COMP0249 - Coursework 2

Evaluating monocular ORB-SLAM2 using the EVO tool on example sequences

Dhyan Shyam, Will Terry, YuChing Kwong
Computer Science
University College London

I. Introduction

In Part 1, we benchmarked the unmodified ORB-SLAM2 system on two widely used datasets: KITTI-07 [1] and TUM fr3 long office-household [2]. We are using a refactored ORBSLAM2 [3] which is built upon [4] which is also built on [5]. The KITTI-07 sequence is a 114.33s stereo video captured from a roof-mounted car camera in Karlsruhe, Germany, covering a ground-truth trajectory of 694.70m and featuring a natural loop closure near the end. The TUM fr3 long office-household sequence is an 87.09s handheld monocular recording through a textured indoor environment; it spans 21.46m and returns to its starting position, creating a large loop closure. These two datasets provide complementary challenges to outdoor urban driving versus indoor handheld motion with ground-truth pose data against which we measured performance.

For each dataset we ran the standard ORB-SLAM2 executables using the provided YAML configuration files. The KITTI-07 run used 2000 ORB features by default, while the TUM run used 1000 features. Next, we conducted a feature count sweep to quantify how the number of extracted ORB features affects both mapping and tracking, with a reduced and increased feature count. Trajectory evaluations were conducted using evo [6], which quantitatively compared the estimated trajectories against ground-truth data.

Originally, monocular measurements were rejected above a 95% confidence interval for a chi-square distribution. We replaced both thresholds with very high values sweeping through 9,999.9 to 9,999,999,999.9, effectively accepting all observations as inliers. This accepts all measurements as valid, which can lead to poor optimisation results because bad measurements are not being filtered out. These bad measurements can pull the optimisation solution away from the true solution. Over time, this can lead to drift and poor tracking because the optimisation is trying to satisfy measurements that should have been rejected. By setting the thresholds to 9999.9, this effectively disables the outlier rejection mechanism entirely. This means all measurements are treated as inliers regardless of their error. No measurements are rejected during optimisation. The optimisation will try to fit all measurements, even the bad ones

Let e_i be the reprojection error (in pixels) of feature i. ORB-SLAM2 tests each match against a chi-squared cutoff τ by checking

$$e_i^2 \;\le\; \tau.$$

By default $\tau \approx 5.99$ (the 95% bound for the monocular), but by replacing the τ with a very large number $M \gg e_i^2$, then

$$e_i^2 \;\le\; M \quad \forall i,$$

So every match is accepted, turning the outlier rejection effectively off. In practice, choosing M larger than the floating-point maximum will make the comparison itself invalid (infinity or NaN), causing the tracker to fail almost immediately.

We then disabled loop-closure altogether by commenting out the calls to DetectLoop() and CorrectLoop() in LoopClosing.cc. This removed bag-of-words loop detection, geometric verification, duplication merging, and global pose-graph optimisation, effectively removing the loop-closure detection.

II. KITTI 07 Sequence

A. KITTI Results

TABLE I: EVO APE results for KITTI dataset

Dataset	Max	Mean	Median	Min	RMSE	SSE	STD
1200	37.2	12.3	11.4	1.55	14.1	202 000	6.91
1500	10.9	4.92	4.93	0.195	5.14	28 900	1.47
1800	7.20	3.31	3.16	0.144	3.51	13 500	1.17
2000	6.46	3.26	3.23	0.269	3.44	13 000	1.10
3000	11.9	4.40	4.25	1.15	4.63	23 500	1.45
5000	10.3	3.74	3.63	0.301	3.98	17 300	1.34
10,000	10.1	2.69	2.47	0.243	2.97	9660	1.25
No Outlier	18.4	7.71	6.46	2.21	8.50	64 900	3.58
No Loop	48.8	15.8	14.5	0.561	18.7	385 000	10.1

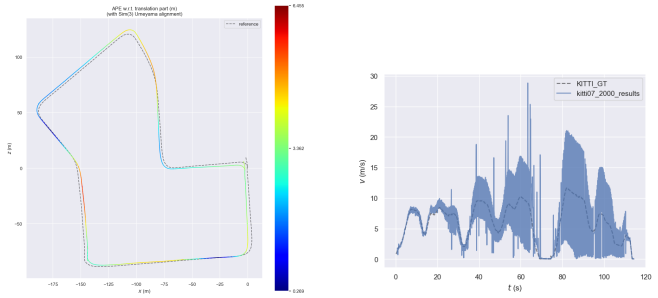
TABLE II: KITTI Path and Timing Comparison

