2012 IEEE 24th International Conference on Tools with Artificial Intelligence

Robust Frame Registration for Multiple Camera Setups in Dynamic Scenes

Xu Zhao[1,2], Zhong Zhou[1,2]*, Ye Duan[3] & Wei Wu[1,2]
1. State Key Laboratory of Virtual Reality Technology and Systems
2. School of Computer Science and Engineering, Beihang University, Beijing, China
3. CGIU Lab, Department of Computer Science, Missouri-Columbia University, USA
Email: {zhaoxu, zz, wuwei}@vrlab.buaa.edu.cn, duanye@missouri.edu

***Abstract*—In this paper, we propose a novel method to register frames from multiple cameras into a consistent global scale. Assuming a moving object is observed in multiple camera setups, we use initial frames to create a global reference structure where the pose variation of each new frame is estimated using a RANSAC-based registration algorithm. We further combine the registration method with other state-of-the-art techniques to build a high quality 3D reconstruction system with a smaller number of cameras than used by more traditional methods. Experimental results show that our method performs better and is more economical than the registration of separate monocular structures from motion methods. 3D reconstruction results on various challenging real-world multi-camera video datasets also illustrate the feasibility and robustness of our method.**

***Keywords*-Frame Registration; Bundle Adjustment; Multiple Camera Setups;**

I. Introduction

Modern multi-camera vision systems have emerged as a popular platform for recording real world dynamic scenes [1], [2]. Contrasted with single camera or stereo camera setups, multiple cameras can be fixed beforehand to cover the measuring space. They are more suitable for solving 3D vision problems from multiple images/videos, such as 3D reconstruction, motion capturing, tele-operation interaction, object tracking and so on. However, real-time applications are often inadequate as most multi-camera vision systems only equip a few cameras due to cost; hence, poor quality is common for reconstructed 3D models in practice.

It is a challenge to reconstruct accurate 3D models using eight or less static camera setups. To improve the reconstruction quality, algorithms usually need more viewpoints to correctly recover the reconstructed points or regions. Classical algorithms like visual hull [3], binocular stereo [4] and voxel coloring (or space carving) [5], [6] used in multi-camera vision systems are not designed for this purpose. On the other hand, multi-view stereo (MVS) algorithms (see Seitz et al. [7] for a survey) are capable of reconstructing accurate 3D models when given a moderate number of calibrated images as input. Multiple videos can be exploited to provide sufficient eligible images for MVS reconstruction in dynamic scenes; however, the position and orientation of each video frame should be first accurately estimated. Therefore, frame registration is one of the key issues which affect reconstruction accuracy.

Working with a virtual camera's dynamic scenes can be problematic when attempting to reproduce high-quality motion. One direct frame registration solution is to perform a traditional monocular structure from motion (SFM) [8] or SLAM algorithm [9] for each camera, and then register all new frames together. However, accumulative errors across multiple cameras will be introduced. Also, it is hard to unify the depth scales of all cameras when lacking reference geometry.

To solve these problems, we propose a novel frame registration method to register multi-video frames into a consistent global scale. One can start with at least 5 of static cameras to create a global reference structure, which also contains much more accurate tracking features than traditional methods. With the help of the reference structure, we formulate the problem into a constrained optimization problem and present a RANSAC-based registration algorithm to estimate the pose variation of each new frame. Then, we combine our registration algorithm with other techniques, like key-frame selection, bundle adjustment refinement, multi-view stereo and surface reconstruction, to build a 3D reconstruction system using a limited number of cameras. Finally, a high quality 3D mesh models will be generated for further applications.

The rest of the paper is organized as follows. In Section II, we review the multi-camera vision systems and discuss the frame registration methods. Section III gives a brief system overview. Section IV introduces the global reference structure in multiple camera setups. Based on the reference structure, we present our frame registration method in section V. We then combine the registration method with other techniques to reconstruct high quality 3D models within a limited number of cameras in section VI. Section VII shows the experimental results. Finally, we discuss and conclude future prospects.

II. Related Work

A. Camera setups

Carnegie Mellon University's Virtualized Reality project was one of the earliest multi-camera vision systems used to

*Corresponding author. Email:zz@vrlab.buaa.edu.cn

1082-3409/12 $26.00 © 2012 IEEE
DOI 10.1109/ICTAI.2012.58

reconstruct real-world events [10]. They constructed a studio, the 3D Dome, which consisted of 51 cameras mounted on a geodesic dome 5 meters in diameter. Multi-baseline stereoscopic reconstructions provided a sense of complete immersion independent of the actual camera positions.

