains 4 views in spatial direction and 4 frames in temporal direction. The numbers associated with each frame indicate the encoding order (this encoding order will be the same throughout the paper.) Within one GOP, the first frames of each view are called "anchor frames" and can only be coded using disparity compensation from the preceding view. View 0 is the "starting view". The frames in this view are coded only by motion compensation. All the remaining frames in a GOP (which we will denote as "regular frames") can use both the temporal and spatial references to exploit the redundancy in two directions. For example, each macroblock in frame 6 can switch between its two corresponding ME/DE blocks from frame 2 and frame 5 respectively.

2.2. Fast predictive search I: Disparity then Motion (DtM)

In standard video coding, fast motion search algorithms can be designed based on the assumption that the motion vectors (MV) are highly correlated among neighboring blocks. We can use causal neighboring blocks to obtain a set of motion vector candidates that can be used to initialize the search and reduce the complexity.[8,9] In a MVC scenario, since there are multiple cameras capturing the same scene, it is possible to predict the motion field (MF) of one view using the other view's motion as reference. We propose that for a given frame, after its disparity field (DF) has been estimated, we will be able to find good search candidates for the motion field with very low complexity. Figure 3 depicts the basic procedure for predicting the motion field from view to view by identifying MV candidates. In Figure 3(a), the vector field depicted with solid arrows is estimated first. When encoding a "regular frame", disparity estimation is performed first. For a given block in a regular frame of current view, Figure 3(b), we track along its disparity vector to its corresponding block in the reference view. Depending on where this block is located, it could be overlapped with at most 4 blocks in the reference view, each having a MV. These can serve as the candidates, denoted MV_1, MV_2, MV_3, MV_4 in Figure 3(b), and provide initialization points for the search. After predicting the motion field of the current view, the same procedure is used to predict the motion field of the next view, using the current view as the reference. We denote this as the Disparity then Motion (DtM) scheme.

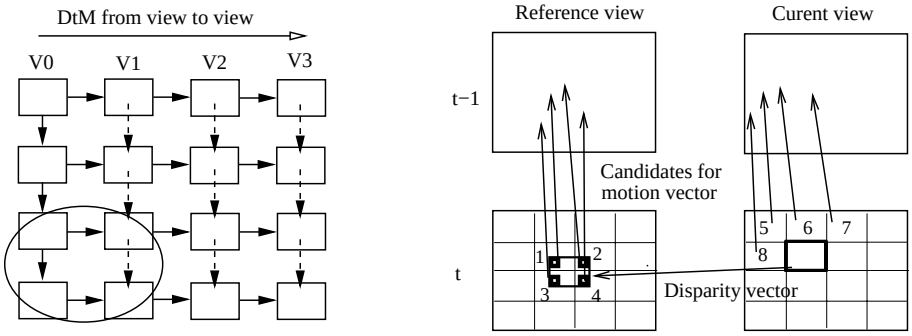

Figure 3. (a)Left: The structure of predicting the motion field from the reference view (DtM) (b)Right: Obtaining MV candidates to predict the motion field

To assess the performance of the additional MV candidates obtained using the DtM approach, we define two sets of candidates: $A = \{MV_5, MV_6, MV_7, MV_8, \vec{0}\}$, which contains the motion vectors selected from blocks in a causal neighborhood (as typically used in many fast motion estimation algorithms) and $B = A \cup \{MV_1, MV_2, MV_3, MV_4\}$, which also includes the additional candidates obtained from the neighboring view. For each block in a regular frame, we choose the candidate that provides the minimum sum of absolute difference (SAD) from each of the sets. The residual image PSNR values are compared in Table 1.

For all three test sequences it can be seen that higher quality is achieved if the additional candidates provided by the DtM scheme are considered. One of the drawbacks of the fast motion search using neighboring blocks is that the search may only identify a motion vector representing a local minimum when the current block happens to have a MV that is different from those of the neighbors. A typical example is in situations where the current

block is located close to the boundary of a moving object. Our DtM approach helps to alleviate this problem by providing a block correspondence in the other view to obtain additional MV candidates.