Features	Path length (m)	Path Diff (m)	Time (s)	Time Diff (s)
Ground Truth	694.7	-	114.3	-
1200	595.0	−99.68	105.6	−8.731
1500	896.8	202.2	113.8	−0.5200
1800	795.0	100.3	113.8	−0.5200
2000	747.8	53.14	113.8	−0.5200
3000	809.2	114.5	113.8	−0.5200
5000	969.7	275.0	113.9	−0.4160
10,000	1143	448.4	113.9	−0.4160
No Outlier	678.2	−16.49	113.8	−0.5200
No Loop	625.5	−69.24	113.8	−0.5200

B. Baseline Test

Here, ORB-SLAM2 is run with the default 2000 feature count. Using evo with Sim(3) alignment, ORB-SLAM2 recovers a 747.837m trajectory, 53.140m longer than the 694.697m

ground truth (Table II) in 113.810s. The largest deviation in the Absolute Pose Error (APE) reaches 6.455m seen in Table I, occurring in the gentle curves and just before breaking for a corner. These APE spikes coincide with the vehicle's highest speeds (30–60s and 75–90s) in the speed profile seen in Figure 1, where rapid translation and abrupt rotation reduce inter-frame overlap and stress the feature matcher. Overall, however, the median and mean errors remain low seen in Table II, demonstrating strong accuracy throughout the whole sequence.



(a) (b)

Fig. 1: Comparing Speeds with the Trajectory for 2000 Features

C. Feature Reduction Tests

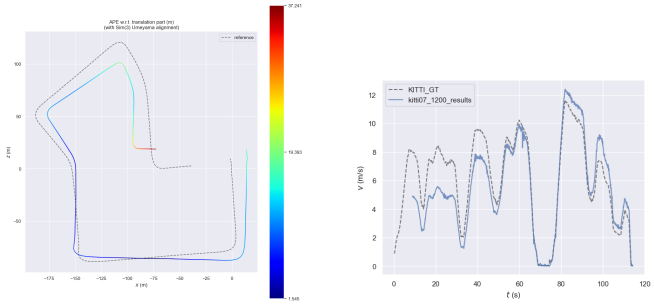
To test the reduction of features and running ORB-SLAM2, we first tried with 500 features. This proved to be too little for any feature detection to even initialise the ORB-SLAM2, even after 10 runs. So this was increased to 1000 features. Here, only twice did the ORB-SLAM2 initialise; however again not even features were captured throughout the whole dataset, and the trajectory was not plotted or provided. So 1000 was deemed the lower bound of features required for it to run. As 1000 was the lower bound, and 2000 was the default value, we settled on using 1200, 1500 and 1800 features as the 3 levels of number of features to run the ORB-SLAM2.

1) 1200: For the 1200 feature configuration, ORB-SLAM2 spends the first eight seconds initialising before processing the full trajectory, seen in Figure 2.

Fig. 2: ORB-SLAM2 trying to initialise for 1200 features

Because evo discards any non-overlapping timestamps before alignment, the startup segment is trimmed from both estimated and ground-truth paths, so the APE plot in Figure 3 appears to start mid-sequence, with no early loop closure. This drives an initial error spike to 37.241m and leaves a residual offset of 1.545m that persists throughout, seen in Table I.

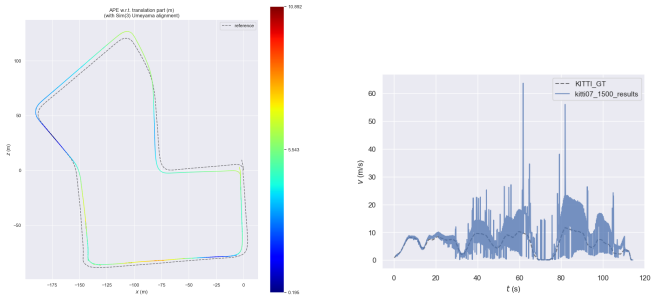
Correspondingly, the recovered path length is 99.678m shorter than the ground truth and total runtime is 105.599s since no poses are recorded during initialisation, seen in Table II. Despite the trimmed start, the overall trajectory still follows the reference outline. Notably, the largest alignment error aligns with the first high-speed segment (0–10s) in the speed profile, showcasing ORB-SLAM's sensitivity to rapid motion before mapping stabilises. Toward the end of the sequence, APE gradually creeps upward again, without a late loop closure to correct drift (ORB-SLAM2 wasn't fully initialised in time to detect the initial loop), the system never resets the accumulated offset.



