Analysis and Implementation of Novel VSLAM Methods

Anonymous CVPR submission

Paper ID Occular Tension

Abstract

The overall problem of Visual Simultaneous Localization and Mapping (VSLAM) and its different aspects have been researched in the computer vision community extensively. This paper critiques three papers on VSLAM and provides an analysis of the innovative contributions of each. Robust Large Scale Monocular Simultaneous Localization and Mapping proposes a framework for implementing VSLAM using monocular cameras. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects proposes a method for VLSAM using prior knowledge of the environment coupled with object recognition. Good Features to Track for Visual SLAM (GF-SLAM) focuses on the aspect of finding good visual features to use in VSLAM based on their temporal observability. This paper also provides an analysis of the implementation of the GF-SLAM paper and a recreation of the paper's key experiment.

1. Introduction

Visual Simultaneous Localization and Mapping (VSLAM) refers to the problem of estimating a 3D model of the environment as sensed by a camera as well as estimating the camera's trajectory in relation to this model. The computer vision community has researched and provided innovative solutions that focus on different aspects of this overall research problem. In this paper, we analyze the following three CVPR conference papers on VSLAM and discuss their approach to particular aspects of the problem. The three papers are:

1. Robust Large Scale Monocular Visual SLAM [3]
2. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects [14]
3. Good Features to Track for Visual SLAM [18]

We also reimplement [18] and recreate its key experiment.

2. Robust Large Scale Monocular Visual SLAM

2.1. Problem Statement

This paper focuses on the problem of using calibrated monocular cameras to perform VSLAM while making the algorithm robust, accurate, and scalable. Monocular VSLAM comes with the challenge of not being able to observe the scale of the scene of the environment. In order to overcome this, loop closures (the return of the camera to a previously observed location) needs to be detected. This is an issue in large environments where many scenes look alike and results in an erroneous 3D model if loop closures are not detected properly. Thus, the paper focuses not only on the general problem of monocular VSLAM but also tackles a key subproblem of dealing with loop closure.

2.2. Innovative Contribution

To solve the problem of monocular VSLAM, the authors of [3] propose a framework based on three innovations: 1) a Structure from Motion (SfM) algorithm based on the *Known Rotation Problem* [12] used to estimate submaps (parts of the camera trajectory and the unknown environment), 2) a loopy belief propagation algorithm used to efficiently align submaps based on a graph of relative 3D similarities to produce a global map that is consistent up to a scale factor, and 3) an outlier removal algorithm that detects and removes outliers in the relative 3D similarity graph used to reject wrong loop closures.

2.3. Proposed Method

The paper proposes a four-part framework to implement the solution to monocular VSLAM: 1) keyframe selection, 2) submap reconstruction, 3) pairwise similarity estimation, and 4) large scale relative similarity averaging.

Keyframe selection: For each frame in the captured video, Harris Points of Interest (PoI) are detected and tracked using a Lucas-Kanade tracker. When the Euclidean distance between the PoI of the current frame and previously selected keyframe is greater than a specified threshold, the frame is selected as a keyframe.

Submap reconstruction: Consecutive keyframes are clustered, and using the *Known Rotation Problem*, a SfM algorithm is applied to each one after extracting the SURF PoI [2] from all member keyframes. Loops are closed inside of each submap by matching these PoI between pairs of keyframes. The epipolar geometry is then calculated using the 5-point algorithm with RANSAC and bundle adjustment [11, 7] between consecutive pairs of images using the SURF matches and tracked Harris PoI. The local 3D orientations are then extracted and used to estimate the global 3D orientation. With this, known tracks of PoI are built and a linear program is used to solve the *Known Rotation Problem*. This is used to estimate the camera pose at each keyframe and the associated 3D point to reconstruct the submap.

Pairwise Similarity Estimation: Loop closures in the reconstructed submaps are detected by first applying a bag of words approach on the SURF descriptors of the 3D points in all submaps to give each submap a unique descriptor. Then, the relative 3D similarities between each keyframe and its 10 nearest neighbors are estimated by matching SURF descriptors with the 3D points of each submap using a k-d tree followed by the 3-point algorithm [17] with RANSAC and nonlinear refinement on those matches.

