Rectification-Free Multibaseline Stereo for Non-ideal Configurations

Hongdong Li and Richard Hartley

Research School of Information Sciences and Engineering,
The Australian National University,
ASSeT, Canberra Research Labs, National ICT Australia

Abstract. SSSD-based linear multibaseline stereo is an efficient implementation of multi-camera stereo vision system. This efficiency, however, vitally relies on the ideal configuration of all cameras. For dealing with non-ideal configurations, conventional stereo rectification algorithms can be used, but the performances are often still not satisfactory. This paper proposes a new algorithm to process non-ideally configured multibaseline stereo system, which not only avoids the rectification procedure but also remains the efficiency of SSSD at the same time. This is fulfilled by using the idea of tensor transfer used in image-based-rendering area. In particular, the multibaseline stereo is reformulated as a novel-view-synthesis problem. We propose a new concept of tensor-transfer to generate novel views as well as compute the depth map.

1 Introduction

It is well recognized that using more than two cameras can significantly improve the performance of stereo vision system, thanks to the redundant information contained in the multiple images [1][4].

In a multibaseline stereovision system, the multiple cameras can be arranged in different ways. This paper focuses on the **linear multibaseline system** proposed originally by Okutumi and Kanade[1]. For such system, the ideal configuration is shown in figure-1(a), where N cameras are arranged in such a way that all camera centers are collinear and all optical axes are in parallel. For such ideal configuration a very simple SSSD (sum of sum of squared differences) computation, based on the fact that the corresponding pixels in multiple views have a same inverse depth $\frac{1}{Z}$, is used to obtain the depth map. The computation is rather efficient, but effectively improves the stereo matching qualities. For this reason it has been popularly applied by many realtime stereo applications, particularly in mobile robot navigation fields[4][2][7]. Another practical reason is that: such linear configuration is easier to be mounted on the roof of an autonomous vehicle than other configurations (e.g., a L-shaped) which are more space consuming.

The efficiency of this SSSD computation, in fact, comes from the ideal configurations of the multiple cameras. Only when the N camera centers are collinear and their optical axes in parallel can we simply sum the SSD curves up. However,

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 810–817, 2005.

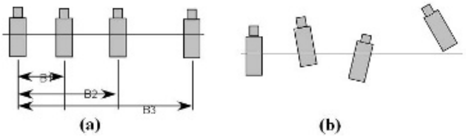


Fig. 1. Linear Multibaseline Stereo System: (a)ideal configurations; (b) non-ideal configurations

in many practical situations it is rather hard to build or maintain the ideality of the configurations. The insufficient precision in mechanization and assembly is one major reason. For example, it is rather hard to simultaneously adjust multiple cameras to make the optical centers precisely collinear and the optical axes parallel. Vibrations during the running of the system also affect the ideality of the camera configurations.

For dealing with generally configured multibaseline stereo, one may use traditional **rectification algorithms**[3][10][9]. However, the performances of these traditional rectification algorithms, when applying to the linear multibaseline system, are not satisfactory. Firstly, the computational efficiency is lost, since every camera needs to be rectified at least twice; Secondly, the matching performance is degraded significantly. The reason is that: every rectification involves image resampling and pixel interpolation, the introduced distortions will successively accumulate across images, which subsequently compromises the matching accuracy.

This paper proposes a new method to deal with the non-ideal linear multibaseline stereo (Figure-1(b)). Our method does not need a rectification processing. We borrow the idea of *image transfer* from image based rendering (IBR) area. By this idea, we re-formulate the multibaseline stereo matching problem as a typical novel-view-synthesis problem. We propose a new concept of inverse trifocal tensor by which a corresponding pixel-chain across multiple views is efficiently established.

Our method can be also used as a new novel-view-synthesis algorithm. Compared with existing methods, our method is featured by that it generates a very realistic novel-view image without resorting to any pre-computed precise depth map.

2 Previous Work

Most rectification algorithms were devised for binocular stereo[16][3]. So far, there exists only a very few papers that intently deal with multi-baseline rectification problem. Although it is commonly thought that one could trivially apply traditional binocular rectification algorithms to multi-baseline stereo in a pairwise manner, as we mentioned before this is not an efficient solution.

