On the study of registration methods for multiple depth cameras in free-viewpoint videos

Francois de Sorbier & Hideo Saito
Graduate School of Science and Technology
Keio University
Yokohama, Japan

Abstract—In this paper, we investigate the best approach for estimating the pose between two or more depth cameras. After an overview and a comparison of the existing methods, we first experiment on the accuracy of the calibration when using features from corresponding colour/depth images. Next, we discuss those approaches by considering the interference produced by the presence of several depth cameras, and computational time issues. Finally, we propose our own implementation resulting from those studies. The main aim of this paper is then to evaluate the colour image based feature detectors for helping in the choice of a calibration method in future developments.

Keywords—depth camera; point cloud; state-of-the-art; alignment; registration; free viewpoint videos;

I. INTRODUCTION

Over the last years, depth cameras, such as Kinect\(^1\), have become quickly popular. The reason of this success is mainly due to the capability of such low cost device to capture in real time a colour image with its corresponding depth map. These advantages have logically led up to several new applications or the improvement of existing ones like free-viewpoint videos’ generation. Although, the captured depth map is often corrupted by missing data due to the spatial disparity between the sensors and emitters of the camera and the properties of the captured surfaces like the shininess. Furthermore, the depth information is estimated based on a structured light pattern [1], which limits the range of the sensor and introduces a structural noise, especially around the edges [2].

Several methods [1-3] have been focusing on the improvement of the depth map or a reconstruction, but lack in some aspects like real-time processing. In the context of free-viewpoint videos, the depth map captured by a single depth camera has a too small resolution for assuming a good quality visualisation. For those reasons, increasing the number of depth cameras in the capture system can help to generate better quality live free-viewpoint videos. This kind of setup also offers the capability to add extra High Definition colour cameras that can compensate the poor quality of the colour sensor embedded in Kinect. This point will not be considered in this paper, and readers should refer to previous works such as [4] for more information.

The use of multiple Kinect-like depth cameras has two constraints [5,6]. The first one is related to the pose estimation of the devices. The second concerns the interference between the infrared light emitters that leads to very noisy depth maps. These constraints are closely related since the noise will consequently interfere on the accuracy of the registration of the 3-D data.

The main goal of this paper is then to investigate the proper method from state-of-the-art researches to register several point clouds generated simultaneously from separate depth cameras. Despite the fact that most of the calibration methods are using fiducial markers [6] or calibration mires [7], we are focusing on feature based approaches that are less intrusive and can be applied to real-time registration.

The study focuses on the comparison among colour images based approaches like SIFT, SURF, ORB, while 3-D based approaches like FPFH, NARF are avoided because of the noise. After extracting characteristic 3D points using the depth or the colour images, we estimate the related rigid transformation between both cameras. We suggest the best method by comparing the results with the ground truth data obtained thanks to a pattern. Besides this, we also evaluate the influence of the noise by performing an offline calibration where the depth cameras capture separately or at the same time the depth of the scene. We finally compare the computational time of these methods to suggest the best approach for an online calibration of the cameras.

Based on this, we propose our implementation for an online calibration of multiple depth cameras. We detect features in the colour images captured by the depth cameras and estimate the correspondences between them. We compute the 3D points associated to those features and check the spatial consistency of the correspondences to exclude the outliers. This implementation has been publicly demonstrated during several events such as CEATEC JAPAN 2012 or the NICT Open House.

II. RELATED WORKS

Existing works can be divided into two parts. The first one describes several common registrations methods between two point clouds, while the second aims at enumerating the methods for reducing the noise caused by the dot pattern interference.

A. Registration of two point clouds

The purpose of the registration is to merge two point clouds into a common coordinate system by finding the rigid transformation that maps one set of 3-D points onto the other. If

\[^1\] www.kinectforwindows.org
correspondences between the point clouds are known, then one of the algorithms compared by Eggert et al. [9] can be applied to compute the 3-D rigid transformation. The main problem is then to determine those correspondences.

A popular approach is the Iterative Closest Points [10] (ICP) algorithm, which, under its initial form, iteratively pairwise each 3-D point from the first point cloud with the closest points from the second point cloud until the distance error is minimized. The iteration process is however well known to be slow and can lead to a local minimum solution, especially if the initial spatial difference is large. In their work on real time reconstruction of arbitrary surface with Kinect, Newcombe et al. [2] proposed a hardware version of the ICP algorithm that works efficiently under the assumption of a small angle between two consecutive captures.

Rusu et al. [11] proposed a point feature descriptor based on the surface’s normal and curvature estimated from the point’s neighbourhood. Even if the feature matching presents a good accuracy, the result can be easily affected by the noise. Others feature descriptors, such as CSHOT [12] and NARF [13], were proposed, but suffer from the same limitation.

