Semi-automatic pose estimation of a fleet of robots with embedded stereoscopic cameras.

CORTÉS, X., SERRATOSA, F. and MORENO-GARCIA, C.-F.

2016

Semi-automatic pose estimation of a fleet of robots with embedded stereoscopic cameras

Xavier Cortés
Universitat Rovira i Virgili
Tarragona, Catalonia, Spain
xavier.cortes@urv.cat

Francesc Serratosa
Universitat Rovira i Virgili
Tarragona, Catalonia, Spain
francesc.serratosa@urv.cat

Carlos-Francisco Moreno-Garcia
Universitat Rovira i Virgili
Tarragona, Catalonia, Spain
carlosfrancisco.moreno@estudiants.urv.cat

***Abstract*—Given a fleet of robots, automatic estimation of the relative poses between them could be inaccurate in specific environments. We propose a framework composed by the fleet of robots with embedded stereoscopic cameras providing 2D and 3D images of the scene, a human coordinator and a Human-Machine Interface. We suppose auto localising each robot through GPS or landmarks is not possible. 3D-images are used to automatically align them and deduce the relative position between robots. 2D-images are used to reduce the alignment error in an interactive manner. A human visualises both 2D-images and the current automatic alignment, and imposes a new alignment through the Human-Machine Interface. Since the information is shared through the whole fleet, robots can deduce the position of other ones that do not visualise the same scene. Practical evaluation shows that in situations where there is a large difference between images, the interactive processes are crucial to achieve an acceptable result.**

Keywords—Homography Estimation, Image Alignment, Interactive Method, Human-Robot Interaction, Cooperative Robotics.

I. INTRODUCTION

In recent years, interaction between robots and humans, and also cooperation between robots have increased rapidly. Applications of this field are very diverse, ranging from robot formations to transport and evacuate people in emergency situations [1] or simply vehicle positioning [2]. Within the area of social and cooperative robots [3], [4], interactions between a group of people and a set of accompanying robots have become a primary point of interest [5], [6].

One of the low level tasks that these systems have to face is automatic pose estimation. If the information of GPS is not available or its accuracy is not enough, one of the usual methods is to localise the robots through detecting landmarks or identifying scenes previously classified. The problem of comparing or aligning two images is usually called image registration in the computer vision research field. Image registration tries to determine which parts of one image correspond to which parts of another image. This problem often arises at the early stages of many computer vision applications, such as scene reconstruction, object recognition and tracking, pose recovery and image retrieval. Therefore, it has been of basic importance to develop effective methods that are both robust in the sense of being able to deal with noisy measurements and also to have a wide field of application. An example of this research field is [7].

We present an interactive and cooperative method to deduce the relative pose of each robot with respect to the rest of the fleet. The method we present is part of a larger project in which social robots guide people through urban areas [5] with tracking abilities [8], [9], [10], [11]. Figure 1 represents three robots performing guiding tasks in an indoor environment. Robots fence the visitor group to force them to follow a specific tour. Therefore, they 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 such that images extracted from the three robots are properly aligned to correctly deduce their relative pose. Moreover, there is a human who, through our Human-Machine Interface (HMI), gives orders to the robots and controls their tasks. Robots have embedded stereoscopic cameras providing 2D and 3D images of the scene (http://www.optitrack.com/) and the human can visualise the 2D images.

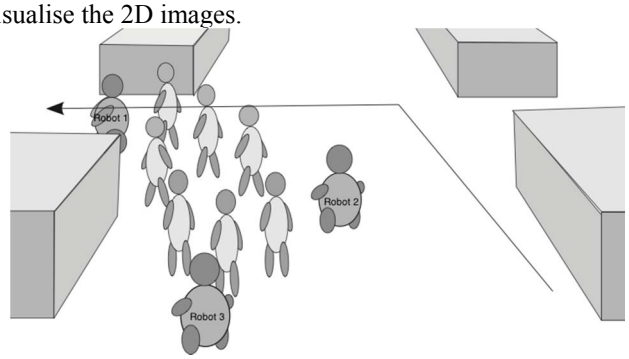

Fig. 1. Three robots performing guiding tasks. Robots are located to fence the group.