Large Scale Similarity Averaging: To align submaps by estimating thier global 3D similarity to the global reference frame, a cost function on relative similarities is minimized by transforming the problem into a graph inference problem. Outliers in the graphs (representing wrong loop closures) are rejected by the *outlier removal algorithm* [3] in which loop closures are incrementally checked by finding the shortest loop of inliers and adding them to an overall graph of inliers if their cycle error and covariance are within specified bounds. Once the outliers are removed, the *loopy belief propagation algorithm* [3] performs graph inference by accumulating the measurements and variances on temporal subgraphs of the original graph as it builds up final average global similarity. This algorithm is parallelized, so that it can be applied on a large scale of submaps.

2.4. Experimental Evaluation

To evaluate the proposed VSLAM framework, the authors compared its performance with to state of the art algorithms in [6] and [9] on the TUM RGB-D [16] and KITTI [1] datasets. They also performed a separate experiment using four cameras with different resolutions on indoor videos they captured. Each experiment used the same optimized parameters for the various parts of the algorithm. When evaluating the results of the algorithms with respect to ground truth, a 3D similarity obtained from the minimum distance between the estimated and actual camera trajectories was used. When compared to the [6] on the TUM RGB-D dataset, the author's approach resulted in a lower RMSE for camera trajectories. When compared to [9] on the KITTI dataset, the author's algorithm estimated camera trajectories that were closer to the ground truth and thus performed better. When compared to [6] on their own videos, the proposed method outperformed it with respect to the ground truth motion of the camera. In addition, the paper discusses the limitations of the framework in not being able to estimate a pure rotation of the camera, the necessity for the sensed environment to be static, and the necessity for consecutive relative similarities to be outlier free. However, the framework still has reasonable performance when applied to datasets that involve some moving objects.

2.5. Subsequent Conclusions

The performance evaluation of the method shows that the authors' proposed monocular VSLAM framework does substantiate their claim. Robust, independent submap generation is achieved by the visual odometry approach based on the *Known Rotation Problem*, and these submaps can be processed and aligned to form the global map and camera trajectory estimates with loop closure through the outlier removal and loopy belief propagation algorithms. Even with the described limitations, the evaluations show the innovative framework does provide a robust, accurate, and scalable solution to loop closure and the overall problem monocular VSLAM.

3. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects

3.1. Problem Statement

This paper focuses on the problem of implementing VSLAM at the level of 3D objects rather than low-level primitives that must be robustly matched (*e.g.* points, lines, etc). Moreover, much of the prior knowledge of objects and the environment could be taken into account "in the loop" of the VSLAM algorithm itself but is currently wasted [14]. Thus, the authors of [14] aim to solve the problem of implementing VSLAM with higher level features that can take into account the semantic labelling of geometry as objects and regions provided by the efficient hardware that exists today.

3.2. Innovative Contribution

To solve the problem of VSLAM approached from the perspective of 3D object recognition, the authors of [14] propose the innovation of building pose graph maps based on an "object-oriented" approach that directly encodes the positions of recognized 3D structures. With each new sensor measurement, the graph is continually optimized and allows for efficient tracking of the camera system based on recognized landmarks. In addition, the algorithms make the assumption that the world has "intrinsic symmetry in the form of repetitive objects" [14] thereby allowing for the the objects in a scene to be identified and segmented as salient repeated elements. KinnectFusion algorithms are used to scan and detect objects matched to previously created models which are then used in camera pose tracking. The algorithm leverages this repetitiveness along with the efficient use of GPU architectures to provide a real time processing system.

3.3. Proposed Method

Creating an Object Database: First, a database of models of repeatedly occurring objects is created via known KinectFusion algorithms [10]. These are objects that are subsequently recognized and used in the VSLAM process.

SLAM Map Representation: The world is represented by a graph where each node stores the 6DOF pose of the discovered object as well as its annotated label from the database.

Real-Time Object Recognition: This portion of the method recognizes objects in the world based on standard mesh recognition algorithms similar to those in [5]. The implementation is parallelized on GPUs to allow the real-time detection of multiple instances of multiple objects. These correspondences are obtained via the use of Point-Pair Features (PPFs) which are 4D descriptors of relative poses and pairs of oriented surface normals for each object.

Camera Tracking and Object Pose Estimation: The iterative closest point (ICP) algorithm [13] is used to to track the pose of the camera model based on the earlier computed object based locations. A Huber penalty function is used in this optimization process. Criteria is developed to ensure successful convergence of the tracking error.

Graph Optimization: The poses of each static object are now viewed as a graph optimization problem which minimizes the sum over all the measurement constraints based on the known features of each object.

