International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:7, 2009

Identifying Blind Spots in a Stereo View for Early Decisions in SI for Fusion based DMVC

H. Ali, K. Hameed, and N. Khan

***Abstract*—**In DMVC, we have more than one options of sources available for construction of side information. The newer techniques make use of both the techniques simultaneously by constructing a bitmask that determines the source of every block or pixel of the side information. A lot of computation is done to determine each bit in the bitmask. In this paper, we have tried to define areas that can only be well predicted by temporal interpolation and not by multiview interpolation or synthesis. We predict that all such areas that are not covered by two cameras cannot be appropriately predicted by multiview synthesis and if we can identify such areas in the first place, we don't need to go through the script of computations for all the pixels that lie in those areas. Moreover, this paper also defines a technique based on KLT to mark the above mentioned areas before any other processing is done on the side view.

***Keywords*—**Side Information, Distributed Multiview Video Coding, Fusion, Early Decision.

I. Introduction

TRADITIONALLY, video encoders have been of a higher complexity as compared to the decoder in case of single source as well as multiple source video encoding. For video transmission of multiple camera, however, Multiview Distributed Coding requires shift of complexity from Encoder to Decoders where decoder has to jointly decode video streams for multiple cameras commonly known as Wyner Ziv Coding [1].In the absence of traditional encoding decoding architecture, Side Information plays an important role in reconstruction of frames that have not been intra decoded; these frames are also known as Wyner Ziv Frames. Side Information can either be constructed from decoded key frames of WZ camera using Motion Compensated Interpolation [2] or through decoded key frames of side cameras using Homogrpahy Compensated Interpolation [2].

Techniques like fusion [3] have been introduced to determine the choice of side information from these Side Information sources. The side information constructed through side camera cannot be fully mapped onto center (WZ) camera and thus leaves some "holes" to be filled either by retransmission or use of side information constructed using MCTI. his paper introduces a technique to determine the holes that would be present in central camera so that process of Homogrpahy should not be applied to these blocks for construction of Inter View Side Information. Although this method intends to save some processing complexity in existing fusion techniques, the results available yet are not comparative in nature to the existing techniques. Moreover, the results will also vary depending upon the type and quality of video.

Authors are with Computer Science Department, Lahore University of Management Sciences, Opposite Sector U, DHA, Lahore Cantt, Pakistan (e-mails: hali@lums.edu.pk, 07060006@lums.edu.pk, nkhan@lums.edu.pk).

II. Defining Identifyable Areas

We know that constructing a depth map [11] of a scene requires at least two images of the same scene. One image provides us the location of a certain point in space mapped upon a two dimensional image plane the depth of which is unknown. Using epipolar geometry, we can map all possible depths of that point in the image plane in space for which the point in the image plane will not change its location in the image plane; depth will be orthogonal to the image plane. When the same line is mapped on to the second image plane that is covering the same scene, we are presented with all the possible locations that that specific point in space can lie at in the second image plane. Fig. 1 shows one such scenario.

If the projection point x_L is known, then the epipolar line e_R - x_R is known and the point X projects into the right image, on a point x_R which must lie on this particular epipolar line. This means that for each point observed in one image the same point must be observed in the other image on a known epipolar line. This provides an epipolar constraint which corresponding image points must satisfy and it means that it is possible to test if two points really correspond to the same 3D point. Epipolar constraints can also be described by the essential matrix or the fundamental matrix between the two cameras. If the points x_L and x_R are known, their projection lines are also known. If the two image points correspond to the same 3D point X the projection lines must intersect precisely at X. This means that X can be calculated from the coordinates of the two image points, a process called triangulation [4].

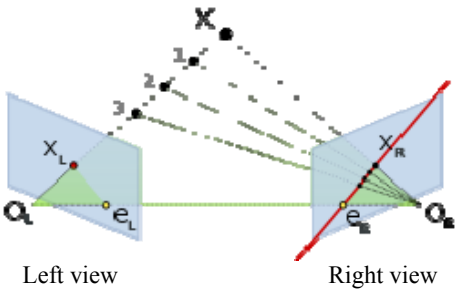

Fig. 1 Sample scenario with epipolar line on right plane and triangulation being done for various positions of X to find its exact depth

If the above mentioned process goes on normally, we can get the required information with a certain exactness to

compute the image values at all pixels in intermediate frames by view synthesis. However, in a 3D scene, there are areas on object surfaces that are not covered by both the cameras at the same time. For those areas, use of epipolar geometry is not a possibility as correspondences of one XL in left image is not found anywhere in the right image. All such areas with no stereo correspondence can not yield results in synthesis that can surpass the results that can be achieved by temporal transformations and hence we mark these areas before the actual fusion mask generation begins [3]. We extend the same concept that we have used in images to the video domain because every frame in a video is an independent image. We call these areas the blind spots in the stereo view because we do not have stereo correspondences for these areas in the images and depth map can not be constructed for these areas. So, an accurate prediction in view synthesis for a center camera for the purpose of homography compensated interpolation can not be obtained [2], [3].