Table 1. Compare different sets of MV candidates, PSNR of the residue image

Aqua	Set A	Set B	ballroom	Set A	Set B	ST	Set A	Set B
V1	29.47	30.65	V1	28.45	29.25	V1	30.70	32.74
V0	29.54	30.79	V2	28.29	29.16	V2	31.14	32.93
			V3	27.94	28.84			

2.3. Fast predictive search II: Motion then Disparity (MtD)

A similar idea can be applied to predict the disparity field after the motion field has been estimated. We can use the disparity at time $t-1$ as the reference to predict the disparity at time t. This approach is illustrated in Figure 4. Again, the field shown with a solid arrow is estimated first. This time when encoding a "regular frame", the motion estimation is performed first. For a given block in a "regular frame" at time t, Figure 4(b), we track along its motion vector to its corresponding block at time $t-1$. This would give us at most 4 different candidates DV_1, DV_2, DV_3, DV_4 to initialize the disparity search. After predicting the DF at time t, this DF will be used as the reference to predict the DF at time $t+1$. We denote this as the Motion then Disparity (MtD) scheme.

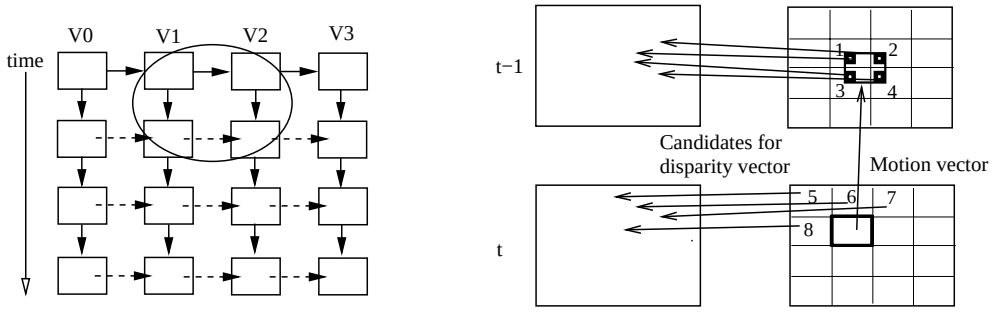

Figure 4. (a)Left: The structure of predicting the disparity field from a time instance (MtD) (b)Right: Obtaining DV candidates to predict the disparity field

2.4. A more efficient search pattern for the proposed DtM, MtD schemes

As described in the above subsections, the main novelty in the proposed algorithms is that we track along the first estimated field (DF/MF) to get candidate vectors for the other field (MF/DF). The additional candidates provide improved prediction in cases where the motion/disparity vector of the current block is not similar that of its neighboring blocks. If the first estimated field is accurate, e.g., because full search with a sufficient range has been used, we will have more confidence about the corresponding matching block and the candidate vectors obtained this way will be more reliable.

In most predictive motion search algorithms for monoscopic video, the mean or median of the candidates is used to initialize the search location. This approach relies on the assumption that the motion field tends to be locally smooth. However, the disparity field is not as homogenous as the motion field and disparity vectors can be seen to exhibit significant variation even across neighboring blocks.[4,5] To see why this is true, consider that in disparity estimation an area within an object that is closer to the camera will have larger disparity than an area in the same object that is further away from the camera. However motion in the two areas is likely to be same unless the object is rotating. Thus blocks that belong to the same moving object could have very different

disparity even though they have similar motion vectors. This suggests that computing the mean/median of a set of disparity vector candidates may not provide as good a predictor as applying the same technique to a set of motion vectors candidates. To tackle this problem, we propose a new search strategy, such that multiple searches are performed around each of the candidates, with each of the searches employing a much smaller search range than what would be typically used in combination a single search window centered at the mean of the candidates. Figure 5 illustrates this search pattern. Note that this is an extension to the disparity case of the technique proposed by Tourapis, Au, and Liou[9] for motion estimation.

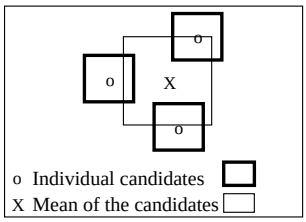

Figure 5. The search pattern that uses all the candidates

There are 9 candidates to be considered: DV_1 to DV_8 and $\vec{0}$. One approach is to order the candidates, with SAD, for example, and only perform search around the top priority candidates. Or we can simply search around all 9 candidates. Since some of the candidates might be the same and their search ranges may overlap, we currently adopt the second approach so that a search is performed for each of the candidates.