Mulligan and Daniilidis[4] proposed a method for processing non-parallel *trinocular* stereo. In their method, the image corresponding to the central camera is rectified twice. A large look-up-table is used to establish the corresponding pixel link across multiple images. The computational overhead will become more serious when more cameras are used. Kimura and Kanade et.al introduced a rectification method also for three cameras[7]. There the three cameras are all rectified so that they are in an ideal orthogonal configuration. A transformation relationship is computed from the intersection points of mutual epipolar lines. However, their method does not work properly for multiple linearly configured cameras, because such configuration is a **degenerate case** of the epipolar relationship. Furthermore, the method can not be extended to more than three cameras. Paper[5] suggested a rectification method for general multibaseline configurations, but it requires prior knowledge of 3D positions of a set of feature points, which is therefore not suitable for many practical applications such as robot navigation.

Novel-View-Synthesis(NVS). The proposed multibaseline algorithm is based on the idea of Novel view synthesis. Using a small set of real input images to generate some novel images as if there were observed at some novel viewing positions, such is the process of novel-view-synthesis, an important application of computer vision. Methods of NVS can be roughly classified as two categories: 3D-reconstruction-based method, and image-based-rendering (IBR) method. Recently, the IBR methods have received more attentions than the reconstruction method, because they avoid much unnecessary and sometimes unstable explicit 3D reconstruction procedure, thus are more efficient.

3 Inverse Tensor Transfer

Trifocal tensor to three views is the same as fundamental matrix to two views. Here the tensor is a 3x3x3 data volume, just as the fundamental matrix is a 3x3 data array. Trifocal tensor encapsulates all projective geometric relationships among the three images.

For three images, denoted by image-1, image-2, and image-3, whose camera matrices are $\mathbf{P}_1$,$\mathbf{P}_2$ and $\mathbf{P}_3$, respectively. We denote the $\mathbf{P}_k^i$ for the i-th row of $\mathbf{P}_k$ matrix, and $\sim \mathbf{P}_k^i$ denote a sub-matrix of $\mathbf{P}_k$ by cancelling out the i-th row, then the trifocal tensor is computed as:

$$\mathcal{T}_i^{qr}\langle 123\rangle = (-1)^{i+1} \det \begin{bmatrix} \sim \mathbf{P_1}^i \\ \mathbf{P_2}^q \\ \mathbf{P_3}^r \end{bmatrix}, \qquad (1)$$

where the subscriptions of $\mathcal{T}$ is called covariant index, the two superscriptions are called contra-variant indices, and the $\langle 123\rangle$ are camera indices.

For NVS application, the task is to synthesize a novel virtual image (denoted by image-0) from two given real image (for example, image-1 and image-2). Conventionally, the tensor transfer method is used to transfer a matched real pixel from image-1 (or image-2) to the virtual image-0. This is the traditional way of doing tensor transfer, where the tensor is computed from two real input images to the third virtual image. We call such tensor an *direct-tensor*, in order to distinguish it from our newly proposed *inverse-tensor*.

For three cameras, there are in total twelve linearly independent trilinear relations. However, only three independent trifocal tensors can be identified. The three tensors are distinguished by different choices of the first image. For example, for image-0, image-1 and image-2, totally there are only three independent tensors can be obtained, say, $\mathcal{T}\langle 012\rangle$, $\mathcal{T}\langle 120\rangle$ and $\mathcal{T}\langle 210\rangle$.

We give the definitions of inverse tensor transfer as: *an **inverse trifocal tensor** is a trifocal tensor transfer relationship $\mathcal{T}\langle ijk\rangle$ where the third camera index k is a real camera.* While, in contrast, *the conventional **direct trifocal tensor** is the transfer relationship from two real cameras to a third virtual camera.* Although there is no new geometric relationship introduced by this inverse tensor transfer, the motivation of especially conceptualize this is that it provides a simple mechanism to establish pixel correspondences across multiple real input images.

There are different ways of doing trifocal tensor transfer, which are distinguished by employing different incident relations. These incident relations include for example the point-line-line, point-line-point or point-point-point[11]. We use the point-line-point relation here in this paper, mainly for its computational simplicity.

4 Rectification-Free Multibaseline Algorithm

Our rectification-free multibaseline stereo algorithm proceeds step-by-step as follows:

1. First we specify the location of the virtual right-eye camera (image-0) with respect to the reference image (image-1) and compute the inverse tensors of $\mathcal{T}\langle 10k\rangle$, $k = 2, 3, 4, \cdots, N$ from the relative position of the cameras. (The two reference cameras' relative position is obtained from any relative orientation algorithm.)