The rest of the paper is organized as follows. In section II, we present our model from a general point of view. In section III, we concretise on the interactive relative pose estimation. In section IV, we experimentally validate our model and show that with few human interactions, the accuracy of the

estimated relative pose drastically increases. We conclude the paper in section V.

II. The Proposed Model

Current automatic methods to extract parts of images and their correspondences in non-controlled environments are far away from having the performance of a human. Figure 2 shows two images extracted from the RESID database (http://www.featurespace.org). In each image 50 salient points have been extracted by method [12]. The outlier detector [13] has considered that 43 salient points were outliers and only 7 where inliers. The correspondence detector has missed 6 of the 7 points (red lines) and only one has been properly matched (green line). This is because of the large differences between both images and more precisely, due to the failure of the initial correspondence detector to find a good initial correspondence.

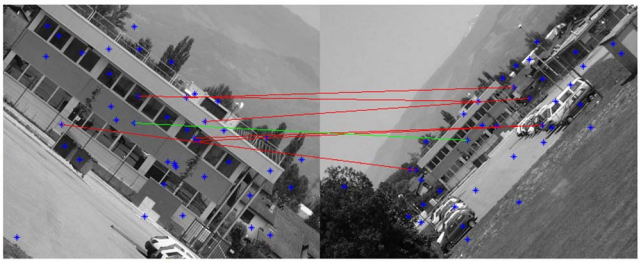

Fig. 2. An example of automatic registration image where only one point has been properly matched (green line) and 6 points have been improperly matched (red lines).

For this reason, in this paper, we propose a semi-automatic method in which a human can interact with the system when it is considered that the quality of the automatically found correspondences is not good enough and then they impose a partial and initial correspondence between some local parts of two scenes. We concretise on how to deduce the relative pose of our robots in an interactive and cooperative manner and the validation section only tests this aspect. The other technical and theoretical aspects of the whole project (http://www.iri.upc.edu/project/show/144) are not commented.

We call this an interactive method since a human aids the image processing modules incorporated on the robots to automatically solve the 3D registration problem when it is necessary [15]. Figure 3 shows part of our HMI. It is possible to visualize the 2D images obtained from robot 1 and 2, and the correspondences imposed by the user. Both robots go on the pavement with Robot 1 following Robot 2. Since the stereoscopic cameras in each robot are calibrated, the imposed correspondence in the 2D domain is translated to a correspondence in the 3D domain and thus, this interaction influences over the obtained 3D alignment, and the relative pose is recomputed.

Figure 4 shows the position of three robots. Robot 1 and Robot 2 can theoretically deduct their relative position through 3D image registration but this is not possible between Robot 1 and Robot 3 since they do not share any part of the 3D image. This problem is solved through the cooperation of the robots. Robot 2 deducts its relative position with respect to Robot 3 and shares this information with Robot 1.

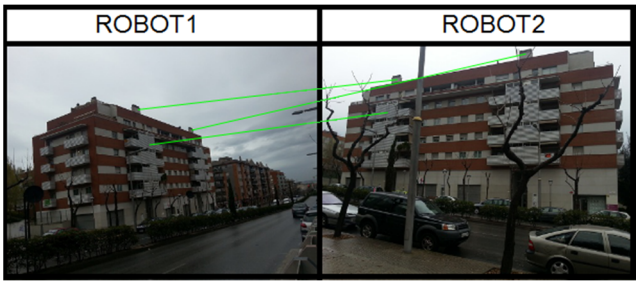

Fig. 3. Screen shot of our Human-Machine Interface with 2D images of robots 1 and 2

