Combining Laser-Scanning Data and Images for Target Tracking and Scene Modeling

Hongbin Zha, Huijing Zhao, Jinshi Cui, Xuan Song and Xianghua Ying

Abstract Working environments of modern robots have changed to unstructured, dynamic and outdoor scenes. There emerged several new challenges along with these changes, mainly in perception of both static and dynamic objects of the scenes. To tackle these new challenges, this research focused on study of advanced perception systems that can simultaneously model static scenes and track dynamic objects. Our research has three features. Multi-view and multi-type sensors, together with machine learning based algorithms, are utilized to obtain robust and reliable mapping/tracking results. In addition, a car-based mobile perception system is developed for exploring large sites. Finally, to improve robustness of the multi-view and mobile perception system, some new camera calibration methods are proposed. This paper presents an overview of our recent study on above mentioned ideas and technologies. Specifically we will focus on multi-sensor based multiple target tracking, simultaneous 3D mapping and target tracking in a mobile platform, and camera calibration.

1 Introduction

During last decade, with rapid developments in robotics technologies and their applications, working environments of robots have changed much. They change from indoor to outdoor, from structured to unstructured, from static to dynamically changing environments.

Hongbin Zha e-mail: zha@cis.pku.edu.cn
Huijing Zhao e-mail: zhaohj@cis.pku.edu.cn
Jinshi Cui e-mail: cjs@cis.pku.edu.cn
Xuan Song e-mail: songxuan@cis.pku.edu.cn
Xianghua Ying e-mail: xhying@cis.pku.edu.cn
Key Laboratory of Machine Perception (MoE), Peking University

To adapt to these changes, and also to achieve more complicated tasks in new working environments, it is necessary to develop advanced robotics perception systems that are more reliable, robust and capable in scene recognition and understanding. Specifically, the challenges emerging with changing environments include three aspects. Firstly, most of the environment is unknown, and contains both static and dynamic objects, with complex relationships and structures among them. Secondly, there are various targets needing to be recognized, and their categories are more complicated. The targets may include static objects for instance architectures, trees and roads, as well as dynamic objects, like pedestrians, bicycles and cars. Finally, in a dynamic and multi-object environment, when multiple dynamic objects interact with each other, the perception task becomes more challenging in object recognition, localization and tracking due to inter-occlusion.

To address above problems, we need to find some new research directions. First of all, it would be necessary to make full use of 3D information of various objects. At the same time, we also have to utilize prior information of specific outdoor environments, gather data from multiple sensors and then fuse them efficiently. Moreover, to cover a wide and complicated area, it would be indispensable to use a distributed sensor network or an actively controlled mobile platform.

According to above mentioned directions, we explored research in multi-target tracking and 3D environment mapping with laser scanners and video cameras. It has three main features: 1) Multi-sensors and machine learning based algorithms are utilized to obtain robust and reliable mapping and tracking. 2) A car-based mobile perception system is developed for 3D mapping of static environment scenes, as well as for recognizing and tracking of dynamic targets. 3) To improve robustness of the multi-view and mobile perception systems, some new camera calibration methods are proposed.

This paper presents an overview of our recent study on above technologies and systems. Specifically we will focus on multi-sensor based multiple target tracking, simultaneous 3D mapping and target tracking in a mobile platform, and camera calibration.

2 Multi-Target Tracking in Dynamic Scene

Multi-target tracking plays a vital role in various applications, such as surveillance, sports video analysis, human motion analysis and many others. Multi-target tracking is much easy when the targets are distinctive and do not interact with each other. It can be solved by employing multiple independent trackers. However, for those targets that are similar in appearance, obtaining their correct trajectories becomes significantly challenging when they are in close proximity or partial occlusions. Previous approaches using joint trackers searching in joint state space requires high time consumptions. Moreover, maintaining the correct tracking seems almost impossible when the well-known ”merge/split” condition occurs (some targets occlude

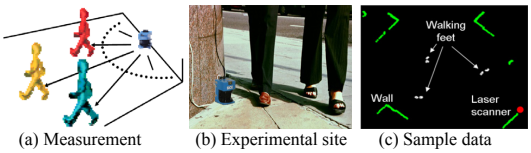
others completely, but they split after several frames). Hence, the goals of our research are: 1) to design a multi-sensor system that will help obtain a better tracking performance with lower time consumption than those obtained from independent or joint trackers when the interactions occur; 2) to make a new attempt to solve the "merge/split" problem in multi-target tracking.