Relocalization: The system accounts for a loss in camera tracking by reinitializing localization based on matching at least three of the objects seen in the previously tracked long-term graph.

3.4. Experimental Evaluation

The authors of [14] reference a video submitted along with it to CVPR as a better description of the advantages of their method.

Loop Closure: Small loop closures are detected and compensated for by the ICP algorithm. Larger loop closures are compensated for by the relocalization method.

Large Scale Mapping: Scaled mapping of a large room (15mX10mX3m) was obtained along with the mapping of 34 different objects around the room. The algorithm uses no priors regarding the original placement of these objects.

Moved Object Detection: The algorithm also displays the ability to track these objects while they, themselves are in motion.

System Statistics: The algorithm displays the amount of storage used as compared to the more traditional KinectFusion algorithm. The given mapped rooms is stored in about 20MB of space as compared to the 1.4GB used by KinectFusion. The resultant compression ratio is 1/70 which is a dramatic improvement.

3.5. Subsequent Conclusions

The authors of [14] make several bold claims with regard to their contributions to the literature. The graph based optimization method does indeed seem novel and the system has a significantly large data compression ratio when compared to KinectFusion algorithms. Unfortunately, the experimentally evaluated conclusions are quite sparse when compared to the dense and well-written introduction and methodology of the paper. The biggest issue pertains to the fact that no standard metric is used to compare the system's advantages to other 3D-object-recognition-based VSLAM approaches. In conclusion, there remains no easy way to validate their claims.

4. Good Features to Track for VSLAM

4.1. Problem Statement

This paper focuses on the problem of excessive eature selection for VSLAM since not all features detected and tracked by conventional methods contribute to the accurate estimation of camera trajectories and the overall map. Only a subset of all tracked and detected features actually produces a good result for VSLAM, and furthermore, there are many different ways of detecting these features. As VSLAM applications grow to include large-scale reconstruction from large sets of features, an efficient ranking algorithm that can be applied to features generated by any detection method is necessary. Thus, this paper focuses on finding the best features out of any given set of previously detected features as a preprocessing step for VSLAM.

In exploring different methods for VSLAM proposed in recent conference papers, we chose to re-implement the methods and key experiment of this paper.

4.2. Innovative Contribution

To solve the problem of finding the best features to use for VSLAM given a set of detected features, [18] proposes a new SVD-based algorithm for more rigorous feature selection to be used during the localization process. The paper develops a mechanism by which such measured features can be ranked with an observability score. The ranked feature set can then be used to generate a reduced feature set with better a better estimation accuracy. The authors call their system the *Good Features* algorithm (GF-SLAM for short). The innovative contribution of this paper is in the three main aspects of GF-SLAM: 1) a system observability measure for features, 2) a rank-k temporal update of observability scores, and 3) a submodular learning method for completing the set of good features if the initial ranking of scores fails to provide the desired quantity of good features for a given application of VSLAM.

System Observability Measure: To counter the use of extraneous features in the localization process of VSLAM, the paper develops a metric of observability scores based on the temporal observability of each detected feature in each captured video frame. Features are represented by their coordinates in the image plane as well as the corresponding world coordinates, allowing this measure to be applied to features extracted from video frames by any detection method. The general idea is that features with higher temporal observability are scored higher and the set of *good features* is created from all features whose scores are above a specified threshold based on the application of VSLAM. These good features are then called *anchors*.

Rank-k Temporal Update of Observability Score: The generation of feature observability scores requires the computation and storage of large observability matrices that increase in size over time. The actual act of creating these matrices is inefficient and can lead to delays in the SLAM process. To counter this, the authors propose a new algorithm to generate the observability matrix based on an incremental SVD of the matrix applied per time step.

Submodular Learning for Feature Grouping: The ranking and thresholding of feature observability scores may not always produce an adequate number of anchors. Thus, this paper proposes a submodular learning algorithm that greedily chooses the best of the remaining features with scores below the threshold to complete the anchor set based on an SVD of the augmented system observability matrix of the anchor set and possible upgraded features.