III. Side Information Generation in DMVC

There are two ways in which side information could be generated in the Distributed Multiview Video Coding paradigm [3]; temporally or homographically. For temporal side information generation, side information is generated using previous and forward frames from the same camera. Interpolation is computed based on the correspondences found in the two frames. For homography based side information generation, we compute the homography projection matrices relating the central view and the side views. Homography is a 3x3 matrix that related one view to another view in the homogeneous coordinate system. This means that the matrix consists of eight parameters other than the homogeneous factor h. We can extract equations for x and y from this matrix and solve for any values of λ. Depending on the model we use we calculate its parameters such that the sum of squared differences between the current frame and the warped side frame is minimized. To compute the model parameters, we use a gradient descent method [5].

IV. Fusion Masking

Now that we have gotten two types of side information, we have to chose between the two or find a technique to use both the techniques simultaneously. This has been dealt with in one way in Ref. [3]. Both the methods are considered for each pixel of every frame for which side information has to be generated. A binary fusion mask is generated for the size of the target side information frame such that 1 indicates that the pixel is taken from the homography based side information and 0 represents otherwise. Previous and Forward frames are used to compare the quality of estimation that is provided by both the techniques at every particular pixel and based on that, the bit in the mask is determined. More specifically, each pixel from the previous frame is considered and the pixel that predicts it better from both side informations is searched. This is done by taking the difference between the current pixel from the previous frame and the pixel at the same position from both side infomations. If the homography-based side information has a smaller error, then, the binary fusion mask bit at this position is set to 1. Otherwise, if the block-based temporal side information has a smaller error, then, the binary mask bit for this pixel is set to 0. The same procedure is done for both the previous frame and the forward frame and or function is used to merge the two masks generated. Afterwards, the fused side information is generated based on the merged mask. This is the side information that is used to estimate the actual WZ frame. However, if we can only determine a few areas before hand that can not provide good results in any case, we will be able to cut down all the above stated procedures for lesser number of pixels in each frame.

V. Early Decision Technique

The purpose of this technique is to identify all such areas that do not have correspondence present in their respective side view and hence can be categorized to lie inside the blind spot of stereo view. The first step is to obtain synchronized side view frames for the particular frame for which we want to identify blind spots. We identify the good features to track [6] in one frame and apply the Lucas Kanade Method [8] for tracking them in the other frame. After obtaining a certain amount of correspondences, we run the iterative voting based RANSAC algorithm [12] to extract the eight best correspondences. Then, the fundamental matrix is computed between the two frames. Notice that this fundamental matrix is not the same as the homography matrix.

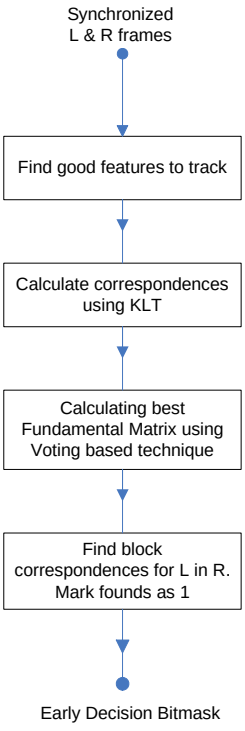

Fig. 2 Flow diagram of proposed technique

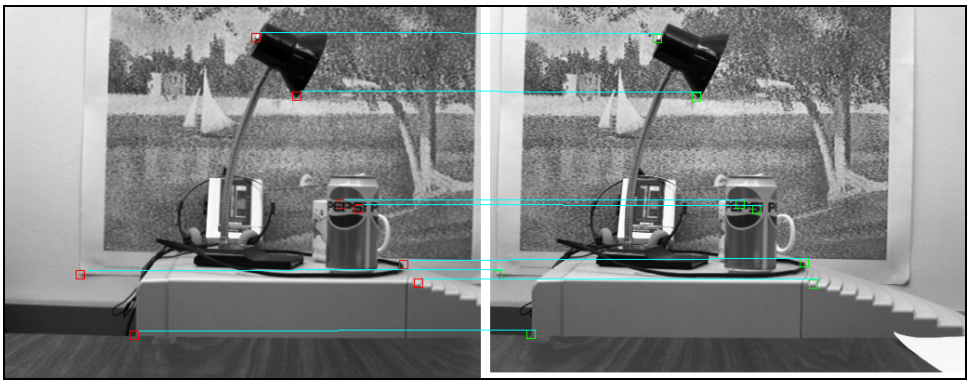

Fig. 3 Sample images. Left image represents the view from a left side camera and Right image represents the view from a right side camera. Camera planes are considered parallel for this particular example however it is not a hard and fast binding or a restriction for application of this technique.Eight best features in the left images and their correspondences that were tracked using a voting based algorithm similar to ransac; fundamental matrix was constructed using these eight correspondences

