Towards Fast and Automatic Map Initialization for Monocular SLAM Systems

Blake Troutman[a] and Mihran Tuceryan[b]
Department of Computer and Information Science, Indiana University - Purdue University Indianapolis, Indianapolis, U.S.A.

Keywords: Simultaneous Localization and Mapping (SLAM), Structure From Motion, Map Initialization, Model Selection, Monocular Vision, Stereo Vision.

Abstract: Simultaneous localization and mapping (SLAM) is a widely adopted approach for estimating the pose of a sensor with 6 degrees of freedom. SLAM works by using sensor measurements to initialize and build a virtual map of the environment, while simultaneously matching succeeding sensor measurements to entries in the map to perform robust pose estimation of the sensor on each measurement cycle. Markerless, single-camera systems that utilize SLAM usually involve initializing the map by applying one of a few structure-from-motion approaches to two frames taken by the system at different points in time. However, knowing when the feature matches between two frames will yield enough disparity, parallax, and/or structure for a good initialization to take place remains an open problem. To make this determination, we train a number of logistic regression models on summarized correspondence data for 927 stereo image pairs. Our results show that these models classify with significantly higher precision than the current state-of-the-art approach in addition to remaining computationally inexpensive.

1 INTRODUCTION

Simultaneous localization and mapping (SLAM) has become a popular approach for estimating the pose of a sensor with respect to its environment in real time, while simultaneously building and improving the map. The SLAM process begins by using measurements from a sensor to initialize a virtual map of the observed environment. After the map is initialized, SLAM systems use the virtual map in correspondence with the sensor measurements to repeatedly estimate the pose of the sensor at regular intervals. This pose can either be used in conjunction with sensor measurements to continually increase the detail of the virtual map or it can be forwarded to an external system that needs to know the current sensor's pose. This process of pose estimation is necessary for both virtual reality (VR) and augmented reality (AR) systems, in which a user moves a device (often a head-mounted display or a view finder) and expects to see a virtual world or virtual objects transformed appropriately, in real time, based on the user's motion. The camera pose is also useful in robotics, as it can be used as a method to localize a robot in its environment.

This paper addresses the problem of automatically determining the best pair of image frames to be used in initializing the visual SLAM process. The means of map initialization depends on the SLAM system's sensor configuration and whether or not markers are used in the environment. SLAM systems that utilize depth-sensing cameras or calibrated stereo camera systems can initialize their virtual maps in a markerless environment relatively quickly and easily compared to markerless monocular (single-camera) SLAM systems. Markerless monocular SLAM systems typically initialize their maps by sampling two image frames at different points in time and matching salient image features between the two images to develop a set of point correspondences. After this, these correspondences can be used to estimate the fundamental matrix, the essential matrix, or a homography between the camera poses of these two images (Hartley and Zisserman, 2004). These matrices can be used to extract the camera pose of the second frame with respect to the first, and then the correspondences can be used with this pose estimate to triangulate the first set of 3D map points.

[a] https://orcid.org/0000-0002-3809-3180
[b] https://orcid.org/0000-0003-3828-6123

Troutman, B. and Tuceryan, M.
Towards Fast and Automatic Map Initialization for Monocular SLAM Systems.
DOI: 10.5220/0010640600003061
In *Proceedings of the 2nd International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS 2021)*, pages 22-30
ISBN: 978-989-758-537-1

Many of the steps of this monocular map initialization process are well known; however, fast automatic selection of the image pairs to use for the initialization process still remains an open problem. If the image pairs do not have a sufficient translation component in the camera movement or if the correspondences are not sufficiently spread across the images, then the pose estimate is likely to be inaccurate and the map initialization will cause the SLAM system to immediately fail. This work presents a fast approach for automatically determining if an image pair is likely to result in an accurate reconstruction by extracting various information from the collection of correspondences and feeding that information to trained logistic regression models.

