2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Interacting Object Tracking in Crowded Urban Areas

Chieh-Chih Wang, Tzu-Chien Lo and Shao-Wen Yang
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
Email: bobwang@ntu.edu.tw,{bright,any}@robotics.csie.ntu.edu.tw

Abstract— **Tracking in crowded urban areas is a daunting task. High crowdedness causes challenging data association problems. Different motion patterns from a wide variety of moving objects make motion modeling difficult. Accompanying with traditional motion modeling techniques, this paper introduces a scene interaction model and a neighboring object interaction model to respectively take long-term and short-term interactions between the tracked objects and its surroundings into account. With the use of the interaction models, anomalous activity recognition is accomplished easily. In addition, move-stop hypothesis tracking is applied to deal with move-stop-move maneuvers. All these approaches are seamlessly integraded under the variable-structure multiple-model estimation framework. The proposed approaches have been demonstrated using data from a laser scanner mounted on the PAL1 robot at a crowded intersection. Interacting pedestrians, bicycles, motorcycles, cars and trucks are successfully tracked in difficult situations with occlusion.**

I. Introduction

Scene understanding is a key prerequisite for making a robot truly autonomous. Establishing the spatial and temporal relationships among the robot, stationary objects and moving objects serves as the basis for scene understanding. In [1], [2], we presented the theory and algorithms to solve the Simultaneous Localization, Mapping and Moving Object Tracking (or SLAMMOT) problem. The experimental results from a ground vehicle at high speeds in crowded urban areas demonstrated the feasibility of SLAMMOT. A wide variety of moving object in crowded urban areas are detected and tracked successfully. The stationary and moving object maps are built incrementally. However, we assumed that the robot and moving objects move independently of each other to reduce the complexity of SLAMMOT enormously. This independence assumption may be unrealistic in human inhabited environments such as crowded urban areas, shopping malls and railway stations. These environments contain a large number of constraints which affect the motions of moving objects. Targets interact both with other moving objects and their surrounding environments. Interactions among moving objects and stationary objects should be of interest for higher level scene understanding.

Without explicitly detecting and modeling interactions, a number of approaches address the filtering and data association issues of interacting object tracking. Veeraraghavan *et al.* [3] addressed a multilevel tracking approach using Kalman filter for tracking pedestrians and vehicles using cameras at intersections. Zhao and Shibasaki [4] accomplished tracking multiple pedestrians using multiple laser scanners where pedestrians' feet are detected and the pattern of the rhythmic swing feet are tracked.

Fig. 1. Left: The PAL1 robot. Right: the robot collecting data at a crowded intersection near National Taiwan University.

To properly address the interacting object tracking problem, detecting and modeling of interactions are critical. Oliver *et al.* [5] described a coupled hidden Markov model framework for recognizing human interactions such as follow, approach+talk+continue, and change direction+approach+talk+continue in a pedestrian plaza. Panangadan *et al.* [6] uses a simple distance-based method for detecting interactions among people crossing a courtyard. Bruce and Gordon [7] proposed a statistical learning method to model interactions between a target and the surrounding environment for better motion prediction. In [8], Khan *et al.* proposed a Markov chain Monte Carlo (MCMC)-based particle filter to track interacting ants in which interactions are modeled through a Markov random field motion prior.

In this paper, both short-term and long-term interactions are defined and integrated into the tracking process. The long-term interactions are modeled with the use of the stationary and moving object map built by SLAMMOT. In addition to the interaction between the tracked object and the stationary objects [9][7], behavior patterns of previous dynamic objects are taken into account. A simple abnormal activity recognition can be accomplished with this approach. The short-term interactions are modeled with the use of neighboring object tracking. As moving objects in urban areas obey the same traffic laws, strong interactions between the tracked object and the neighboring objects should always exist to avoid accidents. The tracking process is fused with the tracked object's own motion and the motion of the neighboring objects. This short-term interactions deal with occlusion issues effectively, provide better move-stop

1-4244-0602-1/07/$20.00 ©2007 IEEE.

switching prediction and achieve better tracking performance and accuracy than traditional methods. Instead of designing more complex motion models, our novel approach is to simply update the target's state estimate using the virtual measurement generated by the interaction models in the update stage of filtering.