(a) (b)

Fig. 3: Comparing Speeds with the Trajectory for 1200 Features

2) 1500: With 1500 features, ORB-SLAM2 begins mapping immediately and processes the entire sequence, recovering an 896.849m trajectory, 202.152m longer than the ground truth while being 0.520s faster as seen in Table II. Because no startup segment is trimmed, Sim(3) alignment introduces an initial APE offset of 6.2m, seen in Figure 4. The overall APE peaks at 10.892m, with mediocre performance as seen in Table I. Again the error spikes twice (at 60s and 78s) that coincide with the vehicle's highest speeds (50–70s and 75–90s). These peaks again show ORB-SLAM2's sensitivity to rapid motion, which temporarily degrades feature matching and grid alignment.

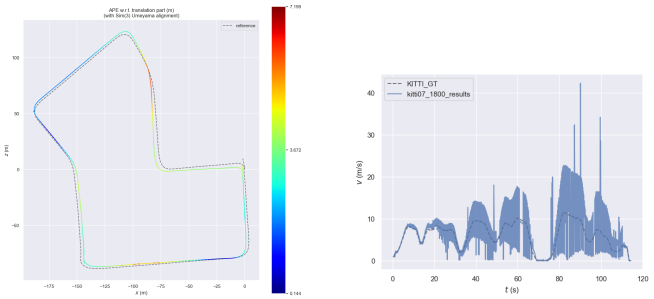


(a) (b)

Fig. 4: Comparing Speeds with the Trajectory for 1500 Features

Between closures, residual drift settles around the mean APE, while the extended path length attests to the uncorrected errors accumulating during high-velocity segments. The fact that the SLAM path is longer than the ground truth arises from micro-jitters and small frame-to-frame drift, each "wiggle" adds distance and from unclosed loops or corner overshoots, which cause the trajectory to arc outside the true route and cover extra meters.

3) 1800: With 1800 features, ORB-SLAM2 processes the entire sequence without any startup trim, recovering a trajectory 100.256m longer than the ground truth and 0.520s quicker, seen in Table II. Sim(3) alignment introduces an initial APE offset of 4.0m seen in Figure 5, and overall the APE peaks at 7.199m and relatively similar error values as found with 2000 features seen in Table I. After the first loop closure around 50s, the error drops to a minimum of 0.144m, confirming tight alignment. The two prominent spikes, 6.1m at 25s and 7.2m at 78s, coincide with the highest velocity segments in the speed profile. Between these events, residual drift remains low, thanks to successful closures. The recovered path still exceeds the true distance due to the accumulated micro-jitters and slight corner overshoots, though the path length difference (+100.256m) is substantially lower than in the 1500-feature run.

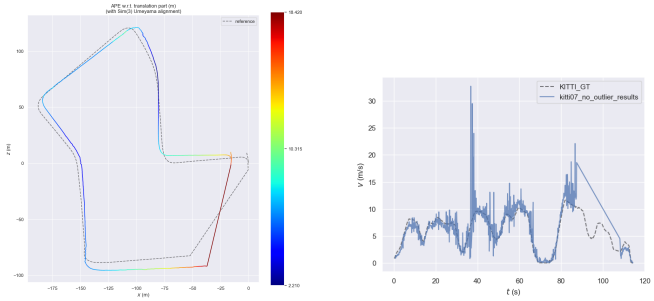


(a) (b)

Fig. 5: Comparing Speeds with the Trajectory for 1800 Features

D. Disabled Outlier Rejection Test

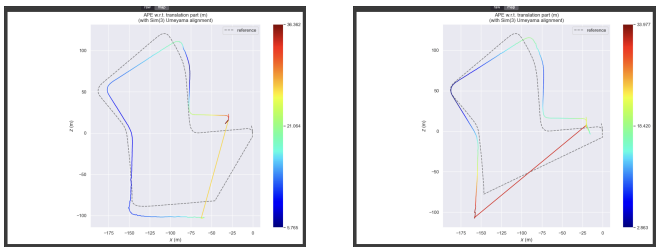
To study the impact of robust outlier rejection, we raised ORB-SLAM2's monocular chi-square thresholds from the default 95% confidence limit to 9,999.9, 9,999,999.9, and 9,999,999,999.9, effectively accepting all feature matches as inliers. With no outlier rejection, every correspondence and mismatches enter the nonlinear least-squares solver. This deteriorates the conditioning and pulls the estimate away from the true trajectory.