To verify the efficiency of the new search pattern, we compared the following two scenarios in the MtD scheme:

- **EachC**: Search with 3×3 windows ($\pm 1 \times \pm 1$) centered at each DV candidate
- **MeanC**: Search with a 9×9 window ($\pm 4 \times \pm 4$) centered at the mean of the DV candidates

Note that the *maximum* number of vectors to search for **EachC** is 81, while MeanC always has to check exactly 81 vectors.

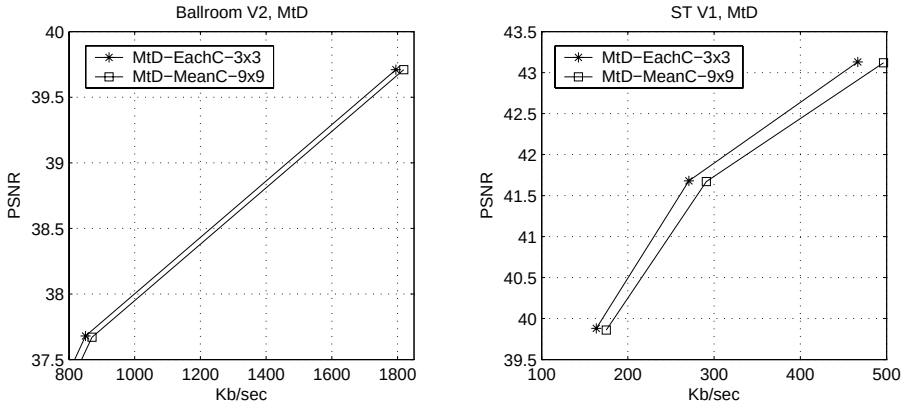

Figure 6. Comparison of different search patterns.

Figure 6 provides the simulation results for two different sequences*. The proposed search pattern achieves higher coding efficiency when we perform the predictive search on disparity estimation. The non-smooth disparity field is better predicted because our search pattern takes all the vector candidates into account instead of adopting the mean of candidates as a single search center.

*For details about the simulation environment, refer to Section 3.2

3. EXPERIMENTAL RESULTS

We performed experiments for the dual full search scheme (full search is used for both ME and DE) and the prediction structures of DtM and MtD. The GOP structure in our simulations has 10 frames in time direction and 3 to 4 views in spatial direction. We perform two sets of simulations, one using the original images, with results provided in terms of residual energy, and one based on a modified version of the H.264/AVC codec,[6] from which rate-distortion results are obtained. The first set of experiments assesses the efficiency and complexity of the proposed algorithms, while the second set provides a comparison of these different schemes in a real codec scenario.

3.1. Residual energy comparisons using original sequences

ME and DE are performed on the original multiview sequences. Each block can switch between its two corresponding matching blocks, whichever gives the lower sum of absolute difference (SAD). There is no widely accepted fast search algorithm for DE and it is known that DE typically requires a much larger search range than ME.[4] To compare different scenarios fairly, in all the experiments we obtain the first estimated fields (those with solid arrows in Figures 2, 3 and 4) with full search. Parameters of the test sequences, such as image dimensions and the full search ranges, are provided for reference in the Appendix. The PSNR is calculated based on the luminance channel of the residue images. They are all compared with the simulcast scheme with same full motion search range. The results are given in Table 2[†].

The dual full search scheme, with both fields full searched, serves as a baseline of the potential gain. The degradation of MtD and DtM, as compared to the dual full search scheme, comes from that one of the field is fast searched with the proposed algorithms. In Table 2 we search surrounding positions of candidates with a 3 $\times$ 3 range, pick the vector with the lowest SAD as the predicted field. We have at most 9 candidates: $V_1 - V_4$ as proposed, $V_5 - V_8$ from the predicted causal blocks, and $\vec{0}$ (refer to Fig. 3(b) and Fig. 4(b)). The maximum number of vectors to check is $9 \times (3 \times 3) = 81$ for the predicted field. Since candidates might be the same, or their search ranges may overlap, the average number of vectors tested ("Vct" in Table 2) is about 40 to 65 for disparity and 30 to 50 for motion. This is a very low cost for the predictive searched field, as compared to full search range of $\pm 32 \times \pm 32 = 4225$ or $\pm 64 \times \pm 64 = 16641$. In Table 2 we compare the number of block matching (BM) operations to be performed for encoding one regular frame. Simulcast has the lowest complexity and is set to be the reference in comparison. For example, MtD for Aqua V1 has full searched motion and predictive searched disparity with on average 42.52 vectors checked. This represents a minimal increase (only 1.02%) over the matching complexity required by simulcast (which requires 4225 operations), as compared to doubling the complexity if full search is used to compute both fields. Even with this reduced complexity MtD provides gains