As move-stop-move maneuvers often occur in crowded urban areas, a soft switching model-set algorithm of the variable structure multiple-model estimator [10], or the move-stop hypothesis tracking approach [2], is applied. The interaction models are seamlessly integrated with this theoretically solid multiple-model estimator framework. The feasibility of the proposed approaches are demonstrated using data collected from a SICK laser scanner mounted on the PAL1 robot at a crowded intersection as shown in Figure 1. The visual images are only for visualization in this work.

II. Background

In this section, we review the theoretical foundations of tracking, describe the variable structure multiple-model estimator and introduce the mathematical notation for describing the proposed approaches.

A. Bayesian Tracking

The tracking problem can be solved with the mechanism of Bayesian approaches such as Kalman filter and Particle filter. Assuming that the true motion mode of a target is known, we can get a simple form of moving object tracking.

$$p(x_k \mid Z_k) \tag{1}$$

where x_k is the true state of the moving object at time k, and $Z_k = \{z_1, z_2, \cdots, z_k\}$ is the perception measurement set leading up to time k. According to the Bayes' theorem and the Markov assumption, Equation 1 is derived and expressed as:

$$p(x_k \mid Z_k) \propto p(z_k \mid x_k) \int p(x_k \mid x_{k-1}) p(x_{k-1} \mid Z_{k-1}) dx_{k-1} \tag{2}$$

where $p(x_{k-1} \mid Z_{k-1})$ is the posterior probability at time $k-1$, $p(x_k \mid Z_k)$ is the posterior probability at time k, $p(x_k \mid x_{k-1})$ is the motion model and $p(z_k \mid x_k)$ is the measurement or perception model.

B. Motion Modeling

The true motion mode is often unavailable in many applications. Online motion modeling is needed. Equation 1 can be modified and formalized in the probabilistic form as:

$$\begin{aligned} p(x_k, s_k \mid Z_k) &\propto p(z_k \mid x_k, s_k) \\ &\sum_{s_{k-1}} \int p(x_k, s_k \mid x_{k-1}, s_{k-1}) p(x_{k-1}, s_{k-1} \mid Z_{k-1}) dx_{k-1} \end{aligned} \tag{3}$$

where s_k is the *true motion mode* of the moving object at time k.

Motion Modeling, or estimation of structural parameters of a system, is called *system identification* in the control literature and *learning* in the artificial intelligence literature. From a theoretical point of view, motion modeling is as important as perception/measurement modeling in Bayesian approaches. From a practical point of view, without reasonably good motion models, the predictions may be unreasonable and cause serious problems in data association and inference.

For online motion modeling, using more models is not necessarily the optimal solution. Additionally, it increases computational complexity considerably. Li [11] provided a theoretical proof that even the optimal use of motion models does not guarantee better tracking performance.

Use of a fixed set of models is not the only option for multiple model based tracking approaches. A variable structure (VS) can be used in multiple model approaches [12]. By selecting the most probable model *subset*, estimation performance can be improved. However, this requires more complicated computation procedures. Not only motion but also other types of information or constraints can be selected and added to the model set. In [13], terrain conditions are used as constraint models and are added to the model set to improve performance of ground target tracking via a variable structure interacting multiple model (VS-IMM) algorithm.

The details of the variable structure multiple-model estimation and the related algorithms are available in [12]. Although our primary contribution is to take both stationary and moving object interactions into account in the update stage instead of in the predication stage, move-stop-move maneuvers are taken care under the variable structure multiple-model estimation framework.

C. Move-Stop Hypothesis Tracking

The move-stop hypothesis tracker follows the variable structure multiple-model estimation, which has two motion model set, the move model set and the stop model set. The model sets of the move-stop hypothesis tracker can therefore be expressed as:

$$\mathbb{Q} = \{q^{(move)}, q^{(stop)}\} \tag{4}$$

where $q^{(move)}$ can consist of common motion models such as the constant-velocity (CV) model, the constant-acceleration (CA) model, the constant-turn (CT) model. Here the interacting multiple model (IMM) approach [14] is applied to integrate all motion models. $q^{(stop)}$ is the *stationary process model* described in Chapter 4.4 of [2].

