Semantic Visual Localization and Mapping Based on Deep Learning in Dynamic Environment

Linhui Xiao*, Jinge Wang, Xiaosong Qiu, Zheng Rong, Xudong Zou

（State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Science, Beijing, China）

***xiaolinhui16@mails.ucas.ac.cn**

INTRODUCTION

By taking advantages of deep learning in object detection, a feature-based visual SLAM system is constructed, which processes the feature points of dynamic objects through a selective tracking algorithm in the tracking thread, thereby significantly reducing the error of pose estimation caused by incorrect matching in dynamic environment. Experiments verified that Dynamic-SLAM has excellent accuracy and robustness in robot localization and mapping.

METHODS

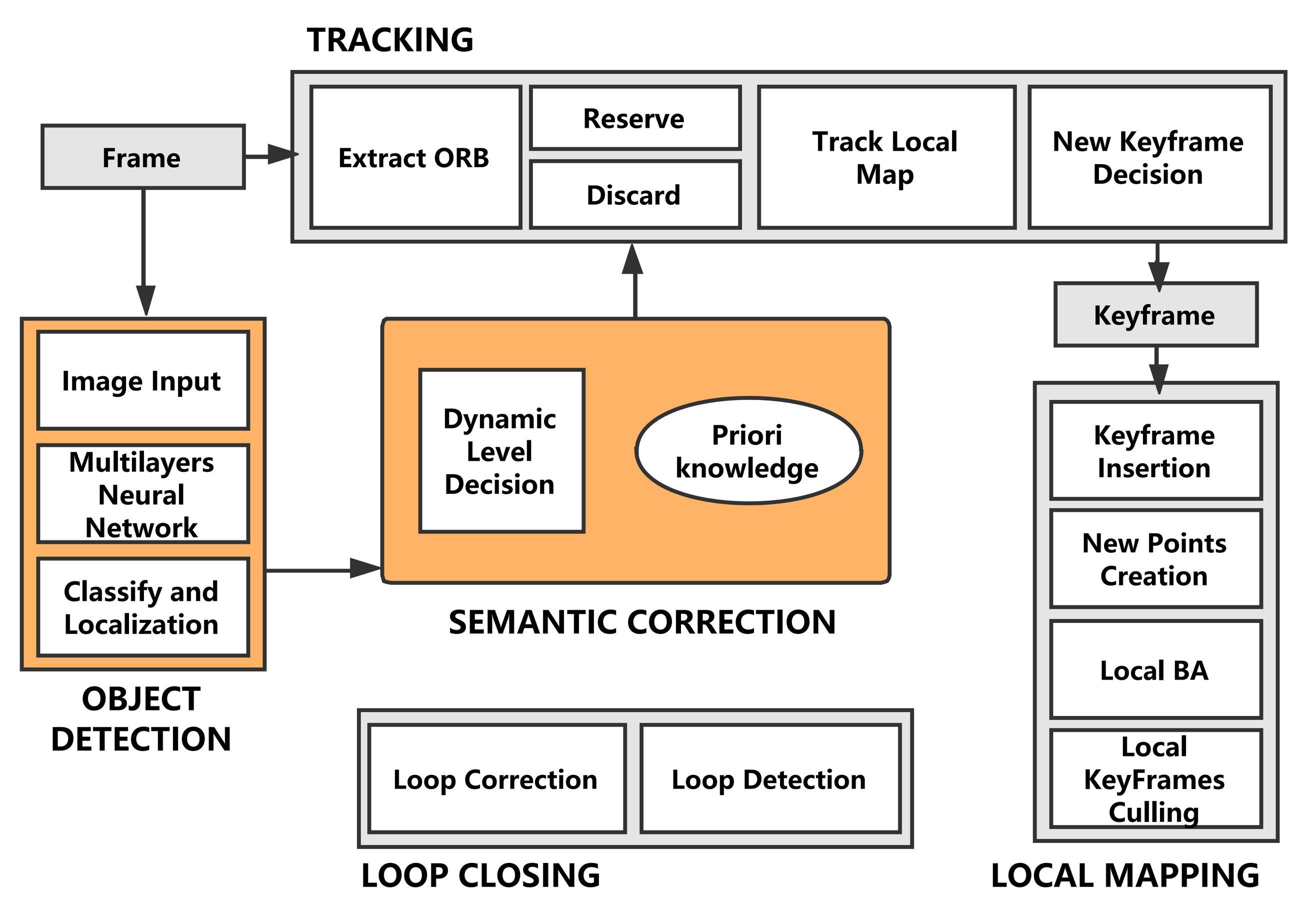

Fig.1 Dynamic-SLAM system overview.

