2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

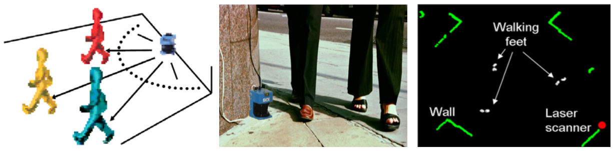
Fusion of Laser and Vision for Multiple Targets Tracking via On-line Learning

Xuan Song[†], Huijing Zhao[†], Jinshi Cui[†], Xiaowei Shao[‡], Ryosuke Shibasaki[‡] and Hongbin Zha[†]

Abstract— **Multi-target tracking becomes significantly more challenging when the targets are in close proximity or frequently interact with each other. This paper presents a promising tracking system to deal with these problems. The novelty of this system is that laser and vision, tracking and learning are integrated and can complement each other in one framework: when the targets do not interact with each other, the laser-based independent trackers are employed and the visual information is extracted simultaneously to train some classifiers for the "possible interacting targets". When the targets are in close proximity, the learned classifiers and visual information are used to assist in tracking. Therefore, this mode of co-operation between them not only deals with various tough problems encountered in the tracking, but also ensures that the entire process can be completely on-line and automatic. Experimental results demonstrated that laser and vision fully display their respective advantages in our system, and it is easy for us to obtain a perfect trade-off between tracking accuracy and time-cost.**

I. Introduction

A robust and efficient multi-target tracking system has become an urgent need in various application domains, such as surveillance, pedestrians flow analysis, intelligent transportation and many others. Compared to the traditional vision-based tracking system, as a new kind of measurement instrument, the laser range scanner has received the increasing attention for solving tracking problems in recent years. In a laser-based tracking system (as shown in Fig.1), the targets are represented by several points, hence the tracking become much easy and it is easy to obtain a much better performance in both accuracy and time-cost when the targets are in far apart. The system [1], [2] has been successfully applied into the JR subway station of Tokyo for the pedestrians flow analysis and reached the 83% accuracy overall.



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

Fig. 1. A typical laser-based tracking system.

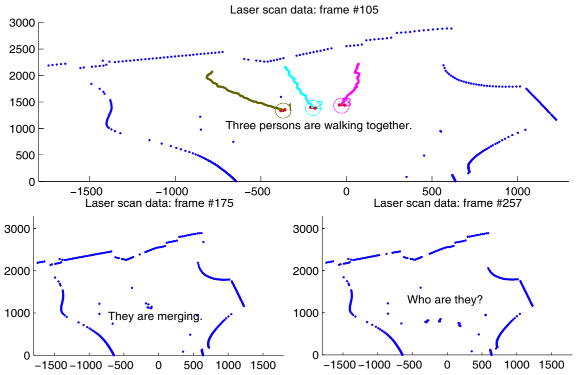


Fig. 2. How can the persons' correct trajectories be maintained under this condition? In frame 105, three persons were walking together. They were merging in frame 175, how could their correct trajectories be maintained when they split?

However, the drawback of a laser-based tracking system is inherent and obvious: it lacks visual information, consequently it is difficult to obtain a set of features that uniquely distinguish one object from another. Hence, when the targets are in close proximity or frequently interact with each other, performing the robust tracking becomes specially challenging. Moreover, when the well-known "merge/split" condition occurs (as shown in Fig.2), maintaining the correct tracking seems to be an impossible mission. It is easy to think of fusing the laser and vision into one framework to solve these problems. Therefore, the core concerns of this research are: (1) How to make the laser and vision fully display their respective advantages in one framework to solve the tough problems encountered in multi-target tracking? (2) How to develop a tracking system that can obtain a perfect trade-off between tracking accuracy and time-cost?

In this paper, we integrate laser and vision, tracking and learning and make them complement each other in one framework to deal with various tracking problems. The key idea of this work can be depicted in Fig.3 and Fig.4. 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 co-operation between laser and vision,

[†]Xuan Song, Huijing Zhao, Jinshi Cui and Hongbin Zha are with the Key Laboratory of Machine Perception (MoE), Peking University, China. E-mail: {songxuan,zhaohj,cjs,zha}@cis.pku.edu.cn.

