Expert Systems With Applications 60 (2016) 258–268

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Online human assisted and cooperative pose estimation of 2D cameras☆

Gaetano Manzo[a], Francesc Serratosa[b,*], Mario Vento[c]

[a] University of Bern, Bern, Switzerland
[b] Universitat Rovira i Virgili, Tarragona, Catalonia, Spain
[c] University of Salerno, Salerno, Italy

ARTICLE INFO

Article history:
Received 27 December 2015
Revised 30 April 2016
Accepted 7 May 2016
Available online 10 May 2016

Keywords:
2D pose estimation
Robot cooperation
Human-robot interaction
Social robots
Structure from motion

ABSTRACT

Autonomous robots performing cooperative tasks need to know the relative pose of the other robots in the fleet. Deducing these poses might be performed through structure from motion methods in the applications where there are no landmarks or GPS, for instance, in non-explored indoor environments. Structure from motion is a technique that deduces the pose of cameras only given only the 2D images. This technique relies on a first step that obtains a correspondence between salient points of images. For this reason, the weakness of this method is that poses cannot be estimated if a proper correspondence is not obtained due to low quality of the images or images that do not share enough salient points. We propose, for the first time, an interactive structure-from-motion method to deduce the pose of 2D cameras. Autonomous robots with embedded cameras have to stop when they cannot deduce their position because the structure-from-motion method fails. In these cases, a human interacts by simply mapping a pair of points in the robots' images. Performing this action the human imposes the correct correspondence between them. Then, the interactive structure from motion is capable of deducing the robots' lost positions and the fleet of robots can continue their high level task. From the practical point of view, the interactive method allows the whole system to achieve more complex tasks in more complex environments since the human interaction can be seen as a recovering or a reset process.

1. Introduction

We present a method to obtain the pose of several 2D cameras that has two novel features. The first is that the human can assist in mapping salient points between pairs of images when one estimates the error is too high. The second is the ability of the method to deduce the poses in a cooperative manner. The system always tries to automatically deduce the poses of the whole cameras meanwhile the human visualises the deduced poses together with the estimated errors in a human-machine interface but, when the human considers appropriate, one can asynchronously interact on the system. Thus, the supervisor can partially modify the point-to-point mapping between two images, which increases the quality of the deduced homography between these images. This interaction decreases not only the poses error of the involved two cameras but also the pose errors of the whole set of cameras.

Fig. 1 shows a schematic view of our method based on a module, which we have called interactive pose estimator, and a human-machine interface. In this example, cameras are embedded on the robots. The input of the general system is a set of 2D images and the output is their relative poses and the estimated errors (as the GPS does it). The human-machine interface receives the 2D-images from the cameras, their current relative poses and the mapped points per pair of images from the interactive pose estimation module. The human-machine interface only outputs the user operation, in other words, the point-to-point mapping impositions to the interactive pose estimation module. Besides, the interactive pose estimation also receives the 2D images and then deduces and sends the relative poses estimation and the regression errors to the system that controls the robots.

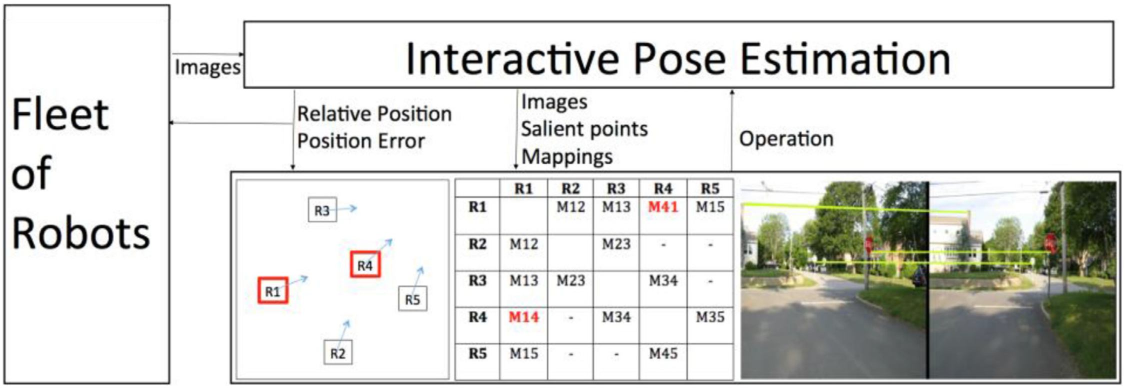
The human-machine interface is as follows. On the left side of it, the user visualises the deduced current poses of the cameras (2D position on the land and robot orientation). In the middle of the interface, the number of mapped points between any pair of cameras is shown. The lowest values of the number of mappings are highlighted in bold to attract the attention of the user. On the

☆ This research is supported by Spanish projects DPI2013-42458-P & TIN2013-47245-C2-2-R.

* Corresponding author at: Departament d'Enginyeria Informàtica i Matemàtiques, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain. Fax: +34977559710.

E-mail addresses: gaetano.manzo@students.unibe.ch (G. Manzo), francesc.serratosa@urv.cat (F. Serratosa), mvento@unisa.it (M. Vento).

http://dx.doi.org/10.1016/j.eswa.2016.05.012
0957-4174/© 2016 Elsevier Ltd. All rights reserved.