The minimum reprojection error:

$$\xi^{*} = \arg\min_{\xi} \frac{1}{2}\sum_{i=1}^{n}\left\| u_i - \frac{1}{S_i} K \exp(\xi^{\Lambda}) P_i \right\|_2^2$$

Selective tracking algorithm:

$$\forall D_i,\quad \left| p_{D_i,K}(u,v) - p_{D_i,K-1}(u,v) \right| \le S(u,v)$$

$$S(u,v) = K \frac{1}{N_L} \sum_{i \in L} \left| \frac{1}{Z_{S_i}} K \exp(\xi_K^{\wedge}) P_{S_i} - \frac{1}{Z_{S_i}} K \exp(\xi_{K-1}^{\wedge}) P_{S_i} \right|$$

Multi-class logistic loss function:

$$L_{conf}(x,c) = -\sum_{i,j,p} x_{ij}^{p} \log(c_i^p) - \sum_{i,p} (1 - \sum_{j,q=p} x_{ij}^{q}) \log(1 - c_i^p)$$

RESULTS

Experiments show that the recall rate of the system is increased from 82.3% to 99.8% compared with the original SSD network. In TUM indoor dynamic dataset, the localization accuracy of Dynamic-SLAM is higher than the state-of-the-art systems. In the KITTI outdoor large-scale dynamic environment, the overall performance is better than state-of-the-art ORB-SLAM2. The system successfully localizes and constructs an accurate environmental map in real dynamic environment of mobile robot, whereas ORB-SLAM2 fails.