In practice, the *minimum detection velocity* (MDV) can be obtained by taking account of the modeled uncertainty sources. For objects whose velocity estimates from the IMM algorithm with the moving models are larger than this minimum detection velocity, the objects are unlikely to be stationary and the IMM algorithm with the moving models should perform well.

For objects whose velocity estimates are less than this minimum detection velocity, tracking should be done with great caution. Instead of adding the stationary process model to the model set, move-stop hypothesis tracking is applied where the move hypothesis and the stop hypothesis are inferred separately.

For move hypothesis inference, tracking is done via the IMM algorithm. For stop hypothesis inference, the stationary process model is used to verify if the system is a stationary process at the moment with a short time period of measurements. The covariances from the move hypothesis and the stop hypothesis are compared. The hypothesis with more certain estimates will take over the tracking process.

Figure 2 demonstrates the performance of move-stop hypothesis tracking. It is clear that move-stop hypothesis tracking correctly tracked a move-stop-move maneuver of the tracked motorcycle. Without move-stop hypothesis tracking, the estimate diverged.

III. Interaction-aided Tracking

In this section, we describe a scene interaction model to represent the long-term interactions and a neighboring object interaction model to represent the short-term interactions. Instead of using complex motion modeling techniques, the interaction models produce virtual measurements to aid tracking via the update stage of filtering.

A. Scene Interaction Model

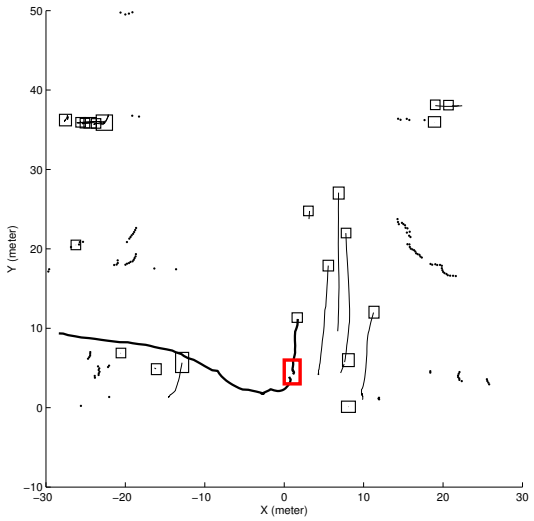
The scene interaction model is designed to represent the long-term interactions between the target and its surroundings. As the temporal and spatial information is embedded in the stationary and moving object map built by SLAMMOT, the scene interaction model uses the map to predict/constrain the possible future motion and pose of the target.

1) Modeling: The environment map built by SLAMMOT previously contains only the occupancy information of stationary and moving objects. Here the map is stored with additional information such as speed and direction of moving objects. Motion directions of tracked targets are discretized into one of nine canonical values, i.e., eight for canonical directions and one for stationary objects. The stationary mode consists of one bin and each of the eight directions consists of b bins which is given as:

$$\beta(v) = \begin{cases} \lfloor \frac{v}{interval} \rfloor & \text{for} \quad 0 \le v < b \cdot interval \\ b-1 & \text{for} \quad v \ge b \cdot interval \end{cases} \tag{5}$$

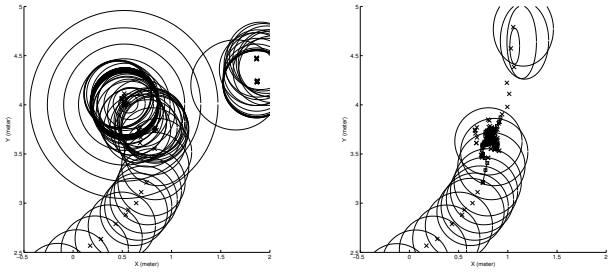
where v is the speed of the occupied object and $interval$ is a pre-determined constant. In our experiments, $interval$ is 10 km/hr and b is 8. Each bin records the occurrence count of each speed value for each direction. Figure 3 illustrates the information contained by a single grid of the built map. Figure 4 shows the built maps in which only the most observed direction of a grid is shown.

In our scenarios, the urban traffics contain strong long-term interactions because of traffic laws. Therefore, the SLAMMOT maps are automatically generated and maintained according to these behavior patterns. The behavior patterns are classified with the use of motion directions of all moving objects in the scene. The scene interaction model will use the corresponding map to predict/constrian the tracked object's motion. Figure 4 shows the SLAMMOT map according to three different behavior patterns.



