Trajectory Visualization of Ego-Motion Videos with Pedestrian Based on Monocular Visual Odometry and Machine Learning

Yifei Zhang　　Hiroshi Watanabe

The Graduate School of Fundamental Science and Engineering，Waseda University

1. Introduction

Monocular camera trajectory visualization of ego-motion videos is the process of the estimating odometry of the moving agent, based on the images taken by the monocular camera fixed on the agent. In the last decade Monocular Visual Odometry (MVO) plays an important role in the area of autonomous navigation. Machine learning has outstanding performance in object detection and classification, e.g, Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) are proved to be efficient in pedestrians detection [1]. There is few machine learning applications in visual navigation and map reconstruction, while the moving human of source input contributes to error propagation during the generation of visual odometry. In this research, a lean trajectory visualization system is proposed. Instead of predicting the motion through a voting scheme and EM algorithm [2], in our approach, the pipeline which combines featured-based visual odometry, Simultaneous Localization and Mapping (SLAM) and HOG + SVM method in order to eliminate moving pedestrians is proposed. According to the result of the experiments, improvements in accuracy are shown. The remainder of this paper is organized as follows. Section 2 provides a review of related works. Section 3 details system pipeline and experiment setups. The result and comparison are elaborated upon in Section 4. Finally, conclusion and relevant future work are proposed in section 5.

2. Related Work

2.1 Feature-based monocular visual odometry

Compared with direct method, such as LSD-SLAM [3], feature-based methods shows more robust and relatively simple to implement [4]. Exclude loop closure part of feature-based SLAM, feature-based visual odometry extracts features and adopts them to the next step of triangulation and estimation. The feature matching is performed with creative liberties by adjusting appropriate parameters. The system accuracy and robustness are influenced by the featured amount, correctness of feature matching, lighting situations, camera rotation, and outlier correction.

2.2 FAST feature detector

As the base of the computer vision problems, FAST [5] feature detector is well-known for its efficiency. The efficient alternative of FAST is the choice of ORB-SLAM [6] which is the one of the most successful to date. FAST is addressed with machine learning steps during the creation of the decision tree. While it is sensitive about high levels noise and dependent on a threshold, it shows good performance in detection speed and feature amount.

2.3 HOG + SVM pedestrian detection [1]

The implementation of Histogram of Oriented Gradients (HOG) combined with Support Vector Machine (SVM) can represent the accurate pedestrian information of images. After detecting and extracting HOG descriptors from positive and negative images, a linear SVM is trained for our test dataset. Non-maximum suppression is adopted to remove redundant and overlapping bounding boxes around the pedestrian. Multiple bounding boxes is detected and keep the largest one while remove the others, which help to decrease the incorrectness.

2.4 Essential matrix estimation

After feature tracking by Kanade-Lucas-Tomasi [7] method, sparse pixel wise correspondences is built. A new detection is triggered when the feature point number is below the threshold. Five point algorithm [8] conjuncted with RANSAC [9] solve the non-linear equations in five degrees of freedom with higher accuracy. Erroneous correspondence outliers are reduced with quantitative iterations. Constructing trajectory is tracked and generated after computing the essential matrix parameters and camera poses.

2.5 Moving object detection and elimination

The camera poses are estimated by computing identical feature points among different images, which assumes that the most part of the environment are still. This will lead to incorrect correspondences and wrong camera estimation when there are moving objects in the still camera scenario.

More types of moving objects are detected and eliminated, the estimation will become more accurate while the process algorithms become complex and inefficient. Human being acts as the main element to transform and effect environment, so pedestrian is one of the main interference factors. For this reason, the pedestrians in the images is chosen to be the main moving object to be detected and eliminated.

3. Experiments

A lean camera trajectory visualization system based on featured-based monocular visual odometry and machine learning is presented, and tested by two publicly available KITTI benchmark sequences with ground truth. The datasets are captured by a monocular camera fixed on a cruise car. Camera intrinsic parameters are available and frames are undistorted.

For the pedestrian detection training and test sets, all the images are resized in 64 pixel * 128 pixel. We divide sequence images into positive and negative parts and combine with existing INRIA human samples. Next, we retrain the negative samples to detect the hard examples using the trained SVM to improve the accuracy.