(a) (b)

Fig. 6: Comparing Speeds with the Trajectory for No Outlier Detection

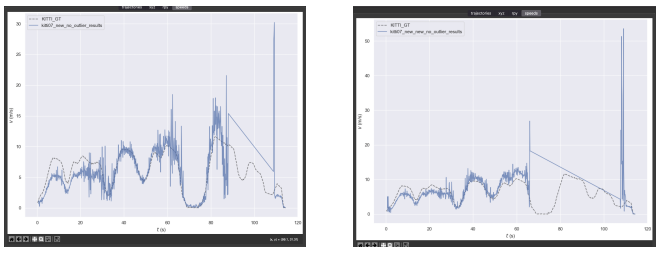
Figure 6 show the resulting map at a threshold of 9,999.9, the system tracks almost to the end (110s) but accumulates marked drift as seen in Table I and a shortened path (678.2 m; –16.5 m path-diff) in Table II.



(a) (b)

Fig. 7: Comparing the Trajectory with chi-square value set at 9,999,999 and 9,999,999,999

Raising the cutoff to 9,999,999.9 and 9,999,999,999.9 makes ORB-SLAM2 fail progressively earlier: the speed curve becomes erratic around 80 s and 70 s, respectively, then falls into a smooth linear descent as the solver effectively "gives up" and holds the last pose, seen in Figures 7 and 8. The recovered path lengths shrink accordingly, and APE statistics worsen. In the absence of any outlier filtering, bad measurements dominate the optimisation, driving large residuals, destabilising loop closures, and eventually aborting tracking altogether.

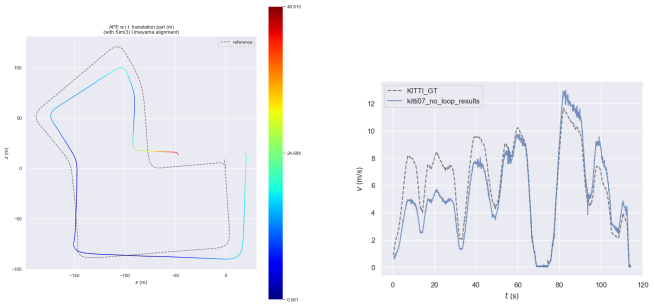


(a) (b)

Fig. 8: Comparing the speed trajectory with chi-square value set at 9,999,999 and 9,999,999,999

E. Disabled Loop Closure Test

After disabling the Loop Closure in the code, ORB-SLAM2 still tracked every frame in the KITTI07 dataset. Although the frame-to-frame tracking and local bundle adjustment still run, the estimated trajectory is only 69.239m shorter than the reference ground truth with the runtime remaining 0.520s quicker, seen in Table II. This is because evo alignment omits the unclosed final segment and Sim(3) alignment now minimises error over an uncorrected pose chain, producing a large initial APE offset of 48.815m that gradually falls to a minimum of 0.561m near the mid-sequence, seen in Figure 9.



(a) (b)

Fig. 9: Comparing Speeds with the Trajectory for No Loop Closure

That mid-sequence valley reflects only local consistency—when the vehicle slows, the front-end bundle adjustment briefly regains accuracy even without any loop closures. However, as speed increases again, drift grows unchecked and without any loop closures to reset drift, the error then climbs back up to 23m at the end, yielding the degraded APE values seen in Table I. The speed profile is identical to the baseline, confirming that the performance drop is purely from removing global corrections rather than any change in vehicle motion.

F. Effect of Increasing Feature Count

Figure 11 plots the APE trajectories for 3000, 5000 and 10,000 features, and the box-plot in Figure 10 summarises their error distributions alongside the 1200–2000 baselines and the "no-loop"/"no-outlier" cases. As the feature budget rises, ORB-SLAM2 leverages more correspondences to improve its front-end accuracy, yielding both lower central error and tighter spread:

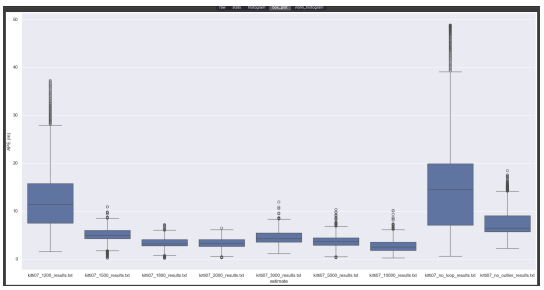


Fig. 10: Box Plots for the KITTI tests

a) 3000 features: 3000 features covers a longer +114.527m path still within 0.5s of the ground truth, but exhibits mediocre APE results compared to 2000 and even 1800 features. Its box plot shows an interquartile range (IQR) broader than the 2000-feature baseline, with more high-error outliers in the sharp-corner segments.

b) 5000 features: 5000 features further reduces bias and variance, while the path length grows to an extra 275.016m. The box plot's whiskers retract and the IQR narrows, indicating fewer extreme APE spikes.

c) 10,000 features: 10,000 features delivers the best accuracy with an extra path length of 448.410m. Its box plot is the most compact and contains the fewest outliers, demonstrating both low systematic error and tight variance.