1.1 Contribution

Given the problem of selecting appropriate image pairs from a video sequence for visual SLAM map initialization, previous works are too slow for resource-limited platforms, lack robustness, only work for non-planar scenes, or struggle in their decision-making when observing pure rotations. This work contributes a solution that attempts to address all of these issues. Namely, this work contributes an algorithm for selecting good image pairs from a video sequence for visual SLAM map initialization that:

- is fast enough to be practical for real time use on resource-limited platforms,
- provides decisions for both planar and non-planar scenes,
- accurately rejects image pairs that demonstrate pure rotational movement,
- and yields higher precision than the current state-of-the-art approach.

2 RELATED WORK

There are a number of visual SLAM systems for real time use that have been developed over the past several years. MonoSLAM (Davidson et al., 2007) is the first successful application of purely visual monocular SLAM, but its map initialization is aided by the use of a known fiducial marker. During initialization, markers can be used to act as a pre-existing map with known 3D point locations. This enables the system to accurately estimate the camera pose each frame before the system has sufficient baseline length to triangulate new points into the map. Markers also enable the system to register an accurate scale for their map points during initialization, as demonstrated in (Xiao et al., 2017). The use of markers is further expanded upon in systems such as (Korah et al., 2011), (Maidi et al., 2011), (Ufkes and Fiala, 2013), (Kobayashi et al., 2013), and (Arth et al., 2015) which utilize fiducials for pose estimation throughout the entire runtime of the system, rather than just using them for map initialization.

Though the use of markers often makes camera pose estimation more reliable and straightforward, it also makes pose estimation systems less versatile, as the systems will break down and have no reference points to measure if the fiducial markers fall out of frame. (Berenguer et al., 2017) proposes a SLAM system that heavily utilizes global image descriptors to perform localization, which frees the system from any use of fiducial markers. However, the global image descriptors used in the aforementioned work depend on omnidirectional imaging and can take more than half of a second to perform localization, making them unsuitable for systems that require rapid, real time tracking (such as in AR/VR systems or in autonomous automobiles). Faster and more robust pose estimation systems exclude the need for markers by tracking somewhat arbitrary patches of the image frame, without any predefined knowledge of what the patches look like or where they should be located in 3D space (Klein and Murray, 2007), (Klein and Murray, 2009), (Sun et al., 2015), (Mur-Artal et al., 2015), (Fujimoto et al., 2016), and (Qin et al., 2018).

Without knowledge of the 3D locations of the tracked points, structure-from-motion (SfM) algorithms are often used to aid in triangulating virtual 3D locations for these points. For example, the PTAM system (Klein and Murray, 2007), (Klein and Murray, 2009) is a real time visual SLAM system that initializes its map by having the user manually select two frames (from different points in time), matching FAST features (Rosten and Drummond, 2006) between the two frames, estimating the pose difference between the frames with the 5-point algorithm (Stewenius et al., 2006) or with a homography decomposition (Faugeras and Lustman, 1988), and finally using the estimated pose to triangulate the point matches into a virtual 3D map. (Sun et al., 2015) adopts the PTAM framework for its map initialization as well. The clear drawback of this approach is that it is not automatic; it requires the user to select the frames that should be used for initialization. (Huang et al., 2017) presents an approach for map initialization that can initialize the map in a single frame without user intervention, but it requires that the system is operating inside of a typical indoor room and can only initialize map points that coincide with the walls of the room.

ORB-SLAM (Mur-Artal et al., 2015) also utilizes SfM techniques to initialize its map, but it additionally implements an automated approach for selecting the frames to use for initialization while maintaining comparable versatility to PTAM. ORB-SLAM achieves this automatic frame selection by estimating both a homography and a fundamental matrix for the current frame against the initial frame using the normalized DLT algorithm and the 8-point algorithm, respectively, both of which are explained in (Hartley and Zisserman, 2004). It then uses a heuristic to assess the success of each model estimate— if the heuristic indicates that neither model estimate is accurate, then the system repeats this process on subsequent frames until at least one of the models appears to provide an accurate pose estimate. This approach, however, is computationally expensive and unsuitable for resource-limited platforms such as mobile phones. There are modified versions of ORB-SLAM, some of which aim to revise the map initialization process (Fujimoto et al., 2016), but stray from monocular SLAM by requiring RGB-D data.