The fundamental matrix that relates any two frames works on the basis of epipolar geometry. If we pick any point on the left frame and take its product with the fundamental matrix, we will obtain a line on the right frame somewhere on which that specific point will be mapped; the point has to be visible in both the frames. The next step is to pick blocks on the first frame, calculate the epipolar line on the second frame by taking the product of the block's center coordinates with the fundamental matrix and searching for correspondence on that epipolar line in the second frame. Multiple methods like diamond search, MAD or Graph Cut based technique [10] could be used for obtaining correspondences. However, we used simple 2D cross correlation along the epipolar line. The block with the highest correlation is considered as the best match; the best match has to lie above a certain threshold.

All such blocks for which the matching is successfully found in the second frame are marked as white or 1 in the early decision bit mask. All the remaining areas are left as 0. All areas that are represented by 0's are the areas with no matching in the corresponding second camera image and hence depth for those regions can not be computed using triangulation [4]. This bit mask is used as the initial state for the first stage in the already present fusion technique described earlier. However, no further calculations for stereo matching or side view synthesis will be done for the pixels marked in the bitmask with 0's. It is sure that a demography based interpolation will not provide good enough results for these regions and no further decision is to be taken for these regions. On the other hand, the already defined procedure [3] will be worked upon all the pixels marked as 1 in this early decision bitmask. This means that it is still possible to have better results from temporal interpolation even though demography compensated side information can be computed. Fig. 2 shows a simple flow of the technique.

VI. Results

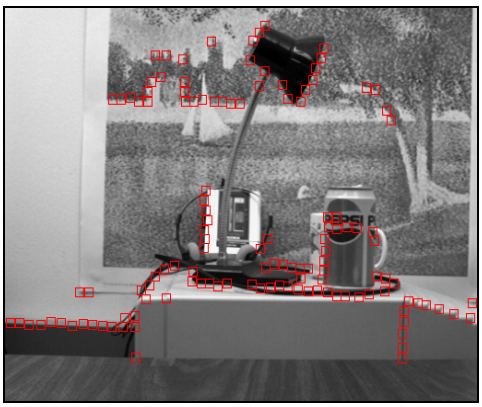

Fig. 4 Red marked boxes on the left side image represent the features extracted as good tracking features and they were further tracked using KLT

Fig. 5 An early decision bitmask that contains pixels marked as black representing the blind spots and pixels marked as white for which decision is pending and will be considered when using the fusion technique

The results obtained from the method show that for videos with higher disparity in cameras, a reasonably large percentage of pixels can be pruned by early decision and a lot of

processing could be saved the frames involved. The results show marking of blind spots based on two different situations; blind spots created due to non-overlapping videos and blind spots created due to the three dimensional nature of objects inside the frames. Although our emphasis has been on the latter one but both types of markings contribute towards saving extra computation that is done on areas with no hope of providing good results through homography compensation techniques. Fig. 3 shows both left and right views of the sample image and the best eight correspondences that were found using the voting based technique for computing the fundamental matrix. Fig. 4 shows the good features that were found for tracking. Fig. 5 displays an early decision bitmask that contains pixels marked as black representing the blind spots and pixels marked as white for which decision is pending and will be considered when using the fusion technique.

Notice that the output in Fig. 5 will be one of the inputs of the existing fusion mask generation procedure. This means that the existing method for generating fusion mask will change very slightly to accommodate the pruning that is done by early decision bitmask generation. In the image shown, the total number of pixels was around 204,800. The overlap between the two images was around 86% and the rest of the areas had no overlap for stereo matching. The output shows that around 79% of the pixels were marked white and correspondences for those pixels were found. However, 7% pixels were found to be lying in blind spots and hence marked so. The fusion mask will only work for the remaining 79% pixels saving computing for around 21% pixels which amount to 43,200 pixels in this image.