2. We then apply the traditional window-based SSD (sum of square difference) stereo matching algorithm for the ideal stereo pair of image-1 and image-0, as if they are real images. The searching is limited within the disparity range of $[D_{min}, D_{max}]$. We assume that all the SSD values are zero, because actually there is no pixel values in the virtual image. Note that no *actual* pixel-similarity matching is performed here.

3. At every hypothesized disparity, we can determine the positions of currently matched pixels at all of the other real images, i.e., image-2, image-3,$\cdots$, using the corresponding inverse tensors, i.e., $\mathcal{T}\langle 102\rangle$, $\mathcal{T}\langle 103\rangle$,$\cdots$. The actual transferring scheme we use is the *point-line-point* formula which is given by: For $x_i \leftrightarrow x^{'}_j$, and $l^{'}_j$ passing through $x^{'}_j$, we have

$$x^{''k} = \hat{x}^i \cdot l^{'}_j \cdot \mathcal{T}^{jk}_i\langle 102\rangle \qquad (2)$$

where the line is chosen to be the vertical line passing through the pixel of the virtual image. In fact, since image-1 and image-0 are in ideal stereo configuration, this line gives the optimal transfer relationship. In addition, the computational overhead is minimized.

4. Now for every disparity hypothesis, we have established a chain of matched pixels. Then we compute the conventional SSSD curve, by the formula:

$$SSSD(d) = \sum_{i=1}^{N} \sum_{(x,y)\in \mathsf{W}} \Phi(I_l^i(x+d,y) - I_r^i(x,y)) \tag{3}$$

where Φ is a robust-estimation kernel function to handle occlusions, w is SSSD window. Some candidates for the robust kernel Φ are:

$$\begin{aligned} \Phi(s) &= |s|, \\ \Phi(s) &= \tfrac{k^2}{2}\log(1+(s/k)^2) \end{aligned} \tag{4}$$

5. Choosing the minimal position as the resulting disparity, and output.

The inverse tensor plays a critical role in this algorithm because it offers a natural and simply way of relating N real corresponding pixels. The relations are guaranteed to be geometrically-valid since they are derived from the trifocal tensor. For continuous video sequence application or real-time robot navigation, the inverse tensor transfer relations need to be computed only once, and can be used later by making a look-up-table. So the computational overhead is not significant.

Our method also presents several contributions to the image based rendering area: Firstly, it alleviates a hidden problem which hinders most transfer based IBR methods, namely, the dense correspondence problem. Our method does not need any pre-computed depth map. We simply test for every possible disparity hypothesis whether all the matched pixels display a consistent color. In this regard, our method looks similar to the voxel-coloring algorithm[6] or image-based photo-hull methods. However, their methods perform in 3D scene domain, while ours fully performs in the image-domain, therefore is more efficient. Secondly, since we avoid dense correspondence, our method also frees from many troublesome problems that associate with the correspondence algorithms, for example, the *aperture problem*, highlight, and *textless regions* problem.

5 Experimental Results

Figure-2(1) gives the experimental setup of our multibaseline stereo system mounted on a prototype autonomous vehicles. The five cameras are arranged approximately linearly with optical axes in parallel. The central camera, which is chosen to be the reference camera, is also used as a time-base-synchronizer to ensure that all five cameras capture images at precisely the same time. The virtual camera is located to the right of the reference camera with a virtual baseline of 10×focal_length. Applying our new multibaseline algorithm, we obtained satisfactory 3D depth map and terrain reconstruction results. Figure-2(2) shows one example image, and figure-2(3) the obtained disparity map displayed in pseudo colors. Figure-2(4) is another example image, and figure-2(5) the corresponding terrain map. To compare the obtained depth map with the ground-truth depth

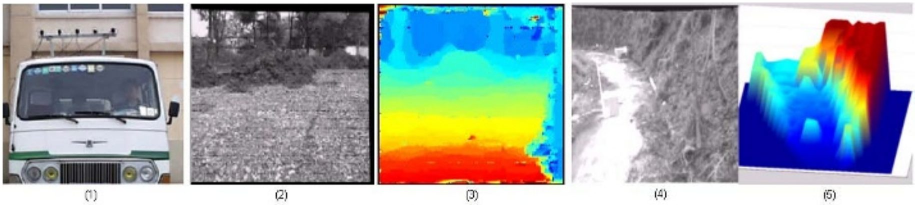


Fig. 2. A Multibaseline stereo system(read from left to right: (1)the camera system;(2) a sample image;(3) the computed depth map by our algorithm (4) another sample image (5) the computed terrain height map by our algorithm.)