[‡]Xiaowei Shao and Ryosuke Shibasaki are with the Center for Spatial Information Science, University of Tokyo, Japan. Email: shaoxw@iis.u-tokyo.ac.jp, shiba@csis.u-tokyo.ac.jp.

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

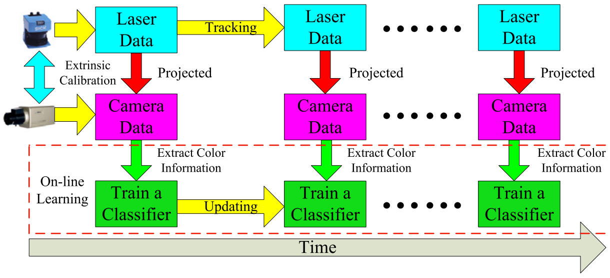


Fig. 3. **Tracking for Learning.** When the targets do not interact with each other, the laser-based independent trackers should be employed and the visual information is extracted simultaneously to perform the on-line learning.

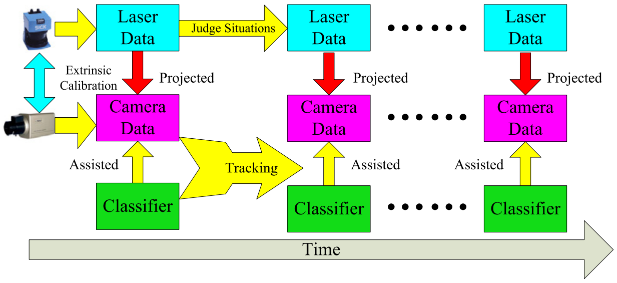


Fig. 4. **Learning for Tracking.** When the targets are in close proximity, the learned classifiers and visual information in turn assist in tracking.

tracking and learning offers several advantages: (1) Laser and vision can fully display their respective advantages (fast of laser and rich information of camera) in this system. (2) Because the "possible interacting targets" are depicted by a discriminative model with a supervised learning process, this model can consider the 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.

II. Related Work

In recent years, several tracking system based on the combination of laser and camera have existed [3], [4], [5], [6], [7], [8], [9]. However, all the above systems can only detect and track one or few persons, which cannot be applied in a real surveillance and monitoring environment. (The system [9] and [6] also utilize the classifiers to assist in tracking, but please note that: their classifiers are the pre-trained detectors and cannot be updated in the tracking, which is completely off-line learning.)

On the other hand, the system proposed by [10], [11] can be utilized to track more targets in the outdoor environment. But once some challenging situations (such as "merge/split") occur, their systems are difficult to maintain the correct tracking. To our knowledge, the proposed system is the only one system that can be used in the wide and open area and robustly track more than 15 targets in the same time as yet.

III. System Overview

The overall tracking system is illustrated in Fig.5. Our tracking system consists four components: Sensor Fusion, Tracking Situation Switch, Non-correlated Targets Tracking and Interacting Targets Tracking.

In the Sensor Fusion part, we utilized a time server to deal with time synchronization problem between different sensors. The extrinsic calibration was conducted by several control points in a box. For details about this part, please refer [11]. On the other hand, in order to switch tracking and learning, we should detect different tracking situations, such as non-correlated targets tracking and interacting targets tracking (correlated targets or merge/split condition). For details about this part, please refer [12].

The two main components that will be described in this paper are the Non-correlated Targets Tracking (tracking for learning) and Interacting Targets Tracking (learning for tracking). In the next two sections, we will provide the details about how the laser and vision, tracking and learning complement each other in one framework.