[†]Note that the MtD results are provided first followed by the DtM results and the same ordering will be followed in Section 3.2

Table 2. Performance comparisons (In MtD and DtM, search 3 $\times$ 3 window of each candidates)

	Simul	Dual Full	MtD (Fig. 4)			DtM (Fig. 3)		
	PSNR	PSNR	PSNR	Vct	BM	PSNR	Vct	BM
Aqua V1	31.49	32.27(+0.78)	32.24(+0.74)	42.52	101.02%	32.23(+0.73)	36.11	100.85%
V0	31.57	32.40(+0.83)	32.26(+0.79)	44.76	101.07%	32.36(+0.79)	33.27	100.79%
BallroomV1	31.92	33.01(+1.09)	32.89(+0.97)	47.13	100.28%	32.65(+0.74)	34.54	100.21%
V2	32.02	33.16(+1.14)	33.09(+1.07)	45.29	100.28%	32.73(+0.71)	29.86	100.18%
V3	31.59	32.85(+1.25)	32.76(+1.17)	44.54	100.27%	32.51(+0.92)	31.08	100.19%
ST V1	35.08	36.59(+1.52)	36.41(+1.33)	60.31	100.36%	36.05(+0.98)	45.86	100.28%
V2	35.24	36.71(+1.47)	36.53(+1.29)	63.35	100.38%	36.18(+0.94)	47.46	100.29%

of 0.7 to 1.3dB as compared to simulcast, with a degradation of only 0.04 to 0.2dB with respect to full search of both fields. DtM results in a 0.7 to 0.98dB gain, with a degradation of 0.05 to 0.55dB as compared to full in search in both fields.

The largest degradation, as compared to the dual full search scheme, comes from DtM on the ST sequence, where there is easily perceivable illumination mismatches cross different views. This mismatches affect the DtM scheme because the first estimated disparity field may not provide a good matching block, which may lead to a suboptimal set of motion vector candidates. In this situation, regular frames can use either a full search DF or a significantly degraded MF, thus leading to degradation in coding efficiency. Similar results can be observed in the Ballroom sequence where DtM also has lower performance than MtD. In both cases the lack of accuracy in the disparity field leads to reduced prediction quality when generating candidates for the motion field. For the test sequences we have used, our results indicate that motion estimation should be performed first and then be used for predictive fast disparity search.

3.2. H.264/AVC-based Simulations

We used the JM9.6 implementation of H.264/AVC[6] to simulate multiview coding with a real codec. To implement the GOP structure illustrated in Figure 2, the multiview sequences are first interlaced so that they are presented to the codec in the required order (again, refer to Figure 2). We employed the multiple-reference function in H.264/AVC so the frames can have both temporal and spatial references. The reference-frame management function has been modified so that it matches the structure in Figure 2. To implement the proposed DtM and MtD schemes, the DF and MF have both to be stored so they can be used as candidates to encode the later frames. This extra memory is required only in the encoder side to initialize the search locations for our predictive scheme. For all three methods as in Figure 2 3 4, the H.264/AVC software will switch for each block between the two reference fields by selecting the one that has the lower cost (including both the vector encoding cost and the residual coding cost).

Comparing Figures 1 and 2, an immediate benefit of MVC can be observed: the anchor frames are now encoded using cross-view prediction, which provides a higher coding efficiency as compared to the simulcast case, where they were independently coded. The quality of these encoded anchor frames is crucial because they serve as the temporal reference for the later frames. Kim, Garcia, and Ortega[7] have studied the bit allocation issue in dependent video coding scenarios such as those arising in MVC. Their results demonstrate that if more bits are spent on anchor frames, a better overall coding efficiency will be achieved, because all the frames following the anchor can be encoded more efficiently. Here we simply use a smaller quantization step size (the QP parameter in H.264/AVC codec) to encode the anchor frames. For example, if the P frames are coded with QP = 28, then the anchor frames will be coded with QP = 26. Figures 7 shows the effect of this QP change. Gains from this bit allocation are observed on all the test sequences and under all three coding schemes (dual full search, DtM, MtD). In all the following results provided in this section, we adopt this smaller QP for anchor frames.