(a) Tracking result of the whole scene. Rectangles denote tracked moving objects. The rectangle with a bold trajectory denotes the tracked motorcycle. The bold rectangle is enlarged in (c) and (d).

(b) The bold rectangle is the tracked motorcycle



(c) Without move-stop hypothesis tracking (d) With move-stop hypothesis tracking

Fig. 2. Move-stop hypothesis tracking: in (c) and (d), $\times$s are the measurements. The distributions of the estimates are shown by 1σ ellipse. The estimates are at the center of the ellipses which are not shown for clarity.

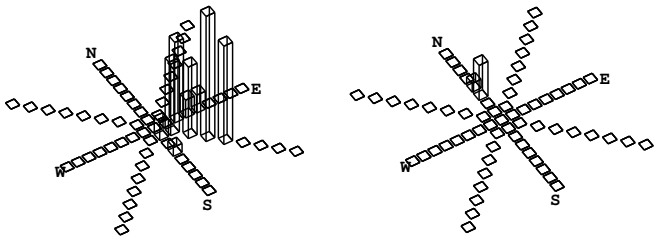
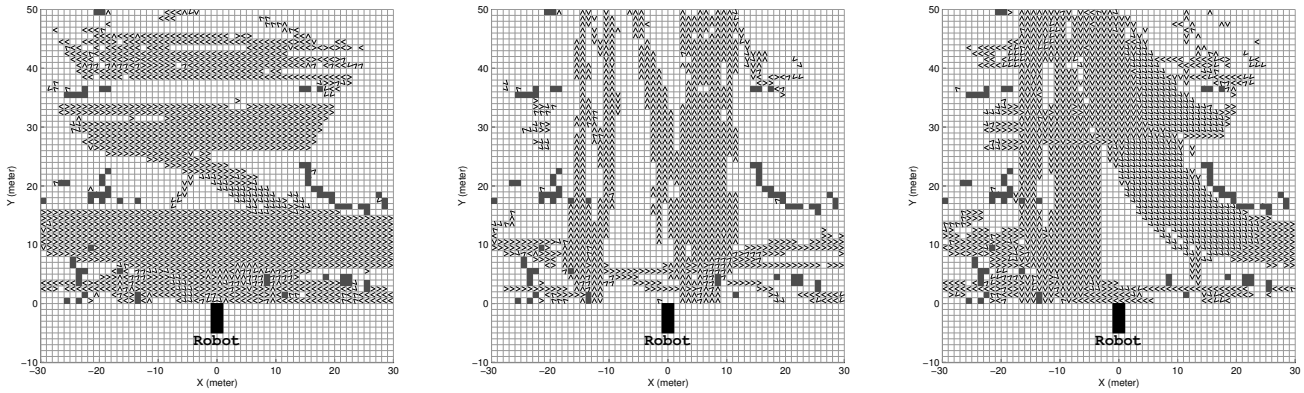


Fig. 3. The SLAMMOT map contains information of occupancy, speed and direction. Two examples are shown in which different motion patterns are embedded into the map. Left is from a road lane and right is from a crosswalk.



(a) Three different behavior patterns of an urban scene. Only the most observed motion direction of a grid is shown by an arrow inside the grid. Black grids are belonging to stationary objects. White grids are unobserved or unoccupied areas. The robot is at the origin $(0,0)$ of the map.

(b) The photos illustrate different behavior patterns of the scene.

Fig. 4. The SLAMMOT map.

2) Prediction: With the use of the SLAMMOT maps and the scene behavior pattern recognition results, we predict the possible motions of a tracked object using a sampling-based method.

Let E_k be a set containing the state vectors of these randomly generated samples $e_k^{[i]}$ at time k where $E_k = \bigcup_i e_k^{[i]}$. The samples are weighted with respect to the SLAMMOT map. If the grid occupied by the sample belonging to stationary objects, the sample's weight $w_k^{[i]}$ is set to zero. If not, the weight $w_k^{[i]}$ of the sample $e_k^{[i]}$ is proportional to the probability of the motion specified by $e_k^{[i]}$ at the occupied grid.