IV. Tracking for Learning

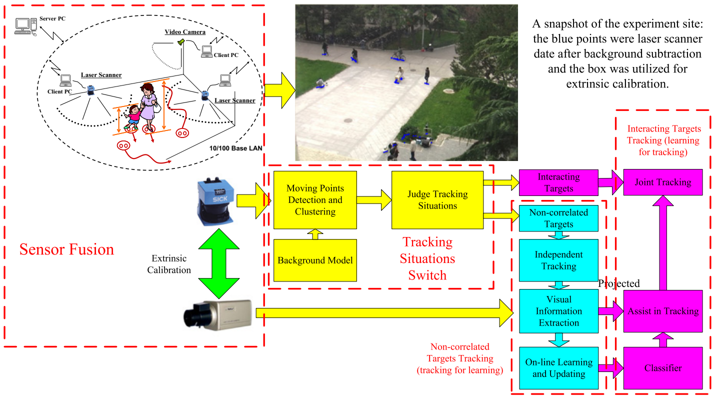
Actually, when the targets are far apart (non-correlated targets), tracking becomes relatively easy since it can be solved through multiple independent trackers. Moreover, laser scanner is insensitive to weather condition and the data is easy to process, therefore it can provide a robust and efficient tracking in this situation. Once we obtain the tracking results, the visual information from camera for this target can be extracted as samples to train a classifier. The positive samples should be from this targets and the negative one are from the "possible confusing targets". Consequently the classifier of each target is a discriminative model that not only depicted the appearance of target, but also consider the information from other targets. An example is shown in Fig.6. In this section, we will provide the details about this.

A. Independent Tracker and Visual Information Extraction

We employed multiple independent particle filter-based trackers to perform the laser-based tracking. We utilized a "two feet walking model" as proposal distribution, and the observation model was similarity between laser point sets and predicted walking style of pedestrians. Please refer [2] for details about this part.

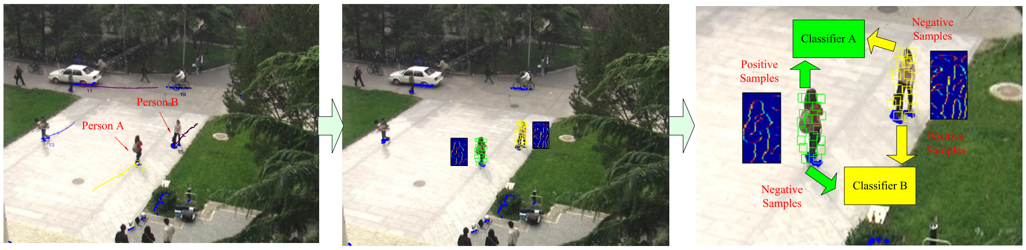
Once we obtained the tracking results of each target, they were projected to the image coordinate (details about extrinsic calibration, please refer [11]), and then a set of random image patches [13] were spatially sampled within the image region of the target. We utilized these random image patches as samples for the on-line supervised learning. In this case, each target was represented by a "bag of patches" model.

Extracting some distinguishable features from the image patches is relatively important for the learning process. There have been a large number of derived features that can be employed to represent the appearance of an image patch,


Server PC
Video Camera
Client PC
Laser Scanner
Client PC
Laser Scanner
10/100 Base LAN
Sensor Fusion
Extrinsic Calibration
Moving Points Detection and Clustering
Background Model
Judge Tracking Situations
Tracking Situations Switch
Interacting Targets
Non-correlated Targets
Independent Tracking
Visual Information Extraction
On-line Learning and Updating
Non-correlated Targets Tracking (tracking for learning)
Interacting Targets Tracking (learning for tracking)
Joint Tracking
Projected
Assist in Tracking
Classifier

A snapshot of the experiment site: the blue points were laser scanner date after background subtraction and the box was utilized for extrinsic calibration.

Fig. 5. The overview of our tracking system.



(a) Target A and B at #8954

(b) Random image patches and visual information extraction

(c) On-line supervised learning

Fig. 6. **Non-correlated targets tracking and the on-line supervised learning.** We employed the laser-based independent trackers to track target A and B and detected that they should be "possible interacting targets" in frame 8954 (Fig.a). Some random image patches were sampling in them and the visual information should be extracted (Fig.b). These image patches were used as positive or negative samples to train the classifier A and B (Fig.c).

