Robust Scale Estimation in Real-Time Monocular SFM for Autonomous Driving

Shiyu Song
University of California, San Diego

Manmohan Chandraker
NEC Labs America, Cupertino, CA

Abstract

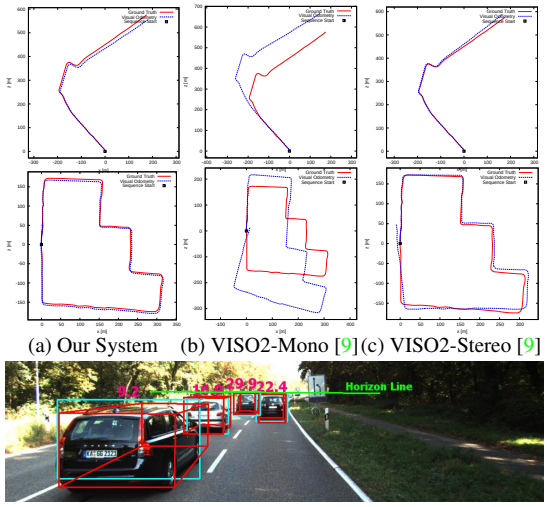
Scale drift is a crucial challenge for monocular autonomous driving to emulate the performance of stereo. This paper presents a real-time monocular SFM system that corrects for scale drift using a novel cue combination framework for ground plane estimation, yielding accuracy comparable to stereo over long driving sequences. Our ground plane estimation uses multiple cues like sparse features, dense inter-frame stereo and (when applicable) object detection. A data-driven mechanism is proposed to learn models from training data that relate observation covariances for each cue to error behavior of its underlying variables. During testing, this allows per-frame adaptation of observation covariances based on relative confidences inferred from visual data. Our framework significantly boosts not only the accuracy of monocular self-localization, but also that of applications like object localization that rely on the ground plane. Experiments on the KITTI dataset demonstrate the accuracy of our ground plane estimation, monocular SFM and object localization relative to ground truth, with detailed comparisons to prior art.

1. Introduction

Vision-based structure from motion (SFM) is rapidly gaining importance for autonomous driving applications. Monocular SFM is attractive due to lower cost and calibration requirements. However, unlike stereo, the lack of a fixed baseline leads to scale drift, which is the main bottleneck that prevents monocular systems from attaining accuracy comparable to stereo. Robust monocular SFM that effectively counters scale drift in real-world road environments has significant benefits for mass-produced autonomous driving systems.

A popular means to tackle scale drift is to estimate height of the camera above the ground plane. We present a data-driven framework for monocular ground plane estimation that achieves outstanding performance in real-world driving. This yields high accuracy and robustness for real-time monocular SFM over long distances, with results comparable to state-of-the-art stereo systems on public benchmark datasets. Further, we also show significant benefits for applications like 3D object localization that rely on an accurate ground plane.

Prior monocular SFM works like [9, 20, 21] use sparse feature matching for ground plane estimation. However, in autonomous driving, the ground plane corresponds to a rapidly moving, low-textured road surface, which renders sole reliance on such feature matches impractical. We overcome this challenge with two innovations in Sec. 4 and 5. First, we incorporate cues from multiple methods and second, we combine them in a framework that accounts for their per-frame relative confidences, using models learned from training data.



(a) Our System (b) VISO2-Mono [9] (c) VISO2-Stereo [9]

Figure 1. Applications of our ground plane estimation. [Top row: Monocular SFM] (a) Scale correction using our ground plane yields monocular self-localization close to ground truth over several kilometers of real-world driving. (b) Our cue combination significantly outperforms prior works that also use the ground plane for scale correction. (c) Our performance is comparable to stereo SFM. [Bottom row: Object localization] Accuracy of applications like 3D object localization that rely on the ground plane is also enhanced.

Accordingly, in Sec. 4, we propose incorporating cues from dense stereo between successive frames and 2D detection bounding boxes (for the object localization application). The dense stereo cue vastly improves camera self-localization, while the detection cue significantly aids object localization. To combine cues, Sec. 5 presents a novel data-driven framework. During training, we learn models that relate the observation covariance for each cue to error behaviors of its underlying variables, as observed in visual data. At test time, fusion of the covariances predicted by these models allows the contribution of each cue to adapt on a per-frame basis, reflecting belief in its relative accuracy.

The significant improvement in ground plane estimation using our framework is demonstrated in Sec. 6. In turn, this leads to excellent performance in applications like monocular SFM and 3D object localization. On the KITTI dataset [8],

our real-time monocular SFM achieves rotation accuracy up to 0.0054° per frame, even outperforming several state-of-the-art stereo systems. Our translation error is a low 3.21%, which is also comparable to stereo and to the best of our knowledge, unmatched by other monocular systems. We also exhibit high robustness directly attributable to accurate scale correction. Further, we demonstrate the benefits of our ground estimation for 3D object localization. Our work naturally complements tracking-by-detection frameworks to boost their localization accuracy – for instance, we achieve over 6% improvement in 3D location error over the system of [1].