Given the samples and their corresponding weights, the effect of the the scene interaction model is represent by the mean and covariance $(\bar{z}^{(scene)})$ of these weighted samples as shown in Figure 5. The SLAMMOT process integrates the previous *real* measurements into the stationary and moving object map. The scene interaction model uses the map to generate the *virtual* measurement about the target to predict or constrain the target's future motion. The rest of fusion is straightforward. The target state is simply updated with this virtual measurement.

With the use of the SLAMMOT map, the scene interaction model may only provide a more uncertain estimate than the prediction from the target's motion models. However, this method effectively takes the constraints/interactions from both stationary and moving objects into account. Figure 6 demonstrates the capability of tracking in a occlusion situation with the use of the scene interaction model. Figure 6(a) shows the tracking result using the IMM model. While the target is occluded, the estimate is predicted without update. The target state estimate uncertainty increases quickly and finally diverges. False data association results in failure in tracking of its surrounding objects. Figure 6(c) shows interacting object tracking with the proposed scene interaction model. The information contained in the SLAMMOT map is employed to predict the target's motion. The occluded ob-

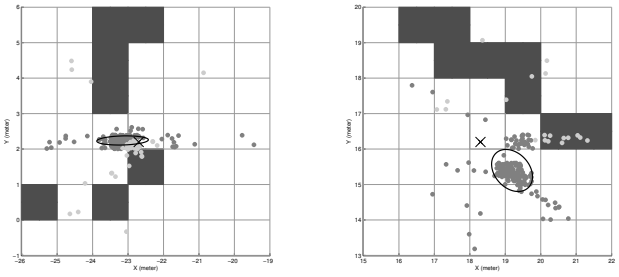
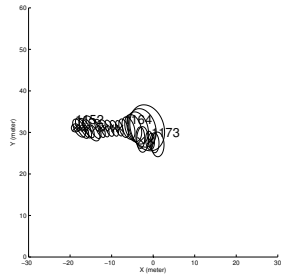
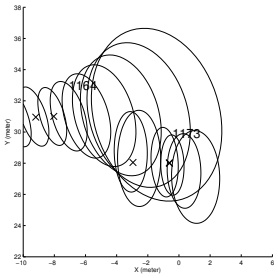


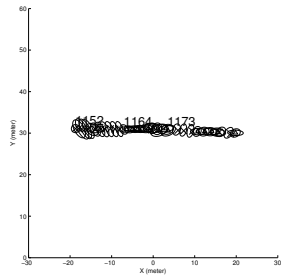
Fig. 5. Sampling-based prediction from the scene interaction model. Left: a motorcycle passing a narrow gate. Right: a car moving near a median strip.



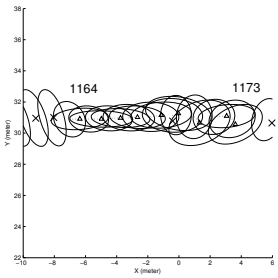
(a) Tracking using the IMM model.



(b) The enlargement of (a).



(c) Tracking with the use of the scene interaction model



(d) The enlargement of (c)

(e) Camera images at scan 1152

(f) Camera images at scan 1164

Fig. 6. The scene interaction model: $\times$s are the measurements. The distributions of the estimates are shown by 1σ ellipse.

ject's state was correctly tracked. The feasibility of tracking is evaluated using the visual images from the onboard camera images as depicted in Figure 6(e). We later will show that this approach also provides a simple way to detect abnormal events.

B. Neighboring Object Interaction Model

The neighboring object interaction model is designed to represent the short-term or immediate interactions between the target and its neighboring objects. As addressed in [5], there are several different types of short-term interactions. In this paper, we only deal with the follow interaction which frequently happens in crowded urban areas. According to the follow interaction assumption, we simply consider interactions between the target and the neighboring objects in front of the tracked object. Note that the proposed framework allows more different types of short-term interactions via multiple hypothesis tracking approaches. However, it is a challenging problem to detect with which objects the target is currently interacting.

With the use of the same virtual measurement technique proposed in the scene interaction model, the neighboring object interaction model generates a corresponding virtual measurement $\tilde{z}^{(neighbor)}$ according to the neighboring object's motion to predict/constraint the target's motion.

