Computer Vision and Image Understanding 114 (2010) 652–666

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Pose estimation from multiple cameras based on Sylvester's equation

Chong Chen, Dan Schonfeld *

Multimedia Communications Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Chicago, 851 South Morgan Street, Room 1020 SEO (M/C 154), Chicago, IL 60607-7053, United States

ARTICLE INFO

Article history:
Received 14 January 2009
Accepted 5 January 2010
Available online 21 January 2010

Keywords:
Pose estimation
Distributed estimation
Sylvester's equation
Multiple cameras
Multiple views

ABSTRACT

In this paper, we introduce a method to estimate the object's pose from multiple cameras. We focus on direct estimation of the 3D object pose from 2D image sequences. Scale-Invariant Feature Transform (SIFT) is used to extract corresponding feature points from adjacent images in the video sequence. We first demonstrate that centralized pose estimation from the collection of corresponding feature points in the 2D images from all cameras can be obtained as a solution to a generalized Sylvester's equation. We subsequently derive a distributed solution to pose estimation from multiple cameras and show that it is equivalent to the solution of the centralized pose estimation based on Sylvester's equation. Specifically, we rely on collaboration among the multiple cameras to provide an iterative refinement of the independent solution to pose estimation obtained for each camera based on Sylvester's equation. The proposed approach to pose estimation from multiple cameras relies on all of the information available from all cameras to obtain an estimate at each camera even when the image features are not visible to some of the cameras. The resulting pose estimation technique is therefore robust to occlusion and sensor errors from specific camera views. Moreover, the proposed approach does not require matching feature points among images from different camera views nor does it demand reconstruction of 3D points. Furthermore, the computational complexity of the proposed solution grows linearly with the number of cameras. Finally, computer simulation experiments demonstrate the accuracy and speed of our approach to pose estimation from multiple cameras.

1. Introduction

Object pose estimation from monocular or multiple views has been one of the most active research topics over the past few decades. In many applications such as human–computer interaction, interactive-conferencing, and virtual reality, it is essential to monitor the rotation of the object out-of-image-plane (namely, pan and tilt).

The 3D rotation angles can be obtained by using various methods for pose estimation. A detailed discussion of many of the existing techniques for pose estimation from single and multiple cameras is provided in the following section. Despite the enormous advances in pose estimation, the problem of efficient and robust pose estimation from 2D image sequences from multiple cameras remains a difficult challenge. In this paper, we provide an extremely efficient and robust solution to pose estimation from 2D image sequences from multiple cameras based on Sylvester's equation.

We first present a solution to pose estimation from a single camera from 2D image sequences based on Sylvester's equation. In particular, we rely on the feature-based approach to directly estimate the 3D pose from 2D image sequences. Scale-Invariant Feature Transform (SIFT) [1] is used to extract corresponding feature points from adjacent images in the video sequence. We introduce a unified method for pose estimation based on Sylvester's equation [2]. This approach can be shown to be equivalent to the classical approach to pose estimation based on Singular Value Decomposition (SVD) for 3D models [3]. However, unlike pose estimation based on SVD methods, our approach is not limited to 3D models and provides an elegant solution to pose estimation directly from 2D image sequences that does not require any training.

We subsequently extend our approach to pose estimation from 2D image sequences from multiple cameras. We extend our approach to pose estimation by forming the problem of centralized pose estimation from multiple cameras as a solution to a generalized Sylvester's equation capturing all of the feature points from all cameras. We then derive a distributed solution to the generalized Sylvester's equation by collaboration among the cameras that relies on iterative refinement of the independent solution to pose estimation based on Sylvester's equation for each camera. We demonstrate that both the centralized and distributed pose estimation from multiple cameras generate superior estimates than the results of pose estimation from any specific camera view. The proposed approach to pose estimation can consequently be used to improve the estimate for any single camera or provide robust pose estimation from a virtual camera view.

* Corresponding author. Address: Department of Electrical and Computer Engineering, University of Illinois at Chicago, 851 South Morgan Street, Room 1020 SEO (M/C 154), Chicago, IL 60607-7053, United States. Fax: +1 312 996-6465.
E-mail address: dans@uic.edu (D. Schonfeld).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.01.002

The remainder of the paper is organized as follows: In Section 2, we provide a brief overview of various methods developed for pose estimation from single and multiple cameras. In Section 3, we introduce the method of pose estimation from 2D image sequences from a single camera based on Sylvester's equation. We extend this approach to pose estimation from 2D image sequences from multiple cameras in Section 4. We first present a centralized approach to pose estimation from 2D image sequences from multiple cameras as a solution to a generalized Sylvester's equation. We then provide a distributed solution to pose estimation from multiple cameras by iterative refinement of the independent solution to pose estimation from each camera. In Section 5, we conduct computer simulation experiments and demonstrate the robustness and efficiency of the proposed approach to pose estimation from multiple cameras. Finally, we present a brief summary and discussion of our results in Section 6.

2. Related work

As our pose estimation approach for multiple camera views is derived from the solution with only one camera, in the following, we will firstly introduce the method of pose estimation from the video sequences of one camera, and then extend to the multi-camera case.