4.3. Proposed Method

Dynamical Measurement Model of VSLAM System: The authors use a constant velocity motion model that describes the motion of a VSLAM system in standard Euclidian three space, $SE(3)$, as the basis for their algorithm. The method uses the position and orientation as well as the linear and angular velocities of the system's camera model:

$$\boldsymbol{x}^W_{R_k} = (\boldsymbol{r}^W_{R_k} \quad \boldsymbol{q}^W_{R_k} \quad \boldsymbol{v}^W_{R_k} \quad \boldsymbol{\omega}^W_{R_k})^T \tag{1}$$

where $\boldsymbol{x}^W_{R_k}$ is the state vector of the camera model, $\boldsymbol{r}^W_{R_k}$ is its position, $\boldsymbol{v}^W_{R_k}$ is its velocity, and $\boldsymbol{\omega}^W_{R_k}$ is its angular velocity. All these quantities are with reference to the world frame as denoted by the superscript W at time k. The state vector is updated with each time segment by the SLAM system.

The algorithm requires access to the following properties of the features that are detected by the VSLAM system:

$$\boldsymbol{p}^{R_k}_i = (p^{R_k}_{x_i} \quad p^{R_k}_{y_i} \quad p^{R_k}_{z_i})^T = R^{R_k}_W((\boldsymbol{q}^W_{R_k})^{-1})(^{(i)}\boldsymbol{p}^W_k - \boldsymbol{r}^W_{R_k}) \tag{2}$$

$$\boldsymbol{h}^{R_k}_i = \begin{bmatrix} h_{x_i} \\ h_{y_i} \end{bmatrix} = distort(\begin{bmatrix} u_0 - fk_u \frac{p^{R_k}_x}{p^{R_k}_z} \\ v_0 - fk_u \frac{p^{R_k}_y}{p^{R_k}_z} \end{bmatrix}) \tag{3}$$

where $\boldsymbol{p}^{R_k}_i$ is the pinhole projection of the features, $\boldsymbol{h}_{x_i}$ is the projection of these coordinates on to the 2D pinhole plane, $R(\boldsymbol{q})$ is the rotation matrix of $\boldsymbol{q}$, and fk_u, fk_v, u_0, and v_0 are the camera's intrinsic parameters. These coordinates are obtained after correcting for distortion which involves the intrinsic parameters of the camera via the $distort(.)$ function from [4].

Computation of System Observability Measure: To compute the observability score for each feature, [18] defines a System Observability Matrix (SOM), Q_{SOM} based on [8]. Q_{SOM} is defined for a fixed number of frames, τ, of the SLAM process. When the current frame iteration grows larger than τ, the oldest elements of the now fixed size Q_{SOM} are replaced directly by the newest feature values. τ is thus an important input parameter to the system as it governs the efficiency of the overall algorithm. Q_{SOM} itself is built as a function of the current frame number, j, where $j < \tau$.

$$\boldsymbol{Q}_{SOM}(j) = [\ Q^T_1 \mid Q^T_2 \mid ... \mid Q^T_j\ ] \tag{4}$$

The replacement process for when $j > \tau$ is described in a later section. Each Q^T_j is built and appended to Q_{SOM} as the frame number increments. Each Q^T_j itself has the following form:

$$Q^T_j = [\ H^T_j \mid (H_j F_j)^T \mid ... \mid (H_j F^{n-1}_j)^T\ ] \tag{5}$$

where F is the process matrix that relates the motion model defined in (1) from one state to the next. H is the measurement Jacobian where the derivatives are arranged as follows (with features and anchors abbreviated as "feat." and "anch." respectively):

$$\begin{bmatrix} \text{image feat. w.r.t. camera state} & \text{image feat. w.r.t. world feat.} \\ \text{image anch. w.r.t. camera state} & 0 \end{bmatrix}$$

Once Q_{SOM} is computed, the temporal observability scores for each feature is computed as:

$$score(f) = \sigma_{min}(Q_{SOM}\ for\ feature\ f) \tag{6}$$

Rank-k Temporal Update of Observability Score: As can be seen from the preceding section, the computation of Q_{SOM} is an extensive operation. To simplify matters the authors present an iterated SVD approach to creating Q_{SOM} when $j > 3$ and $j < \tau + 1$.

Fundamentally, Q_{SOM} is created via (4) for the first three frames. For subsequent frames, the SVD of Q_{SOM} is combined with the corresponding calculated Q_j in (6) to form the new Q_{SOM}. This process involves the redefinition and diagonization of several new matrices, the details of which are too lengthy to include here. For exact details, please refer to section 2.3 of [18].