Given the estimate $\hat{x}_k$ of the target at time k, let $\hat{y}_k$ be the state estimate of the nearest neighboring object in front of the target. The virtual measurement from the neighboring object interaction model and its corresponding covariance can be computed as:

$$\tilde{z}_k^{j(neighbor)} = H_k^j \left(\hat{x}_k^j + \left(F_k^j - I \right) \hat{y}_k^j \right) \qquad \forall m_j \in \mathbb{Q} \tag{6}$$

$$\tilde{R}_k^{j(neighbor)} = H_k^j \left(F_k^j T_k^j F_k^{j^T} + Q_k^j \right) \qquad \forall m_j \in \mathbb{Q} \tag{7}$$

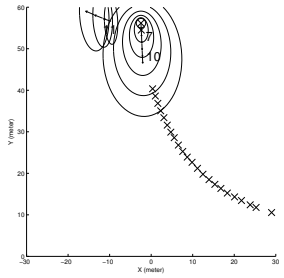
where F_k^j is the process model under the motion model m_j of the tracked target at time k, I denotes the identity matrix, H_k^j is the measurement model under the motion model m_j at time k, T_k^j is the covariance of the neighboring object $\hat{y}_k$ under the motion model m_j at time k, and Q_k^j is the motion noise model under the motion model m_j of the tracked target at time k.

The estimate of the tracked object's pose is then updated with this virtual measurement straightforwardly. Figure 7 demonstrates that tracking using the neighboring object interaction model perform well in the occlusion situation. Figure 7(a) shows the tracking result using the IMM model. While the target is occluded, the estimate is predicted without update. The target state estimate uncertainty increases quickly which results in wrong data association in this crowded scene. The track is lost in this case. Figure 7(c) shows interacting object tracking with the use of neighboring object interaction model. The motion of the target's neighboring object is applied to predict the target's motion. The occluded object's state was correctly tracked. The correctness of tracking is evaluated using the visual images from the onboard cameras, as shown in 7(e).

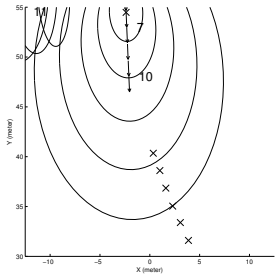
IV. Experimental Results

A couple of interacting object tracking results have been shown in the previous sections. Figures 8 demonstrates the tracking results of pedestrians, bicycles, motorcycles, cars and trucks.

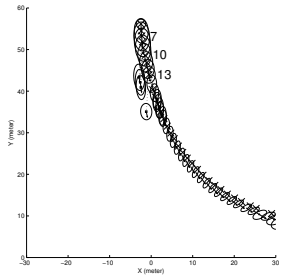
Anomalous event detection can be easily accomplished using the interaction models. Here anomalous events are defined as that objects act differently from predictions of the scene interaction model and the neighboring interaction model. Figure 9 demonstrates an example of abnormal event recognition in which a bicyclist disobeyed the traffic laws. As there is no other object around this bicycle, the neighboring object interaction model was not activated but the scene interaction model quickly showed that the bicycle's motion is very different from the prediction. The attached video shows



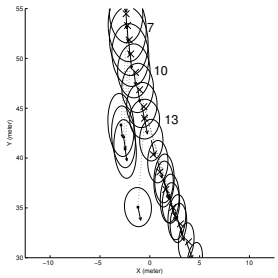
(a) Tracking using the IMM model.



(b) The enlargement of (a).



(c) Tracking with the use of the neighboring object interaction model



(d) The enlargement of (c)

(e) Camera images at scan 6

Fig. 7. The neighboring object interaction model: $\times$s are the measurements. The distributions of the estimates are shown by 1σ ellipse.

