Stereo Camera Tracking for Mixed Reality-Based PreViz of Stereoscopic 3D Cinema Using ICP Algorithm

Shohei Mori, Fumihisa Shibata, Asako Kimura, and Hideyuki Tamura
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, Japan

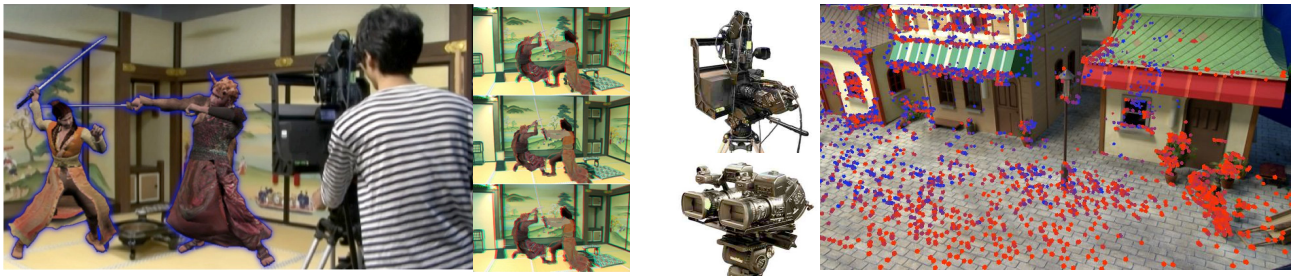


Figure 1. The left figure shows an image of S3D MR-PreViz in a studio and S3D still-shots of S3D MR-PreViz movies with various depths. Acting virtual characters are superimposed onto stereo frames in real-time for shooting PreViz movies on site. The middle figure shows cinematographic stereo cameras, called stereo rig, used in S3D MR-PreViz system (Top: Beamsplitter rig. Bottom: Parallel rig). Two Sony PMW-EX3 are mounted on each. The right figure shows reprojected multiple sub-maps onto a frame, which are coded in color (red to green) to show multiple sub-maps are registered in an entire map. Sub-maps are built and are extended along a rehearsal path determined in advance.

Abstract

This paper presents a method of real-time 6-DOF stereo camera tracking using ICP algorithm used in our mixed reality based previsualization (PreViz) system for stereoscopic 3D filmmaking. We eliminate the time-consuming preparation processes required for improving the accuracy in our previous work. On the assumption that the camera repeatedly passes along a path roughly determined at the rehearsal, multiple maps is created along the path using stereo vision technology and is continuously added and refined during a PreViz shooting. Experiments with real data in a miniature set and full scale set for filmmaking were conducted.

1. Introduction

The number of augmented reality (AR) and mixed reality (MR) applications in art and entertainment have been increasing [1, 2]. This rapid growth in AR/MR is due to enormous contributions and active developments being made, particularly in that of vision-based monocular camera tracking methods [3] in the last decade. One of the practical applications based on AR/MR and robust tracking technologies is our MR-PreViz [4]. PreViz, a short term of Pre-Visualization, refers to simple computer-generated shots created in the preproduction phase in filmmaking. Well-planned PreViz shots are used as shooting templates and performances in PreViz are re-used in actual shootings. We applied MR technologies to PreViz for supporting better preparation in the real world. After fetching geometric information and storing it to a database, called Landmark Database (LMDB), for tracking, MR-PreViz enables cinematographers to shoot PreViz in unknown indoor and outdoor scenes in the real world by superimposing acting virtual characters onto the scene captured by a cinematographic camera.

The system is actually used in several filmmaking and we are constantly improving it. We also added support for cinematographic stereo cameras, called stereo rig (The middle of Fig. 1), to the system for stereoscopic 3D (S3D) movie production [5]. This system, called S3D MR-PreViz, provides high definition (HD) S3D PreViz in real-time (The left of Fig. 1). It achieves MR composition on the left and right images based on camera pose obtained from the above monocular camera tracking on the left camera, but the right image is ignored for the tracking. We purposely selected the tracking style; after we fetch geometric information of a scene for the monocular tracking, the baseline and the convergence of a rig can be modified to change the looks of S3D by trial and error. It is, however, troublesome to construct LMDB for PreViz although it offers stable tracking results. In this paper, we propose a vision-based tracking method for achieving stereo camera pose estimation without the LMDB construction, and discuss the applicability of the tracking method for PreViz in S3D filmmaking.