For $j > \tau + 1$, Q_{SOM} is updated via replacement. The newest Q_j in (6) is computed and placed in the positions occupied by the oldest elements of the matrix. As an example, if at time k, $Q_{SOM}(k) = \left[\; Q_1^T \mid Q_2^T \mid ... \mid Q_k^T \; \right]$, then at time $k + 1$, the oldest (*i.e.* the first element) is replaced, $Q_{SOM}(k+1) = \left[\; Q_{k+1}^T \mid Q_2^T \mid ... \mid Q_k^T \; \right]$.

Submodular Learning for Feature Grouping: If the set of anchors created after the thresholding of the feature observability scores is smaller than desired, a greedy submodular learning algorithm upgrades features with with the next best scores to anchors to complete the anchor set. This is done by creating an SOM of just the anchors and candidate rows for each remaining feature. The algorithm iterates through each candidate feature, appending it's row to the end of the anchor SOM and then taking the SVD of that matrix to get the observability score for that feature via (6) with the modification that Q_{SOM} is now the anchor and candidate row SOM. The feature whose row produces the maximum observability score is upgraded to be an anchor and removed from the set of features. This greedy algorithm recomputes the anchor and candidate SOMs and repeats the process until the the desired number of anchors is obtained.

4.4. Experimental Method

The feature reduction algorithm was implemented on readily available simulated VSLAM data that was created for [15]. The robot is made to move in a circle in a simulated environment of dimensions 12mX12m with 72 landmarks forming a square. The VSLAM process uses an Extended Kalman Filter (EKF) to estimate the locations of each landmark and subsequently estimates its own position in the map. In the original EKF algorithm, the simulated robot detects as many landmarks that it can detect. The authors apply their SVD ranking approach to select a certain subset of the number of features located. Experimentally, the authors parameterize the number of anchors used as well as the noisiness of the system, and compare their generated results to that of the original system. They simulation is run over 40 trials and count the number of times their algorithm outperforms the original. The noise values used are $\sigma = \{0.5, 1.0, 1.5, 2.0, 2.5\}$. The number of features used are $K = \{3, 4, 5, 6, 7, 8, 10, 12\}$. The results are provided in Table 1.

4.5. Our Implementation

We chose to re-implement the method and key experiment of [18]. The first step was to run the original EKF-VSLAM algorithm in [15] that was used by the authors in their experimentation. The code for [15] was readily available to download and use online.

Algorithm Structure: The algorithm simulates a robot moving in a circular trajectory surrounded by 72 detectable landmarks. A monocular camera is placed on the robot for visual sensing purposes. Broadly the algorithm is divided in to the following sections:

1. *Data Generation*: The trajectory to be followed by the robot is generated. Each landmarks's position is also generated. Care is taken to seed the random number generator at a fixed point so as to ensure consistency across measurements.

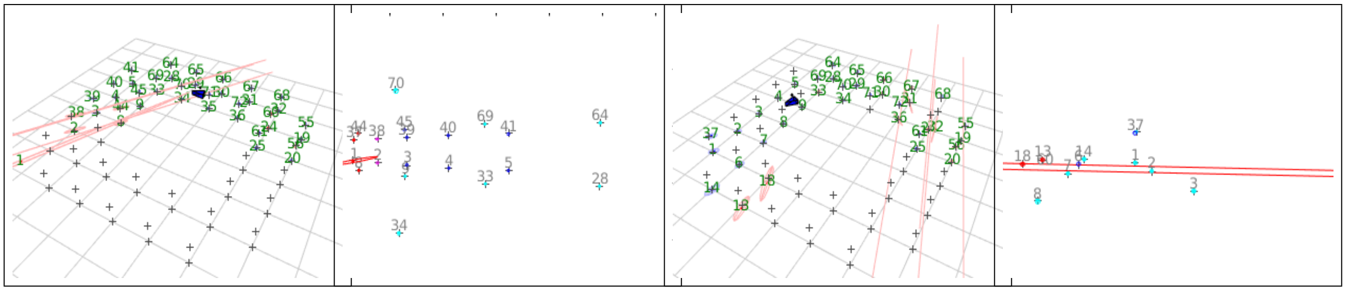
2. *Simulation*: The simulated robot is moved ahead on the predefined trajectory. The measurement vectors are generated based on the location of the landmarks located by the robot. In the terms of our implementation, we consider the features to be these landmark positions.

3. *Localization*: The robot's current position and orientation is localized by sending all the detected features to an Extended Kalman Filtering Algorithm.