	R1	R2	R3	R4	R5
R1		M12	M13	M41	M15
R2	M12		M23	-	-
R3	M13	M23		M34	-
R4	M14	-	M34		M35
R5	M15	-	-	M45	

Fig. 1. Basic scheme of our method composed of our interactive pose estimation module and the human-machine interface.

right side of it, the user visualises the 2D images of two manually selected cameras together with the imposed point-to-point correspondences. The user can visualise any pair of 2D images by selecting one of the cells in the middle of the human-machine interface. Thus one can update the imposed correspondence by erasing or creating mappings between points. The two robots (or cameras) in the left panel and the pose errors in the central panel that correspond to the current images in the right panel are highlighted in red.

In robotics or camera surveillance, properly estimating the pose of cameras is considered to be a crucial low-level task. In the first case, cameras are embedded on the robots and therefore have mobile positions. By contrast, in the second case, we have static cameras. Nevertheless, in both cases, without properly setting the camera poses, the method is not able to deduce the position, direction, speed or acceleration of the objects or humans that are in the surroundings.

The method we present is part of a larger project in which social robots guide people through urban areas (http://www.iri.upc.edu/project/show/144). We have analysed the relation between humans and robots and also the behaviour of humans when social robots move closer to them (Garrell & Sanfeliu, 2012). We have also presented a tracking method that follows people, which allows occlusions and mobile cameras (Serratosa, Alquézar, & Amézquita, 2012) and a robot navigation system (Ferrer & Sanfeliu, 2014). Moreover, we have presented some results on structure from motion. In other words, given several 2D cameras, the method reconstructs the 3D position of the cameras (Rubio, 2015).

One important aspect of this project is the human-machine interaction. Human interaction has been applied to classify objects (Ferrer, Garrell, Villamizar, Huerta, & Sanfeliu, 2013) and also to deduce the pose of the robots.

Several levels of interaction could be considered to deduce the pose of the robots. The highest level could be to impose the position of a camera resulting from this knowledge has been acquired through another method. Our proposal is related to the lowest level. A human is very good and fast at mapping points on two different scenes, independently of the intrinsic or extrinsic characteristics of the images. Thus, what the user is asked to do is simply to select a salient point on one of the images and map this point on another image. Note this action is performed asynchronously to the process of deducing the pose and the supervisor tends to perform it when robots have to stop because the system is unable to deduce the pose in a completely automatic way. Therefore, the human interaction is a mechanism which allows the robot to continue in extreme situations.

Two papers have been presented that shoud be seen as necessary previous research in order to achieve the current paper. In the first, the homography between 2D images is computed (Cortés & Serratosa, 2015) and in the second, the 3D positions of the robots are deduced given 3D cameras (Cortés & Serratosa, 2016). In (Cortés & Serratosa, 2015), the influence of some human point-to-point mapping impositions while deducing the homographies that convert one image into the other was analysed for the first time. That paper concluded that with very few impositions, the end-point error generated by the semi-automatically deduced homographies is much lower than the error committed by the automatically deduced homographies. With this knowledge, in (Cortés & Serratosa, 2016), the poses of some robots were deduced in a semi-automatic way. Nevertheless, to avoid the process of estimating 3D poses from 2D images, known as structure-from-motion (Xu, Tao, & Xu, 2015), the paper assumes robots have 3D and 2D cameras. The 3D images are used to compute the 3D point-to-3D point correspondences and the 2D images are shown to the human. Since both cameras have been calibrated, when the human imposes 2D point-to-2D point correspondences, these are easily converted to 3D point-to-3D point correspondences and used to deduce the final poses in a semi-automatic way.

In this paper, we move one step further since the robots only have 2D cameras and the method obtains the 3D pose of the cameras performing structure from motion with human interaction. Note that this is the first time that an interactive structure from motion method has defined.

Fig. 2 (left) represents three robots performing guiding tasks. Robots fence the visitor group to force them to follow a specific tour. Robots need to work in a cooperative manner to keep a triangular shape in which people have to be inside. In these cooperative tasks, it is crucial to have a low-level computer vision task so that images extracted from the three robots and some static cameras in the surrounding are properly aligned to correctly deduce their relative poses. In this environment, there is a human that, through our human-machine interface, gives orders to the robots and controls their tasks. What we propose in this paper is that the human can also visualise the images of the cameras and interact in a low level task through asynchronously imposing point-to-point mappings. Fig. 2 (right) shows the images taken from the first two robots and three mappings that the human has imposed. Note that the system deduces the poses automatically but, when the human asynchronously imposes a mapping, then it takes into consideration this mapping and it continues its process automatically.

Besides, we say it is a cooperative model since the relative pose of two robots is deduced through the other robots when these robots do not share any part of the scene they visualise. Fig. 3 shows the pose of three robots. Robot 1 and Robot 2 can deduce their relative pose but this is not possible between Robot 1 and Robot 3 since they do not share any part of their images. This problem is solved through the cooperation of the robots. Robot 2