Our recent research on these goals covers three aspects: 1) To solve partial occlusion and interaction in a joint space, a detection-driven MCMC based particle filter framework is proposed. In this approach, MCMC sampling is utilized to increase search speed with detection maps providing searching directions. 2) To address the problem of severe occlusion and interaction, we use learning and classification loops for data association. 3) To reduce time consumption and to improve tracking performance, we fuse laser and vision data. Compared to traditional vision-based tracking systems, the laser range scanner can provide directly 3D depth information of targets, which is absent in visual images. In a laser-based tracking system (as shown in Fig.1), the targets are represented by several points, and hence the tracking become much easy with good performance in both accuracy and time-cost.

2.1 Detection-Driven MCMC Based Particle Filter

In the visual tracking area, to keep track of multiple moving objects, one generally has to estimate the joint probability distribution of the state of all objects. This, however, is intractable in practice even for a small number of objects since the size of the state space grows exponentially in the number of objects. An optional solution of the problem is to use the detection based data association framework, which is originated from radar tracking techniques. Most of existing laser based tracking systems used this framework. In these systems, a clustering/segmentation based detection algorithm provides locations of potential targets. Then, the measurements are associated with previously estimated target trajectories in a data association step. Above detection driven tracking schemes greatly rely on the performance of the detection algorithms. Only observation at locations with high detection responses are considered as potential measurements. This will incur that false alarms and non-detections significantly influence performance of the tracker.

In our recent research, we construct our novel observation with two types of measurements, including a foreground image frame given by background subtraction and a detection map on the single frame. For inference, we proposed a detection



(a) Measurement (b) Experimental site (c) Sample data

Fig. 1 A typical laser based tracking system

incorporated joint particle filter, considering the both types of measurements. First, data association for the targets to detected measurements is incorporated to the state proposal, to form a mixture proposal that combines information from the dynamic model and the detected measurements. Then, we utilize a MCMC sampling step to obtain an efficient multi-target filter.

We applied this idea to laser scan frames [1]. Four laser scanners are used for scanning on the height of 16cm from horizontal ground. Fig.2 is a screen copy of trajectories of tracked persons using our MCMC particle filter, where green points represent laser points of background (doors, tables, walls, etc.); white points represent the laser points of moving feet. Colour lines are trajectories. Red circles are position of people at current time. The numbers denotes the trajectory indices. The standard particle filter can track an individual person very well using 100 particle samples, if he/she is quite far away from other persons. However, if two persons walk closely, it is very common that one person's trajectory "hijacks" another person's, since there is not a joint likelihood to handle the interaction situation. Our MCMC particle filter benefits from the feature detection and the mixture transition proposal. It tracks 28 persons simultaneously and nearly in real-time, and gives a robust tracking result, even using only 100 particle samples.

We also applied this idea to vision based tracking systems. In [2],we proposed a Probabilistic Detection-based Particle Filter (PD-PF) for multi-target tracking. In our method, we incorporate possible probabilistic detections and information from dynamic models to construct a mixed proposal for the particle filter, which models interactions and occlusions among targets quite effectively.

2.2 On-line Learning for Data Association

To address long-time occlusion and severe interactions, we proposed an online learning based approach for data association. The core idea of our research is il-

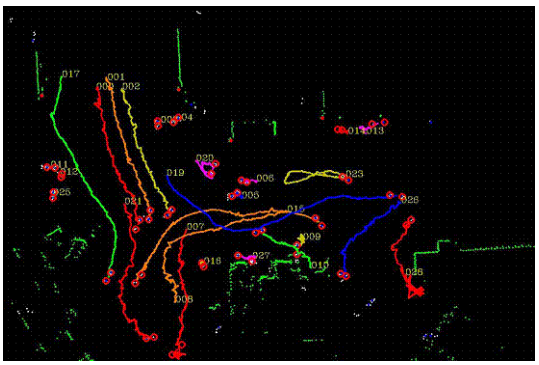


Fig. 2 Laser based tracking results using detection-driven MCMC based PF.

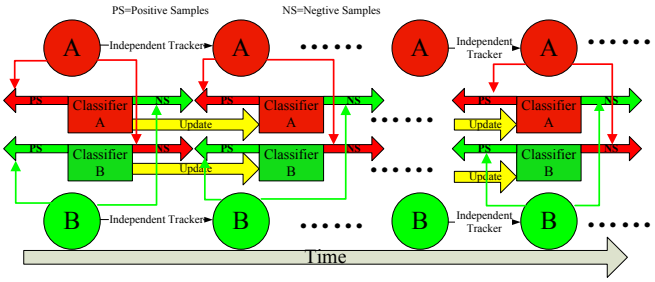


Fig. 3 Tracking for learning.