4. *Visualization*: The corresponding position of the landmarks, detected features, actual robot's position and orientation and estimated robot's position and orientation are visually updated on a three dimensional plot.

Calculating $\boldsymbol{Q_{SOM}}$: The first step in generating the observability matrix is mapping features detected by the SLAM algorithm to their corresponding $\boldsymbol{p}_i^{R_k}$'s from (2) and $\boldsymbol{h}_{x_i}$ from (3). This is done between step 2 and 3 by extracting the corresponding elements from the algorithms pinhole perspective calculations. The camera's corresponding position and orientation, $\boldsymbol{x}_{R_k}^W$, is also extracted by (1). These features are used to create a corresponding H matrix which in turn is used to iteratively build up Q_{SOM} as described in (4) and (6). Q_{SOM} is created to contain information from 10 iterations of the EKF algorithm at a time, *i.e.* $\tau = 10$. As described previously, Q_{SOM} is initially created iteratively, then via the iterative SVD and finally via the replacement method.

Feature Scoring: Step two provides us with a list of features that has been detected by the robot. Each of these features have corresponding rows in the Q_{SOM} matrix. For each feature, the corresponding rows are iteratively extracted and concatenated to form a feature matrix. The SVD is computed for each of these matrices and the minimum positive singular value is extracted as a corresponding score



(a) VSLAM (all features) (b) Pinhole View (all features) (c) VSLAM (10 ranked feat.) (d) Pinhole View (10 ranked feat.)

Figure 1: Experimental Results: (a) VSLAM is performed using all the features. All the landmarks (+ signs) are detected by the robot and used for localization. (b) Shows the pinhole view of the robot in the first scene. These pixel coordinates form the $\boldsymbol{h}_{x_i}$ that (3) used in the algorithm. (c) VSLAM is performed with the the 10 highest ranked features. In the figure this is apparent from the fact that all the landmarks in the vicinity of the robot are not labelled. (d) Corresponding pinhole view. As expected, fewer features are being used for the localization process.

of each feature by (4). The resultant scores are then used to sort the original features as well as the corresponding $\boldsymbol{p}_i^{R_k}$'s and $\boldsymbol{h}_{x_i}$'s. A subset of the scored features with scores above an application-specified threshold is designated as the anchor set and is passed on to the localization step and the algorithm proceeds as normal.

Submodular Learning: We implemented submodular learning to complete the anchor set as described by the paper's proposed greedy algorithm. This portion of the implementation also runs only if the resultant set of anchors after thresholding the feature scores is less than two below the desired number of anchors.

Experimental Implementation: The experiment described in [18] is recreated exactly. We run the experiment over 40 trials for different values of pixel noise and a chosen number of features. The following values are used to generate the 40 trials: $\sigma = \{0.5, 1.0, 1.5, 2.0, 2.5\}$ and $K = \{3, 4, 5, 6, 7, 8, 10, 12\}$. Each trial runs for 200 video frames. The positional and Euclidian L2-norms are used to compare the the generated results to that of the orignal EKF method.

Experimental Results: Figure 1 shows the results of the experiment. Table 1 displays the number of times the authors were succesful in beating the EKF method using their approach. Table 2 displays the positional norm achieved by the EKF method as compared to our algorithm over noise values. Table 3 displays the Euclidian norm achieved by the EKF method as compared to our algorithm over noise Our. values implementation performed better then the original in 8/40 cases (20%) in terms of the positional norm. In terms of the Euclidian norm, it performed better in 14/40 cases (35%). These results are tabulated in Table 4.

Metric Used	Num. Successful	Accuracy
Translation Error $\sum_k \lVert\Delta \boldsymbol{r}^W_{R_k}\rVert_2$	37/40	92.5%
Rotation Error $\sum_k \lVert\Delta \boldsymbol{\theta}^W_{R_k}\rVert_2$	33/40	82.5%

Table 1: Author's experimental results.