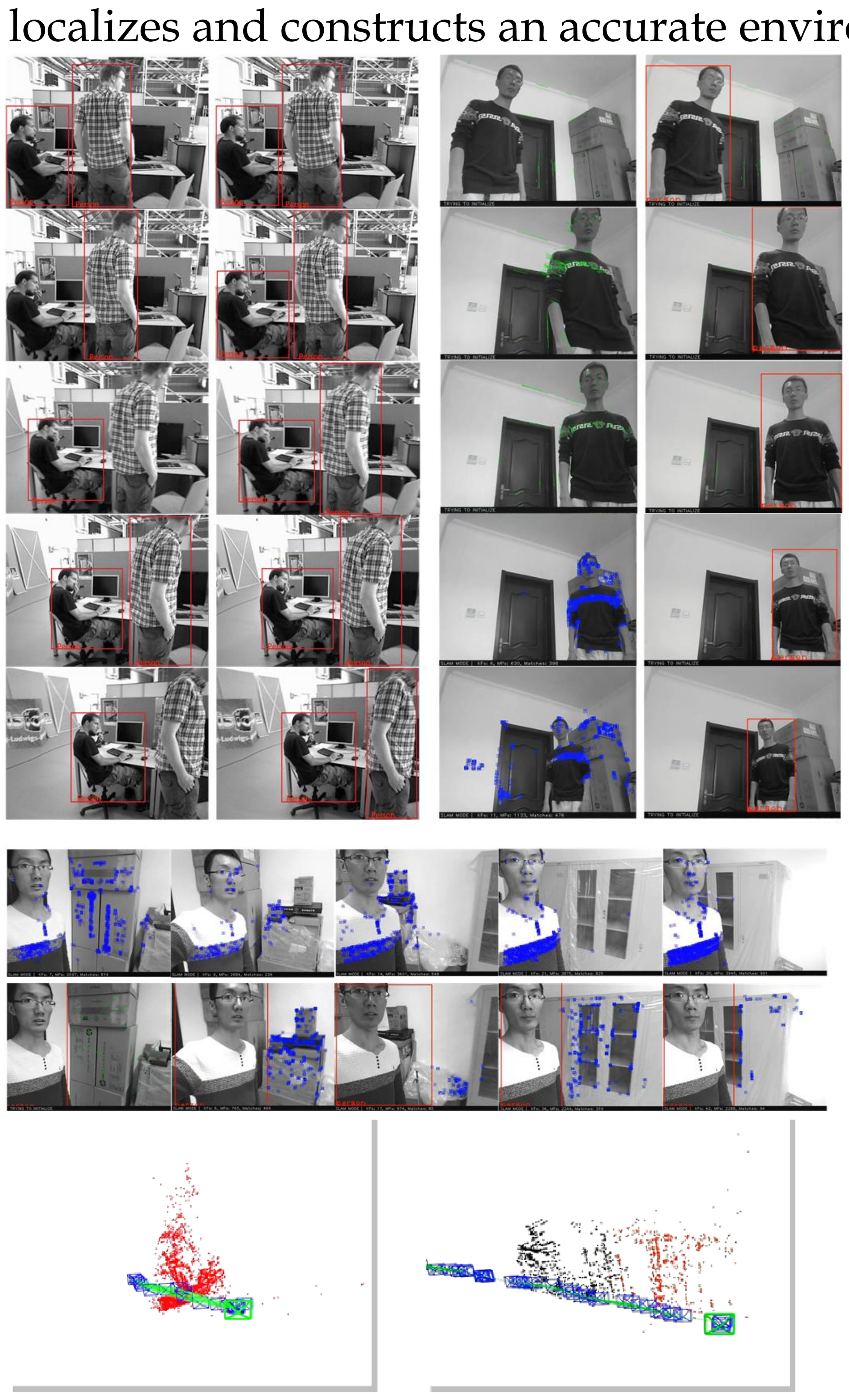

Fig. 2 A series of test experiments.