4. Performance Evaluation

4.1 Error metrics

We use Root Mean Square Error (RMSE) and the Standard MSE (SMSE) to evaluate the performance of the proposed system and simple geometric monocular visual odometry with ground truth in every parameters. Ground truth data includes 12 parameters of the camera position about rotation and translation. The Table 1 show the result. From Table 1, computed on KITTI 05 dataset, we confirm that pedestrians feature points elimination reduce the error of camera pose estimation and trajectory generation.

4.2 Path generation

The dataset of KITTI 05 sequence include pedestrians, its reconstructed trajectories are shown in Fig 1.

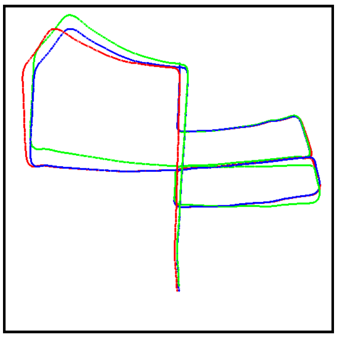

Fig. 1 Trajectories computed on the KITTI 05 set

The red path is ground truth, and Green one is simple MVO, and the blue one is the result of this paper.

5. Conclusion

In this paper, we proposed MVO system combined with machine learning. Both methods had accumulated error problem. However compared with geometric MVO, the performance of MVO combined with machine learning showed higher accuracy and precision. It indicates that the moving objects effect poses estimation. The pipeline in this paper provides a baseline for future work to detect more moving objects. From this experiment, it also indicates that the result of machine learning also depends on good labelled training and test data, and computational burden is another issue with the increase of error correlations.

References

1. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” IEEE Trans on CVPR, San Diego, CA, USA, pp. 886-893 vol. 12005.
2. R. Roberts, C. Potthast and F. Dellaert, "Learning general optical flow subspaces for egomotion estimation and detection of motion anomalies,” IEEE Conference on CVPR, Miami, FL, pp. 57-642009.
3. J. Engel, T. Schops, D.Cremers, “LSD-SLAM: Large-Scale Direct Monocular SLAM,” ECCV, Springer, Cham, vol.8690, 2014.
4. A. Huletski, D. Kartashov and K. Krinkin, "Evaluation of the modern visual SLAM methods," AINL-ISMW FRUCT, St. Petersburg, pp. 19-25, 2015.
5. T.Drummond, E.Rosten, “Machine laerning for high speed corner detection,” ECCV, Graz, Austria, vol.1, pp. 430-443, 2006.
6. R. Mur-Artal, J. M. M. Montiel, J. D. Tardos, “ORB-SLAM: A versatile and accurate monocular SLAM system,” IEEE Trans on Robotics, vol.31, no.5, pp. 1147-1163, 2015.
7. C.Tomasi, T.Kanade, “Detection and tracking of point features,” Carnegie Mellon University Technical Report on CS, pp.91-132,1991.
8. D.Nister, “An efficient solution to the five-point relative pose problem,” IEEE Trans on PAMI, vol.26, no.6, pp.756-770, 2004.
9. M.A.Fischler, R.C.Bolles,“RANSC sample consensus: A paradigm for model fitting with application to image analysis and automated cartography,” Communication ACM, vol.24, no.6, pp. 381-395, 1981.

Table 1 Performance Comparison For Pose Estimation Parameters

	ML-VO				Geometric VO			
	Rotation			Translation	Rotation			Translation
SMSE	0	0.1124	0.0039	0.0483	0.0001	1.0198	0.0721	317.3621
	0.11152	0	0.082	5.4936	1.02	0	1.0009	4.6331
	0.0036	0.1405	0	**0.1799**	0.0721	1.001	0.0001	**1.0887**
RMSE	0	0.0007	0.0009	0.0027	0.0002	0.0022	0.004	0.2184
	0	0	0.007	0.0204	0.0022	0	0.0026	0.0187
	0.009	0.001	0	**0.1252**	0.0024	0.0026	0.0002	**0.3081**