Thanks to this interaction, the accuracy of the relative pose estimation of the whole fleet of robots increases. This type of interaction is completely different from the ones presented in [1], [5] since in those cases, the human interaction is performed in a higher level. For instance, in [1], the interaction is based on imposing orders such as “move straight ahead” or “go upstairs”. In [5], the orders are “follow this person” or “bring me to the exit”. In [6], the interaction is used to learn the matching process. Nevertheless, our experience has shown us that in most of the cases, robots cannot perform these orders due to they cannot solve the low-level registration problem. Therefore, human interaction in this low-level task, which is easier for humans, frequently makes unnecessary the interaction on higher level tasks in which the interaction is more complicated, due to the need of having more knowledge of the current situation, such as position of other robots, automatically built map of the environment, or current position of other objects.

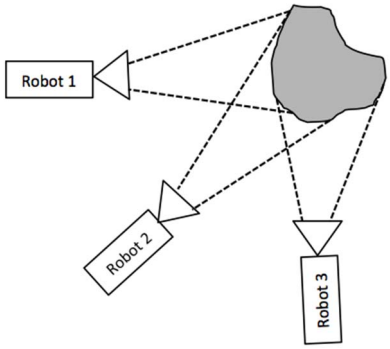

Fig. 4. Two robots visualising the same scene.

Figure 5 shows a schematic view of our method based on an Interactive Pose Estimation module and a Human-Machine Interface. The inputs of the general system are the 2D and 3D images of all robots and the output is their relative pose and the regression error (as done in the GPS). We are interested on minimising the pose error, but in a real application, we cannot deduce this error since the ground truth pose is not available. However we can obtain the regression error between aligned images knowing the lowest is the regression error and the best is the pose estimation. Then, we assume there is a direct dependency between the regression error and the pose error. Thus, the regression error is reduced, and the system also tends to reduce the pose error.

On the one hand, the HMI receives from the fleet of robots the 2D-images and the current relative poses of the robots in

the fleet and the deduced error of these relative poses from the Interactive Pose Estimation module. The Human-Interactive Interface outputs the user impositions to the Interactive Pose Estimation module and does not output any information to the fleet of robots. On the other hand, the Interactive Pose Estimation receives from the fleet of robots the 3D images and returns to it the relative poses estimation and the regression errors.

The HMI is composed as follows. On the left side, the user visualises the deduced current relative pose of the fleet (2D position on the land and robot orientation). The (0,0,0) position is assigned to the centre of the window. On the centre of the interface, the regression error between any combinations of robots' images is shown. This matrix is maid symmetric since we introduce in cell [i,j] and in [j,i] the same values, which are the last regression error while deducing the position of the i^{th} robot with respect the j^{th} robot or vice versa. The maximum regression errors are highlighted on bold to attract the attention to the user. On the right side, the user visualises the 2D images of the two manually selected robots together with the imposed correspondences. The user can visualise any of the combinations of 2D images by selecting one of the cells in the centre of the HMI. Then, the user can update the imposed correspondence by erasing or creating mappings between points. The two robots in the left panel and the regression errors in the central panel that correspond to the current images in the right panel are highlighted in red. If the regression error between two robots is higher than a threshold, then it is assumed that the corresponding images do not share any salient points and then, the regression error is automatically hidden (marked with a hyphen). Nevertheless, if the user considers that these robots really visualise the same part scene, one can select those robots. Note the 2D images are not an input of the interactive pose estimation module but they are used to be visualised by the user to impose the point's correspondences.

A preliminary version of this method was presented in [6], where, we presented a simple interactive method to estimate the homography between two 2D images. No 3D images were available and thus, the relative poses of the robots were not deduced. When this system was put into practice, we realised most of the cases where robots did not correctly react to the humans orders occurred due to they did not solve appropriately the low-level image registration problem. Therefore, we believe that by putting the human interaction into the image alignment problem, most of these non-correct robot reactions are solved. Nevertheless, technology tends to make systems run as much autonomously as possible. For this reason, the main weakness of our method is that robots are less autonomous. Nevertheless we believe better registration methods will surely to be discovered, and then, less need of human interaction is going to be needed. A similar interaction method was presented in [17]. In that case, there is only one robot and the human decides if a selected part of the image is a human's face and in the case that it is, imposes the name of the person.