Given the high computational cost of ORB-SLAM's approach to automatic frame selection, it would be desirable for a SLAM system to include some algorithm that can select an appropriate pair of frames for its SfM-based map initialization much more quickly. There are relatively few works that have presented such an algorithm. A work by Tomono (Tomono, 2005), which predates many of the visual SLAM systems in the literature, presents an approach for detecting degeneracy in a potential fundamental matrix estimate before a system would need to perform the 8-point algorithm. However, this approach is only applied to fundamental matrix estimates, it does not distinguish between degeneracy caused by insufficient translation and degeneracy caused by planar scenes (which could still be useful for a homography estimate), and the threshold parameter used is dependent upon the number of correspondences, making this approach less robust for real use. VINS-Mono (Qin et al., 2018), a visual-inertial SLAM system, utilizes a simpler frame selection approach in which the system chooses to attempt initialization with a pair of frames if it is able to track more than 30 features between them with each correspondence yielding a disparity of at least 20 pixels. A later SLAM system (Butt et al., 2020) implements a similar approach to that used in VINS-Mono, but utilizes the median and standard deviation of correspondence pixel disparities to additionally select whether a homography estimate is appropriate or if a fundamental matrix estimate is appropriate. However, both of these systems are still dependent on an evaluation of the estimated model accuracy to reject erroneous initial estimates and they also give an incorrect decision under pure rotational movements.

3 METHODS

3.1 Structure-from-Motion (SfM) Approaches

The goal of stereo SfM methods is to both derive the transformation between two camera poses and to project image data back into 3 space using the derived pose transformation. The cameras are modeled using a pinhole camera model,

$$\mathbf{x} = \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{X} \tag{1}$$

where $\mathbf{x}$ is a homogeneous 2D projection of the homogeneous 3D point $\mathbf{X}$, $\mathbf{R}$ is the 3×3 rotation matrix of the viewing camera, $\mathbf{t}$ is column vector of the translation of the viewing camera, and $\mathbf{K}$ is the camera intrinsics matrix, defined by

$$\mathbf{K} = \begin{bmatrix} f & s & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where f is the camera's focal length, s is the skew (zero in our implementation), and c_x and c_y are the image center coordinates. It is assumed that the first camera is at the origin with no rotation, $[\mathbf{I}|\mathbf{0}]$, so $[\mathbf{R}|\mathbf{t}]$ represents the extrinsic camera parameters of the second viewing camera, i.e., the transformation of the pose from the first camera to the second camera. These extrinsic parameters are derived by estimating the fundamental matrix, the essential matrix, or a homography between the two views.

Given two views of the same scene from different camera viewpoints, epipolar geometry can be used to determine the transformation from the first camera's pose to the second camera's pose. Epipolar geometry refers to the property of stereo image projection in which 3D points observed in one view can be constrained to unique epipolar planes in 3 space, each of which intersects with the baseline of the cameras. These planes can then be projected into the other view as epipolar lines. Each of these lines intersects with the corresponding projection of the respective 3D point that was observed in the first view. The fundamental matrix of two views is a 3×3 matrix which encodes this constraint for all points projected in the two views. Specifically, this constraint takes the form of

$$\mathbf{x'^T F x} = 0 \tag{3}$$

where $\mathbf{F}$ is the 3×3 fundamental matrix of the two views, and $\mathbf{x}$ and $\mathbf{x}'$ are homogeneous 3×1 column vectors of the projections of a 3D point in the first view and second view, respectively.

If the intrinsic camera parameters are known for both views, then the fundamental matrix can be transformed into the corresponding essential matrix with

$$\mathbf{E} = \mathbf{K'^T F K} \tag{4}$$

where $\mathbf{K}$ and $\mathbf{K}'$ are the camera intrinsic matrices for the first view and second view, respectively.

The extrinsic camera parameters $[\mathbf{R}|\mathbf{t}]$ of the second view are extracted from the essential matrix by using the singular value decomposition of $\mathbf{E}$ and a single correspondence pair to test for cheirality (the constraint that dictates that points viewed by a camera must fall in front of the camera). This process is explained in detail in (Hartley and Zisserman, 2004).