2. Related Work

The tracking method employed in S3D MR-PreViz is called Rehearsal Path Method (RPM). This Structure from Motion (SfM) based approach uses a monocular camera. RPM achieves real-time 6-DOF tracking under the following constraint: the camera path at the rehearsal is roughly known in advance. Therefore, geometric information observed from the path is stored as LMDB. The LMDB allows for robust camera tracking in real-time. However, empirical data shows that the whole operation takes about fifteen minutes to complete. In this paper, we propose a method based on stereo vision and Iterative Closest Points (ICP) algorithm [6] without the

time-consuming preparation by mapping and extending the map during the tracking (The right of Fig. 1).

Stereo vision [7] has been employed for 6-DOF localization in robotics. Simultaneous Localization and Mapping (SLAM) using stereo vision is often referred to as stereo SLAM [8 - 10]. However, the speed of such approaches is less than 24 [FPS], which is preferable for PreViz, owing to computationally expensive features such as SIFT [8, 9], edge-points [10], etc. SLAM considers maps building and tracking of wide areas such as long looping corridor in a building or one entire room. However, what we are trying to do with S3D MR-PreViz is in a much smaller scale. Thus, the complicated processes mentioned above could be simplified for real-time operations.

Popular real-time tracking methods actively utilize hardware accelerators. Parallel Tracking and Mapping (PTAM) separates mapping process, including the time-consuming bundle adjustment and tracking process, into multi-threads processed on a multi-core CPU [3]. KinectFusion tracks a scene densely described in voxels using a rangefinder, Microsoft Kinect sensor, and GPGPU [11]. We also utilize hardware accelerators (multi-core CPU) and some constraints on a stereo rig described below in detail.

3. Settings of the Stereo Rig

A stereo rig has mechanisms of changing the baseline length and the convergence manually or mechanically to change the looks of S3D. One of the useful guidelines used for the setup for safe S3D is the "1/30th rule," which refers to *"the interaxial distance should be 1/30th of the distance from the camera to the first foreground object"* [12]. We apply the constraints for stereo rigs settings to simplify determining the disparity search range of stereo matching ($d_{\min}$, $d_{\max}$), which varies when the settings are changed and is bothersome to be set manually.

Stereo Matching

We assume that a stereo rig is calibrated and rectified by the procedure in S3D MR-PreViz system [5] in advance. Therefore, the 3D point $\mathbf{p} = (X, Y, Z)$ used for the camera pose estimation is calculated from point (u_L, v_L) on the left image and point (u_R, v_R) on the right image by using a well-known formula (*Eq.* 1).

$$\mathbf{p} = \left(\frac{B(u_L - c_u)}{2d}, \frac{f(v_L - c_v)}{d}, \frac{fB}{d} \right)^T \quad (1)$$

Here, B is baseline, f is focal length, (c_L, c_R) is position of optical axis, d is disparity (= u_L - u_R), and v= v_L = v_R.

Native Pixel Parallax (NPP)

In stereography, the maximum disparity is calculated by native pixel parallax (NPP) to avoid painful retinal rivalry area and to create safe S3D movies [12]. This means that no object with larger disparity than the NPP is observed on the input image. Thus, we utilize the NPP as the maximum disparity $d_{\max}$ (*Eq.* 2).

$$\mathrm{NPP} = d_{\max} = \left(\frac{W_{interocular}}{W_{screen}} \right) W_{px} \quad (2)$$

Here, $W_{\mathrm{interocular}}$ is interocular distance (= 2.5 [inches]), W_{screen} is screen size in inches, and W_{px} is width of input image in pixel. Using $d_{\max}$, we calculate the minimum depth, $Z_{\min}$ using the following formula.

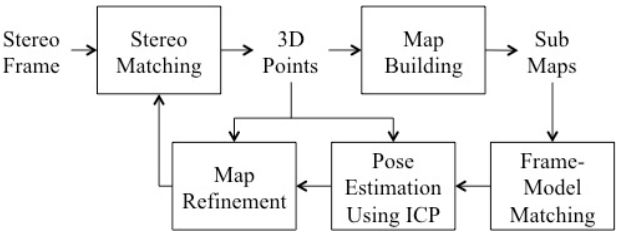


Figure 2. Procedure of the proposed method.

$$Z_{\min} = \frac{fB}{d_{\max}} \quad (3)$$

Stereo Uncertainty

The depth estimation by stereo matching has uncertainty often referred to as stereo uncertainty (ΔZ) calculated using the following formula,

$$\Delta Z = \frac{Z^2 \partial d}{fB} \quad (4)$$

where ∂d is subpixel precision.