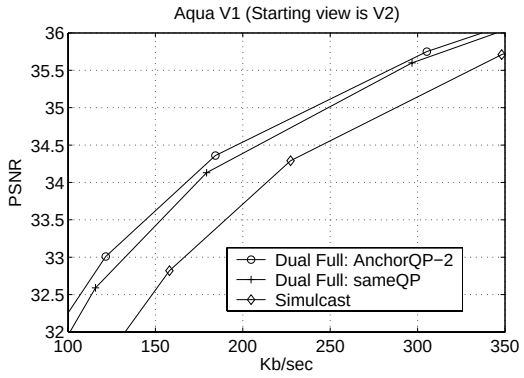

Figure 7. The effect of changing QP for the anchor frames

In Figure 8 9 and 10, we present the rate-distortion curves (R-D curves) of our simulation results. The MtD approach achieves very good performance for all three test sequences, even with a small search window (3×3) around each candidate. The R-D curves are very close to the curves obtained by the dual full search coding scheme. The performance of DtM predictive search varies among different test sequences. The Aqua sequence has the most dense camera setting among our three test sequences: 15 cameras with about 1.8cm spacing. There is significant cross-view redundancy, thus disparity estimation finds reliable matching block. Figure 8 shows that the DtM approach provides almost the same coding efficiency as MtD. For the Ballroom sequence, the R-D performance of DtM exhibits some degradation (0.1 to $0.2dB$) with respect to the dual full search scheme. Using our modified H.264/AVC, this is the sequence with the highest gain in coding efficiency (up to $1.5dB$) when comparing MVC with simulcast (Figure 9). DtM preserves much of this gain with a low complexity for predictive motion search. The worst case of DtM appears to be on the ST sequence (Figure 10). As addressed in Section 3.1, the cross-view illumination mismatch reduces the accuracy of the first estimated disparity field. With a 3×3 search range for MV candidates, the DtM's R-D curves lie about halfway between the dual full scheme and simulcast. We provide one more set of simulation results as the search range is increased to 5×5. The corresponding R-D curves move to about 0.1 to $0.2dB$ below the dual full search scheme.

Once again we see that the key to the performance of the proposed predictive algorithm is that most reliable estimation should be performed first, so that the fast predictive searh on the second field can make use of good candidate vectors. For MVC, since the camera view settings vary among different applications, it is likely that computing the motion field first will be in general more efficient, so that MtD should in general be chosen over DtM in MVC.