The fundamental matrix is solved with eight or more image point matches by re-formulating the constraint in Equation 3 into a set of linear equations. This approach is known as the 8-point algorithm. The essential matrix is solved using a more complex approach known as the 5-point algorithm, described in (Nistér, 2003).

Additionally, the transformation from one pose to another can be described with a homography, $\mathbf{H}$, which is a 3×3 matrix that represents the projective transformation from one image point to its projection in another image. Namely, a homography is a matrix for a pair of views such that

$$\mathbf{H x} = \mathbf{x}' \tag{5}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are homogeneous vectors representing 3D points in the first view and second view, respectively. A homography between two views is estimated with four or more image matches using the direct linear transformation method, explained in (Hartley and Zisserman, 2004). The extrinsic camera parameters are then extracted from the homography using the approach described in (Faugeras and Lustman, 1988).

Visual SLAM systems often estimate either the fundamental matrix, the essential matrix, or a homography between two frames of the video feed to initialize their maps. The 8-point algorithm can accurately estimate the fundamental matrix in image pairs where there is complex structure in the corresponding feature points; however, the 8-point algorithm struggles when the projected points are coplanar in 3 space. In contrast, homographies can be estimated with coplanar points to yield highly-accurate poses, but their accuracy degrades if the points are not coplanar. The 5-point algorithm that estimates the essential matrix can be accurate for both coplanar points and non-coplanar points, but it often yields poses that are less precise than those that are achieved with the 8-point algorithm or 4-point DLT algorithm. To utilize these SfM approaches, we use the corresponding OpenCV (Bradski, 2000) implementations for each algorithm.

3.2 Model Configuration

This approach utilizes four logistic regression models for decision-making. Each model takes in correspondence data as input and answers a different question regarding the data's likelihood of resulting in a good map initialization. Specifically, the models aim to answer the following questions:

- Model F: Would estimating the fundamental matrix (via 8-point algorithm) be a good approach for initializing the SLAM map with this correspondence data?
- Model E: Would estimating the essential matrix (via 5-point algorithm) be a good approach for initializing the SLAM map with this correspondence data?
- Model H: Would estimating a homography (via 4-point DLT algorithm) be a good approach for initializing the SLAM map with this correspondence data?
- Model R: Is this correspondence data the result of a pure rotation (no translation in camera movement)?

Each model provides an output of 1 if the answer to its respective question is "yes" and 0 if the answer is "no." In practice, a SLAM system could utilize these models for map initialization in real time by continuously evaluating each frame against a reference frame taken at a previous point in time. For example, the first frame that is registered upon the system's activation could be considered the reference frame, and then, each frame afterward could be paired with this reference frame and the pair could be tested against these models. When at least one of the models predicts that the two frames in question are sufficient for SfM estimation, the corresponding SfM approach is used to initialize the map with the related correspondences.

Additionally, all four models take the same 23 features as inputs. The features consist of data pertaining to the correspondences between the image pairs being evaluated. A correspondence is defined as a 2D-to-2D keypoint match between two images and consists of two image points, defined as $p_0 = (x_0, y_0)$ and $p_1 = (x_1, y_1)$, where p_0 is a point in the first image

and p_1 is its matching location in the second image. We define the pixel disparity of a correspondence as the euclidean distance between its p_0 and p_1 values.

Specifically, the features used include the number of correspondences, the mean and standard deviation of the pixel disparities, the minimum and maximum values of each component of the correspondences (x_0, x_1, y_0, and y_1), and the range of each component of the correspondences. Additionally, we include eight features that each represents the proportion of correspondences that approximately yield a specific angle. We calculate a correspondence's angle, θ, as

$$\theta = \arctan\left(\frac{y_1 - y_0}{x_1 - x_0}\right) \tag{6}$$

where (x_0, y_0) is the point in the first image and (x_1, y_1) is the corresponding point in the second image. The angles included in these eight features simply cover 8 evenly distributed angles between 0°and 360°.