Recent similar multi-camera setups include ETHZ Blue-C system [11], INRIA GrImage system [2] and Tsinghua University's multi-camera Dome [12]. The application areas are further extended to motion capture, markerless interaction, light field modeling, and so on. According to the number of cameras, we classify camera setups into four categories: 1) single camera setup [13], 2) stereo camera setup [14], 3) setup with a small number of cameras [2], [11] and 4) setup with a great number of cameras [10], [12]. A comparison of these setups is given in Table I.

Table I
A COMPARISON OF DIFFERENT CAMERA SETUPS

	Main purpose	Advantage	Disadvantage
Single camera setup (hand-held)	Urban or interior scene modeling	Cheapest	It needs user interaction and can only reconstruct static scene
Stereo camera setup (2-3 cameras)	Stereoscopic 3D production	Simple but practical in industry	It can only reconstruct rough depth image
Multi-camera setup (5-8 cameras)	Motion capture	More common in real scenes and systems	It can only reconstruct rough 3D models
Multi-camera setup ($>$ 20 cameras)	Light field modeling	High quality 3D reconstruction	Expensive and for research purpose only

This paper focuses on the multi-camera setup, especially the setup with a limited number of cameras. This kind of camera setup is more common in real scenes and systems, like indoor/outdoor multi-camera monitoring systems, motion capture systems for film production, tele-operation training or immersion systems. However, due to the number of limitations, these systems usually generate rough to poor 3D models, which can negatively affect the user experience.

Multi-camera video based reconstruction method aims to accurately reconstruct 3D models using a limited number of cameras. Tung and Matsuyama [15] achieved accurate and complete reconstruction results by combining narrow and wide baseline stereo and then fusing in a probabilistic framework. Temporal cues are introduced to overcome the limitations of MVS reconstruction using 14 cameras. However, instead of developing any video-based reconstruction algorithm or mounting 20 or more cameras in setup, our focus was on selecting eligible frames from multi-camera videos and accurately registering them into one unified coordinate system, which is a more affordable approach and more appropriate when creating any MVS algorithms.

B. *Frame registration algorithms*

Given a set of images depicting a number of 3D points from different viewpoints, bundle adjustment algorithms are usually used to simultaneously refine the 3D coordinates describing the scene geometry as well as the parameters of the relative motion and the optical characteristics of the cameras. Intuitively, there are two ways to apply the straightforward bundle adjustment to register frames in multiple camera setups.

The first method is to select several key frames separately from multiple cameras' videos and then perform a global bundle adjustment [16] for all images. However, matched 3D features are too sparse and noisy to estimate camera parameters, especially when the object to be reconstructed is textureless and small. Cheng et al. [17] manually initiated the feature correspondences across adjacent views. In contrast, our frame registration method will automatically register multi-video frames and ensure registration accuracy as much as possible.

The second method is to perform a traditional structure from motion (SFM) [8] or mono-SLAM algorithm [9] for each camera, and then register all selected frames together [18]. However, this kind of method will introduce accumulative errors by not considering the relationship across multiple cameras. Moreover, it is hard to unify the depth scales of multiple cameras when lacking reference geometry. Unlike the traditional SFM/mono-SLAM algorithm approach, our method creates a global reference structure for multiple cameras. New frames from these cameras are subsequently registered into one unified coordinate system, which not only avoids accumulative errors but also registers multi-video frames into a consistent global scale.

A similar idea of using a sparse 3D model as a reference structure was also proposed by Imre et al. [19]. With the help of the reference structure constructed by a set of static cameras, they tried to estimate the pose of another moving camera by using unscented Kalman filters. However, their method cannot register all frames from multiple cameras simultaneously. Our frame registration method creates a more accurate and denser reference structure. This allows the user to estimate the frame pose variation of all cameras with a RANSAC-based registration algorithm. Most notably, combining our frame registration method with other state-of-the-art techniques allows users to build a high quality 3D reconstruction system with a limited number of static cameras.

III. SYSTEM OVERVIEW

Our main goal is to provide a robust frame registration for multiple camera setups in dynamic scenes. To demonstrate the effect of our method, we built a multi-camera reconstruction system, which can also achieve a high quality 3D reconstruction result with only a limited number of cameras