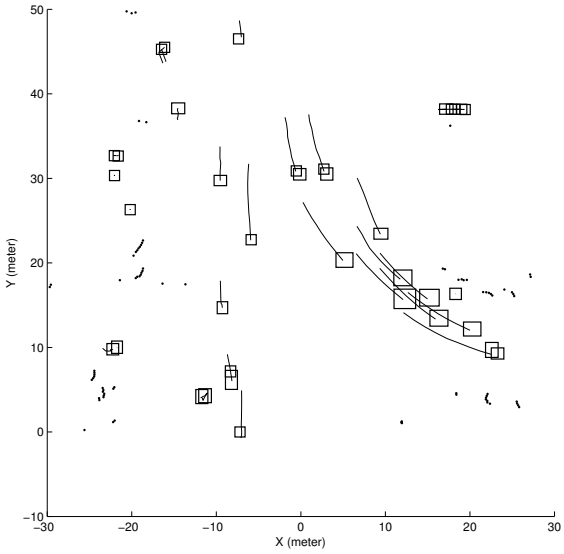


Fig. 8. Experimental Tracking results of pedestrians, bicycles, motorcycles, and cars.

the whole sequence of this abnormal event recognition and interacting object tracking results.

Another example of anomalous event detection is shown in Figure 10. A car broke down and stopped in the road. As the car was stationary, the neighboring object interaction model was not activated but the scene interaction model quickly showed that the car's motion is very different from the prediction. Traffic accidents can be easily detected by employing the proposed interaction models.

V. Conclusion

Tracking a wide variety of interacting moving objects in crowded urban areas is difficult. Based on our previous contribution to SLAMMOT, the primary contribution of this paper is to introduce the scene interaction model and the neighboring object interaction model for taking both long-term and short-term interactions into account. These interaction models and move-stop hypothesis tracking are seamlessly integrated using the variable-structure multiple model estimation framework. Fusion of these interaction models is simply accomplished using the virtual measurements in the update stage of filtering. The ample experimental results using data from a laser scanner collected at a crowded urban intersection have demonstrated the feasibility and effectiveness of the proposed algorithms.

Future work will further add more short-term interactions such as passing to deal with more complicated scenarios, and collect more data to analyze the statistical properties of the scene interaction models in different urban areas. It would be of interest to study the effectiveness of the proposed algorithms in areas with weaker interactions such as offices and homes.

VI. Acknowledgments

This work was partially supported by grants from Taiwan NSC (#94-2218-E-002-077, #94-2218-E-002-075, #95-2221-E-002-433), Quanta Computer, Australia's CSIRO, and Intel.

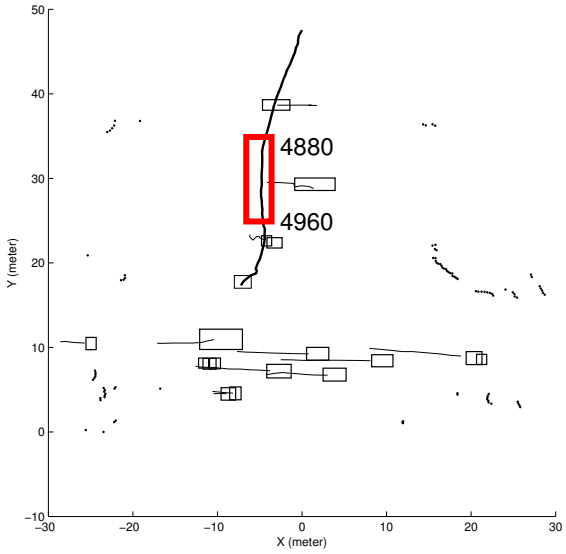
References

[1] C.-C. Wang, C. Thorpe, and S. Thrun, "Online simultaneous localization and mapping with detection and tracking of moving objects: Theory and results from a ground vehicle in crowded urban areas," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Taipei, Taiwan, September 2003.

[2] C.-C. Wang, "Simultaneous localization, mapping and moving object tracking," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 2004.

[3] H. Veeraraghavan, O. Masoud, and N. P. Papanikolopoulos, "Computer vision algorithms for intersection monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 2, pp. 78–89, June 2003.