3.3 Model Training

Models are trained on data that is derived from the *fr3/structure_texture_far* RGB-D dataset from the Technical University of Munich (Sturm et al., 2012). This dataset consists of an RGB video of textured papers on the ground with a poster sitting upright behind the papers, bent in a zig-zag fashion to produce structure. Throughout the video, the camera demonstrates both rotational and translational movements, mostly moving to the left and rotating to keep the objects in frame. For model training, the video is split up into batches of 90-frame sequences. For each batch, every frame is paired with the first frame in the batch to form a stereo image pair. We use the KLT optical flow algorithm (Shi and Tomasi, 1994), (Lucas and Kanade, 1981) to track approximately 150 to 300 features each frame. This optical flow approach allows the resulting correspondences to be free of significant outliers. After collecting these correspondences on each stereo image pair, we estimate the pose of the second camera with respect to the first using the 8-point algorithm, 5-point algorithm, and 4-point DLT algorithm. This provides us with three different pose estimates, each of which is used to judge the success of the respective approach on the stereo image pair by using the ground truth pose data provided in the dataset.

We find that two error metrics are good descriptors of successful pose estimates. These metrics include the normalized translational chordal distance and the median scaled reconstruction error. The normalized translational chordal distance, E_t, is defined by

$$E_t = \|\mathbf{t}_g - \mathbf{t}_e\|_F \tag{7}$$

where $\mathbf{t}_g$ and $\mathbf{t}_e$ are the normalized translation vectors of the ground truth pose and the estimated pose, respectively, and E_t is the Frobenius norm of their difference. The median scaled reconstruction error, E_R, is defined by

$$E_R = \frac{\sum_i^n \|\mathbf{P}_{g_i} - s\mathbf{P}_{e_i}\|_F}{n} \tag{8}$$

where $\mathbf{P}_g$ is a 3D ground truth point, $\mathbf{P}_e$ is the corresponding 3D point that was triangulated from the estimated pose, and n is the number of correspondences. Additionally, this metric includes a scale factor, s, which is defined by

$$s = \frac{\|\mathbf{P}_{g_0} - \mathbf{P}_{g_1}\|_F}{\|\mathbf{P}_{e_0} - \mathbf{P}_{e_1}\|_F} \tag{9}$$

where s can be described as the ratio of the distance between the first two points of the ground truth to the distance between the corresponding triangulated points. This scale factor allows us to correct for scale ambiguity in the map generation.

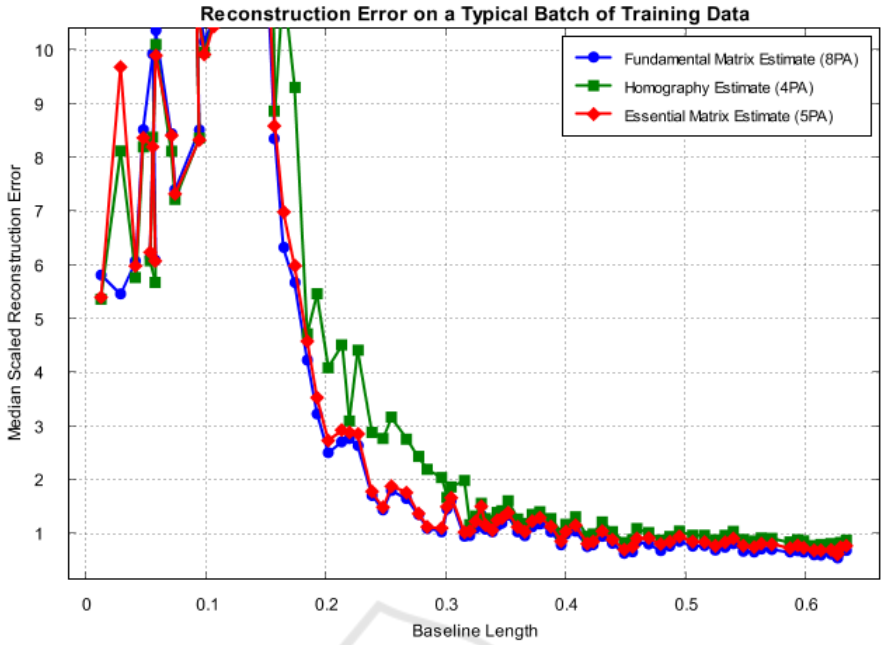
These error metrics are used to aid in generating labels for the training data. As the baseline between the cameras increases in these tests, both of these error metrics show a rapidly decreasing error followed by a plateau at some baseline length on most of the batches, as shown in Figure 1. Using this trend, thresholds for each error metric are selected empirically. We determine that a pose estimate could be considered suitable for SLAM initialization if its normalized translational chordal distance was less than 0.8 and its median scaled reconstruction error is less than 1. This results in the following labelling criteria:

$$\text{label}(I) = \begin{cases} 1 & E_t < 0.8 \text{ and } E_R < 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

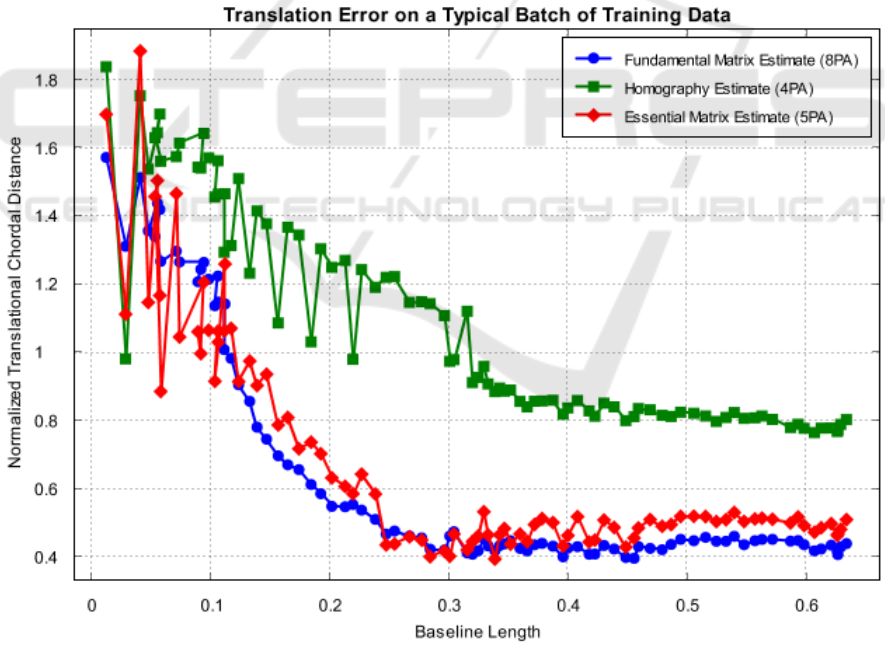
where I is a set of correspondences for a stereo image pair.

For Model R, the classifier that predicts when correspondences are the result of a pure rotation, we simply use the ground truth data to determine the baseline length and consider the label to be 0 when the baseline length exceeds 0.05 and 1 otherwise.

In total, the training data consists of the correspondence data from 927 different stereo image pairs.



(a) Median Scaled Reconstruction Error



(b) Normalized Translational Chordal Distance

Figure 1: Error metrics for pose estimates of SfM approaches on a typical batch of training data generated from fr3/structure_texture_far. (a) is the median scaled reconstruction error as the baseline increases throughout the batch. (b) is the normalized translational chordal distance as the baseline increases throughout the batch. As the ground truth baseline increases, both error metrics show a rapid decrease followed approximately by a plateau on each SfM approach. Correspondence sets are considered sufficient for SLAM initialization with a given SfM approach if both errors fall on the plateaus. Note that the median scaled reconstruction error in (a) is unstable at low baselines and may have an error upwards of 10^3 (which is cut off here to show the more general pattern).