III. Interactive Pose Estimation Module

Figure 6 shows our Interactive Pose Estimation general module bounded inside the dashed rectangle, which is the upper module in Figure 5. The aim of the scheduler is to keep updated the consistency of the relative position information between the robots. To achieve this, it is responsible for selecting the pairs of robots to deduce their relative pose using an implementation of the weighted round robin algorithm [18]. This selection depends on the time elapsed since the last relative pose update, the known current relative pose, the regression errors and also the correspondences imposed by the user. The scheduler also provides to the Robot Interactive Homography Estimation module the homography $H^{*}_{i,j}$, deduced from the user imposed correspondences. Then, the Robot Interactive Homography Estimation module deduces the relative homography $H_{i,j}$ of the ith robot with respect to the j^{th} one. It takes into consideration the imposed correspondences between these robots if there are any, and also the 3D-image alignment $H^{*}_{i,j}$. Moreover, the module returns a regression error of the obtained projection $Error_{i,j}$. As we have seen in Figure 5, this regression error is visualised at the HMI to help the user to decide which pairs of robots need some point mappings to be manually imposed.

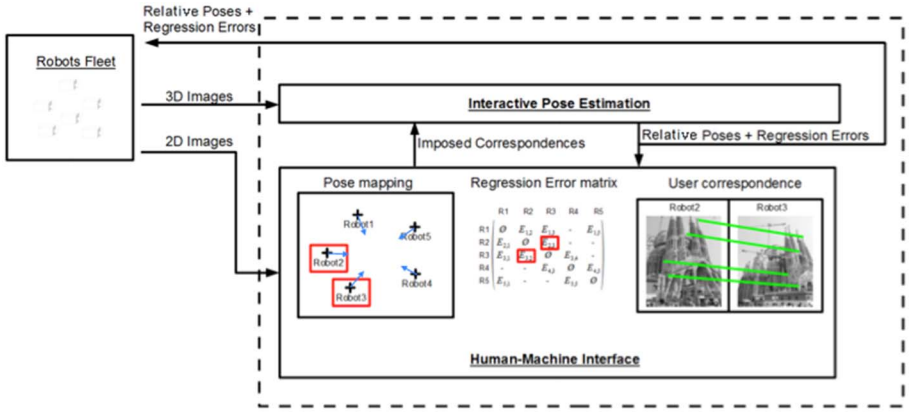

Fig.5. Basic scheme of our method composed of an Interactive Pose Estimation module and the Human-Machine Interface.

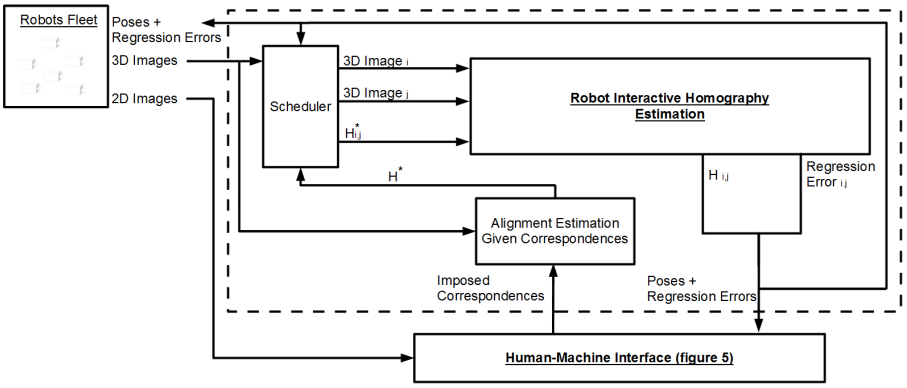

Fig. 6. Interactive Pose Estimation module bounded by the dashed rectangle.