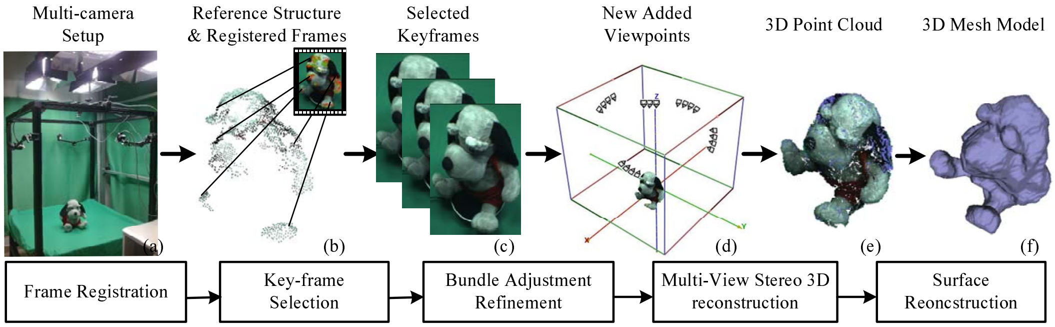


Figure 1. Framework of our multi-camera reconstruction system. Multi-camera videos are captured with five static cameras. During video tracking, frames are registered into one unified coordinate system based on a global reference structure. A number of eligible key frames are then selected. New added viewpoints are further refined using bundle adjustment. Accurate 3D point cloud model and mesh model are finally reconstructed from these key frames.

(e.g. 5 cameras in our setup). Figure 1 shows the whole system pipeline.

Frame registration is the first step in our multi-camera reconstruction system. Unlike the traditional structure from motion method, we firstly construct a global reference structure from multiple cameras (Section IV). Then we project the visible features to the initial frame and track the salient ones using optical flow. Given the 3D features and their reliable track pairs, we present a RANSAC-based Levenberg-Marquardt method to estimate frame poses (Section V). Finally, we introduce a heuristic key-frame selection method and combine other state-of-the-art techniques, including bundle adjustment refinement, multi-view stereo and surface reconstruction, to reconstruct high quality 3D models with only a limited number of cameras (Section VI).

To sum up, the main contributions of this paper include:

(1) A robust frame registration method for multiple camera setups, based on a global reference structure, which can register all frames of all cameras into one unified coordinate thereby preventing accumulative errors.

(2) A high-quality 3D reconstruction system which can select a moderate number of calibrated key-frames from multi-camera videos and combine both the narrow and wide baseline stereo for more accurate reconstruction.

(3) A more economic approach to recording motion due to a more efficient use of fewer cameras than what is commonly used by practitioners in the video recording field.

IV. Reference Structure

A reference structure is a consistent 3D geometry which helps to implement a frame registration algorithm. In traditional structures built from motion (SFM) methods, 3D feature points are first estimated through the initial two frame reconstruction. These 3D points, together with corresponding 2D features, are then used to compute the camera motion parameters. Generally, with more 3D features and less tracking errors, higher registration accuracy is achieved.

However, in multiple camera setups, the traditional SFM methods are inappropriate. We find that the dynamic scenes observed by multiple camera setups usually contain less texture. In this situation the number of classical tracking features, such as KLT, FAST and SURF etc., may be insufficient for the camera parameter estimation. Moreover, due to the unknown depth scales during the initial two frame reconstructions, the depth scales estimated from different camera video sequences will not match.

Therefore, to improve the frame registration accuracy in multiple camera setups, we developed a global and accurate reference structure, which can also provide the maximum number of tracking points possible. To meet these requirements, we used the state-of-the-art multi-view stereo (MVS) algorithm to construct the reference structure from multiple cameras. Although the generated 3D points are usually sparse when using a small number of cameras, the number of feature points is much higher than the number found in classical tracking features.

We define the reference structure as a set of 3D patches. The representative equation follows:

$$\{\alpha | \alpha := (p, v, s)\} \tag{1}$$

where p is the position, v is the visibility set and s is the feature salience measure set of all visible 2D features. Here we use patch-based multi-view stereo (PMVS) [20] with initial frames to generate the 3D patches as well as their positions and visibility sets. Then we project them to the corresponding images to obtain candidate 2D features. However, not all features are good enough to track. So we use the Shi-Tomasi score [21] to measure the salience of candidate features and save these values to the reference structure.

V. Frame Registration

The goal of the frame registration algorithm is to calibrate the external parameters of each frame. Unlike other frame registration algorithms in most multiple camera setups, multiple static cameras can be accurately pre-calibrated using a calibration object (e.g. chessboard). This means that the internal and initial external parameters of multiple cameras can be determined before registering new frames.

As mentioned in Section IV, a global reference structure can be explicitly constructed beforehand. Then we project 3D features to the corresponding visible images, and use the robust Lucas-Kanade-Pyramid optical flow method [22] to track these 2D salient features. The external parameters of new frames can be updated during tracking. Figure 2 shows an example of a reference structure and the corresponding tracking features. The green points are the tracking features which project onto five static cameras. The red lines indicate the tracking trajectory.

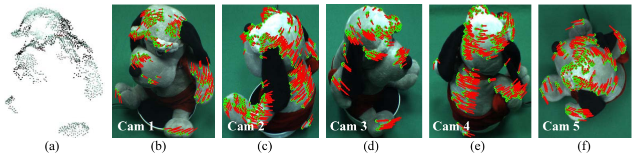


Figure 2. Video tracking. 3D features (reference structure, cf. Left) are projected into visible image planes as 2D features (in green) to track. If the 2D features are not salient, they will be discarded in all subsequent frames during optical flow tracking.

