Robotics and Autonomous Systems 62 (2014) 1717–1730

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Evaluation of non-geometric methods for visual odometry

Thomas A. Ciarfuglia [a,*], Gabriele Costante [a], Paolo Valigi [a], Elisa Ricci [a,b]

[a] Department of Engineering, University of Perugia, Perugia, Italy

[b] Fondazione Bruno Kessler, Povo, Trento, Italy

HIGHLIGHTS

- We stress the advantages of non-geometric (learned) VO as an alternative or an addition to standard geometric methods.
- Ego-motion is computed with state-of-the art regression techniques, namely Support Vector Machines (SVM) and Gaussian Processes (GP).
- To our knowledge this is the first time SVM have been applied to VO problem.
- We conduct extensive evaluation on three publicly available datasets, spanning both indoor and outdoor environments.
- The experiments show that non-geometric VO is a good alternative, or addition, to standard VO systems.

ARTICLE INFO

Article history:
Received 12 December 2013
Received in revised form
31 July 2014
Accepted 5 August 2014
Available online 13 August 2014

Keywords:
Autonomous robots
Visual odometry
Ego-motion estimation
SLAM
SVMs
Gaussian processes

ABSTRACT

Visual Odometry (VO) is one of the fundamental building blocks of modern autonomous robot navigation and mapping. While most state-of-the-art techniques use geometrical methods for camera ego-motion estimation from optical flow vectors, in the last few years learning approaches have been proposed to solve this problem. These approaches are emerging and there is still much to explore. This work follows this track applying Kernel Machines to monocular visual ego-motion estimation. Unlike geometrical methods, learning-based approaches to monocular visual odometry allow issues like scale estimation and camera calibration to be overcome, assuming the availability of training data. While some previous works have proposed learning paradigms to VO, to our knowledge no extensive evaluation of applying kernel-based methods to Visual Odometry has been conducted. To fill this gap, in this work we consider publicly available datasets and perform several experiments in order to set a comparison baseline with traditional techniques. Experimental results show good performances of learning algorithms and set them as a solid alternative to the computationally intensive and complex to implement geometrical techniques.

1. Introduction

In the last decade Visual Odometry (VO) has become one of the main building blocks of autonomous robot and vehicle navigation and mapping systems, integrating and often superseding classical odometry measurements and other means of computing robot trajectory online, such as scan matching with laser scanners, GPS positioning and IMU measurements integration. The advantages of VO are closely related to the low cost of camera devices and to their highly informative data streams that allow for high precision estimates. The high precision obtained by VO increases the performances of full Simultaneous Localization and Mapping (SLAM) systems where VO output can be checked for consistency with loop closing algorithms [1,2].

In the last few years a widely accepted framework for computing motion from a sequence of images has been defined [3]. At its core is the computation of optical flow between subsequent frames from which, when the camera parameters are known, the geometric computation of the rotation and translation parameters of a frame with respect to the previous is possible. Recently a VO based on Machine Learning (ML) techniques is emerging, where camera calibration is not necessary and the motion model is learned from many labelled sample pairs of optical flow-ground truth movement.

Both approaches have their advantages and disadvantages. Geometric VO is well established and several algorithms exist [4,5]. Its precision depends on the correctness of feature matching between consecutive frames which in turn is influenced by robot speed, feature tracking, motion blur, visual similarity or degenerate configurations [6,7,3]. Moreover all these algorithms are very brittle to the

* Correspondence to: DI, via Duranti 93, 06125 Perugia, Italy. Tel.: +39 0755853677.

E-mail addresses: thomas.ciarfuglia@gmail.com, thomas.ciarfuglia@unipg.it (T.A. Ciarfuglia).

http://dx.doi.org/10.1016/j.robot.2014.08.001
0921-8890/

presence of outlier matches in the optical flow computation. For this reason almost always a more robust approach is adopted, using a RANSAC incarnation [8] or filtering techniques [9,10]. These correction steps make the geometric VO algorithm not only very successful, but also a very computationally intensive, process.

Learning-based VO on the other hand does not need to have a model of the camera image acquisition, but relies on having enough labelled data to train a regression model capable of predicting movement from some feature extracted from video sequences. Until now the features used have all been quantizations of the optical flow input, but we hypothesize that other kinds of features, for example extracted from an image segmentation process [11], could be used as well. The process of learning a VO model is a multi-output problem and each estimated variable involved (for example linear speed and angular speed) is usually learned independently from the others. In any case learned VO shares the advantages of any learned model of being robust to noise and outliers when data used for training is subjected to the same kind of noise of test data, as it usually is. Moreover to our knowledge, given enough data, training a learned VO is a much easier endeavour than coding a geometric one. Despite this fact only a few examples of this second line of approach exist in the literature [12,13].