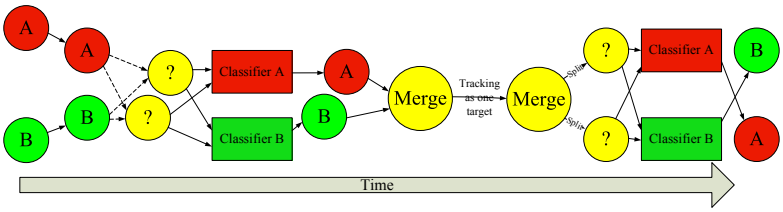


Fig. 4 Learning for tracking.

lustrated in Fig.3 and Fig.4. When two targets do not interact with each other (see Fig.3), tracking becomes very easy and multiple independent trackers are employed. Due to the reliability of these tracking results, they are used as positive or negative samples to train a classifier for each target. When the two targets are in close proximity (see Fig.4), the learned classifiers are used to assist in tracking. Specifically, when the two targets merge, we assign a new state space and track this "merging target" as one target. When they split, their classifiers are used again to specify a correct identification.

We applied this idea respectively to vision based systems [3] and laser based systems [4]. In a vision based tracking system, when the targets do not interact with each other, we utilize independent trackers to perform the tracking. Once we obtaine the tracking results of each target, a set of random image patches are spatially sampled within the image region of each target. We utilize these random image patches as samples for the online supervised learning. In this case, each target is represented by a "bag of patches" model.When the targets are in close proximity or present partial occlusions, a set of random image patches are sampled within the interacting region of the detected map, and the feature vectors of these image patches are input to the classifiers of interacting targets respectively. The outputs of these classifiers are scores, which are used to weight the observation model in the particle filter. The overview of the process is shown in Fig.5.

Sometimes, several targets occlude another target completely. Maintaining the correct tracking of targets seems quite impossible. Once it occurs, we initialize the state of the "merging targets" and track it as one target. If we detect that this

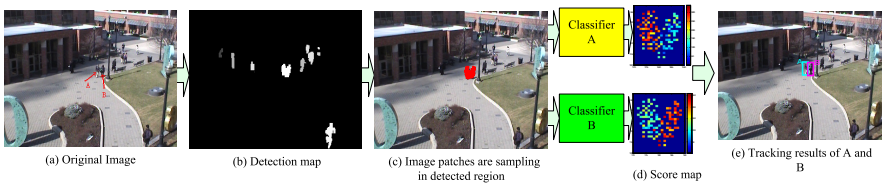


Fig. 5 Correlated targets tracking.

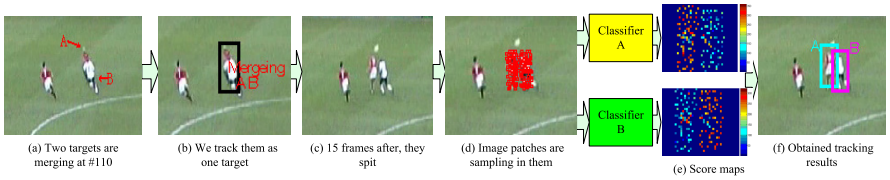


Fig. 6 Merge/Split condition.

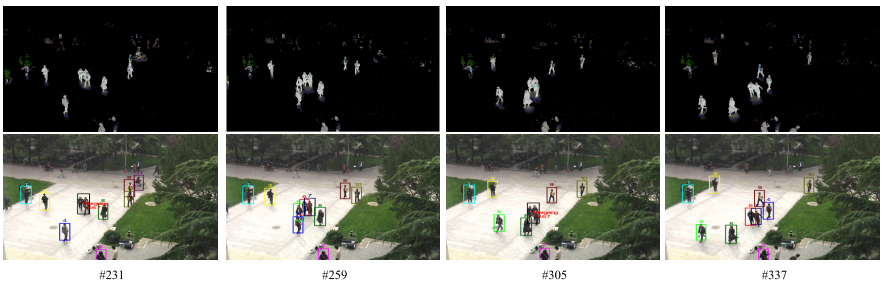


Fig. 7 Tracking results of surveillance video.

"merging target" splits and becomes an interacting condition or a non-correlated condition, we utilized the classifiers of these targets to identify them (as shown in Fig.6). Hence, we can link the trajectories of these targets without difficulty. Some experimental results were shown in Fig.7.

We have applied the same idea to laser based tracking systems [4]. Although it performs better than previous approaches in the situations of interactions and occlusions, due to the missing appearance information, it is very hard to obtain a set of features that uniquely distinguish one object from another.

2.3 Fusion of Laser and Vision

The work on fusion of laser and visual data is motivated from pursuing a reliable and real-time multi-target tracking system, which is difficult to achieve via only laser or only visual data. We have tried several strategies for the fusion of two modes. In [5],

a Kalman Filter based approach is utilized for decision-level fusion of two independent tracking results respectively from a laser-based sub-system and a vision-based subsystem. Although the time-consuming of this approach is rather high, it can provide reasonable tracking results. In [6], the fusion is processed in the detection stage at first to improve detection rates and decrease false-alarm rates. Then, an observation model that combines visual and laser features is utilized for filtering. The fusion in detection-level saves a lot time since no more window scanning is necessary in image frames. However, it still request much time for gathering measurements from both modes in filtering process. Moreover, it is difficult to decide the confidence coefficients of laser data and vision data when complex interaction situations occur.

In our most recent research, we proposed a fusion strategy that tries to make these two modes to fully display their respective advantages in one framework. The key idea of this work is illustrated in Fig.8. When the targets do not interact with each other, the laser scanner can perform the efficient tracking and it is easy for us to extract visual information from the camera data. Due to the reliability of these tracking results, they are used as positive or negative samples to train some classifiers for the "possible interacting targets". When the targets are in close proximity, the learned classifiers and visual information will in turn assist in tracking. This mode of cooperation between laser and vision, and between tracking and learning, offers several advantages: (1) Laser and vision can fully display their respective advantages (fast measurements of laser scanners and rich information of cameras) in this system. (2) Because the "possible interacting targets" are represented by a discriminative model with a supervised learning process, the method can employ information from the "confusing targets" and can sufficiently exploit the targets' history. Through these discriminative models, we can easily deal with some challenging situations in the tracking. (3) This "tracking-learning adaptive loop" ensures that the entire processes can be completely on-line and automatic.

3 Omni-Directional Sensing of a Dynamic Environment Using an Intelligent Vehicle

This research focuses on the sensing technologies of intelligent vehicles. We intend to develop a car of omni-directional eyes looking at the environment of both static and dynamic objects, where the car detects the moving objects in surrounds, and tracks their states, such as speed, direction, and size, so that dangerous situations can be predicted. Moreover, the car can be also used to generate a 3D copy of the dynamic urban scenery that contains both stationary objects, e.g. buildings, trees, road, and mobile objects, e.g. people, bicycles and cars. Here, we need to consider the following issues: finding the vehicle's pose as it moves around, detecting and tracking the moving objects in surroundings, and generate a 3D representation of the whole environment.

An intelligent vehicle system has been developed as shown in Fig. 9, where five single layer laser scanners (briefly noted by "L") are mounted on the car to profile

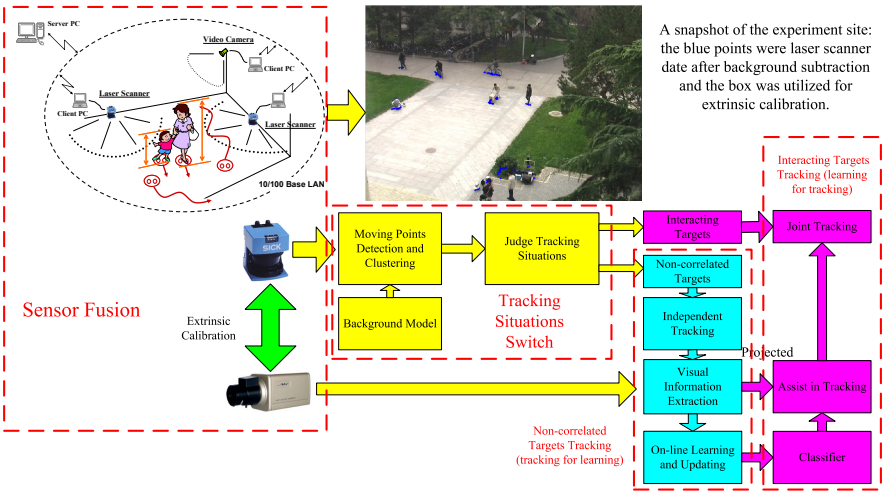


Fig. 8 Fusion of laser and vision

object geometry along the streets from different viewpoints and with different directions; a video camera is also integrated to monitor the front of the vehicle and obtain textures; a GPS (Global Positioning System)/IMU (Inertial Measurement Unit) based navigation unit is applied to give outputs of the vehicle pose. Sensor layouts might be varied according to applications. However, a common and important issue here is how to fuse such a large number of sensors, so that the multi-modal sensing data can conduct the above missions, while a comprehensive perception that overcomes the shortages of each singular sensor is achieved.

Fig. 9 A picture of the intelligent vehicle.

3.1 System Architecture

Normally, localization of the host vehicle is conducted by using the positioning sensors such as GPS, IMU, and VMS (Vehicle Motion Sensor). Thus the vehicle pose that contains the position (x,y,z) and orientation (ω, ϕ, κ) of the host vehicle is estimated in a high frequency. With it as input, the environmental sensors such as laser scanners, radars, video cameras are exploited, so that the sensing of local environments as the intelligent vehicle running along streets can be integrated to generate a global knowledge of the whole environment.

Such an approach is widely accepted in existing intelligent vehicles and mobile mapping systems. Localization and environmental perception are conducted individually using different sensing technologies. The system architecture is straightforward. However, disadvantage of such an approach is that the environmental perception is heavily dependent on outputs of the localization module. For example, erroneous localization outputs yield displacements between the environmental sensing data to the same static objects, and the slow motion objects such as pedestrians might not be reliably detected due to localization errors.

In this research, the system architecture is designed as shown in Fig. 10, where localization of the host vehicle is formulated as a SLAM with MODT (Simultaneous Localization And Mapping with Moving Object Detection and Tracking) by fusing both the positioning sensors (GPS/IMU) and the environmental sensors (laser scanners/cameras). We proposed a method of integrating both positioning (GPS/IMU) and environmental (laser scanners) sensors to improve localization accuracy, meanwhile, to conduct a 2D mapping and moving object detection/tracking. With these as inputs, other environmental sensors can perform an advanced (e.g. 3D) environmental perception with high accuracy.

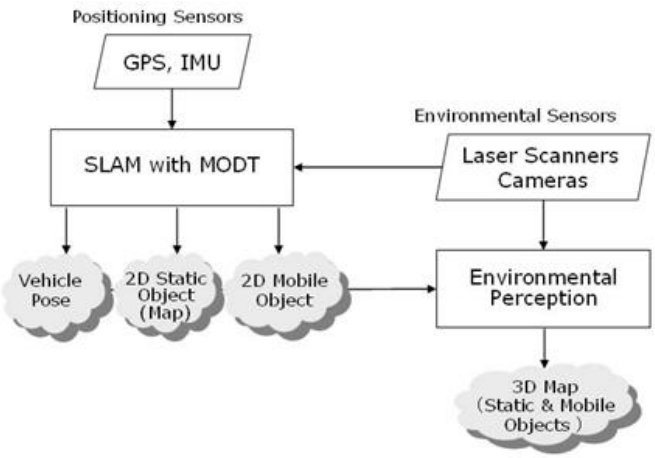


Fig. 10 System Architecture.

In order to achieve the purpose, a sensor fusion strategy and software implementation is proposed as shown in Fig. 11. There are four processors inside the vehicle, for sensor data logging, SLAM with MODT, 3D mapping, moving object (briefly "MO") recognition, respectively. All the processors are connected through Ethernet, so that the data processing and transferring are conducted in an online mode. The processor 1 controls all the sensors, records GPS/IMU/laser scan data, and distribute them to other processors through Ethernet. The processor 2 will receive the GPS/IMU data and the horizontal scanning laser data (L1) to conduct SLAM with MODT [7],[8]. The laser data of moving objects are forwarded to the processer 4, where each object is recognized as a person, a group of people, a bicycle or a car by fusion with video images [9] . On the other hand, the vehicle pose estimated by processor 2 is forwarded to processor 3, as well as the other processing results on static and moving objects. A 3D mapping is conducted by processor 4, which with the input of vehicle poses and sensor geometry parameters, integrates all laser scan data in a global coordinate system [10] .