The Alignment Estimation Given a Correspondence module deduces a homography $H^*_{i,j}$ between these images taking into consideration the human's correspondence proposal. It is based on the Direct Linear Transformation algorithm [19] that solves a set of variables from a set of similarity relations. It obtains a matrix homography (or linear transformation) $H^*_{i,j}$, which contains the unknown parameters to be solved. In [6], authors used this algorithm to obtain a homography between two sets of 2D points. In this scheme, the same method is used, but using 3D points.

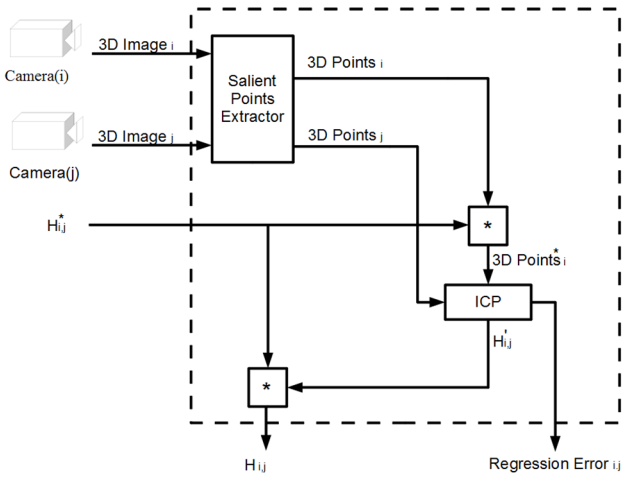

Fig. 7. Robot Interactive Homography Estimation module.

Figure 7 shows the Robot Interactive Homography Estimation module. The Salient Point Extractor module obtains a pair of sets of 3D-points [16] given a pair of 3D-images. The set of points $3Dpoints_i$ are projected towards $H^*_{i,j}$ to obtain the same set of points but referenced in the same coordinate system than the set $3Dpoints_j$, considering the human's correspondence proposal. We call this set of points $3Dpoints^*_i$. This process provides a better initial alignment than the Iterative Closest Point algorithm (ICP) [14] to find a final alignment of both sets of points through homography $H'_{i,j}$. Moreover, the regression error $Error_{i,j}$ is defined as the sum of the square distances between points in $3Dpoints^*_i$ and the projected points in $3Dpoints_j \cdot H^*_{i,j}$. The module returns homography $H_{i,j} = H^*_{i,j} \cdot H'_{i,j}$.

IV. Practical Evaluation

A. The Database

The database employed was created as follows. We used a sequence of 360 2D-images taken from the "Sagrada Familia" church in Barcelona (Spain). This sequence of pictures has been manually taken around the church by pointing the camera at the centre of it. Each image is taken with a separation of approximately one degree with respect to the centre of the church. The average distance between two consecutive shots is 1.1 meters. Given the whole sequence, we used the Bundler method [20] to extract 100,532 3D-points of the church and the information of which 2D-images visualise these 3D points. Each image has captured from 4,000 to 40,000 3D-points. Moreover, the Bundler method returns the relation between the 3D-points and the position in pixels in the images. Then, the positions of the cameras were deduced by the pose estimation method presented in [21].

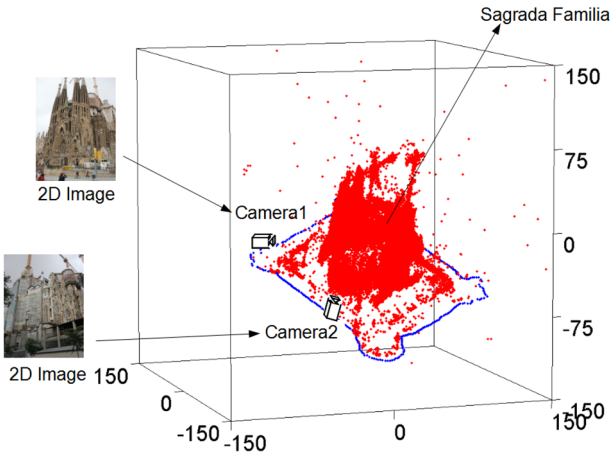

Fig. 8. "Sagrada Familia" point cloud (red points) and the camera poses (blue points) in a 3D coordinate system (in meters) in which the origin of the axes is the centre of the church.