The purpose of this work is twofold: first we want to stress the advantages of learning-based VO to uphold it as a viable alternative, or a useful addition, to the consolidated approaches to visual ego-motion estimation. Past proposed ML approaches to VO have shown the feasibility of these techniques, but were tested only on self produced datasets, not on a comparable basis using publicly available datasets. This also limits the possibility to compare different ML solutions to the same problem. To overcome this limit we conduct extensive evaluation on three publicly available datasets spanning both indoor and outdoor environments.

The second contribution is to introduce a comparison between two of the most successful regression algorithms to the problem of ego-motion estimation. These two state-of-the-art techniques are *Support Vector Machines* (SVM) and *Gaussian Processes* (GP). Both these algorithms share the kernel mapping technique to handle non-linear relations between features and for this reason they are known as *Kernel Machines*. To our knowledge this is the first time SVM have been applied to the VO problem, while GP were proposed in [14].

This paper is organized as follows: Section 2 presents related works on both geometric-based and learning-based VO. Section 3.1 describes optical flow processing and feature vector construction, while Section 3.2 introduces the two adopted regression techniques, namely SVMs and GP. Section 4 describes our experimental set-up and shows the results achieved on the three datasets considered. In Section 5 we provide conclusions and discuss future developments.

2. Related works

One of the first real time VO approaches is presented in [4]. In this work a stereo system managed to navigate in an outdoor environment outperforming classical wheeled odometry systems and setting the trend for future research. Most geometric VO approaches can be divided into *stereo* VO, using two cameras pointing in the same direction and with synchronized acquisition, and *monocular* VO, using only one camera to compute motion.

These two frameworks to VO use similar techniques but have different problems to tackle. Stereo systems compute feature matchings in space, *i.e.* between two images taken from two different points in space looking at the same scene. From these matchings and knowing the distance between the camera centres of the two shots, it is possible to triangulate the 3D positions of all the matched points [8]. When the stereo rig moves and another pair of images is taken, the newly computed 3D features are matched with the previous ones, allowing motion computation. Examples of this approach are described in [15–17].

Monocular systems instead use two different images of the same scene, but taken at different times during camera motion. Some of the most successful early systems were based on filters like EKF and Particle filters as in [18,19,10] but were only suitable for small indoor environments, like office areas. Some critical issues for monocular VO are the parallax problem, *i.e.* the difficulty of tracking distant features since their relative motion is confused with pixel noise, and the problem of scale recovery. One of the works that addressed the parallax problem is [9] where an inverse-depth parametrization allows an uncertainty for each feature spotted to be computed. The problem of scale recovery depends on the lack of any metric information in the motion reconstruction process, so that each reconstruction step is computed in its own scale, with no explicit relation to other steps. To recover scale many approaches have been suggested. In [20] the scale is recovered using the extra information from an IMU, while in [21] an optimization approach using loop closing information is able to recover scale errors. In [22] the height from the plane of motion is used as a metric information to recover scale and in [23] the dimension of known features, like walls, is used to recover absolute scale factor from an EKF failure to the next. In all cases, to address the problem of scale uncertainty, an extra information is needed in the form of an extra sensor, loop closing data or metric set up information.

In general, VO precision depends mostly on scene illumination and image texture. When the images are dark and without enough visual elements to track, VO performances drop drastically. Even when illumination and texture are good the correctness of feature matching between consecutive frames is influenced by robot speed, motion blur, visual similarity or degenerate configurations [7,3,6]. To make the algorithm robust to the presence of outlier matches in the optical flow computation often an outlier rejection scheme (*e.g.* RANSAC) is used instead to remove false matches [9,10]. All these layers of computation make the geometric VO algorithm very successful, at the expense of a very computationally intensive process.

On the other hand machine learning approaches to VO are not yet common, although they do exist and have proven to be effective. Learning-based approaches to VO use the same input data as the geometric VO but learn the relation automatically from input data to ego motion without any explicit geometric computation. To do so they require some consistent data to learn the motion model, but once this is learned it can be used to predict the motion in any case where the input is similar to the training case. This approach has several underlying advantages. While geometrical VO requires that a full and precise calibration of camera parameters has been made before operating, learning-based VO on the contrary does not need camera calibration parameters and is able to learn them, as shown in [14]. Since it is trained on metric ground truth, learned VO in monocular settings reconstructs trajectories with correct scale. Finally, learned approaches are robust to noise and outliers, if enough training data are provided.

The first example of learning-based VO is the one of Roberts et al. [24] where they divide each frame into cells and compute an average optical flow for each block, then train a knn regressor for each of them. The motion prediction is obtained through a voting scheme between different blocks. Again Roberts proposes a learning approach in [12]. This time it is shown how the optical flow field can be approximated with a linear sub-space when the environment where the robot moves has some scene depth regularity [25]. The work leverages this property to learn the sub-space through an EM algorithm. Guizilini and Ramos [14,13] use a similar feature parametrization of optical flow and propose Coupled Gaussian Processes (CGP) as a regression algorithm. These works were