σ	Avg. $(\sum_k \lVert\Delta \boldsymbol{r}^W_{R_k}\rVert_2)_{all}$	Avg. $(\sum_k \lVert\Delta \boldsymbol{r}^W_{R_k}\rVert_2)_{rnk}$
0.5	6.3591	7.0524
1.0	4.8382	8.9654
1.5	6.1305	10.1870
2.0	6.5744	7.0839
2.5	6.4082	8.5952

Table 2: Position Based Results

σ	Avg. $(\sum_k \lVert\Delta \boldsymbol{\theta}^W_{R_k}\rVert_2)_{all}$	Avg. $(\sum_k \lVert\Delta \boldsymbol{\theta}^W_{R_k}\rVert_2)_{rnk}$
0.5	0.7082	7.2744
1.0	0.9237	5.8175
1.5	0.8389	8.5750
2.0	7.9474	4.7680
2.5	8.0217	6.1176

Table 3: Euclidian Norm Based Results

4.6. Subsequent Conclusions

As can be seen, our results do not agree with those in [18] though the experiment (including all procedures, parameters, and metrics) was reproduced exactly. We believe the reason for this an incorrect creation of the Q_{SOM} ma-

Metric Used	Num. Successful	Accuracy
Translation Error $\sum_k \|\|\Delta \boldsymbol{r}^W_{R_k}\|\|_2$	8/40	20.0%
Rotation Error $\sum_k \|\|\Delta \boldsymbol{\theta}^W_{R_k}\|\|_2$	14/40	35.0%

Table 4: Our experimental results.

trix. This paper turned out to be particularly difficult to re-implement. The main reason behind this was the convoluted mathematical definitions and notations used by the authors especially with regard to the contents and creation of Q_{SOM}, including each feature's temporal contribution of a row to the matrix and the correspondence of the matrix's singular values to individual features. We tried to contact the authors via email but are unfortunately yet to hear back from them. They also do not provide any of the software or code used.

Based on the metrics they define and the empirical results they provide, the author's feature scoring algorithm appears to successfully aid VSLAM processes. Unfortunately, their results do not appear to be easy to reproduce. For this algorithm to achieve more success in the community it is our suggestion that the authors clean up their mathematical notation as well as release commented software and code which performs the algorithm successfully.

5. Conclusion

The vision community has done much research in the area of VSLAM and its various aspects. Through this paper, we present an analysis of three papers' innovative methods: large-scale monocular VSLAM that is robust to observed data size in [3], KinnectFusion-based VSLAM that relies on 3D object recognition and prior models of those objects in [14], and the selection of a subset of detected features to use for VSLAM based on their temporal observability in [18]. We also present our implementation of GF-SLAM from [18] and the results and conclusions drawn from our attempt to recreate the main experiment it was originally tested on.

References

[1] G. A., P. Lenz, S. C., and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[2] H. Bay, T. Tuytelaars, L. Van Gool, and R. Hartley. Surf: Speeded up robust features. In *Computer Vision - ECCV 2006*, pages 401–417. Springer, 2006. 2

[3] G. Bourmaud and R. Megret. Robust large scale monocular visual slam. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1638–1647, June 2015. 1, 2, 7

[4] A. J. Davison, Y. G. Cid, and N. Kita. Real-time 3d slam with wide-angle vision. In *IFAC/EURON Symposium on Intelligent Autonomous Vehicles*, 2004. 4

[5] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 3

[6] J. Engel and D. Schops, T. nad Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV, Lecture Notes in Computer Science*, pages 834–849, 2014. 2

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[8] D. Goshen-Meskin and I. Bar-Itzhack. Observability analysis of piece-wise constant systems. i. theory. *IEEE Transactions on Aerospace and Electronic Systems*, 28(4):1056–1067, 1992. 4

[9] H. Lim, J. Lim, and H. J. Kim. Real-time 6-dof monocular visual slam in a large-scale environment. In *ICRA*, 1991. 2

[10] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davision, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinnectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 3

[11] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):756–777, 2004. 2

[12] C. Olsson, A. Eriksson, and R. Hartley. Outlier removal using duality. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1450–1457, June 2010. 1

[13] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *Proceedings of the IEEE International Workshop on 3D Digital Imaging and Modeling (3DIM)*, 2001. 3

[14] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, June 2013. 1, 2, 3, 7

[15] J. Sola, T. Vidal-Calleja, J. Civera, and J. M. M. Montiel. Impact of landmark parametrization on monocular ekf-slam with points and lines. *International Journal of Computer Vision*, 97(3):339–368, 2012. 5

[16] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgbd slam systems. *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ Interanational Conference on.*, pages 573–580, 2012. 2

[17] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 2

[18] G. Zhang and P. A. Vela. Good features to track for visual slam. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1373–1382, June 2015. 1, 3, 4, 5, 6, 7