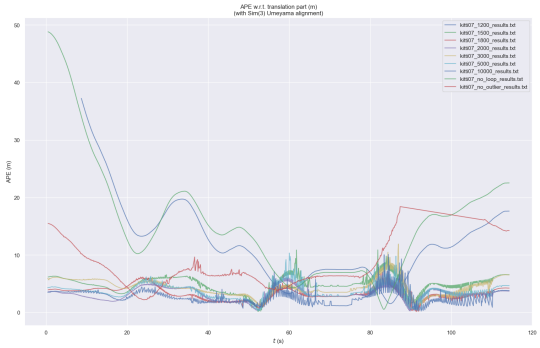


Fig. 11: APE Translation comparison for the KITTI tests

These graphs seem noisy; however, this is not because too many features inherently degrade accuracy, but because the high temporal resolution of the error metric (one APE sample per frame at 10Hz) exposes small frame-to-frame jitters. Each additional feature adds one more residual into the bundle-adjustment, so even minute mismatches, features on low contrast regions, slight corner misalignments, or marginal edge points introduce tiny corrections that show up as high-frequency oscillations in the APE curve.

Across all runs, timing remains essentially unchanged, showing that the back-end solver scales nicely with more features. These results confirm that increasing the feature count beyond the 2000 feature points default yields diminishing but real gains in trajectory accuracy, visible both in the raw APE traces and in the shrinking boxes of the box plot. In essence, more features reduce bias and variance in the mean statistics, yet amplify the visibility of per-frame jitter in the APE time series.

III. TUM FR3 Long Office Household Sequence

A. TUM Results

TABLE III: evo APE Results for TUM Dataset (Values $\times 10^{-2}$)

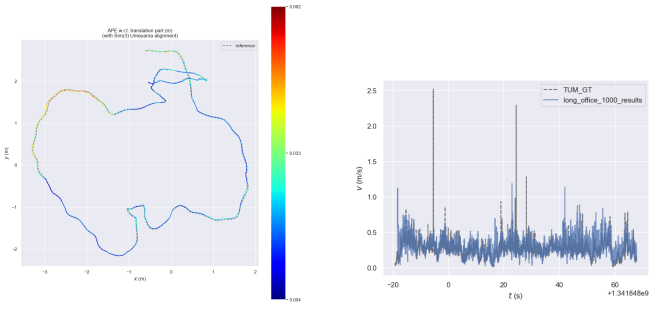
Dataset	Max	Mean	Median	Min	RMSE	SSE	STD
500	4.6	1.4	1.3	0.1	1.6	49.7	0.7
660	13.7	2.9	2.4	0.4	3.3	240.1	1.5
820	4.1	0.9	0.9	0.0	1.1	28.6	0.5
1000	6.2	2.2	2.0	0.4	2.4	152.8	1.0
2000	4.3	1.1	1.0	0.0	1.2	35.6	0.5
5000	153.5	43.0	27.9	4.6	57.2	83 386.4	37.7
10,000	8.1	2.1	1.9	0.2	2.4	151.1	1.2
No Outlier	154.6	23.7	17.4	1.0	33.1	28 087.4	23.2
No Loop	9.2	3.0	2.5	0.1	3.7	345.7	2.1

TABLE IV: TUM Path and Timing Comparison

Features	Path length (m)	Path Diff (m)	Time (s)	Time Diff (s)
Ground Truth	22.197	-	87.089	-
500	19.80	-2.4	86.24	-0.853
660	22.18	-0.016	76.32	-10.773
820	26.33	+4.132	86.17	-0.921
1000	25.99	+3.797	86.24	-0.853
2000	25.13	+2.928	86.14	-0.953
5000	52.20	+29.999	85.97	-1.121
10,000	26.91	+4.711	85.97	-1.121
No Outlier	98.83	+76.636	86.24	-0.853
No Loop	25.29	+3.089	86.24	-0.853