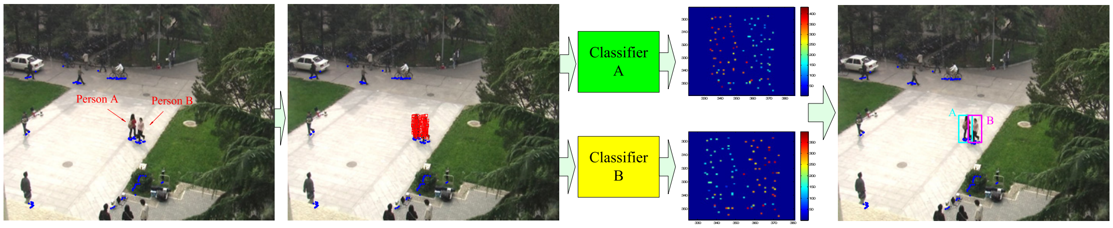
such as color information (RGB intensity vector), texture information (Haralick feature), edge information (edge orientation histogram) and so on. Since we utilized image patches as samples, the feature vector should contain local information. On the other hand, by conducting some experiments, we found that the distinguish-ability of texture was not good for our data. Hence, we employed the RGB intensity + edge orientation histogram (EOH) [14] to extract features from image patches (as shown in Fig.6-b). We adapted an d-dimensional feature vector to represent each image patch. Therefore, these feature vectors can be utilized as samples for the learning or testing.

B. On-line Learning and Updating

For the "possible interacting targets", the strong classifiers should be trained, which represent the appearance of targets. Let each image patch be represented as a d-dimensional feature vector. For target k in frame t, $\{\mathbf{s}^i_{t,k}, l^i_{t,k}\}_{i=1}^N$ denote N samples and their labels, where $\mathbf{s} \in \Re^d$ and $l \in \{-1, +1\}$. The positive samples are the image patches come from region of target k, while the negative samples are the image patches that come from some "possible confusing targets". In this work, we employed Classification and Regression Trees [15] as weak classifiers. Once the new samples are available, the strong classifier should update synchronously, which would make the classifier stronger and reflect the changes

```
Learning Algorithm
Input: Feature vectors of image patches and their labels $\{\mathbf{s}^i_{t,k}, l^i_{t,k}\}_{i=1}^N$, $t = 1, ..., T$
Output: The strong classifier $H(\mathbf{s}_{t,k})$ of target $k$ at time $t$
Train a Strong Classifier (for frame 1)
  1) Initialize weights $\{w_i\}_{i=1}^N$ to be $1/N$ .
  2) For $j = 1...M$ (train $M$ weak classifiers)
      a) Make $\{w_i\}_{i=1}^N$ a distribution.
      b) Train a weak classifier $h_j$.
      c) Set $err = \sum_{i=1}^N w_i |h(\mathbf{s}^i_{1,k}) - l^i_{1,k}|$.
      d) Set weak classifier weight $\alpha_j = 0.5 \log(1 - err)/err$.
      e) Update example weights $w_i = w_i e^{\alpha_j |h_j(\mathbf{s}^i_{1,k}) - l^i_{1,k}|}$
  3) The strong classifier is given by $sign(H(\mathbf{s}_{1,k}))$, where
     $H(\mathbf{s}_{1,k}) = \sum_{j=1}^M \alpha_j h_j(\mathbf{s}_{1,k})$.
Update the Strong Classifier (for new frame $t$ is coming in)
  1) Initialize weights $\{w_i\}_{i=1}^N$ to be $1/N$ .
  2) For $j = 1...K$ (choose $K$ best weak classifiers and update their weights)
      a) Make $\{w_i\}_{i=1}^N$ a distribution.
      b) Choose $h_j(\mathbf{s}_{t-1,k})$ with minimal $err$ from
         $\{h_1(\mathbf{s}_{t-1,k}), ...h_M(\mathbf{s}_{t-1,k})\}$.
      c) Update $\alpha_j$ and $\{w_i\}_{i=1}^N$.
      d) Remove $h_j(\mathbf{s}_{t-1,k})$ from $\{h_1(\mathbf{s}_{t-1,k}), ...h_j(\mathbf{s}_{t-1,k})\}$.
  3) For $j = K + 1...M$ (add new weak classifiers)
      a) Make $\{w_i\}_{i=1}^N$ a distribution.
      b) Train a weak classifier $h_j$.
      c) Compute $err$ and $\alpha_j$ .
      d) Update examples weights $\{w_i\}_{i=1}^N$.
  4) The updated strong classifiers is given by $sign(H(\mathbf{s}_{t,k}))$ , where
     $H(\mathbf{s}_{t,k}) = \sum_{j=1}^M \alpha_j h_j(\mathbf{s}_{t,k})$.
```