2.1. Pose estimation from monocular camera

In the monocular view case, many approaches have been proposed during the past years, and most can be classified into two major categories: (i) feature-based methods and (ii) appearance-based methods. Appearance-based methods often rely on some training to obtain templates for object pose estimation. Feature-based methods rely on corresponding features from different images. Solutions with redundant data can be classified into two classes: (i) linear methods and (ii) nonlinear methods [4]. Nonlinear solutions are generally more robust to noise; however, they suffer from heavy computation, usually require a good initialization, and most cannot guarantee convergence. While linear methods require less computation, the results are often inferior due to lack of accuracy and corruption by noise.

Ji et al. [4] develop a linear least-squares framework for multiple geometric features including points, lines, and ellipse–circle pairs. Orthonormality constrains due to rotation are approximately imposed within the linear framework. In [5], the transform between feature matches is computed with a hierarchical RANSAC approach. The object pose estimation from corresponding points based on SVD techniques have been well established [3,6–8]. Horn's algorithm computes the eigensystem of a derived matrix and is similar to the SVD approach [9]. Derivation of a closed form solution to this problem can be simplified by using unit quaternions to represent rotation as shown in [10]. Olsson et al. [11] generalize the method in [9] by incorporating point, line and plane features in a common framework for finding the globally optimal solution to the problem of pose estimation, but it can only deal with a known object and also parameterize rotations with quaternions. Moreover, a reformulation of the same problem can be represented using the essential matrix [3,12].

The solutions discussed above can be shown to provide the optimal estimate for pose estimation from 3D points while satisfying the orthogonality of the rotation matrix, e.g. SVD-based pose estimation. Unfortunately, these methods cannot be used for solution from 2D projected feature points since in this case the orthogonality constraints are not satisfied. Instead, in the 2D case, one must incorporate pseudo-orthogonality constraints which capture the projection of the 3D rotation on the image plane. In the existing solutions to these problem, the pseudo-orthogonality constrains are weakly enforced and often ignored. Many methods rely on an unconstraint solution to the pose estimation problem and subsequently adjust the solution to satisfy the pseudo-orthogonality constraints for the 2D model.

We employ the feature-base approach and rely on corresponding 2D points from images to directly estimate the 3D pose while incorporating the pseudo-orthogonality constraints. We demonstrate that pose estimation can be obtained as a solution of Sylvester's equation, which can be solved with many methods such as Kronecker Product [13] and Bartels–Stewart approach [14]. This solution is proved to be equivalent to the SVD-based methods for 3D–3D pose estimation, yet it can also be used for the 2D cases.

2.2. Pose estimation from multiple cameras

The reconstruction from multi-view stereo has received a large amount of attention over the past few decades. In [15], the authors provide a quantitative comparison of several multi-view stereo reconstruction algorithms on their datasets with ground truth. Further research that could be used to recover the 3D structure, including 3D shape and 3D motion, from 2D motion in the image plane is generally referred to as structure-from-motion (SFM) [16]. A framework for various camera models is introduced in [17], which provides nonlinear methods based on point correspondence across views. SFM methods firstly determine the 2D motion in the image plane and then estimate the 3D shape and 3D motion. 3D reconstruction and shape analysis are beyond the scope of this paper.

In [18], a method is proposed for arbitrary view synthesis from an uncalibrated multiple camera system. In [19], the multiple view matching is achieved by a combination of image invariants, covariants, and multiple view relations. Sturm [20] models cameras as possibly unconstrained sets of projection rays and introduces a hierarchy of general camera models. He also establishes the foundations for a multi-view geometry of general (noncentral) cameras, analogously to the perspective case. Yu and McMillan [21] present a General Linear Camera model to describe many camera models. In [22], structured light patterns are used to extract the raxel parameters of an imaging system, to present a general imaging model. In this framework, Press [23] derives the discrete SFM equations for generalized cameras, and illustrates the connections to the epipolar geometry.

Chang and Chen [24] conclude that pose estimation for a multiple camera system is usually solved by perspective-n-point (PnP) methods or Least Square approaches. Kahl [25] presents a framework to solve geometric structure and motion parameters based on L_∞-norm, which can be applied for efficient computation of global estimates. In [26], the fractional programming and the theory of convex underestimators are relied to unify the framework for minimizing the standard L_2-norm of reprojection errors. In this paper, we develop two numerical algorithms with the observations from all cameras, and also rely on the information from other cameras to update the first camera, to make its pose estimate more robust.

Rother and Carlsson [27], based on a reference plane, develop a linear algorithm for computation of 3D points and camera positions from multiple perspective views by finding the null-space of a matrix built from image data using SVD. In [28], a system, consisting of six cameras, is built to remove the inherent ambiguities of confusion between translation and rotation. However, this system does not use one pose estimation for all information from all cameras simultaneously. Frahm et al. [29] combine all information of all cameras to estimate the pose of a multi-camera system. In contrast, we are estimating the object's pose, and by using all information from all cameras to give one pose estimate, we also