All in all, we have generated a database of 360 registers. Each register is composed of a 2D-image, a sparse 3D-image

(the whole 3D-image is not stored but instead a cloud of points from 4,000 to 40,000 3D-points per image) and the pose of the camera (we consider this as the position of the robot). In average, the separation between two consecutive robots is 1.1 meters.

Figure 8 shows the obtained 3D model of "Sagrada Familia" (red points), and the different poses of the camera that has captured the images of the model (blue points). Axes are expressed in meters and the centre of the church is the origin of the coordinate system. Note that, due to the process used to generate this database, there are a lot of noisy points.

This database was used since facilitates the situation presented in the introduction of this paper where there are several robots around a group of people looking at the centre of the group (Figure 1). In the experiments, we suppose there are from 8 robots (45 degrees between consecutive robots and an average distance of 50 meters between them) to 72 robots (5 degrees between consecutive robots and an average distance of 6 meters between them). Note that from the structure of the database, we can deduce the sequence of robots and which are the ones that are the closest to the others. Nevertheless, in the experiments, this information is not used and robots do not know which is the closest one until the system deduces their relative position. Finally, it is not the aim of the method to perform path planning. This is a high level task considered in the general project when the robot positions are deduced through the method we present.

B. Experiments

Figure 9 shows the robot position error in meters of the system with respect to the number of interactions and the degree of separation between images (from 5 to 45). As it is supposed to be, the farther away the images are, the larger the error is since less 3D-points are shared and also the larger the distortion is between images. Moreover, when only one or two interactions are imposed, only translations can be deduced in the alignments. For this reason, the robot position error reduction is not so important. It is clear that with three interactions, the error is drastically reduced independently of the level of separation between paired images. This is because an affine homography can be deduced. Finally, when more than three interactions are imposed, the error is only slightly reduced.

Given the results of Figure 9, we could recommend to the user to interact a maximum of three times per pair of robots through all pairs of robots instead of interacting more than three times in some few pairs of robots and keeping some pairs of robots without interaction. Nevertheless, since we assume robots are moving and therefore, images are constantly updated, we only have the reliable information of the current regression error. From Figure 9, we can see that it is better to interact on the pairs of robots that have the largest regression errors because in these cases, the human interaction accentuates the regression error's decrease. We highlight the need of this human's interaction through painting in bold the pairs of robots with the largest regression errors in the HMI.

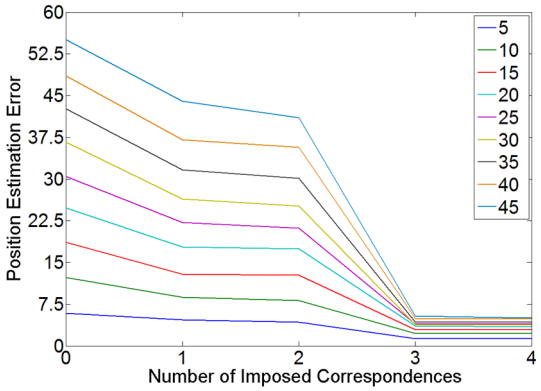

Fig. 9. Mean robot position error in meters with respect to the number of interactions and the level of separation.

A crucial aspect of the cooperative robotic systems is the ability to keep the information updated thorugh time. In this case, we need to know the pose of the fleet of robots in real time. The costliest process is the ICP algorithm (Figure 7) wich is performed each time a new image arrives, independently if there is a human interaction or not. Another costly process is the alignment estimation given the correspondences (Figure 6), which is performed through the Direct Linear Transform algorithm. Nevertheless, this step is only performed with the points that the human has interacted and, as explained before, we recommend a maximum of three interactions. For this reason, the runtime of this algorithm is almost negligible.