A depth map can be captured with its corresponding colour image. The mapping function is estimated thanks to a calibration phase [1, 8, 14] or obtained from parameters hard coded in the device, like with Kinect. Such correlation can bring 2-D features detection as an initial stage for estimating the 3-D correspondences, since a depth value is defined for each colour pixel. This approach has been often used in the context of Simultaneous Localisation And Mapping applications with consecutively captured images using a single sensor like in [3, 15] and [16].

B. Reducing interference

Since interference between the infrared dot patterns is the main origin of the noise with Kinect based device, several solutions have been proposed to reduce it. The first category proposes to place a shuttering system [1] in front of the infrared emitter of the device. When a pattern is projected by one camera, the other ones are occluded. This kind of approach has still many constraints related to the synchronization and the capturing frame-rate, and the complexity of the setup increases with the number of devices.

A motion based approach was also proposed to reduce the noise [17, 18]. A small vibration motor attached onto the Kinect induces a small shaking movement of the sensor. When the dot pattern projected by this moving sensor is observed from another device, it appears blurry reducing consequently the interference. However, the colour image may also become blurry as mentioned by [17], which can decrease the overall quality during the keypoint detection for instance. Kainz et al. [19] proposed to add an integration of the captured data into a voxel grid in order to improve the quality of their scene reconstruction with multiple Kinetc.

The noise can also be reduced by accumulating the depth values over the time as described in [2]. However, this approach is not really efficient with dynamic scenes and only reduces a small quantity of noise as experimented by Maimone and Fuchs [20].

Note that this interference problem can be overcome with “Time Of Flight” technology based cameras by selecting a different frequency of the emitted light for each device.

![Figure 1: The two pictures at the top present the colour images. The two corresponding depths map located at the middle row are generated under interference, while the two others at the bottom are without.](image)

III. COMPARISON OF REGISTRATION METHODS

Our goal is to compare several methods for the point cloud registration using the colour information with the depth data together. For this investigation, we design a setup made of two Kinects sharing a part of a same scene, but located at very different places as noticeable in Fig. 1. We use OpenNI to process the streams and to map the depth image onto the colour image viewpoint. Note that more accurate calibration methods such as [14] could have been applied.

The relative pose of both devices, defined as the ground truth, is estimated thanks to a fixed pattern. For dealing with the interference, we differ the capture of the pattern for each camera: when a device is generating its depth map, the other ones are not active. At each detected corner of the planar pattern in the first view exists a 3-D value for which we can easily deduce a correspondence in the second view based on the known pattern. The transformation is computed by using the least squares fitting described by Arun [21] minimised with RANSAC.

Prior to the calibration stage, we apply a bilateral filter [22] on the captured raw depth map that will reduce the effect of the structural noise while edges will be preserved.

During our experiment, we compared six common feature detectors. We selected SURF [23], BRISK [24], FAST [25], MSER [26], ORB [27] and SIFT [28].
A. Experiment with separated captures

In this experiment, we compared the accuracy of the evaluation of the 3-D rigid transformation by capturing the images at two different times. When one sensor is capturing the scene, the other one is not active, which avoids the interference. However, the scene needs to be static for avoiding matching problems. We captured several samples of the scene, estimated the rigid transformation by capturing the images at two different times. When one sensor is capturing the scene, the other one is not active, which avoids the interference. How-
formation, but it still contains many incorrect matches that will strongly influence the accuracy of the result. For removing these outliers, we need to apply an iterative estimation of the rigid transformation based on a limited number of pairs with the well-known RANSAC approach. For each iteration, we check the number of inliers and keep the corresponding transformation as a result if this number is higher than the previous one. As presented in the previous section, the rigid transformation can be correctly estimated with a high number of iterations since we do not consider computational time issues in the case of static devices.

Eventually, the 3-D rigid transformation can be refined by applying the ICP algorithm as a final stage. We set this first transformation as an initial solution and iteratively process all the data of the point clouds until minimization.

B. Moving devices

In the case of moving devices, we need to consider about real-time calibration. In the previous section, we discussed about the computational time issues of each feature detector for calibrating multiple depth cameras. We deduced that the best detectors based on the accuracy with and without interference, the repeatability and the processing time. We finally proposed two implementations depending on if the calibration had to be performed in real-time or not.

V. CONCLUSION

In this paper, we presented a comparison of several key-point detectors for calibrating multiple depth cameras. We deduced the best detectors based on the accuracy with and without interference, the repeatability and the processing time. We finally proposed two implementations depending on if the calibration had to be performed in real-time or not.

ACKNOWLEDGMENT

This work is supported by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCES