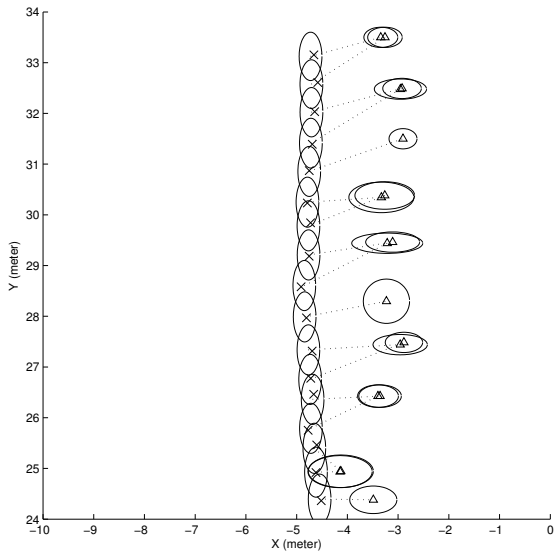
[4] H. Zhao and R. Shibasaki, "A novel system for tracking pedestrians using multiple single-row laser-range scanners," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 2, pp. 283–291, March 2005.



(a) An abnormal event: a bicyclist disobeyed the traffic lights. The bicycle is denoted by a black rectangle with a bold and longer (6.67 second) trajectory. The other moving objects are shown with 1.33 second trajectories. The bold rectangle area is enlarged in (d).

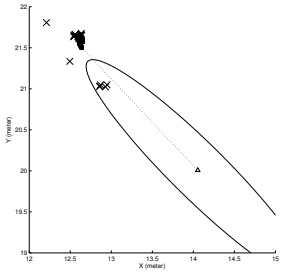
(b) Visual images at scan 4880. The bold rectangle indicates the tracked bicycle.

(c) Visual images at scan 4960. The bicycle is occluded.

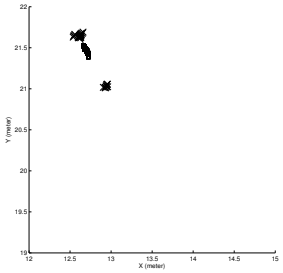


(d) Enlargement of the bold rectangle area in (a): the symbol $\times$s are the measurements, $\triangle$s indicate the virtual measurements from the scene interaction model. The inconsistency between the tracked bicycle and the predictions from the interaction model is clearly shown.

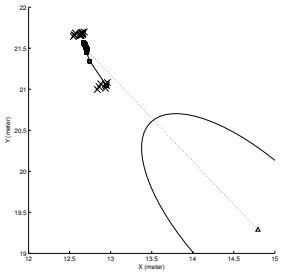
Fig. 9. Anomalous event detection



(a) The state is updated with the map pattern 1.



(b) The state is updated with the map pattern 2.



(c) The state is updated with the map pattern 3.

(d) Visual images at scan 183.

Fig. 10. Anomalous event detection: a car disobeyed the traffic law and stopped in the road. The symbol $\times$s are the measurements, $\triangle$s indicate the virtual measurements from the scene interaction model.

[5] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, August 2000.

[6] A. Panangadan, M. J. Matarić, and G. S. Sukhatme, "Detecting anomalous human interactions using laser range-finders," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE Press, Sep 2004, pp. 2136–2141.

[7] A. Bruce and G. Gordon, "Better motion prediction for people-tracking," in *IEEE International Conference on Robotics and Automation (ICRA)*, New Orleans, LA, USA, April 2004.

[8] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 27, no. 11, pp. 1805–1918, November 2005.

[9] T. Kirubarajan and Y. Bar-Shalom, "Tracking evasive move-stop-move targets with an MTI radar using a VS-IMM estimator," in *Proceedings of the SPIE Signal and Data Processing of Small Targets*, vol. 4048, 2000, pp. 236–246.

[10] X. R. Li, X. Zwi, and Y. Zhang, "Multiple-model estimation with variable structure. iii. model-groupswitching algorithm," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 35, no. 1, pp. 225–241, January 1999.

[11] X.-R. Li and Y. Bar-Shalom, "Multiple-model estimation with variable structure," *IEEE Transactions on Automatic Control*, vol. 41, no. 4, pp. 478–493, April 1996.

[12] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. part v. multiple-model methods," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1255– 1321, October 2005.

[13] T. Kirubarajan, Y. Bar-Shalom, K. Pattipati, I. Kadar, E. Eadan, and B. Abrams, "Tracking ground targets with road constraints using an IMM estimator," in *Proceedings of the IEEE Aerospace Conference*, Snowmass, CO, March 1998.

[14] H. A. P. Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with markovian switching coefficients," *IEEE Trans. On Automatic Control*, vol. 33, no. 8, Aug. 1988.