4 EXPERIMENTAL RESULTS

To evaluate model performance, we construct a test set from the *fr3/structure_texture_near* dataset, also provided by (Sturm et al., 2012). This dataset is similar to the *fr3/structure_texture_far* dataset, as it is a video of the same scene; however, it is a much more challenging dataset for SfM as the camera is closer to the objects. This causes keypoints to move off-frame more quickly and also leads to a far less homogeneous distribution of keypoints in many parts of the video. The test set is derived in the same fashion as the training set, where the video is split into several batches and each frame is paired with the first frame of its respective batch. Due to keypoints moving off-frame more quickly in this set, we only use batches of 30 frames instead of 90. The resulting test set consists of correspondence data for 1,062 different stereo image pairs.

Each classifier is able to run inference on the entire test set in approximately 1 millisecond, which demonstrates the high speed of this approach and its real time viability, even for resource-limited platforms. The accuracy, precision, recall, and F1 score for each of the classifiers is shown in Table 2.

For comparison, we also evaluate the current state-of-the-art approach, which involves purely evaluating the mean/median and standard deviation of pixel disparities in the correspondences to make the classification. In this comparison, we utilize the thresholds used in (Butt et al., 2020), where the correspondences are considered suitable for homography estimation if the median disparity is above 50 pixels and are considered suitable for fundamental matrix estimation if, additionally, the standard deviation of the disparities is above 15 pixels. The results for the state-of-the-art approach are shown in Table 1.

The precision of the classifiers is the most important metric here, as false positives would lead to a SLAM system attempting to initialize under poor conditions and may cause the system to perform a high amount of wasted computation and/or initialize with a corrupted map. The logistic regression models for the fundamental matrix approach and homography approach both yield higher precision than the current state-of-the-art classification algorithm. Specifically, the precision for the fundamental matrix approach increased from 0.0811 in the state-of-the-art approach to 0.3639 in the logistic regression model, and the precision for the homography approach increased from 0.2852 in the state-of-the-art approach to 0.3725 in the logistic regression model. The logistic regression model for the homography approach has a significantly worse recall than the state-of-the-art approach (dropping from 0.4601 to 0.1166), though this is less crucial as it indicates a higher proportion of false negatives rather than false positives.

Additionally, we evaluate the models' performance after filtering the samples with the pure rotation classifier (Model R). When a sample is detected to be the result of a pure rotation, its label is automatically set to 0, i.e., unsuitable for map initialization. Otherwise, the sample is passed on to the other models for prediction. The results of this approach are displayed in Table 3.

Using the rotation classifier as a preliminary model to immediately reject samples that do not yield sufficient baseline increases the precision of Model F from 0.3639 to 0.4370 and increases the precision of Model E from 0.4290 to 0.5041. This approach also decreases the precision of Model H from 0.3725 to 0.3600. In practice, however, this approach of using the rotation classifier for early rejection does not need to be used for all models—it can be utilized for only Model F and Model E if those are the only models that are improved by this alteration.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an approach for quickly predicting the success of SfM techniques on an arbitrary set of image point correspondences through the use of logistic regression models. The models used in this paper yield significantly higher precision than the current state-of-the-art approach without having to perform the computationally-expensive SfM techniques before making their determinations. The model features used are also very quick to calculate, making this approach viable for real time performance on resource-limited platforms.

This work may further be validated by implementing it in a real time visual SLAM system and evaluating the system's stability. Additionally, the models may be improved by labelling the training data using a more robust evaluation criterion. Rather than using empirical thresholds for the normalized translational chordal distance and median scaled reconstruction error, perhaps ORB-SLAM's (Mur-Artal et al., 2015) heuristic approach could provide a more accurate scheme for labelling the training/testing data. The models' performance may also be improved if they are expanded into neural networks and trained on more scenes.

Table 1: Performance of the current state-of-the-art classification approach on the fr3/structure_texture_near test data.

Classification	Accuracy	Precision	Recall	F1 Score
Fundamental Matrix	0.7589	0.0811	0.0133	0.0229
Homography	0.7401	0.2852	0.4601	0.3521

Table 2: Performance of each logistic regression model on the fr3/structure_texture_near test data.