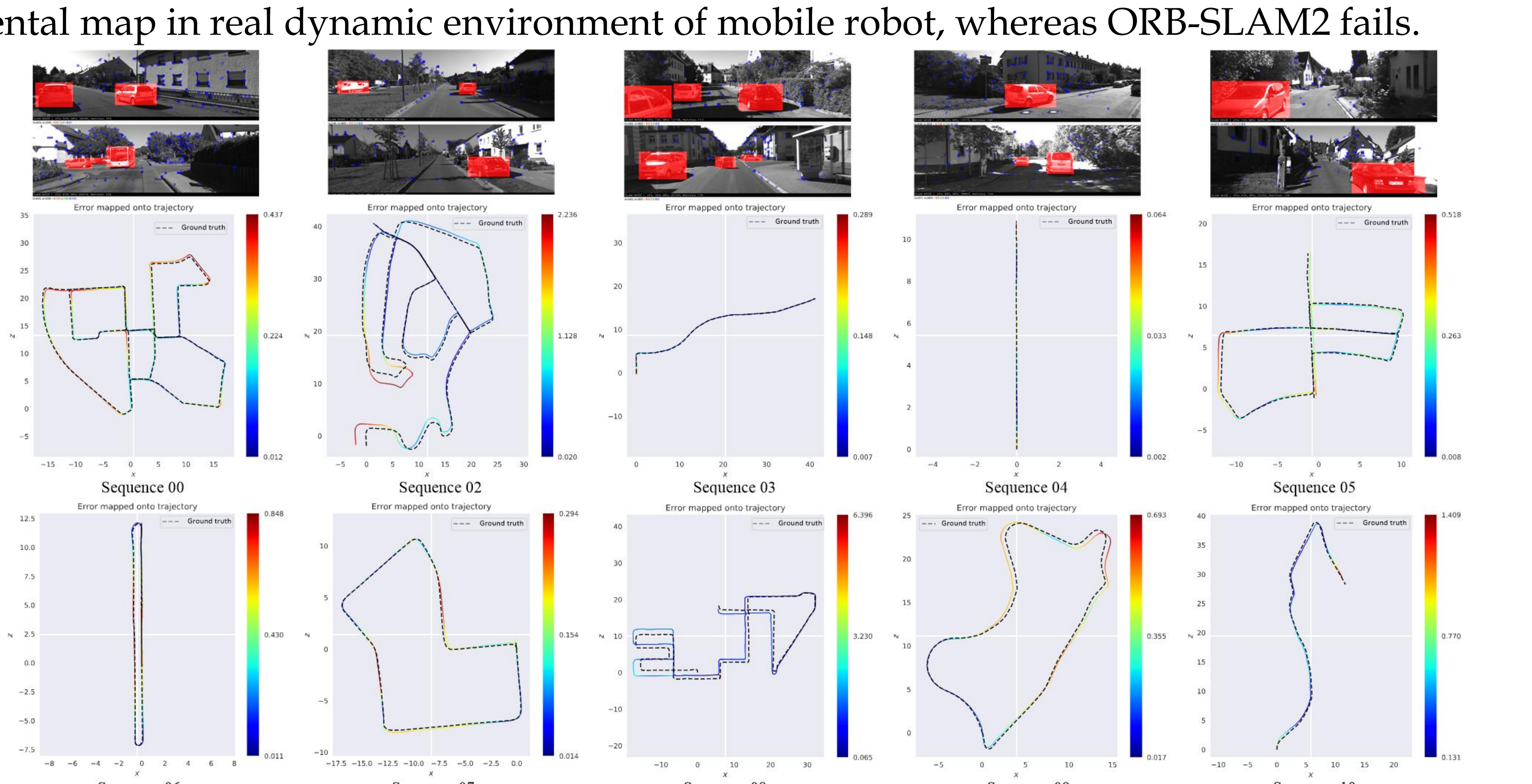

TABLE 1

TUM DYNAMIC DATASET. COMPARISON OF LOCALIZATION ACCURACY

Dynamic Sequence	Length (m)	RMSE (cm)				
		PTAM	LSD -SLAM	SVO (+BA)	ORB-SLAM2 (Mono)	Dynamic-SLAM
fr2_desk_ps	17.044	×	31.73	17.14	4.85	**1.87**
fr3_sit_xyz	5.496	0.83	7.73	6.03	0.60	**0.56**
fr3_sit_half	6.503	×	5.87	12.06	1.99	**1.88**
fr3_sit_rpy	1.110	-	-	7.89	**1.84**	2.45
fr3_walk_xyz	5.791	×	12.44	9.30	2.17	**1.68**
fr3_walk_half	7.686	×	×	11.25	**2.14**	2.71
fr3_walk_rpy	2.698	-	-	18.91	6.53	**4.03**

TABLE 2

COMPARISON OF ACCURACY IN THE KITTI DATASET

Sequence	Dimension (m × m)	RMSE (m)			Scale	Improve (%)
		ORB-SLAM (Monocular)	ORB-SLAM2 (Mono)	Dynamic-SLAM		
00	564×496	6.68	**4.44**	4.83	17.16	-8.78
01	1157×1827	×	×	×	×	×
02	599×946	21.75	20.37	**20.01**	21.71	1.77
03	471×199	1.59	1.08	**0.83**	11.42	23.15
04	0.5×394	1.79	1.15	**1.13**	26.63	1.74
05	479×426	8.23	5.73	**5.62**	21.70	1.92
06	23×457	14.68	14.25	**12.13**	21.18	14.88
07	191×209	3.36	1.82	**1.76**	10.65	3.30
08	808×391	46.58	30.29	**27.49**	10.52	9.24
09	465×568	**7.62**	9.37	9.29	21.91	0.85
10	671×177	8.68	8.91	**8.52**	17.38	4.38

Fig. 3 Results of Dynamic-SLAM in the dynamic environment by using TUM dataset and KITTI dataset.

CONCLUTION

This framework has three major innovations. First, based on deep learning, an SSD object detector which combines prior knowledge is constructed to detect dynamic objects at the semantic level. Second, in view of low recall rate of the existing SSD object detection network, a missing detection compensation algorithm based on the speed invariance in adjacent frames is proposed, which greatly improves the recall rate for detection. Finally, a feature-based visual SLAM system is constructed, which processes the feature points of dynamic objects through a selective tracking algorithm in the tracking thread, thereby significantly reducing the error of pose estimation caused by incorrect matching.

FUNDING

This work was supported by the Young Thousand Talents Program.