Fig. 7. On-line learning algorithm



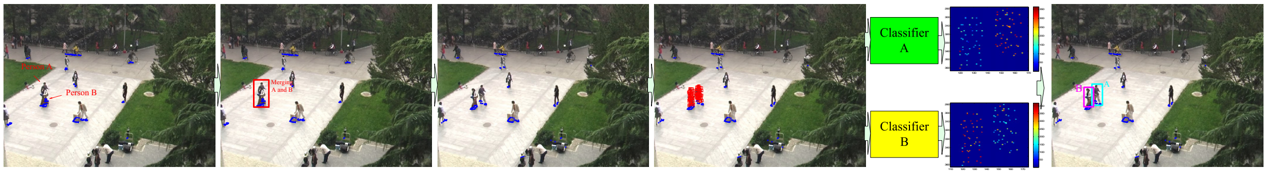
(a) Target A and B were in close proximity

(b) Random image patches were sampling in the interacted region

(c) Score maps

(d) Tracking results of A and B

Fig. 8. **Correlated targets tracking.** We detected that target A and B were correlated targets (Fig.a). Some random image patches were sampling in them (Fig.b). We used their classifiers to obtain the score maps (Fig.c). After the particle filtering process, we acquired their tracking results (Fig.d).



(a) Two targets were merging at #5427

(b) We tracked them as one target

(c) 55 frames after, they spit

(d) Image patches were sampling in them

(e) Score maps

(f) Obtained tracking results

Fig. 9. **Merge/Split condition.** In frame 5427, we detected that A and B were merging (Fig.a); we track A and B as one target (Fig.b). After 55 frames, we detected that they split (Fig.c), and some random image patches were sampling in them (Fig.d). We used their classifiers to obtain their score maps (Fig. e). After the particle filtering process, we obtained the tracking results (Fig.f).

in the object appearance. Therefore, poor weak classifiers are removed and newly trained classifiers are added, which is motivated by *Ensemble Tracking* [16]. The whole learning algorithm is shown in Fig.7.

V. Learning for Tracking

When the targets are in close proximity or interact with each other, laser-based independent tracker is difficult to maintain the correct tracking. Specifically, when the "merge/split" conditions occur, associating the identities of the targets becomes a significantly challenging problem. In this case, the visual information and the learned classifiers should help us to deal with them. In this section, we will provide details about how these classifiers assisted in tracking to deal with some challenging situations encountered in the tracking.

A. Correlated Targets Tracking

Once we detected that some targets were in close proximity from the laser data, we concluded that they were correlated targets. When this condition occurred, a set of random image patches were sampled within the interacting region on image, and the feature vectors of these image patches were imputed to the classifiers of interacting targets respectively. The outputs of these classifiers are scores. Hence, we could obtain the score maps of these interacting targets effortlessly.

