Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Efficient hundreds-baseline stereo by counting interest points for moving omni-directional multi-camera system

Tomokazu Sato *, Naokazu Yokoya

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan

ARTICLE INFO

Article history:
Received 31 July 2009
Accepted 1 February 2010
Available online 16 February 2010

Keywords:
Multi-baseline stereo
Depth estimation
Interest point
Omni-directional camera system
Multi-camera system
Efficient algorithm
Spatial consistency
TNIP-based stereo

ABSTRACT

In this article, we propose an efficient method for estimating a depth map from long-baseline image sequences captured by a calibrated moving multi-camera system. Our concept for estimating a depth map is very simple; we integrate the counting of the total number of interest points (TNIP) in images with the original framework of multiple baseline stereo. Even by using a simple algorithm, the depth can be determined without computing similarity measures such as SSD (sum of squared differences) and NCC (normalized cross correlation) that have been used for conventional stereo matching. The proposed stereo algorithm is computationally efficient and robust for distortions and occlusions and has high affinity with omni-directional and multi-camera imaging. Although expected trade-off between accuracy and efficiency is confirmed for a naive TNIP-based method, a hybrid approach that uses both TNIP and SSD improve this with realizing high accurate and efficient depth estimation. We have experimentally verified the validity and feasibility of the TNIP-based stereo algorithm for both synthetic and real outdoor scenes.

1. Introduction

Depth map estimation from images is a very important topic in the field of computer vision because depth information can be used in several different applications such as 3D modeling, object recognition, surveillance, and novel view synthesis. In the past decades, a lot of methods for stereo algorithm are developed by many researchers, and most of these works were designed for a pair of standard camera units [1]. On the other hand, like the Google Street View, we can now easily acquire an omni-directional image sequence for large outdoor environments by moving a vehicle where the camera is mounted. However, for such an omni-directional image stream, most of conventional works designed for two-frame images do not work well due to large image distortion and large baseline. In this paper, in order to realize efficient and accurate depth estimation for omni-directional image sequence, we extend the conventional multi-baseline stereo framework that is proposed by Okutomi and Kanade [2]. This method has good feature in that an arbitrary number of images can be simultaneously used for depth estimation. This increases the accuracy of depth estimation and decreases the ambiguity in stereo matching. Using recent developments in camera calibration techniques, the multiple-baseline stereo framework have been employed for a freely moving video camera [3–7]. A freely moving video camera is suitable for the 3D modeling of a large-scale environment because it easily makes a long-distance baseline between cameras. However, when we employ the long-baseline omni-direcitonal images, following weaknesses of original multi-baseline stereo becomes the critical problem in practice.

1.1. Image distortion

In omni-directional video sequence, image patterns around the physical 3D point is easily distorted and resolutions of these patterns are not uniform due to both the characteristic of omni-directional vision and the large motion of the camera system. Because depth information should be estimated for any directions, rectification techniques cannot resolve this problem.

1.2. Occlusion

In large-baseline stereo in outdoor environment, there are much occluders than standard short-baseline stereo. When a point on an object where the depth is to be estimated is occluded by other objects in a part of an input video, the occluder gives a negative score to the score function of the multiple baseline stereo: SSSD (sum of SSD). This negative score prevents the algorithm from obtaining a correct estimation of the depth map around occlusions.

1.3. Computational cost

Although the utilization of multiple input images increases the accuracy of depth estimation, it consumes a large amount of

* Corresponding author. Fax: +81 743 72 5299.
E-mail address: tomoka-s@is.naist.jp (T. Sato).

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.02.006

memory and computational resources. Some patches for the distortion and the occlusion problems need additional computational cost.

In order to solve these problems, we propose a novel approach that estimates depths by using interest points, as shown in Fig. 1, that are corners and cross points of edges in images. The framework of our depth estimation method is basically the same as the original multiple baseline stereo except for a newly employed objective function: TNIP (total number of interest points). The concept is based on the very simple assumption that the corners of objects and cross points of texture edges in the 3D space (3D interest points) will appear in video images as 2D interest points at the projected positions of the 3D interest points. By searching a depth that maximizes the total number of 2D interest points under epipolar constraint, the depth can be determined as the position of a 3D interest point. It should be noted that the proposed method assumes that camera parameters of the input videos are pre-calibrated and the camera calibration problem is beyond the scope of this research.