Model	Accuracy	Precision	Recall	F1 Score
F	0.6959	0.3639	0.5822	0.4479
E	0.7326	0.4290	0.6453	0.5154
H	0.8343	0.3725	0.1166	0.1776
R	0.8380	0.6091	0.9302	0.7362

Table 3: Model performance on the fr3/structure_texture_near test data after using Model R as a preliminary classifier to aid in classification.

Model	Accuracy	Precision	Recall	F1 Score
F	0.7598	0.4370	0.4622	0.4492
E	0.7815	0.5041	0.5256	0.5146
H	0.8399	0.3600	0.0552	0.0957

REFERENCES

Arth, C., Pirchheim, C., Ventura, J., Schmalstieg, D., and Lepetit, V. (2015). Instant Outdoor Localization and SLAM Initialization from 2.5D Maps. *IEEE Transactions on Visualization and Computer Graphics*, 21(11):1309–1318.

Berenguer, Y., Payá, L., Peidró, A., and Reinoso, O. (2017). Slam algorithm by using global appearance of omnidirectional images. In *Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO,*, pages 382–388. INSTICC, SciTePress.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Butt, M. M., Zhang, H., Qiu, X. C., and Ge, B. (2020). Monocular SLAM Initialization Using Epipolar and Homography Model. *2020 5th International Conference on Control and Robotics Engineering, ICCRE 2020*, pages 177–182.

Davidson, A. J., Reid, I. D., Molton, N. D., and Stasse, O. (2007). MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067.

Faugeras, O. and Lustman, F. (1988). Motion and Structure From Motion in a Piecewise Planar Environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 02.

Fujimoto, S., Hu, Z., Chapuis, R., and Aufrère, R. (2016). Orb slam map initialization improvement using depth. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 261–265.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition.

Huang, J., Liu, R., Zhang, J., and Chen, S. (2017). Fast initialization method for monocular slam based on indoor model. In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2360–2365.

Klein, G. and Murray, D. (2007). Parallel Tracking and Mapping for Small AR Workspaces. *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234.

Klein, G. and Murray, D. (2009). Parallel tracking and mapping on a camera phone. *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 83–86.

Kobayashi, T., Kato, H., and Yanagihara, H. (2013). Novel keypoint registration for fast and robust pose detection on mobile phones. *Proceedings - 2nd IAPR Asian Conference on Pattern Recognition, ACPR 2013*, pages 266–271.

Korah, T., Wither, J., Tsai, Y. T., and Azuma, R. (2011). Mobile augmented reality at the hollywood walk of fame. *2011 IEEE Virtual Reality Conference*, pages 183–186.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, page 674–679, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Maidi, M., Preda, M., and Le, V. H. (2011). Markerless tracking for mobile augmented reality. *2011 IEEE International Conference on Signal and Image Processing Applications, ICSIPA 2011*, pages 301–306.

Mur-Artal, R., Montiel, J. M., and Tardos, J. D. (2015). ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5):1147–1163.

Nistér, D. (2003). An Efficient Solution to the Five-Point Relative Pose Problem. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2:II–195.

Qin, T., Li, P., and Shen, S. (2018). VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. *Computer Vision – ECCV 2006*, pages 430–443.

Shi, J. and Tomasi, C. (1994). Good Features to Track. *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Stewenius, H., Engels, C., and Nister, D. (2006). Recent Developments on Direct Relative Orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60:284–294.

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

Sun, L., Du, J., and Qin, W. (2015). Research on combination positioning based on natural features and gyroscopes for AR on mobile phones. *Proceedings - 2015 International Conference on Virtual Reality and Visualization, ICVRV 2015*, pages 301–307.

Tomono, M. (2005). 3-D localization and mapping using a single camera based on structure-from-motion with automatic baseline selection. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 3342–3347.

Ufkes, A. and Fiala, M. (2013). A markerless augmented reality system for mobile devices. *Proceedings - 2013 International Conference on Computer and Robot Vision, CRV 2013*, pages 226–233.

Xiao, Z., Wang, X., Wang, J., and Wu, Z. (2017). Monocular ORB SLAM based on initialization by marker pose estimation. *2017 IEEE International Conference on Information and Automation, ICIA 2017*, (July):678–682.