Once we obtained the score maps of the interacting targets, we employed the particle filter technique [17] to obtain the positions of these targets. The likelihood for updating in the particle filter was

$$P_{scores}(\mathbf{y}_t|\mathbf{x}_{t,k}) = \frac{1}{\sqrt{2\pi/\sigma}} \sum_{i=1}^{N} \beta_i \exp\left(\frac{(\mathbf{d}(\mathbf{x}_{t,k}) - \mathbf{d}^i_{t,k})^2}{\sigma^2}\right) \tag{1}$$

where β_i was the normalized score of image patch i, $\mathbf{d}(\mathbf{x}_{t,k})$ the center position of candidate target k, $\mathbf{d}^i_{t,k}$ the center position of image patch i, and σ was the covariance which depended on the size of the image patch. For each target, the observation was further peaked around its real position. As a result the particles were much focused around the true target state after each level's re-weighting and re-sampling. Subsequently, we obtained the new position of these interacting targets. The overview of the process is shown in Fig.8.

B. Merge/Split Condition

Sometimes, it is difficult to obtain the separate detections by the laser-based clustering algorithm. Moreover, the targets in image may occlude each other completely. Hence, the solution described above is not available. Once we detected that such situation occurred, we dealt with it as a "merge/split" condition.

If some targets were merging together, we initialized the state of the "merging targets" and tracked it as one target. If we detected that this "merging target" split and became an interacting condition or non-correlated condition, we utilized the classifiers of these targets to identify them (as shown in Fig.9). Hence, we can link the trajectories of these targets without difficulty.

With the help of the classifiers and visual information, our method is able to deal with various complex situations in the tracking. In addition, the tracking and learning, laser

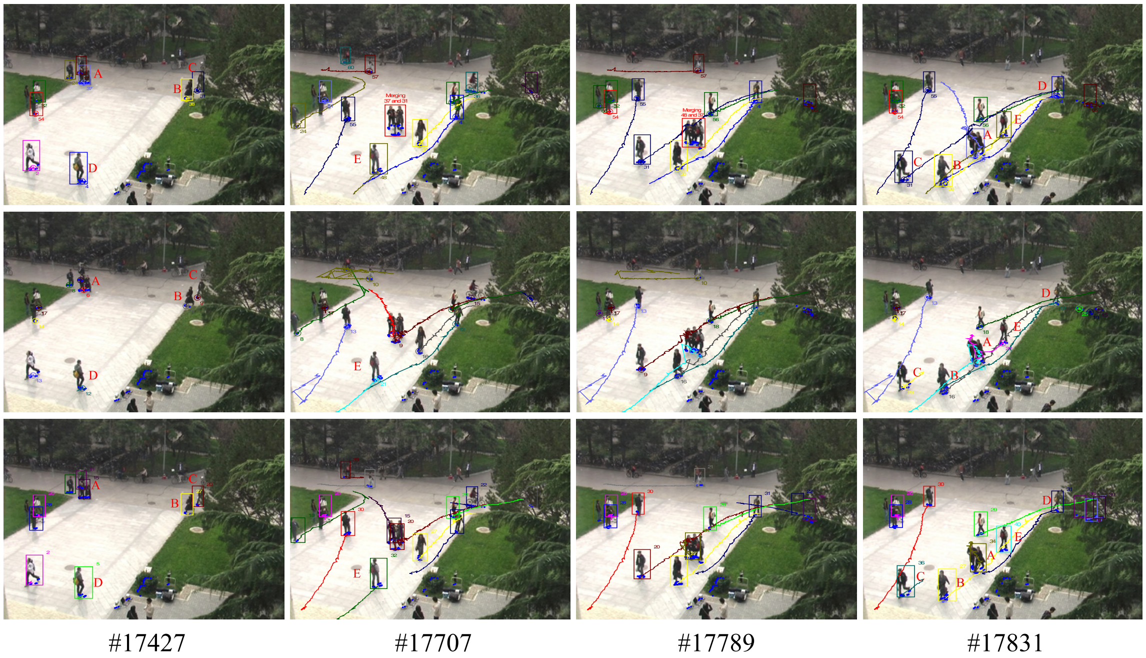


Fig. 10. **Tracking results.** The first row is the results of our system, the second and the third are the results of laser only and camera only respectively. Please note the key target A, B, C, D and E (E appeared in 17707). Target A and C, A and E were merging in frame 17707 and 17789 respectively. 42 frames after (17831), only our tracking system maintained their correct trajectories.

and vision complement each other in the proposed method, consequently becoming an adaptive loop, which ensures that the entire process can be completely on-line and automatic.