Based on this formula, the maximum depth $Z_{\max}$ and the minimum disparity $d_{\min}$ are calculated by the following equations respectively,

$$Z_{\max} = \sqrt{\frac{2\partial Z fB}{\partial d}} \quad (5)$$

$$d_{\min} = \sqrt{\frac{\partial d fB}{2\partial Z}} \quad (6)$$

where ∂Z is $\Delta Z/2$.

These constraints can reduce mismatches and computational cost by ignoring pixels between 0 to $d_{\min}$ in stereo matching. For simplicity, convergence is assumed to be 0 degree, and these issues will be addressed in the future works.

4. Stereo Camera Tracking and Mapping

We propose to utilize ICP algorithm with the input being reasonably selected 3D points for stereo camera pose estimation and to build maps using stereo matching during S3D MR-PreViz shooting. Fig. 2 shows the procedure of proposed method. In this section, each step is explained in detail.

4.1. Stereo Matching

We utilize 3D points extracted from consecutive input stereo frames by simple stereo matching for map building and camera pose estimation between frames. Limiting disparity search range described in the previous section would reduce computational cost.

While shooting S3D MR-PreViz, the baseline tends to be shorter than expected because of the 1/30th rule for achieving safe stereoscopy during camera tracking. Therefore, subpixel precision is necessary for tracking in practical situations.

We employ subpixel refinement introduced by OpenCV on the left image after feature point detection using FAST feature detector [13]. Next, after stereo matching of Normalized Cross Correlation (NCC) criterion with 11x11 [px] window within the limited range, NCC values in the range is used to find the subpixel $\mathbf{p}_{\mathrm{subR}}$ on the right image by fitting a parabola (*Eq.* 7).

$$\mathbf{p}_{subR} = \begin{bmatrix} \dfrac{R(u_R-1,v)-R(u_R+1,v)}{2R(u_R,v)-4R(u_R-1,v)+2R(u_R+1,v)} \\ v \end{bmatrix} \quad (7)$$

Here, R(u, v) is a calculation of NCC at pixel (u, v). The above processing is considered as one set and each set is assigned to a thread in OpenMP (multi-threading) to reduce computing time.

4.2. Frame-Model Matching

We propose to manage one large map as multiple small maps, which we refer to sub-maps. This allows that some closest sub-maps are matched simultaneously using multi-threading technology. Whole points in each sub-map are reprojected and points within the image plane are considered as candidates for matching between the sub-map and the current frame. A criterion of selecting sub-maps is the distance between pose in a sub-map and of the current frame. This multiple sub-maps approach is more computationally efficient rather than reprojecting whole points within the entire map especially when the map becomes large.

We apply motion model in [3] for the point reprojection to the current frame. Newcombe *et al*. [14] suggest that this approach is rather weak at fast motion but it would not be a problem because we can assume that camerawork is smooth in filmmaking owing to equipment such as stable head, dolly, rails, stable camera, etc.

We applied perspective transformation to templates for NCC template matching because simple template matching is very weak at perspective view, scale change, and rotation, especially when the camera is far from where a sub-map was built. Each template image $I_{template}$ is extracted from a stored image in the i^{th} sub-map I_i, given four corners of a window (x_j, y_j, f) (j = 0, 1, 2, 3) in n^{th} camera coordinates (*Eq.* 8).

$$\mathbf{I}_{\text{template}} = \mathop{\mathbf{H}}_{j}\left(\phi_{n,i}\left(\mathbf{P}\left(x_j, y_j, f\right)\right), \left(x_j, y_j, f\right)\right)\mathbf{I}_i \quad (8)$$

Here $\mathbf{P}(\mathbf{p})$ is a projection matrix of $\mathbf{p}$, $\varphi_{n,i}(x, y)$ is reprojection of a point to n^{th} frame to i^{th} sub-map, and $\mathbf{H}(\mathbf{p}, \mathbf{p}')$ is perspective transformation matrix from $\mathbf{p}$ to $\mathbf{p}'$. This procedure is also applied for the multi-threading.

4.3. Pose Estimation Using ICP Algorithm

We apply point-to-point ICP algorithm formulated below by using singular value decomposition (SVD) proposed by Arun *et al*. [6] (*Eq.* 9).

$$\min_i\left(\left\|\mathbf{p}'_i - \mathbf{p}_i\right\|\right) \quad (9)$$

Input points are selected from frame-model correspondences based on heuristics described below.