VII. Future Research and Workarounds

As metioned, this technique is supposed to save computation and improve the performance of the decoder in side information generation by pruning certain regions that can not amount to good estimations. Although the results show reasonable amount of pruned pixels for a certain medium level of disparity containing video, this has been done using extra steps for pre-processing. More implementation is pending that will establish more specific comparisons in terms of processing cycles, memory requirements and time consumed for the fusion based technique alone and the fusion based technique combined with the early decision technique introduced in this paper.

Moreover, there can be ways which, if used, can fill the blind spots in stereo in various ways. A few of such methods have been devised as formal research papers such as one in Ref [13]. Our technique helps identifying the blind spots that can either be marked as useless for further use in simpler fusion techniques, but it is still possible to somehow fill these areas with appropriate material. The reason for that is even though correspondences are not available, deductions from simple observations can still be applied and using those, we can fill those holes with reasonably suitable material. For instance, we know that if an object is occluding another object, the occluded object will lie further away from the camera as compare to the occluding object. If we can detect objects using some techniques, techniques that are also available currently but somewhat complex, we can use data from only one camera as well to fill the areas marked as blind spots in our technique.

The break point at which it gets better to use the technique introduced in this paper has to be carefully defined for various kinds of videos. Moreover, based on these results, alterations can be made to decoder and encoder designs. One such idea can be that the encoder keeps a buffer of the frames it has captured and not sent and if the previous and forward frames are not available for a certain frame and the stereo vision can not provide for enough required pixels, the decoder could request for a certain frame from the encoder and use it to temporally decode the targeted frame based on motion compensation techniques. Various other modifications could be thought of.

Acknowledgement

We are thankful to Mr. Ijaz Akhter from the Computer Vision Lab at LUMS for assistance in the steps we planned for improved results and reduced complexity. Without his help, it would not have been possible to come up with a complete solution for our problem statement in such a short duration. We also acknowledge the consistent support of Dr. Nadeem A. Khan throughout the span of this research; his guidance streamlined the focus of this research.

References

[1] Girod, B.; Aaron, A.M.; Rane, S.; Rebollo-Monedero, D., "Distributed Video Coding," Proceedings of the IEEE , vol.93, no.1, pp.71-83, Jan. 2005

[2] Guol, X.; Lu, Y.; Wu, X.; Gao, W.; and Li, S. "Distributed Multi-View Video Coding," Proc. SPIE, Jan. 2006, Vol. 6077, 60770T (2006); DOI:10.1117/12.642989

[3] Ouaret, M.; Dufaux, F.; Ebrahimi, R., "Fusion-based multiview distributed video coding," Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, October, 2006, Santa Barbara, California, USA DOI:10.1145/1178782.1178803, 2006

[4] [http://en.wikipedia.org/wiki/Epipolar_geometry], "Epipolar geometry," Wikipedia.

[5] Dufaux, F. and Konrad, J. "Efficient, Robust, and Fast Global Motion Estimation for Video Coding," IEEE transactions on image processing, vol. 9, no.3, March 2000.

[6] Jianbo Shi; Tomasi, C., "Good features to track," Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on , vol., no., pp.593-600, 21-23 Jun 1994

[7] Tommasini, T.; Fusiello, A.; Trucco, E.; Roberto, V., "Making good features track better," Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on , vol., no., pp.178-183, 23-25 Jun 1998

[8] Singh, M.; Mandal, M.; Basu, A., "Robust KLT tracking with Gaussian and Laplacian of Gaussian weighting functions," Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on , vol.4, no., pp. 661-664 Vol.4, 23-26 Aug. 2004

[9] K. Wojciechowski, B. Smolka, H. Palus, R.S. Kozera, W. Skarbek and L. Noakes, "A voting strategy for high speed stereo matching." Computer Vision and Graphics International Conference, ICCVG 2004, Warsaw, Poland, September 2004, Proceedings, 10.1007/1-4020-4179-9_27, 2004

[10] Kolmogorov, V.; Zabih, R., "Computing visual correspondence with occlusions using graph cuts ," Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on , vol.2, no., pp.508-515 vol.2, 2001

[11] Ikeuchi, K., "Constructing A Depth Map from Images," Massachusetts Institute of Technology, A. I. Memo No. 744. 1983

[12] Torr, P.H.S, Murray, D.W., "The development and comparison of robust methods for estimating the fundamental matrix," International Journal of Computer Vision 24(3), 271-300, 1997

[13] Shungang Hua; Jing Zhang; Zongying Ou, "Novel view generation from two reference images based on the same optical axis," Computer and Information Technology, 2004. CIT '04. The Fourth International Conference, pp. 801-806, 14-16 Sept. 2004.