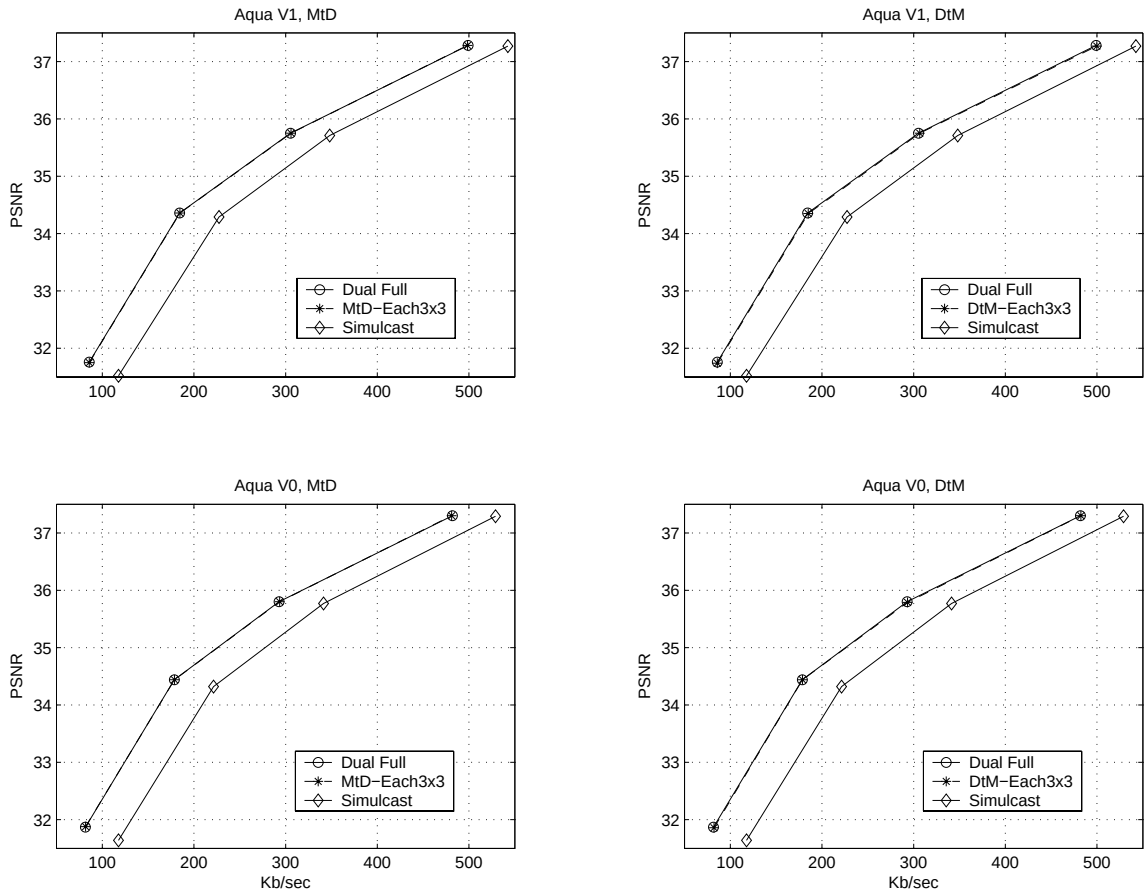

Figure 8. Simulation result for Aqua sequence