3.2 SLAM with MODT

A laser-based SLAM is proposed in our previous research [7] , where the problem is formulated as SLAM with object tracking and classification, and the focus is on managing a mixture of data from both dynamic and static objects in a highly cluttered environment. Here people and cars might get very close to each other, and their motion patterns have much variability and are always unpredictable. Thus,

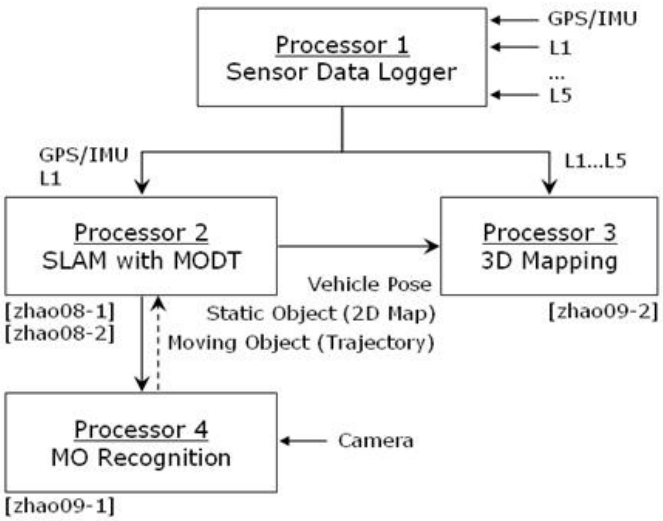


Fig. 11 Sensor fusion and software modules.

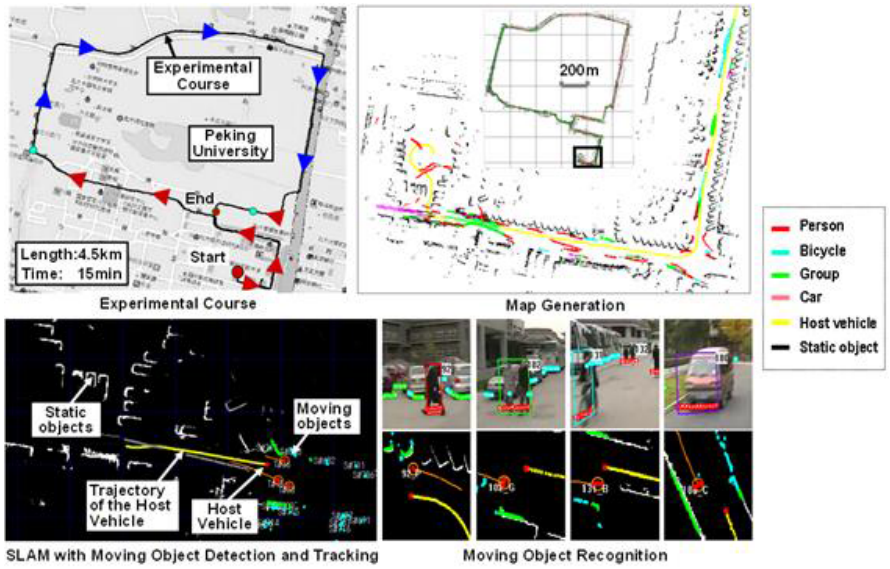


Fig. 12 Experimental results on SLAM with MODT and MO recognition.

it is risky to discriminate moving or static objects just by buffering an area using the data from a previous measurement. Also, it is risky to judge based only on an instance measurement, as many objects might have similar data appearance due to limited spatial resolution, range error, partial observation, and occlusions. The general idea behind our system is that the detected objects should be discriminated in a spatial-temporal domain. In this way, after an object is detected, it is tracked until the system can classify the object into either a static or moving object with certainty. On the other hand, in order to achieve a localization of global accuracy and robustness, especially when the vehicle makes a non-cyclical measurement in a large outdoor environment, a GPS/IMU assisted strategy is also developed. The sporadically-available GPS measurements in urban areas are used to diagnose errors in the vehicle pose estimation, and the vehicle trajectory is then adjusted to close the gap between the estimated vehicle pose and the GPS measurement. An extensive experiment is presented in [8] ,where the host vehicle ran a course about 4.5km in a highly dynamic environment, and a map is generated containing the data of both the static and moving objects observed along the road. Some of the results are demonstrated in Fig. 12.

3.3 Sensor Alignment and 3D Mapping

A method is developed in [10] for calibrating the sensor system of multiple single-layer laser scanners. The problem is formulated into registration of laser point clouds

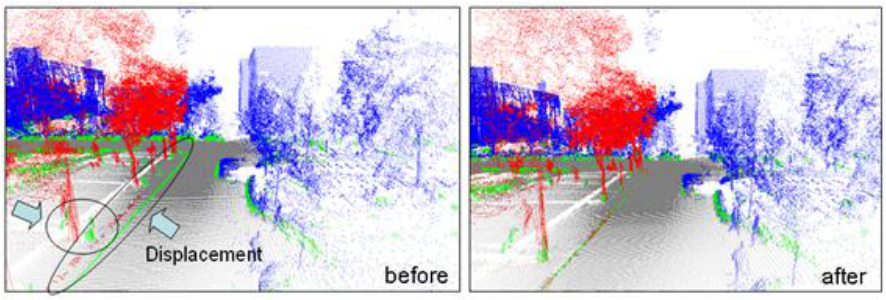


Fig. 13 Experimental results of sensor alignment.