VI. Experiments and Results

We evaluated our tracking system in the real scene at a library. In order to deal with the occlusions among targets, we utilized two single-row laser scanners. In addition, one camera was set on the third floor of the building. We tested our system with 30-minute long data (about 45000 frames). When the distance between targets was less than 1.5 meters, we started to train the classifiers, and once the distance between them was less than 0.3 meters, we considered it as the challenging situations. In this section, we will present our tracking results and the perform some quantitative evaluations.

A. Tracking Results

Fig.10 shows an example of the tracking results. The first row is our tracking results, the second is the results with laser only, and the third is the results with camera only (we utilized PD-PF [18] to perform the visual tracking). We found that once the interactions occurred, laser-based independent trackers frequently made the false labeling. On the other hand, the vision-based trackers were difficult to deal with "merge/split" problem. Please note the key target A, B, C, D, and E. Target A and C, A and E were merging in frame 17707 and 17789 respectively, 42 frames after, only our tracking system maintained their correct trajectories.

TABLE I
Disposal of Challenging Situations

	Correlated Targets		Merge/Split	
	Total/Disposal	Disposal Rate	Total/Disposal	Disposal Rate
Our System	3832/3396	**88.62**%	537/439	**81.75**%
Laser Only	3832/1821	47.52%	537/145	27.00%
Vision Only	3832/2566	66.96%	537/189	35.20%

We selected 6000 continuous frames which interactions frequently took place and made a statistical survey about how many challenging situations (such as correlated targets or "merge/split" condition) we could deal with. Once one of these conditions occurred, but no failed tracking was caused by it, a successful disposal was counted. The details for this are shown in Table 1. From this table, we can see that our tracking system can deal with most correlated targets or merge/split conditions, which is difficult for the laser or vision only.

B. Performance Evaluation

We evaluated the performance of our tracking system from two aspects: tracking accuracy and time-cost (cpu computational time). We made a statistical survey of 3000 continuous frames and the details of them are shown in Fig.11. Please note that the ground truth was obtained by a semi-automatic way (trackers+manual labeling). By comparing the tracking results with ground truth, it was easy for us to recognize different failed tracking situations, including target missed, false location and identity switch. In addition, the time-cost was normalized into 0 to 1. Moreover, once we started to train the classifiers, the interactions should be counted in

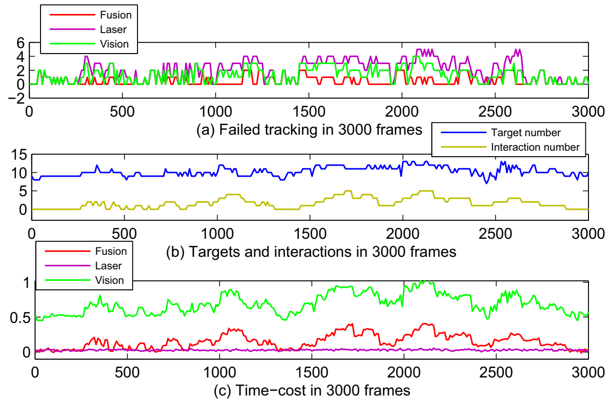


Fig. 11. **Performance evaluation.** (a) shows the failed tracking of three systems in these frames. (b) shows the target and interaction number in 3000 frames, and (c) shows the time-cost of three systems per frame (Note that the time-cost is normalized into 0 to 1).

TABLE II
OVERALL TRACKING PERFORMANCE

	Average Accuracy	Average Time-cost
Our System	**89.53%**	**0.1511**
Laser Only	73.37%	0.0296
Vision Only	80.95%	0.6906

this experiment.

From Fig.11, we can see that when there was no interaction (from 1-273 frames), the time-cost of our system was same to the laser-based one and much more efficient than the vision-based. When the interactions occurred, although our system needed some extra computation, it could deal with most interactions which was difficult to the laser-based one. In addition, the time-cost of our system was still less than the vision-based because only a limited targets needed the visual processing. The overall tracking accuracy and time-cost are shown in Table 2. In the condition of average 10.26 targets per frame, 16.35% interactions, our system obtained the highest tracking accuracy and was much faster than the vision-based. Actually, with the increasing number of tracking targets and interactions, our system can obtain a much better performance than the other two systems.