Note that although we adopted a Shi-Tomasi threshold to determine the salient features, we cannot set this value too high in order to retain a sufficient amount of features (considering the outliers). Also the tracking errors must be considered in the whole process. Once the Shi-Tomasi score of a feature is below the threshold, we discard it in all subsequent frames. Silhouette constraint is used to avoid background noises as well.

According to the above conditions, we formulate the frame registration process. Let us assume $\Phi = \{\alpha_i \mid i \geq 1\}$ is the set of 3D patches, and let m_{ij}^n denote the point where the 3D patch α_i projects onto the camera j at the keyframe n. (m_{ij}^n, m_{ij}^{n+1}) is a track pair given by the video tracking algorithm. At first, the projection matrix P_j^0 of the 0th frame is known and all camera intrinsic matrices $\{K_j \mid j \geq 1\}$ are the same all the time. Then, the succeeding frame pose variation can be estimated through the rotation matrix R_j^{n+1} and translation vector T_j^{n+1}, as shown in Figure 3.

With the help of 3D patches and their track pairs, we can derive the new frame pose by minimizing the following function:

$$[R_j^n, T_j^n]^* = \arg \min_{[R_j^n, T_j^n]^*} \sum_{\alpha \in \Phi} (f(\alpha_i, K_j[R_j^n, T_j^n]), m_{ij}^n)^2 \qquad (2)$$

$$s.t. \det(R) = 1, R^T = R^{-1}$$

where function $f(\alpha_i, P_j^n)$ denotes projecting 3D patches α_i to the nth frame of the camera j using projection matrix P_j^n. The optimization constraint is to ensure that matrix R is a rotation matrix. Instead of using rotation matrix, we use quaternion representation, which not only reduces the parameter DOFs from 12 to 7, but also helps us to convert this problem to unconstrained optimization. Moreover, the quaternion vector and translation vector vary continuously during tracking. We use the Levenberg-Marquardt method to solve equation (2) and use the RANSAC method to improve the robustness when considering tracking errors and small non-rigid motion. The overview of our frame registration algorithm for multiple camera setups is as shown in Figure 4.

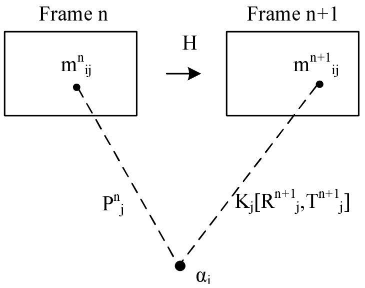


Figure 3. Notation. 3D patches α_i projects to frame n at point m_{ij} by known projection matrix P_j. (m_{ij}^n, m_{ij}^{n+1}) is a track pair of 3D patch α_i, which will be used to recover the pose variation of subsequent frame $n+1$.

```
PROCEDURE FrameRegistrationForMultipleCameraSetups()
  recontructReferenceStructure()
  FOR each camera j
    captureFrame(j)
    FOR each frame i
        IF i=0 THEN
          determineTrackingFeatrues(i)
        ENDIF
          pyramidOpticalFlowTracking(i)
          estimateFramePoseRANSAC(i)
          //* used in the reconstruction system
        IF isKeyframe(i) = TRUE THEN
          saveFrameAndPose(i)
        ENDIF
    ENDFOR
  ENDFOR
  //* used in the reconstruction system
  bundleAdjustmentRefinement()
END
```

Figure 4. Pseudo-code of our frame registration algorithm.

VI. Multi-camera reconstruction system

After new frames are registered into one unified coordinate system, multiple view geometry approaches can then be used to interpret the dynamic scenes. One of the important applications is to reconstruct 3D models using multiple camera setups. It is essential for most multi-camera vision system, such as tele-immersion system, markerless 3D interaction system and others. However, the quality and the appearance of the reconstructed 3D models are still poor when only using a limited number of cameras. In this section, we combine our frame registration algorithm with other state-of-the-art techniques to obtain a high-quality initial 3D models with only a limited number of cameras.

Figure 1 illustrates the framework of our multi-camera reconstruction system, which basically consists of five phases: (1) frame registration, (2) key-frame selection, (3) bundle adjustment refinement, (4) multi-view stereo 3D reconstruction and (5) surface reconstruction.