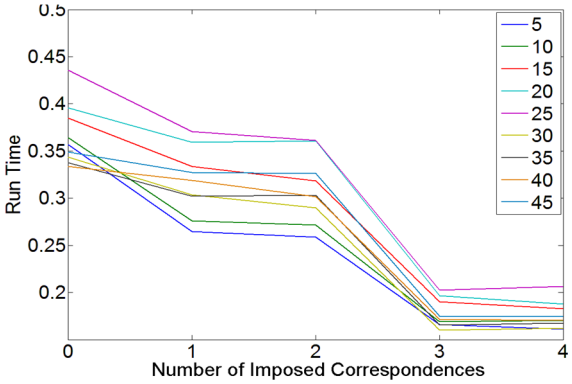

Fig. 10. Runtime of the ICP module in seconds with respect to the number of interactions and the level of separation (Matlab, i7 950, 3.07 GHz, 6 GB RAM, Windows 7).

Figure 10 shows the runtime in seconds of the ICP with respect to the number of interactions and the level of separation between images (from 5 to 45). ICP is a regression method that finds the solution more accurate and faster when the initial alignment is more congruent to the affine matrix. Considering the human's interactions, more interactivity leads to a more accurate homography matrix $H^*_{i,j}$ deduced in the Alignment Estimation Given a Correspondence module and so, easier is the task of the ICP to find the homography $H'_{i,j}$.

Finally, it is important to consider that previous human interactions positively influence on the pose accuracy and

runtime of the following non-interactive estimations. That is, the following estimations tend to be deduced in a more accurate and faster manner.

V. CONCLUSIONS

When a fleet of robots has to perform a task in a collaborative way, one of the most important low-level tasks that they have to face is the pose estimation of all robots. Clearly, the imminent reaction of each robot directly depends on its pose and the relative pose of the other robots with respect to them. Besides, an inaccurate pose estimation of one of the robots influences not only at the low-level behaviour of this robot, but also on the high-level task of the whole fleet. Our experience has shown us that in most of the cases, the high-level task has not been accurately carried out is because pose estimation was not properly solved.

When GPS or landmarks cannot be used, it is usual to solve the pose estimation through image alignment. Nevertheless, depending on the environment, this is a really difficult task. For this reason, we have advocated for incorporating human interactivity, which is only demanded when the automatic image alignment solver is not able to achieve a good estimation. Our policy is based on the fact that it is worth to ask to a human to solve a low level task than not to achieve a high level task, although the human response could be slower than the fully automatic system. Nevertheless, note that while the human interacts on a pair of robots, the other ones can automatically solve the image alignment and the pose estimation.

The novelty of this paper is to use human interaction to improve the pose accuracy of a fleet of robots in a cooperative robotics framework. The system automatically aligns the 3D-points but manually aligns the 2D-points when it is needed. Stereoscopic cameras are embedded in the robots. In this paper, we have only explored how the interactive image alignment is solved to achieve the pose estimation. We have depicted the whole framework and we have concretised on the modules that specifically deduces the pose of the fleet given the human interaction. The method we have presented is part of a larger project in which social robots guide people through outdoor or indoor scenes. Experimental section shows that it is clear that pose accuracy increases with only few human interactions. Moreover, the runtime of the alignment module (based on ICP algorithm) is reduced when the initial estimation is close to the solution. Therefore, an initial human interaction does not only convey the automatic system to achieve a better pose estimation, but also reduces the runtime of the subsequent image alignments.

Since we consider that the methodology has been validated through the current database, as a future work, we intend to integrate this method into a real-time robot environment. To do so, we first have to decide which part of the method has to be run in the robots and which part has to be run in a central server. Moreover, we have to analyse how to parallelise the functions. We want to code the robots' software in ROS and the server software in C. Finally, we intend to integrate this system into the social robots guide people project.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Casper J.& Murphy RR., (2003),Human–robot interactions during the robot-assisted urban search and rescue response at the world trade centre, IEEE Transactions on Systems, Man, and Cybernetics, Part B. 33: 367–385.