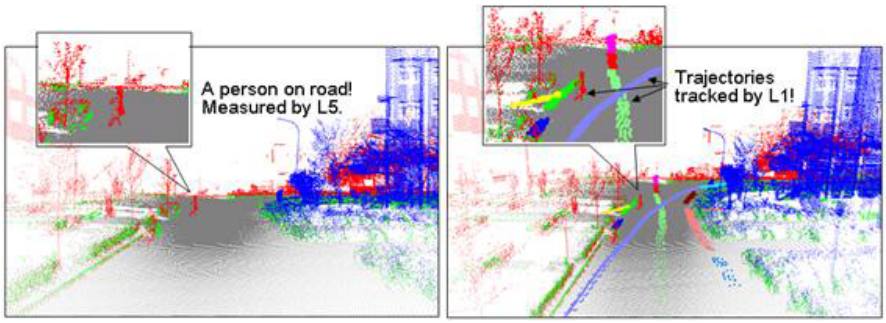


Fig. 14 An omni-directional view of the dynamic environment.

of different laser scanners in a short run where the relative vehicle poses are considered of necessary accuracy. The registration is conducted in two steps: horizontal and vertical matching, using the laser points of vertical objects and ground surfaces respectively. An experiment is shown in Fig. 13, demonstrating the results before and after the sensor alignment. After the sensor alignment, the laser points from different laser scanners can be integrated into a global coordinate system with more consistency. In Fig. 14, a result is shown for integrating the laser points from laser scanners L1 (colored trajectories), L2 (on-the-road in gray, off-the-road in green), L4 (blue) and L5 (red). In the center of the view, a person is captured by the laser scanner L5 (red). It is easy to know that the person is walking on the road, with the laser points measured by the laser scanner L2 nearby its feet being colored with gray. A trajectory of the person is also known, which is captured by the laser scanner L1 (colored trajectories), so that we can predict where the person is walking ahead, etc.

4 Camera Calibration

Camera calibration is a process of modeling the mapping between 3D objects and their 2D images, and this process is often required when recovering 3D information from 2D images, such as in 3D reconstruction and motion estimation. The parameters of a camera to be calibrated are divided into two classes: intrinsic and extrinsic. The intrinsic parameters describe the camera's imaging geometric characteristics, and the extrinsic parameters represent the camera's orientation and position with respect to the world coordinate system. Many approaches to camera calibration have been proposed and they can be classified into two categories: using calibration objects and self-calibration. We have investigated those using spheres, conics and circles [11], [12],[13]. Here we describe some results given in [12] and [13].

4.1 Calibration from Unknown Principal-Axes Aligned Central Conics

Conics are one of the most important image features like points and lines in computer vision. The motivation to study the geometry of conics arises from the facts that conics have more geometric information, and can be more robustly and exactly extracted from images than points and lines. In addition, conics are very easy to be produced and identified than general algebraic curves, though general algebraic curves may have more geometric information. In our research [13], we discovered a novel useful pattern, principal-axes aligned. Moreover, the properties of two arbitrary principal-axes aligned conics with unknown or known eccentricities are deeply investigated and discussed.

These properties are obtained by utilizing the generalized eigenvalue decomposition of two principal-axes aligned conics. We define the absolute points of a conic in standard form, which is analogy of the circular points of a circle. Furthermore, we define the dual conic formed by two absolute points, which is analogy of the dual conic consisted of circular points. By using the dual conic formed by two absolute points, we proposed a linear algorithm to obtain the extrinsic parameters of the camera. We also discovered a novel example of the principal-axes aligned conics, which is consisted of a circle and a conic concentric with each other while the parameters of the circle and the conic are both unknown, and two constraints on the IAC can be obtained from a single image of this pattern.

4.2 Calibration from a Circle and a Coplanar Point at Infinity

In a large scene, like football or basketball courts, very large calibration patterns may be required. In most of such cases, however, setting a large planar pattern on

Fig. 15 Some calibration results

the site is infeasible. One may suggest that the lines in the courts may be used for calibration. However, it requires that the locations of these lines must be given in advance. Another problem is that, the field of view (FOV) of a camera usually involves a portion of the whole court, and this often makes the number of lines in FOV not sufficient to estimate any intrinsic or extrinsic parameters of the camera. In the case of calibrating multiple cameras distributed in the football or basketball scenes, cameras with common FOVs are often required. It is not difficult to achieve this by changing the orientations of all cameras in the scenes to view the center circle and the midfield line.