In Section IV and V, we have presented the frame registration algorithm for multiple camera setups. Additionally, we select several eligible key frames for 3D reconstruction during video tracking. In literature, numerous key-frame selection methods are proposed for the structure from motion recovery [23], [24]. However, these methods are not suitable for multi-view stereo reconstruction, which usually requires a proper baseline between adjacent key-frames for stereo matching. To combine the narrow and the wide baseline stereo reconstruction in multiple camera setups, we define a simple but effective key-frame selection scheme illustrated in Figure 5.

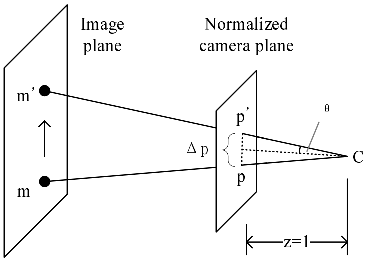


Figure 5. Basic idea of our key-frame selection scheme. Turn angular offset θ is used to measure the motion.

Given track pairs, we use turn angle offset to measure the camera motion. Assume that the camera C observes the traced feature m moving to a new position, m', in the image plane. Since the camera intrinsic matrix is known, we project these two pixels onto the normalized camera plane (depth $z = 1$). We approximately use the tangent function to define the turn angular offset θ:

$$\theta \approx \arctan\left(\frac{\Delta p}{2z}\right) = \arctan\left(\frac{|p' - p|}{2}\right) \qquad (3)$$

Considering the tracking noise, we sort all the tracked features, and take the median value to compute and compare with the threshold. For stereo matching, the baseline between two adjacent frames cannot be too narrow or too wide. We set 2θ equals to 5 degrees as threshold. For each video sequence, we selected four to six key frames, which is enough to accurately reconstruct parts of the object from one camera viewpoint.

Once all selected key frames are available, we used the patch-based bundle adjustment method [25] to further refine the calibration result. We iteratively performed this procedure four times to achieve a better viewpoint calibration results for multi-view stereo reconstruction. Then we again used PMVS algorithm [20] to reconstruct the 3D point cloud model. Users can optionally extract silhouette information from the first frames to avoid background noise. Surface reconstruction algorithms like Poisson surface [26] or Touch-expand algorithm [27] can generate the final watertight mesh model. Our multi-camera reconstruction system can also use other algorithms instead, like the detail feature preserving surface reconstruction method [28] etc.

VII. Experimental Results

A. System Implementation

A multi-camera acquiring platform shown in Figure 1(a) was set up. Five Flea2 cameras provided by PointGrey, Inc. are fixed on the platform at five different positions which can basically cover the measuring space. The target object(s) should be placed in the measuring space to make sure they can be observed by all cameras. A synchronized multi-camera acquiring system is developed and deployed on PCs. Multiple cameras are firstly calibrated by Bouguet's geometric calibration toolbox [29]. And the color calibration of multiple cameras is done by our radiance-based method [30].

We experiment the running time of our proposed system in a PC with 3.0Hz Intel Core 2 Duo CPU. Except manually calibrating static cameras and capturing dynamic scenes, all other phases can be performed automatically. The PMVS reconstruction, including initializing reference structure and generating 3D point cloud model, needs about 3 to 10 minutes according to the number of input images. Our frame registration and selection method can achieve nearly 7.5 fps. The bundle adjustment refinement and final surface reconstruction phases cost about 10 minutes. Therefore, the total computation time of our system is under thirty minutes.

Table II
Summary of parameters used in our system.

Parameters	Value
Initial PMVS (L/C/T/N/S)	1/2/0.7/2/11
Turn angular offset threshold	0.4
Shi-Tomasi threshold	10-200
Optical Flow (L/S/I)	5/5/20
PBA (L/E/S)	2/7-4/7
Final PMVS (L/C/T/N/S)	1/2/0.7/3/7
Touch-expand (V/R)	200/2

Parameters used in our system are summarized in Table II. In practice, we set PMVS level $L = 1$, density csize $c = 2$, photo consistent threshold $T = 0.7$, minimal visible image number $N = 2$ and window size $S = 11$ to get as dense as possible stable 3D features. In the final PMVS reconstruction stage, we change minimal visible image number $N = 3$

Table III
A COMPARISON WITH THE GROUND TRUTH

	Position diff.(mm)		Principal axis diff.(°)	
	Mean	Std. Dev.	Mean	Std. Dev.
Our method	1.3880	0.7698	0.0699	0.0380
SFM method	12.951	9.0739	0.6995	0.3764

and window size $S = 7$ for efficiency. We set optical flow pyramid level $L = 5$, window size $S = 5$, and the iterative number $I = 20$. Most parameters are fixed except for the Shi-Tomasi threshold and PBA expected reprojection error E. Shi-Tomasi threshold depends on the number of initial reconstructed 3D patches, and the PBA expected reprojection error E decreases from 7 to 4 in the iterative bundle optimization phase. The meaning of other PBA parameters is the same for PMVS parameters. In the Touch-expand algorithm, a voxel resolution of $V = 200$ and Gaussian filter radius at $R = 2$ was set to generate final watertight mesh models.