[2] Ifthekhar, Saha, Jang (2015). Stereo-vision-based cooperative-vehicle positioning using OCC and neural networks, Optics Communications, 352, pp: 166–180

[3] Kim, S. Taguchi, S. Hong, S. Lee, H. (2014). Cooperative behavior control of robot group using stress antibody allotment reward, Artificial life and robotics 19 (1), pp: 16-22.

[4] Garcia,C., Cena, P. F. Cardenas, R. Saltaren, L. Puglisi, R. Santonja, A (2013). Cooperative multi-agent robotics system: Design and modelling, Expert Systems with Applications, 40,pp: 4737–4748.

[5] Garrell, A. Sanfeliu, A. (2012). Cooperative social robots to accompany groups of people, International Journal of Robotics Research, 31(13): 1675-1701.

[6] Cortés X. & Serratosa, F., (2015). An Interactive Method for the Image Alignment problem based on Partially Supervised Correspondence, Expert Systems With Applications 42 (1), pp: 179 - 192.

[7] Hou, Sun, Jia, Zhang, (2012). An Autonomous Positioning and Navigation System for Spherical Mobile Robot, Procedia Engineering, 29, pp: 2556–2561

[8] Serratosa, F. Alquézar R. & Amézquita, N. (2012). A Probabilistic Integrated Object Recognition and Tracking Framework, Expert Systems With Applications, 39, pp: 7302-7318.

[9] Montiel, O.Orozco-Rosas, U. Sepulveda, R. (2015). Path planning for mobile robots using Bacterial Potential Field for avoiding static and dynamic obstacles, Expert Systems with Applications, 42,pp: 5177–5191.

[10] G. Manzo, F. Serratosa & M. Vento, Online Human Assisted and Cooperative Pose Estimation of 2D-cameras, Expert Systems With Applications, , pp:, 2016.

[11] X. Cortés & F. Serratosa, Cooperative Pose Estimation of a Fleet of Robots based on Interactive Points Alignment, Expert Systems With Applications, 45, pp: 150–160, 2016.

[12] Harris C. & Stephens M. (1988). A combined corner and edge detector. Proceedings of the 4th Alvey Vision Conference. pp. 147–151.

[13] Fischler, M.A. Bolles, R.C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6), pp: 381–395.

[14] Zhang, Z. (1994). Iterative point matching for registration of free-form curves and surfaces, Int. J. Comput. Vision 13 (2), pp: 119–152.

[15] Correal, R. Pajares, G. Ruz, J.J., (2014). Automatic expert system for 3D terrain reconstruction based on stereo vision and histogram matching, Expert Systems with Applications, 41,pp: 2043–2051.

[16] Thirion, J. (1996). New feature points based on geometric invariants for 3D image registration, International Journal of Computer Vision 18 (2), pp: 121-137.

[17] Villamizar, M., Andrade-Cetto, J. Sanfeliu, A. Moreno-Noguer, F. (2012). Bootstrapping Boosted Random Ferns for discriminative and efficient object classification, Pattern Recognition 45(9) pp: 3141-3153.

[18] Katevenis, M. Sidiropoulos, S. Courcoubetis, C. (1991). Weighted round-robin cell multiplexing in a general-purpose ATM switch chip, IEEE Journal on Selected Areas in Communications, 9(8).

[19] Hartley, R.& Zisserman, A.(2003). Multiple View Geometry in computer vision. Cambridge University Press.

[20] Snavely, N. Todorovic, S. (2011). From contours to 3D object detection and pose estimation, International Congress on Computer Vision.

[21] Rubio A. (2015). Efficient monocular pose estimation for complex 3D models, accepted for publication in International Congress on Robotics and Automation.