By using the objective function TNIP for depth estimation, the problems mentioned earlier can be solved: (1) Detected position of the 2D interest point is ideally not affect by image rotation and distortion. (2) The score function TNIP is not significantly affected by occluders and the position of corners indicates the unique position in 3D space. (3) The computational cost of depth estimation is very cheap because the depth can be determined by only counting the interest points. However, depths for non-interest points cannot be estimated by TNIP; this is not a critical problem for 3D modeling and some other applications because 3D interest points contain the corners of the 3D models. Further it should be noted that the TNIP-based method estimates the depth for the 3D corners rather than that for the target pixel. Thus, the accuracy of the depths obtained from the raw TNIP function is a little lower than that obtained by SSSD; however, TNIP drastically decreases the computational cost. In this research, in order to resolve the weaknesses of the raw TNIP function, we also suggest a hybrid approach in which both SSSD and TNIP are used. In the hybrid approach, first, TNIP is used to roughly and quickly determine the depth for each interest point. For a limited searching range by TNIP, the depth value is then re-searched by SSSD with very small window size.

The reminder of this paper is organized as follows. First, related stereo algorithms and 3D reconstruction methods are reviewed in Section 2, and the contribution of the proposed method is also explained. In Section 3, the original multi-baseline stereo method for a moving video camera is described. Then, the new score function TNIP for multiple baseline stereo is proposed in Section 4. Each process for estimating a dense depth map is detailed in Section 5. Experimental results with simulation and a real scene are used to demonstrate the validity and feasibility of the proposed method in Section 6. Finally, Section 7 presents the conclusion and outlines of future studies.

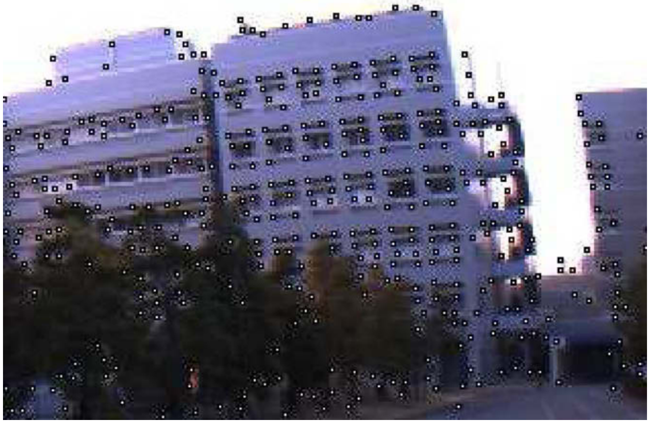

Fig. 1. Example of interest points.

2. Related works

2.1. Multi-view reconstruction

In the field of traditional stereo reconstruction, 3D information has been estimated as depth maps by assuming camera parameters are pre-calibrated. Although they have conventionally been designed for binocular and trinocular stereo imaging, recent works tend to use multi-view images [8–13]. In these multi-view approaches, in addition to the traditional depth map based 3D representation [2,12,13], voxel based [6,8,10] and polygon mesh based [11] 3D representation are employed for 3D reconstruction. However there are many combinations for 3D representation and modeling approaches in these conventional works, one of the common problems for the multi-view stereo reconstruction is how to corresponds pixels between multiple images. In order to correspond pixels between images, the photo-consistency measure have commonly been used. The photo-consistency measure is a similarity measure that correlates the pixels on multiple images based on variance of image pixels for projected position of the unique 3D position. The key difference between these multi-view works and the proposed method is that we do not need to utilize any intensity-based similarity measure for making correspondences between images. Instead of similarity measures, the spatial-consistency of the positions of feature points is employed, and the method efficiently tests the spatial consistency by simply counting the interest points along epipolar lines.

2.2. Feature-based 3D reconstruction

Another work that closely related to our approach is conventional feature-based stereo [14–16]. The feature-based stereo method uses feature points in images, such as intensity edges on epipolar lines, as positions of matching candidates in image pairs. In this approach, for these matching candidates, a similarity measure such as SSD or NCC is computed and the corresponding points between images are determined based on pattern similarity. In the same manner as conventional feature-based stereo algorithms, the proposed method also employs feature points to realize the efficient and robust determination of corresponding points. However, our algorithm basically determines corresponding points without using intensity-based similarity measures.

Similar to the case of feature-based stereo algorithms, EPI(epipolar plane image)-based 3D reconstruction [17] uses motion of the edge features along the epipolar line to recover the 3D information. In this method, video is first captured using a video camera that moves along a direction vertical to the optical axis at a constant speed. The line images of the corresponding epipolar plane are then collected and expanded vertically to generate the EPI. In the EPI, corresponding points can be determined easily because they lie along a line in the EPI. In conventional studies, these lines are detected by Hough transformation. Okutomi et al. [18] applied the EPI-based method for rotating camera motion. This method is used for an object on a turning table and it can detect the sin curve in the EPI image instead of the lines. The problem faced in the EPI-based methods [17,18] is that only a steady camera motion is allowed. Although some deviation from the steady camera motion can be compensated, it is difficult to process the images that include camera motion along the optical axis.