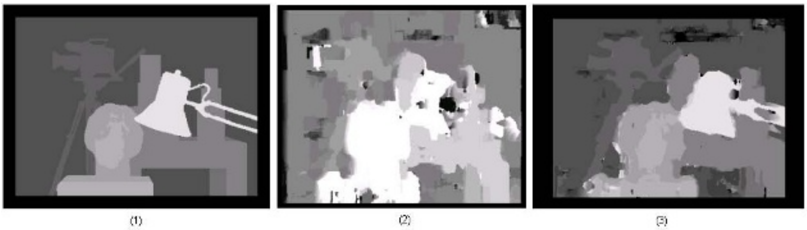
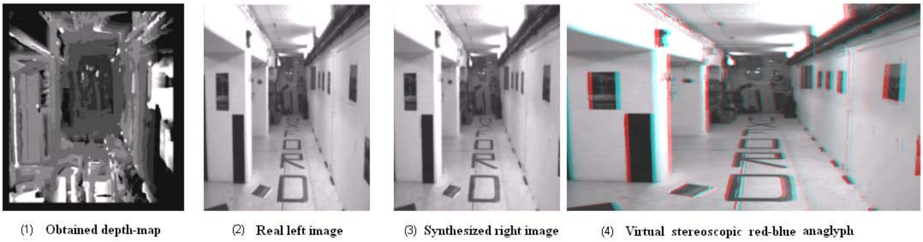


Fig. 3. Stereo correspondence results (read from left to right: (1) ground truth depth map; (2) result by conventional two-view SSD method; (3) result by our new multi-baseline method.)

map, we performed experiments on the Tsukuba stereo images. Prior to our algorithm, we intently introduce some arbitrary homographies to the original Tsukuba images (i.e., to make them to be un-rectified) in order to demonstrate our algorithm. The obtained depth map by our algorithm is shown in figure-3(3). For comparison purpose, figure-3(1) and 3(2) display the ground-truth depth map and the depth map obtained by a simple two-view SSD algorithm. It is seen that our algorithm produces good result at fairly efficient computation. We test our algorithm on more complex imageries. The camera is basically moving forward with a left side pan. So the images are in very general configurations, say, neither collinear nor parallel. In our experiments we do not use any calibration information. Even from the image contents itself, it is also a rather complex scenario for stereo matching: the white-wall and most part of the floor are all texture-less, and the ceiling light causes some intensity saturations at the right corner of the images. We have tested our algorithm on this sequence. The SSSD window size was set to 11x11. Figure-4(1) shows the obtained depth-map corresponding to frame-1. Though there contain some matching errors, the essential 3D scene structure is revealed. Remember that we actually started from a sequence which does not have any real stereo pair.

More interesting result is that: when we use this algorithm to generate novel views, even when the computed depth map is not so accurate, the synthesized image still looks very photo-realistic. Note the texture-less regions and highlight regions in the original image, which would destroy most conventional stereo



(1) Obtained depth-map (2) Real left image (3) Synthesized right image (4) Virtual stereoscopic red-blue anaglyph

Fig. 4. The results by our virtual stereoscopic generation algorithm(read from left to right: (1) the obtained depth map;(2)the original input reference image; (3) the generated virtual right-eye image;(4) display the images of (2) and (3) using stereoscopic red-blue anaglyph.)

algorithms. Figure-4(2) and 4(3) show the generated stereo pair, where the left image is the reference image and the right image is the virtual right-eye image. We display them again in the red-blue anaglyph format in figure-4(4), from which the computed stereo disparities can be easily verified.

6 Conclusions

The purpose of stereo rectification is no other than establishing a simple indexing mechanism for stereo matching. This paper proposes an alternative method to accomplish this goal. Our method makes use of a new *inverse tensor transfer* technique. Although geometrically this inverse tensor follows the same idea of conventional direct trifocal tensor, it does provide a natural and easier way to retrieval corresponding pixels across multiple views.