B. Baseline Test

Here, ORB-SLAM2 is run with the default 1000 feature count. ORB-SLAM2 reconstructs a 25.994m path, overshooting the ground truth by 3.797m, as seen in Table IV in 86.236s. Despite the noisy, jittery camera motion inherent in a handheld recording, the system achieves sub-decimeter accuracy, with very good APE statistics found in Table III. The map trace oscillates at the tens-of-millimetres level, reflecting small frame-to-frame localisation jitter as the map is continually refined, while the speed profile shows the characteristic stop-and-go behaviour of an office walkthrough, seen in Figure 12. The speed profile also remains below 1 m/s throughout, confirming that these oscillations stem from the inherent shakiness of the handheld sequence rather than rapid motion. Overall, the baseline run demonstrates that ORB-SLAM2 can deliver highly accurate, loop-consistent trajectories indoors, even under handheld "shaky" conditions.



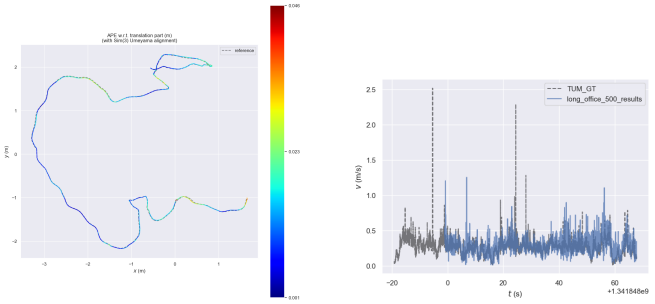
(a) (b)

Fig. 12: Baseline for 1000 Features in TUM

C. Feature Reduction Tests

To test the reduction of features and running ORB-SLAM2, we first tried with 300 features. This proved to be too little for any feature detection to get ORB-SLAM2 to initialise. So this was increased to 500 features. Here, the ORB-SLAM2 worked half of the time, throughout the whole sequence. This was deemed good enough to be considered the lower bound for the number of features. With 1000 being the upper bound, the 3 levels of features considered are 500, 660 and 820.

1) 500: At 500 features—the lower bound for a successful run—ORB-SLAM2 often takes several seconds to detect and triangulate enough keypoints before initialisation completes. As a result, evo alignment discards the first 20s of both estimated and ground-truth poses, so the APE map "starts" after this unseen startup, seen in Figure 13. The system then reconstructs a path, 2.400m shorter than the reference, found in Table IV, again quicker than the ground truth by 0.853s.

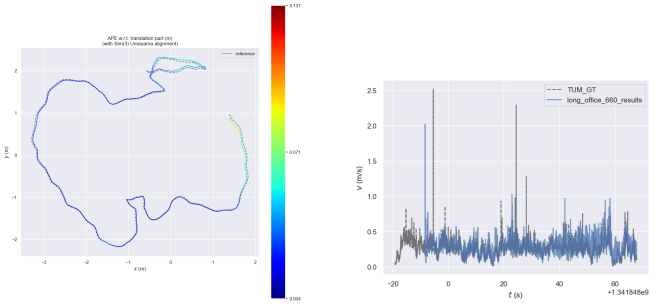


(a) (b)

Fig. 13: 500 Features in TUM

Once initialised, the front-end maintains sub-decimeter accuracy, found in Table III. The APE time-series oscillates around 0.01–0.02m, reflecting handheld jitter rather than rapid motion, since the speed profile remains below 1m/s throughout. Compared to the 1000-feature baseline, the initial error spike and trimmed startup segment highlight the cost of too few features: more frames are needed to bootstrap the map, but once running, tracking quality remains comparable.

2) 660: At 660 features, ORB-SLAM2 requires roughly 15s to detect and triangulate enough keypoints before initialisation, so evo alignment again trims the startup, and the system runs for 76.316s, seen in Table IV. Despite the reduced feature budget, it reconstructs a path, only 0.016m shorter than the ground truth, in that time.

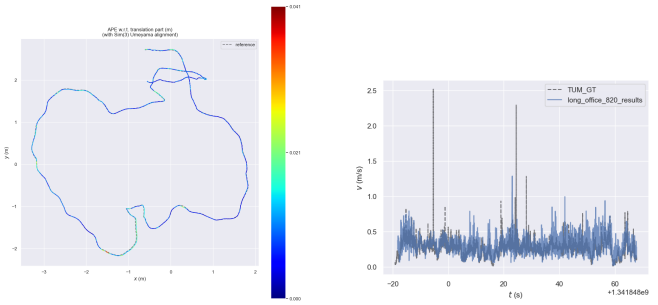


(a) (b)

Fig. 14: 660 Features in TUM

Once initialised, the APE rapidly falls to a steady band around 0.01–0.03m with good statistics, seen in Table III. The speed profile remains below 1m/s, seen in Figure 14, confirming that all residual oscillations reflect handheld jitter rather than rapid motion. Compared to the 500-feature run, initialisation is not as delayed, and the central APE rises slightly, but post-boot accuracy and robustness remain high.