B. Frame Registration

To test the effectiveness of our frame registration method, we use a chessboard dataset to build a ground truth with Bouguet's toolbox [29]. Figure 6 shows one example image and reconstruction result of the chessboard dataset.

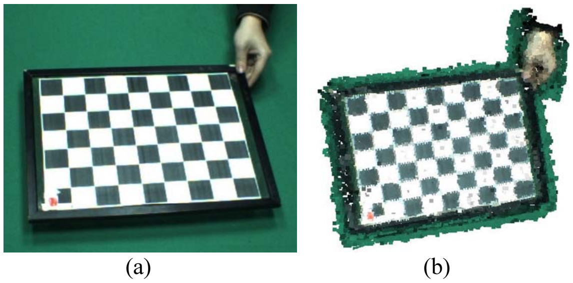


Figure 6. Example image (a) and reconstruction result (b) of the chessboard dataset, which are used as the ground truth to evaluate our frame registration algorithm.

Then, we compute the position difference as well as the principal axis difference between the estimations and the ground truth, as shown in Table III. For SFM solution, we use Boujou [31] to calibrate all frames of each camera and then register them to the reference frame of the static cameras.

Table III shows that the calibration of our method is consistent with the ground truth, while the SFM method deviates from the ground truth due to the accumulative errors.

However, features on the chessboard are usually easy to track. More generally, we take the Dog dataset to measure the valid feature number and reprojection error of each frame. As shown in Figure 7, the feature number is reducing during tracking because of the track errors. But overall, our method can find more salient features and result in smaller reprojection errors. Thanks to the reference structure, our method can achieve a better calibration result than the registration of separate monocular SFM method.

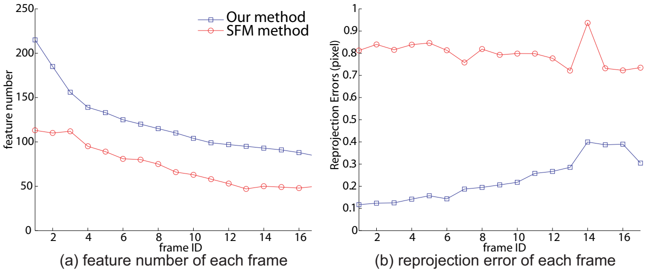


Figure 7. Comparison between our registration method and SFM method. Our method can find more salient tracking features and achieve higher accuracy in terms of reprojection error.

We further use epipolar geometry to check the consistency between pairs of calibrated images. For a pair of images, we draw pairs of epipolar lines which pass through the same feature points in two images. Take keyframe 1 and 5 from camera 2 in Dog dataset for example. Shown in Figure 8, the initially recovered pose contains errors (approximately 2-4 pixels in some places). After bundle optimization, the same color epipolar lines accurately pass through the same features.

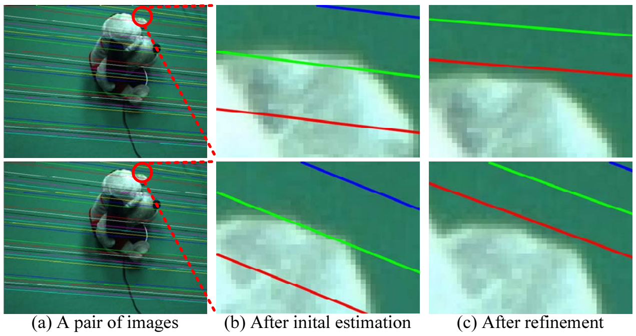


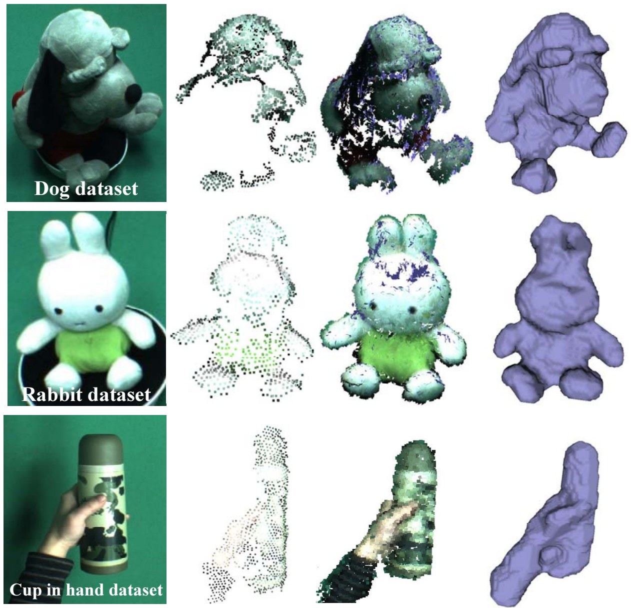
Figure 8. Epipolar lines are used to evaluate our frame registration method. Areas with red circle in the image pair (a) are zoomed in (b) to show that pairs of epipolar lines pass through the same feature points as well as with bundle adjustment refinement (c).