Our currently SSSD computation is justified by the popularity of the approximate linear multi-baseline configurations. For more general configurations where for example the visual occlusions and projective distortions are apparent, we consider incorporating some recent techniques of wide-baseline matching or affine invariant matching(eg, SIFT, or [22][21]to improve the performance.

Acknowledgments

National ICT Australia is funded through the Australian Government's Backing Australia's Ability Initiative, in part through the Australian Research Council. The authors are grateful to the anonymous reviewers' valuable comments.

References

1. M. Okutomi and T. Kanade, A multiple baseline stereo. IEEE-trans on PAMI, Vol.15, No.4, pp.353-363, 1993.
2. T. Williamson, C. Thorpe, A Specialized Multibaseline Stereo Technique for Obstacle Detection, Proc IEEE-CVPR-98, pp.238-244, 1998.

3. C. Loop, Z.Y. Zhang, Computing rectifying homographies for stereo vision. Proc. IEEE-CVPR-99,v1, pp.125-131,1999.
4. J. Mulligan, K. Kaniilidis, Trinocular Stereo for Non-parallel Configurations, Proc. ICPR-2000, 1:567-570,2000.
5. T. Sato, M. Kanbara, N.Yokoya, H.Takemura, Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-Baseline Stereo Using a Hand-Held Video Camera, Int. J. Comput. Vision, 47(1-3), pp.119-129, 2002.
6. S. M. Seitz and C. R. Dyer,Photorealistic scene reconstruction by voxel coloring. In Proc. IEEE-CVPR, pp. 1067-1073, 1997.
7. M.Kimura, H.Saito, T.Kanade, Stereo Matching between Three Images by Iterative Refinement in PVS, IEICE Trans. on Information and Systems, Vol.E86-D, No.1, pp.89-100, 2003.
8. Y. Li, S.Lin, H. Lu, S.B. Kang and H.Y. Shum, Multibaseline Stereo in the Presence of Specular Reflections. In Proc. ICPR-2002, 2002.
9. A. Fusiello, E. Trucco, and A. Verri, Rectification with unconstrained stereo geometry, Proc BMVC-1997, pp.400-409,1997.
10. J.M. Gluckman and S.K. Nayar, Rectifying Transformations That Minimize Resampling Effects, Proc. IEEE-CVPR-2001,2001.
11. R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd edition, Cambridge University Press, 2004.
12. D. Papadimitriou and T. Dennis. Epipolar line estimation and rectification for stereo image pairs. IEEE trans on Image Processing, 5(4):672-676, 1996.
13. M. Pollefeys, R. Koch, and L. Gool. A simple and efficient rectification method for general motion. Proc. of the 7th ICCV-1999, 1999.
14. D.Scharstein and R.Szeliski, A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, IJCV 47(1-3), pp.7–42, 2002.
15. Y. Li, S. Lin, H. Lu, SB Kang, and HY Shum, Multibaseline stereo in the presence of specular reflections, ICPR-2002,, vol.3, pp. 573-576, 2002.
16. Lim, Mital, Davis and Paragios, Uncalibrated stereo rectification for automatic 3D surveillance, Proc. IEEE-ICIP-2004, 2004.
17. A. W. Fitzgibbon, Y. Wexler and A. Zisserman, Image-based rendering using image-based priors, Proc. ICCV-2003, 2003.
18. S. Avidan, A. Shashua. Novel View synthesis by Cascading Trilinear Tensors, IEEE-Trans on Visualization and Computer Graphics Vol.4, Iss.4 ,1998.
19. P. Mordohai, G. Medioni, Dense Multiple View Stereo with General Camera Placement using Tensor Voting, Proc. of 3DPVT'04, 2004.
20. J. Xiao, M. Shah, From Images to Video: View Morphing of Three Images. Vision, Modeling, and Visualization, 2003. (VMV2003),pp.495-502, 2003.
21. P. Pritchett, A.Zisserman, Wide baseline stereo matching, Proc. ICCV-1998,pp.754-760, 1998.
22. V. Ferrari, T.Tuytelaars, L. Van Gool, Wide-baseline muliple-view Correspondences, IEEE-CVPR-2003, pp711-718, 2003.