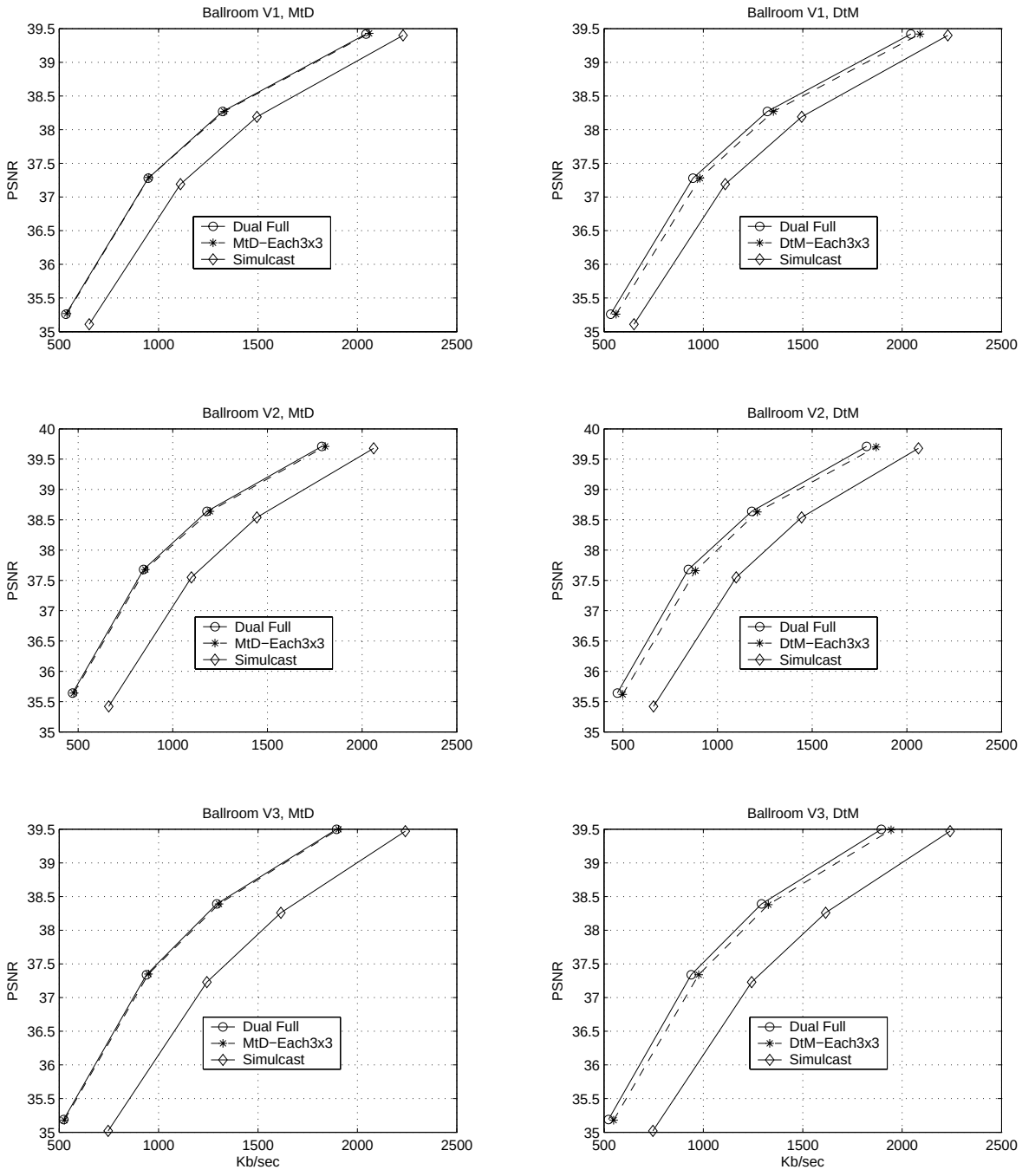

Figure 9. Simulation result for Ballroom sequence

.