C. 3D Reconstruction

We also evaluate the proposed system on three real-world multi-camera video datasets. These datasets are captured synchronously at 15 fps on our multi-camera acquiring platform. Our datasets are more challenging than other static image datasets for two reasons. First, the image resolution is quite low to ensure a guaranteed frame rate. Second, the new selected key-frame may contain motion blurs which will affect the matching process in MVS reconstruction. Besides, the target objects are well chosen to study the application gamut of the proposed algorithm. In summary, the datasets contain several dimensions of properties: objects with fine details but complex surface, objects with weak texture, and objects which can move freely in hand. Three datasets are separately identified as Dog, Rabbit and Cup in Hand. Their properties are listed in TableIV.

Table IV
REAL WORLD MULTI-CAMERA VIDEO DATASETS.

Dataset	Dog	Rabbit	Cup in Hand
# Camera	5	5	5
# Image	30	30	40
Image size	640 × 480	640 × 480	640 × 480
Object size	$25 * 25 * 30$	$10 * 10 * 20$	$8 * 8 * 25$
Dimension	complex	textureless	moving free



(a) example images (b) reference structure (c) 3d point cloud (d) mesh model

Figure 9. Example images (a), reference structure (b), 3D point cloud (c) and mesh model (d) illustrate reconstruction results of the Dog dataset (cf. Top), the Rabbit dataset (cf. Middle) and the Cup in Hand dataset (cf. Bottom). Our proposed system can reconstruct high quality 3D models with a limited number of static cameras.

The reconstruction results and example images of Dog, Rabbit and Cup in Hand are shown in Figure 9. After selected key frames are added to input images, the reconstructed 3D point cloud models (Figure 9(c)) grow much denser than the initial reference structure (Figure 9(b)). The final reconstructed mesh models (Figure 9(d)) can recover concave regions and ensure the reconstruction completeness which other methods like visual hull or voxel coloring cannot achieve. All the datasets are well reconstructed using the proposed algorithms regardless of self-occlution (e.g. Dog dataset), textureless region (e.g. Rabbit dataset) or motion blurs (e.g. Cup in Hand dataset). One of the important reasons is that our frame registration algorithm is effective and robust. It ensures that our system is not seriously affected by the structure of the object to be reconstructed. Therefore, our proposed system can reconstruct high quality 3D models with only a small number of static cameras.

VIII. CONCLUSION

This paper presents a robust frame registration algorithm for multiple camera setups in dynamic scenes. Benefit from the global reference structure is realized through the frame registration of all cameras combined into one unified coordinate system. This new frame regiistration system efficiently avoids accumulative errors which is a notable improvement over the results obtained by the separate SFM/mono-SLAM algorithm registration method. A high quality 3D reconstruction system which uses a limited number of static cameras, and thereby demonstrates the effectiveness of our frame registration algorithm. One limitation of our system is that our Levenberg-Marquadt method depends on the initialization of the parameters to be estimated. Future research plans include the study of dynamic reference structure based registration algorithm to support large-scale non-rigid motion in dynamic scenes.

ACKNOWLEDGMENT

This work is supported by the National 863 Program of China under Grant No. 2012AA011803, the Natural Science Foundation of China under Grant No. 61170188, and the National 973 Program of China under Grant No. 2009CB320805.

REFERENCES

[1] Q. Zhao, "A survey on virtual reality," *Science China Ser F-Inf Sci*, vol. 52, pp. 348–400, 2009.

[2] J. Allard, C. Menier, B. Raffin, E. Boyer, and F. Faure, "Grimage: markerless 3d interactions," in *SIGGRAPH'07*, San Diego, California, 2007, pp. 9–12.

[3] J.-S. Franco and E. Boyer, "Exact polyhedral visual hulls," in *British Machine Vision Conference (BMVC)*, Norwich, UK, 2003, pp. 329–338.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision (IJCV)*, vol. 47, no. 1, pp. 7–42, 2002.

[5] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Computer Vision and Pattern Recognition (CVPR)*, San Juan, Puerto Rico, 1997, pp. 1067–1073.

[6] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision (IJCV)*, vol. 38, no. 3, pp. 199–218, 2000.

[7] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 519–528.