3) 820: With 820 features, ORB-SLAM2 completes initialisation within 1-2s, making evo alignment have no trims in the startup segment, and the system then runs 0.921s quicker, seen in Table IV. It reconstructs a path, overshooting the reference by 4.132m.



(a) (b)

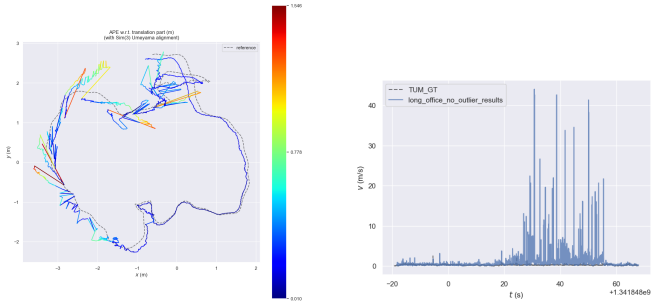
Fig. 15: 820 Features in TUM

Figure 15 shows the speed profile stays below 1m/s, confirming that the sub-centimetre oscillations found in the statistics shown in Table III stem from handheld jitter rather than fast motion. Compared to the 660-feature run, the startup completes more quickly, and the central APE tightens slightly.

D. Disabled Outlier Rejection Test

When outlier rejection is disabled, the estimated trajectory is completely drifted and, together with speed, gradually drifts more as time passes, as shown in Figure 16. This indicates a pattern of scale drift, which is usually considered a common issue in monocular ORB-SLAM2. This is expected as if there is no absolute scale information, things like noise will create inaccurate relative estimates, and it accumulates over time.

Disabling outlier rejection forced every feature match, even gross mismatches, into the optimiser. This resulted in the reconstructed path to balloon to 98.833m, which is 76.636m more than the reference in a similar time frame of 86.236s, seen in Table IV.



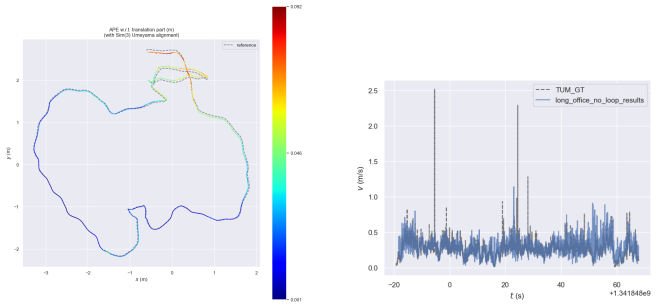
(a) (b)

Fig. 16: Disabled Outlier Rejection in TUM

Initial APE remains low (0.02m) while front-end tracking holds, but from 30s onwards, the solver takes in spurious correspondences and the error climbs in repeated spikes up to 1.546m. The APE time series shows these high-frequency excursions, even though the speed profile stays below 1m/s throughout, in Figure 16, confirming that the drift is algorithmic rather than motion-induced. The overall statistics also degraded, found in Table III, illustrating that without any outlier filtering, bad matches dominate the bundle-adjustment, destabilising the map and causing large drifts.

E. Disabled Loop Closure Test

When the loop closure is disabled, ORB-SLAM2 still tracks every frame but never applies global pose-graph corrections. The reconstructed path is 3.089m longer compared to the ground truth in the similar 86.236s, see in Table IV. The Sim(3) alignment over an uncorrected pose chain produces an initial APE of 0.0919m that falls to a minimum of 0.0008m around 40s, reflecting only local bundle-adjustment consistency. However, without loop closures to reset drift, APE then creeps back up to 0.06m by the end. The overall error statistics degrade by a small margin seen in Table III. The speed profile in Figure 17 remains below 1m/s throughout, confirming that this late-sequence drift is due entirely to the absence of loop-closure corrections rather than any increase in motion dynamics.