On the assumption that 3D points are static between frames, a square measure of a triangle composed of three 3D points is the same between frames. Therefore, we randomly select three points from matched correspondences and compare its square measure between matched frames. If the change is large, then the triangle is ignored, and if it is small, we use the center of gravity as one input 3D point for the ICP algorithm.

This random selection continues until statistically reliable number of points S is extracted. The statistically reliable number is calculated by the following well-known formula in statistics,

$$S = \frac{N}{\left(\dfrac{e}{k}\right)^2 \dfrac{N-1}{P(100-P)}+1} \quad (10)$$

where N is a population given by ${}_nC_3$ when n is number of matching points, e is the range of sampling error set to 0.05, k is reliability coefficient set to 1.96, and P is the population proportion set to 50. This removes outliers generated by mismatches in both stereo matching and in frame-model matching.

The previous camera pose is set as the initial state of the loop in ICP algorithm. This makes the loop finishes faster and gives a reliable result. In addition, 2D-3D correspondences described in section 4.2 are also used as input of ICP algorithm: 2D points are projected into the world coordinates and 3D points are reprojected onto the current frame in the world coordinates.

4.4. Mapping

Based on the camera pose $\mathbf{T}_i$ in the i^{th} frame, we transform points $\mathbf{p}^c{}_i$ obtained from stereo matching at the frame to register it to the entire map as a sub-map (*Eq.* 11). This map building is performed when the number of matched points is smaller than the threshold (number of matched points / number of observed points < 0.5).

$$\mathbf{p}_i = \mathbf{T}_i\mathbf{p}_i^c \quad (11)$$

We assume that our rig passes along the rehearsal path or where is close to it repeatedly. Thus, assigned sub-map is refined using the result of the stereo matching. A new location of a point $\mathbf{p}_i$' in a sub-map is written as

$$\mathbf{p}_i' = \frac{n_{ref}\mathbf{p}_i^{map} + \mathbf{p}_i}{n_{ref}+1} \quad (12)$$

where n_{ref} is the number of times of the point being referred and $\mathbf{p}_i^{map}$ is a point in i^{th} sub-map. n_{ref} is also used for deleting useless points, which has extremely low value even after the sub map referred to for a certain number of times.

A camera pose is also registered to the map, which is used for recovering from lost condition described in the next section. It is possible that multiple sub-maps are observed in a view (the right of Fig. 1), but duplicated points are not registered because only newly detected feature points are registered by eliminating tracked feature points from candidates of the registration.

4.5. Recovering from "Lost"

The camera tracking is determined as being "lost" when enough number of points is not matched or camera pose estimation obtained from ICP algorithm is far from that of motion model. These metrics are determined empirically; It is practically impossible to move and rotate a beamsplitter rig on a sachtler head on a camera rail or a dolly at 1,000 [mm/sec] and by 180 [deg/sec] respectively.

When the camera tracking is lost, previously tracked camera pose and multiple camera poses that were stored previously in sub-maps are used for recovering. All points in each sub-map are reprojected and frame-model matching is performed as described in section 4.2 with larger search window size. The matching of each sub-map is performed simultaneously using multi-threading.



(a) Miniature set

(b) Full scale set

Figure 3. Several frames from the input sequence

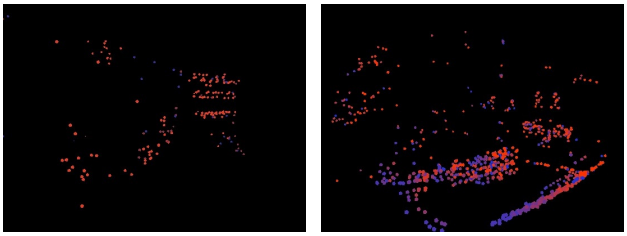


Figure 5. Result of continuous map building in the miniature set (left) and full scale set (right).