[8] M. Pollefeys, L. van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, “Visual modeling with a hand-held camera,” *International Journal of Computer Vision (IJCV)*, vol. 59, no. 3, pp. 207–232, 2004.

[9] G. Klein and D. Murray, “Parallel tracking and mapping for small ar workspaces,” in *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, Nara, Japan, 2007, pp. 1–10.

[10] T. Kanade, P. Rander, and P. J. Narayanan, “Virtualized reality: constructing virtual worlds from real scenes,” *IEEE Multimedia*, vol. 4, no. 1, pp. 34–47, 1997.

[11] M. Gross, S. W, rmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt, “blue-c: a spatially immersive display and 3d video portal for telepresence,” in *SIGGRAPH’03*, San Diego, California, 2003, pp. 819–827.

[12] Y. Liu, Q. Dai, and W. Xu, “Point cloud based multi-view stereo for free-viewpoint video,” *IEEE transaction on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 407–418, 2009.

[13] M. Pollefeys, D. Nistr, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewnius, R. Yang, G. Welch, and H. Towles, “Detailed real-time urban 3d reconstruction from video,” *International Journal of Computer Vision (IJCV)*, vol. 78, no. 2, pp. 143–167, 2008.

[14] S. Heinzle, P. Greisen, D. Gallup, C. Chen, D. Saner, A. S-molic, A. Burg, W. Matusik, and M. Gross, “Computational stereo camera system with programmable control loop,” *ACM Trans. Graph.*, vol. 30, no. 94, pp. 1–10, 2011.

[15] T. Tung, S. Nobuhara, and T. Matsuyama, “Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo,” in *International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009, pp. 1709–1716.

[16] M. I. A. Lourakis and A. A. Argyros, “Sba: A software package for generic sparse bundle adjustment,” *ACM Transactions on Mathematical Software*, vol. 36, no. 1, pp. 1–30, 2009.

[17] C.-M. Cheng, S.-F. Wang, C.-H. Teng, and S.-H. Lai, “Image-based three-dimensional model reconstruction for chinese treasure–jadeite cabbage with insects,” *Computers & Graphics*, vol. 32, no. 6, pp. 682–694, 2008.

[18] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H. P. Seidel, “Markerless motion capture with unsynchronized moving cameras,” in *Computer Vision and Pattern Recognition(CVPR)*, Miami, FL, 2009, pp. 224–231.

[19] E. Imre, J. Y. Guillemaut, and A. Hilton, “Moving camera registration for multiple camera setups in dynamic scenes,” in *British Machine Vision Conference (BMVC)*, Aberystwyth, UK, 2010, pp. 1–12.

[20] Y. Furukawa and J. Ponce, “Accurate, dense, and robust multi-view stereopsis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 8, pp. 1362–1376, 2010.

[21] J. Shi and C. Tomasi, “Good features to track,” in *Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 1994, pp. 593–600.

[22] J.-Y. Bouguet, “Pyramidal implementation of the lucas kanade feature tracker description of the algorithm,” in *USENIX Technical Conference*, 1999.

[23] Y.-H. Seo, S.-H. Kim, K.-S. Doo, and J.-S. Choi, “Optimal keyframe selection algorithm for three-dimensional reconstruction in uncalibrated multiple images,” *Optical Engineering*, vol. 5, no. 47, pp. 53–65, 2008.

[24] M. T. Ahmed, M. N. Dailey, J. L. Landabaso, and N. Herrero, “Robust key frame extraction for 3d reconstruction from video streams,” in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, Angers, France, 2010, pp. 231–236.

[25] Y. Furukawa and J. Ponce, “Accurate camera calibration from multi-view stereo and bundle adjustment,” in *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, 2008, pp. 1–8.

[26] M. Kazhdan, M. Bolitho, and H. Hoppe, “Poisson surface reconstruction,” in *Symposium on Geometry Processing (SGP)*, Aire-la-Ville, Switzerland, Switzerland, 2006, pp. 61–70.

[27] V. Lempitsky and Y. Boykov, “Global optimization for shape fitting,” in *Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, USA, 2007, pp. 1–8.

[28] X. Zhao, Z. Zhou, Y. Duan, and W. Wu, “Detail-feature-preserving surface reconstruction,” *Computer Animation Virtual Worlds*, vol. 23, pp. 407–416, 2012.

[29] J.-Y. Bouguet, “Camera calibration toolbox for matlab,” 2008, http://www.vision.caltech.edu/bouguetj/calib_doc/.

[30] X. Zhao, Z. Zhou, and W. Wu, “Radiance-based color calibration for image-based modeling with multiple cameras,” *Science China Ser F-Inf Sci*, vol. 55, no. 7, pp. 1509–1519, 2012.

[31] 2d3 Inc., “Boujou software,” 2011, http://www.2d3sensing.com/.