(a) (b)

Fig. 17: Disabled Loop Closure Test

F. Effect of Increasing Feature Count

Figure 11 plots the APE trajectories for 3000, 5000 and 10,000 features, and the box-plot in Figure 10 summarises their error distributions alongside the 1200–2000 baselines and the "no-loop"/"no-outlier" cases. As the feature budget rises, ORB-SLAM2 leverages more correspondences to improve its front-end accuracy, yielding both lower central error and tighter spread:

G. Effect of High Feature Counts on TUM "Long Office"

Finally, we evaluate the impact of increasing the ORB-SLAM2 feature budget from the 1000-feature baseline up to 2000, 5000 and 10,000 features. Table III shows that median APE decreases slightly as feature count grows, falling from 0.0198m at 1000 features to 0.0097m at 2000, then hovering around 0.0279m (5000) and 0.0187m (10,000). RMSE and mean errors exhibit the same diminishing-returns trend. The corresponding time values reveal that the raw APE curves found in Figures 19 and 18 become a bit "noisier" with more features, due to the higher temporal resolution of small residuals, but these fluctuations remain below 0.05m. The path lengths (Table IV) and timing, around 86s, change negligibly, confirming that the back-end solver scales to 10,000 features without cost.

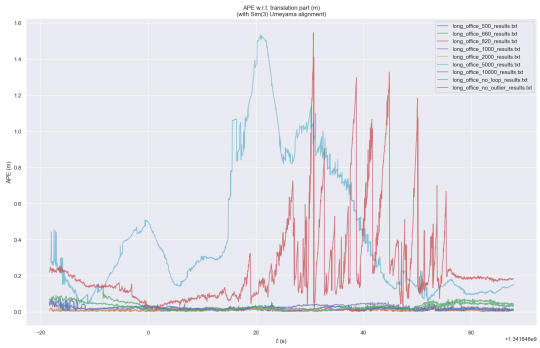


Fig. 18: APE Translation comparison for the TUM tests

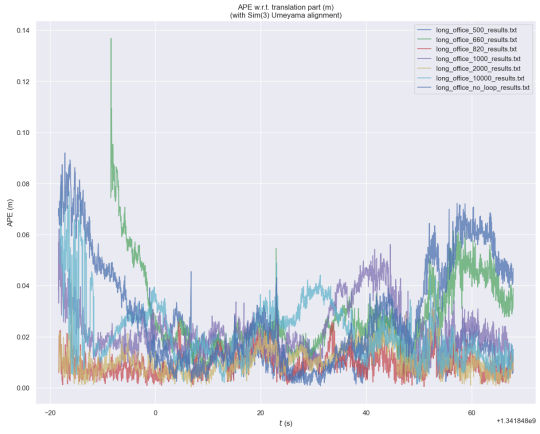


Fig. 19: APE Translation comparison for the TUM tests (Without Extreme Failures)

The box-plots in Figures 21 and 20 condense these observations, as feature count increases, the interquartile range shrinks only modestly, and the whiskers retract slightly, indicating smaller worst-case errors.

The filtered box-plot in Figure 21, which omits extreme spikes, makes this convergence even clearer beyond 2000 features, where the error distribution is essentially saturated. In other words, adding more than 2000 features yields only marginal gains in indoor accuracy, while exposing the optimiser to more low-level jitter that shows up as high-frequency oscillations in the APE trace.

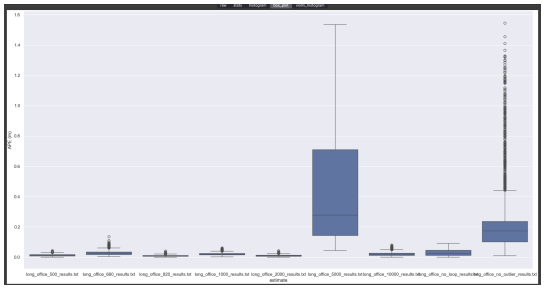


Fig. 20: Box Plots for the TUM tests

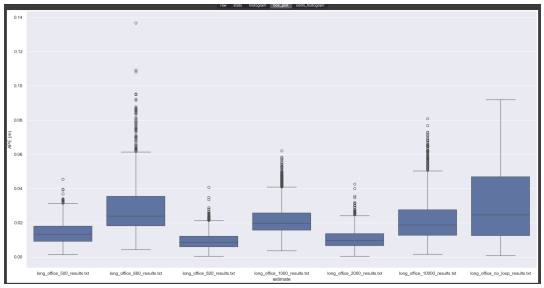


Fig. 21: Box Plots for the TUM tests (Without Extreme Failures)

There are two separate box plot images and comparison APE translation images, as there is a filtered version that takes out the extreme values, to make the graphs more readable and visually appealing.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[3] S. Julier, "Refactored orb slam2," https://github.com/sjulier/Refactored_ORB_SLAM2, 2021.

[4] M. J. M. M. Mur-Artal, Raúl and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.