VII. Conclusion

In this paper, we present a novel multi-target tracking system using laser and vision in the open and wide area. By combining the laser and vision, tracking and learning, the proposed system can easily deal with various challenging situations in the tracking. In addition, experimental results show that our system fully incorporates the respective advantages of the two sensors and obtain a superior performance in both tracking accuracy and time-cost.

However, we found that sometimes we should face different targets (such as pedestrians, bicycles, cars and etc.) that have quite different dynamic models. For the present system, we utilized the same tracking model for them, which decreases our tracking performance. In the future, a classification module will be added to deal with this problem.

VIII. Acknowledgments

This work was supported in part by the NSFC Grant (No.60605001), the NHTRDP 863 Grant (No.2007AA11Z225) and NKBRPC Grant (No. 2006CB303100). We specially thank High-Level University Building Program of China for their supporting.

References

[1] 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*, pp. 283–291, 2005.

[2] X. Shao, H. Zhao, K. Nakamura, K. Katabira, R. Shibasaki, and Y. Nakagawa, "Detection and tracking of multiple pedestrians by using laser range scanners," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2174–2179, 2007.

[3] K. Arras, N. Tomatis, and R. Siegwart, "Multisensor on-the-fly localization using laser and vision," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 462–467, 2000.

[4] P. Chakravarty and R. Jarvis, "Panoramic vision and laser range finder fusion for multiple person tracking," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2949–2954, 2006.

[5] M. Scheutz, J. McRaven, and G. Cserey, "Fast, reliable, adaptive bimodal people tracking for indoor environments," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1347–1352, 2004.

[6] J. Blanco, W. Burgard, R.Sanz, and J. L. Fernandez, "Fast face detection for mobile robots by integrating laser range data with vision," *Proc. IEEE International Conference on Robotics and Automation*, pp. 625–631, 2003.

[7] D. Schulz, "A probabilistic exemplar approach to combine laser and vision for person tracking," *Proc. Robotics: Science and Systems Conference*, pp. 362–367, 2006.

[8] N. Bellotto and H. Hu, "Vision and laser data fusion for tracking people with a mobile robot," *Proc. IEEE International Conference on Robotics and Biomimetics*, pp. 7–12, 2006.

[9] G. Cielniak, T. Duckett, and A. J. Lilienthal, "Improved data association and occlusion handling for vision-based people tracking by mobile robots," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3436–3441, 2007.

[10] X. Song, J. Cui, H. Zhao, and H. Zha, "Bayesian fusion of laser and vision for multiple people detection and tracking," *Proc. of IEEE International Conference on Instrumentation, Control and Information Technology*, pp. 3014–3019, 2008.

[11] J. Cui, H. Zha, H. Zhao, and R. Shibasaki, "Multi-modal tracking of people using laser scanners and video camera," *Image and Vision Computing*, pp. 240–252, 2008.

[12] X. Song, J. Cui, X. Wang, H. Zhao, and H. Zha, "Tracking interacting targets with laser scanner via on-line supervised learning," *Proc. IEEE International Conference on Robotics and Automation*, pp. 2271–2276, 2008.

[13] L. Lu and G. Hager, "Dynamic foreground/background extraction from images and videos using random patches," *Proc. Neural Information Processing Systems*, pp. 351–358, 2006.

[14] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 53–60, 2004.

[15] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," *Chapman Hall, New York: Wadsworth*, 1984.

[16] S. Avidan, "Ensemble tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261–271, 2007.

[17] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 1, pp. 197–208, 2000.

[18] X. Song, J. Cui, H. Zha, and H. Zhao, "Probabilistic detection-based particle filter for multi-target tracking," *Proc. of British Machine Vision Conference*, pp. 223–232, 2008.