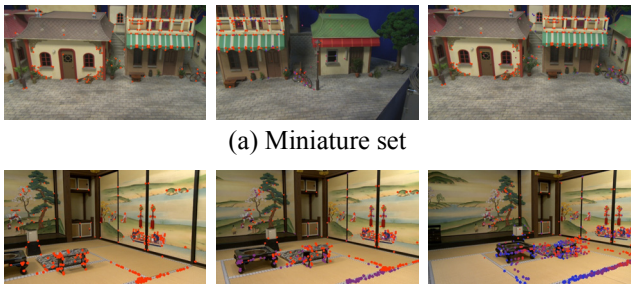
5. Experimental Results

We conducted experiments in two types of environments, a miniature set and a full scale set (Fig. 3). The baseline distance was set to 60 [mm] and 100 [mm] respectively based on the 1/30th rule. Images were captured using S3D MR-PreViz frameworks [5] while a beamsplitter rig moved on a rail. The size of the images were 720x405 [px]. The frame rate was 15.0 (miniature) and 11.8 (full scale) [FPS] and changed from 6.2 to 19.6 (miniature) and 6.0 to 13.2 (full scale) [FPS] acording to the number of detected points by FAST. The most time-consuming process was filtering for input images (bilateral filter: 35.1 [ms], without witch 29.9 and 19.6 [FPS]). Therefore, it is necessary to integrate this process into GPGPU processing in S3D MR-PreViz or to alternative other filters.

The map building and refinement continued during the tracking (Fig. 4 and 5). Fig. 6 shows result of S3D MR-PreViz and the proposed system works properly. We believe that the maps can be stored in a certain form and applied to bundle adjustment for more accuracy while checking the results of MR-PreViz movies.

6. Conclusions

This paper has presented a method of real-time 6-DOF stereo camera tracking using ICP algorithm. Our previous work required for time-consuming preparation and it is accomplished in real-time using stereo vision in proposed method. We continuously build new sub-maps along with a rehearsal path and refine them at frames observed from the path during S3D MR-PreViz shooting for robust tracking. Future work will state the convergence settings on a stereo rig.



(a) Miniature set

(b) Full scale set

Figure 4. Reprojected sub-maps onto input frames



Figure 6. Result of S3D MR-PreViz in color anaglyph.

References

[1] C. E. Hughes, *et al.*: "Mixed reality in education, entertainment and training," IEEE Computer Graphics and Ap-Applications, Vol. 25, No. 6, pp. 24-30, 2005.

[2] J. M. Gauthier: "Building interactive worlds in 3D: Virtual sets and pre-visualization for games, film and the Web," Focal Press, 2005.

[3] G. Klein and D. Murray: "Parallel tracking and mapping for small AR Workspaces," Proc. Int. Symp. Mixed and Augmented Reality (ISMAR), pp. 225 - 234, 2007.

[4] R. Ichikari, *et al.*: "Mixed reality pre-visualization for filmmaking: On-set camera-work authoring and action rehearsal," The Int. J. Virtual Reality, Vol. 7, No. 4, pp. 25 – 32, 2008.

[5] S. Mori, *et al.*: "Enabling on-set stereoscopic MR-based previsualization for 3D filmmaking," SIGGRAPH ASIA 2011, Technical Sketch, 2011.

[6] P. Besl and N. McKay: "A method for registration of 3-D shapes," IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. 14, No. 2, pp. 239 - 256, 1992.

[7] M. Z. Brown, *et al.*: "Advances in computational stereo," Proc. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 25, No. 8, pp. 993 – 1008, 2003.

[8] S. Se, D. Lowe, and J. Little: "Local and global localization for mobile robots using visual landmarks," Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.414 – 420, 2001.

[9] P. Elinas, *et al.*: "σSLAM: Stereo vision SLAM using the Rao-Blackwellised particle filter and a novel mixture proposal distribution," Proc. IEEE Int. Conf. on Robotics and Automation, pp.1564 – 1570, 2006.

[10] M. Tomono: "Robust 3D SLAM with a stereo camera based on an edge-point ICP algorithm," IEEE Int. Conf. on Robotics and Automation, pp. 4306 – 4311, 2009.

[11] R. A. Newcombe, *et al.*: "KinectFusion: Real-Time Dense Surface Mapping and Tracking," Proc. ISMAR 2011, pp. 127 – 136, 2011.

[12] B. Mendiburu: "3D movie making: Stereoscopic digital cinema from script to screen," Focal press, 2009.

[13] E. Rosten and T. Drummond: "Machine learning for high-speed corner detection," Proc. 9th European Conf. on Computer Vision, pp. 430 – 440, 2006.

[14] R. A. Newcombe, *et al.*: "DTAM: Dense tracking and mapping in real-time," Proc. ICCV, pp. 2320 – 2327, 2011.