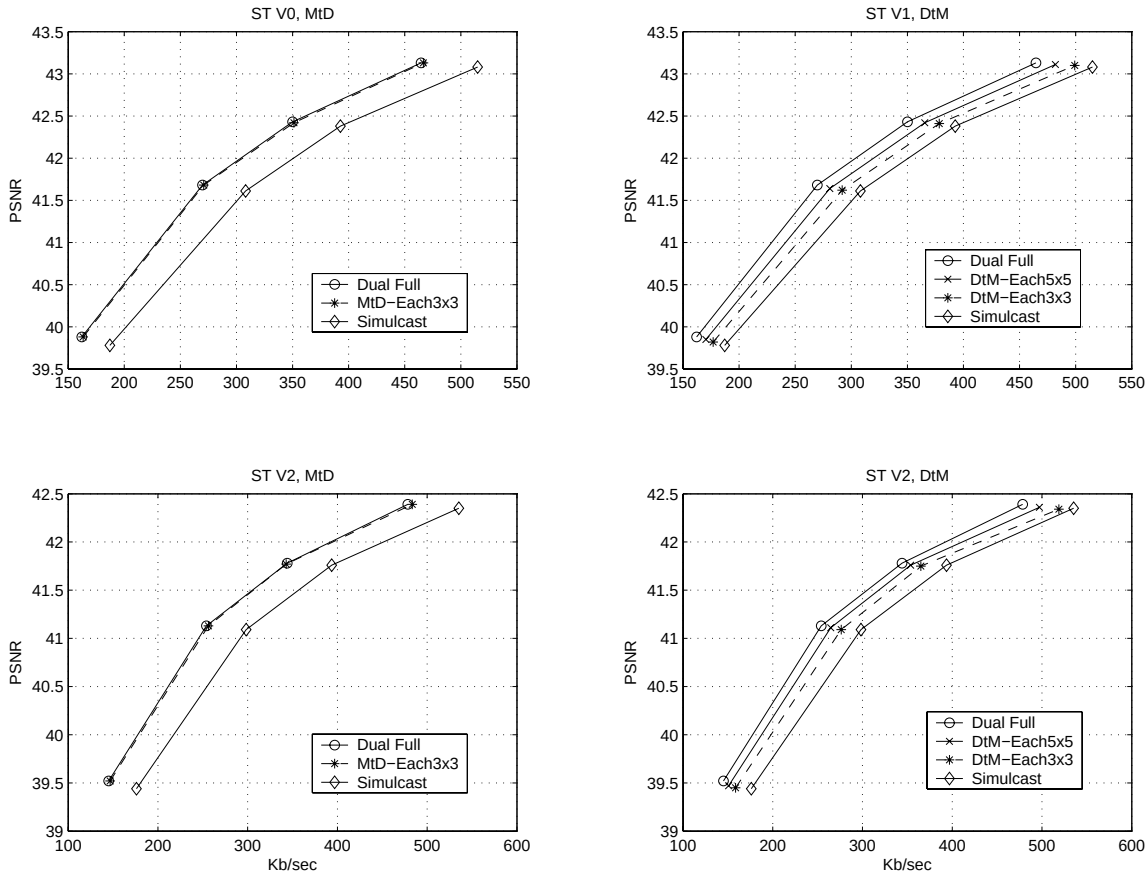

Figure 10. Simulation result for ST sequence

4. CONCLUSIONS

Higher coding efficiency can be achieved in MVC that exploits both temporal and cross-view redundancy. In this paper we propose a novel predictive fast search algorithm to reduce the complexity for MVC. After one of the motion/disparity fields is estimated, the proposed algorithm obtains good candidate vectors to perform the estimation on the other field with very low complexity. A more efficient search pattern employing the candidate vectors is also proposed. The new vector candidates can provide additional prediction information if the first estimated field is accurate. Since motion estimation generally provides better block matching than disparity estimation, MtD generates very consistent coding efficiency among different test sequences, as compared to DtM. Simulation results on the original images and with H.264/AVC both show that MtD can achieve coding efficiency that is very similar to the dual full search scheme, while the complexity is reduced significantly. These two simulations also verify that in general MtD should be chosen over DtM.

APPENDIX A. MULTIVIEW TEST EQUENCE PROFILE

Aqua
Tanimoto Laboratory in Nagoya University, Japan
http://www.tanimoto.nuee.nagoya-u.ac.jp/english/index.html

Ballroom
Mitsubishi Electric Research Laboratories, Cambridge, MA
ftp://ftp.merl.com/pub/avetro/mvc-testseq

ST
Institute of Communication Theory and Signal Processing, Univ. of Hannover, Germany
ftp://ftp.tnt.uni-hannover.de/pub/3dav/3DAV_Test_Data/EE2/

Table 3. Parameters of the sequences and simulation setting

Sequence	Dimension	Frame Rate	No.Views	GOP in simulation	Full search range
Aqua	320×240	10 fps	15	V2→V1→V0	±32 × ±32
Ballroom	640×480	25 fps	8	V0→V1→V2→V3	±64 × ±64
ST	640×480	15 fps	6	V0→V1→V2	±64 × ±64

REFERENCES

1. ISO/IEC-JTC1/SC29/WG11, "Call for proposals on multi-view video coding," *MPEG Document N7327* , Poznan, Poland, Jul 2005.
2. J. Ellinas and M. Sangriotis, "Stereo video coding based on interpolated motion and disparity estimation," in *Image and Signal Processing and Analysis, 2003, Proc. IEEE 3rd International Symposium on* **Vol.1**, pp. 301–306, 2003.
3. Y. Luo, Z. Zhang, and P. An, "Stereo video coding based on frame estimation and interpolation," *IEEE Trans. on Broadcasting* **Vol.49, Issue 1**, pp. 14–21, Mar 2003.
4. G. Li and Y. He, "A novel multi-view video coding scheme based on H.264," in *2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia, Proc. IEEE* **Vol.1**, pp. 493–497, Dec 2003.
5. W. Yang, K. N. Ngan, and J. Cai, "MPEG-4 based stereoscopic and multiview video coding," in *Intelligent Multimedia, Video and Speech Processing, Proc. IEEE 2004 International Symposium on*, pp. 61–64, Oct 2004.
6. "http://iphome.hhi.de/suehring/tml/index.htm," *Image Processing Research Department, Fraunhofer-Institute for Telecommunications, Heinrich-Hertz-Institut, Germany* , JM 9.6 is released Jul 2005.
7. J. H. Kim, J. Garcia, and A. Ortega, "Dependent bit allocation in multiview coding coding," in *Image Processing, Proc. IEEE 2005 International Conference on*, Sep 2005.
8. J. Chalidabhongse and C.-C. Kuo, "Fast motion vector estimation using multiresolution-spatio-temporal correlations," *IEEE Trans. on Circuits and Systems for Video Technology* **Vol.7, Issue 3**, pp. 477–488, Jun 1997.
9. A. M. Tourapis, O. C. Au, and M. L. Liou, "Highly efficient predictive zonal algorithms for fast block-matching motion estimation," *IEEE Trans. on Circuits and Systems for Video Technology* **Vol.12, Issue 10**, pp. 934–947, Oct 2002.