The main contributions of our another research [13] is: It is the first work to deal with the problem of camera calibration just using the information in the midfield, i.e., the center circle and the midfield line. It shows that the camera calibration using one circle and one line passing through the circle's center, and that using one circle and one point at infinity in the support plane of the circle, are equivalent. From the latter one, the derivations may become very clear. It may be used for multiple camera calibration or camera network calibration in football or basketball scenes.

We performed a number of experiments, both simulated and real, to test our algorithms with respect to noise sensitivity. In order to demonstrate the performance of our algorithm, we capture an image sequence of 209 real images, with resolution $800* 600$, to perform an augmented reality task. Edges were extracted using Canny's edge detector and the ellipses were obtained using a least squares ellipse fitting algorithm. One of the examples is shown in Fig. 15 to illustrate the calibration results.

5 Discussion

When dealing with unknown, unstructured and dynamic working environments, robots need to recognize both static and dynamic objects of the scene. This paper gives an overview of our recent research on moving target tracking and 3D scene

modeling via multi-sensor perception systems in a multi-view or mobile platform. 3D information from laser scanning data complements the missing depth in video images. Therefore 3D scene modeling and target detection/tracking become much easy and robust. On the other hand, assistance from visual data improves the recognition accuracy by allowing many machine learning algorithms to be put into use easily. The results showed that the proposed methods can find many potential applications in robotics and other fields such as intelligent transportation and surveillance.

Acknowledgements This work was supported in part by NKBRPC (No.2006CB303100), NSFC Grant (No.60333010), NSFC Grant (No.60605001), the NHTRDP 863 Grant (No.2006AA01Z302) and NHTRDP 863 Grant (No.2007AA11Z225).

References

1. Cui J., Zha H., Zhao H., Shibasaki R.: Fusion of detection and matching based approaches for laser based multiple people tracking. In: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 642–649 (2006)
2. Song X., Cui J., Zha H, Zhao H.: Probabilistic detection-based particle filter for multi-target tracking. In: Proc. of British Machine Vision Conference, 223–232 (2008)
3. Song X., Cui J., Zha H, Zhao H.: Vision based mutilple interacting targets via on-line supervised learning. In: Proc. European Conference on Computer Vision, 642–655 (2008)
4. Song X., Cui J., Zha H, Zhao H.:Tracking interacting targets with laser scanner via on-line supervised learning. In: Proc. IEEE International Conference on Robotics and Automation, 2271–2276 (2008)
5. Cui J., Zha H., Zhao H., Shibasaki R.: Multimodal tracking of people using laser scanners and video camera. Image and Vision Computing (IVC), 240–252 (2008)
6. Song X., Cui J., Zhao H., Zha H.: A Bayesian approach: fusion of laser and vision for multiple pedestrains detection and tracking. International Journal of Advanced Computer Engineering (IJACE) (2009)
7. Zhao H., Chiba M., Shibasaki R., Shao X., Cui J., Zha H.: SLAM in a dynamic large outdoor environment using a laser scanner. In: Proc. IEEE Int. Conf. on Robotics and Automation, 1455–1462 (2008)
8. Zhao H., Chiba M., Shibasaki R., Katabira K., Cui J., Zha H.: Driving safety and traffic data collection - a Laser scanner based approach. In: Proc. IEEE Intelligent Vehicles Symposium, 329–336 (2008)
9. Zhao H., Zhang Q., Chiba M., Shibasaki R., Cui J., Zha H.: Moving object classification using horizontal laser scan data. In Proc. IEEE Int. Conf. on Robotics and Automation (2009)
10. Zhao H., Xiong L., Jiao Z., Cui J., Zha H.: Sensor alignment towards an omni-directional measurement using an intelligent vehicle. In Proc. IEEE Intelligent Vehicle Symposium, 292–298 (2009)
11. Ying X., Zha H.: Geometric interpretations of the relation between the image of the absolute conic and sphere Images. IEEE Trans. on Pattern Analysis and Machine Intelligence(PAMI), 28(12), 2031-2036 (2006)
12. Ying. X, Zha H.: Camera calibration using principal-axes aligned conics. In: Proc. 8th Asian Conf. on Computer Vision, 138–148 (2007)
13. Ying. X, Zha H.: Camera calibration from a circle and a coplanar point at infinity with applications to sports scenes analyses. In: Proc. 2007 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 220–225 (2007)