To summarize, our main contributions are:

- A novel data-driven framework that combines multiple cues for ground plane estimation using learned models to adaptively weight per-frame observation covariances.
- Highly accurate, robust, scale-corrected and real-time monocular SFM with performance comparable to stereo.
- Novel use of detection cues for ground estimation, which boosts 3D object localization accuracy.

2. Related Work

Stereo-based SFM systems routinely achieve high accuracy in real-time [2, 15]. Several monocular systems have also demonstrated good performance in smaller indoor environments [3, 11, 12]. Successful large-scale monocular systems for autonomous navigation are less extant, primarily due to the challenge of scale drift. Strasdat et al. [22] propose a large-scale monocular system that handles scale drift with loop closure. However, autonomous driving requires real-time scale correction on a per-frame basis.

Prior knowledge of the environment is used to counter scale drift in several monocular SFM systems, such as non-holonomic constraints for wheeled robots [19] or geometry of circular pipes [10]. Like ours, other systems also handle scale drift by estimating camera height above the ground plane [9, 20, 21]. However, they rely on triangulation or homography decomposition from feature matches that are noisy for low-textured road surfaces and do not provide unified frameworks for including multiple cues. In contrast, we achieve far superior results by combining cues from sparse features, plane-guided dense stereo and object detection, in a data-driven framework whose observation covariances are weighted by instantaneous visual data.

To localize moving objects, Ozden et al. [16] and Kundu et al. [13] use simultaneous motion segmentation and SFM. A different approach is that of multi-target tracking frameworks that combine object detection with stereo [5] or monocular SFM [1, 23]. Detection can handle farther objects and together with the ground plane, provides a cue to estimate object scales that are difficult to resolve for traditional monocular SFM even with multiple segmented motions [17]. We note that the utility of our accurate ground plane estimation is demonstrable for any object tracking framework. Indeed, this aspect of our contribution is complementary to existing sophisticated localization frameworks like [1, 23], as established in Sec. 6.4.

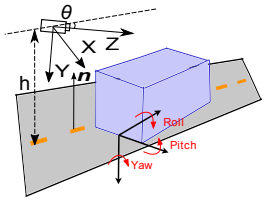


Figure 2. Geometry of ground plane estimation and object localization. The camera height h is the distance from its principal point to the ground plane. The pitch angle is θ and $\mathbf{n}$ is the ground plane normal. Thus, the ground plane is defined by $(\mathbf{n}^\top, h)^\top$.

In contrast to most of the above systems, we present strong monocular SFM results on publicly available real-world driving benchmarks over several kilometers [8] and accurate localization performance relative to ground truth.

3. Background

Notation We denote a vector in $\mathbb{R}^n$ as $\mathbf{x} = (x_1, \cdots, x_n)^\top$. A matrix is denoted as $\mathbf{X}$. A variable x in frame k of a sequence is denoted as x^k.

Monocular SFM Our contributions are demonstrable for any monocular SFM system – as a particular choice, we use the real-time system of [21]. It also uses the ground plane for scale correction, however, relies purely on sparse feature matching. We demonstrate a vast performance improvement by incorporating our novel cue combination framework.

Ground Plane Geometry As shown in Fig. 2, the camera height (also called ground height) h is defined as the distance from the principal center to the ground plane. Usually, the camera is not perfectly parallel to the ground plane and there exists a non-zero pitch angle θ. The ground height h and the unit normal vector $\mathbf{n} = (n_1, n_2, n_3)^\top$ define the ground plane. For a 3D point $(X, Y, Z)^T$ on the ground plane,

$$h = Y \cos\theta - Z \sin\theta. \tag{1}$$

Scale Correction in Monocular SFM Scale drift correction is an integral component of monocular SFM. In practice, it is the single most important aspect that ensures accuracy. We estimate the height and orientation of the ground plane relative to the camera for scale correction.

Under scale drift, any estimated length l is ambiguous up to a scale factor $s = l/l^*$, where l^* is the ground truth length. The objective of scale correction is to compute s. Given the calibrated height of camera from ground h^*, computing the apparent height h yields the scale factor $s = h/h^*$. Then the camera translation $\mathbf{t}$ can be adjusted as $\mathbf{t}_{\text{new}} = \mathbf{t}/s$, thereby correcting the scale drift. In Section 4, we describe a novel, highly accurate method for estimating the ground height h and orientation $\mathbf{n}$ using an adaptive cue combination mechanism.

Object Localization through Ground Plane Accurate estimation of both ground height and orientation is crucial for 3D object localization. Let $\mathbf{K}$ be the camera intrinsic calibration matrix. As [1, 5, 23], the bottom of a 2D bounding

box, $\mathbf{b} = (x, y, 1)^\top$ in homogeneous coordinates, can be back-projected to 3D through the ground plane $\{h, \mathbf{n}\}$:

$$\mathbf{B} = (B_x, B_y, B_z)^\top = -\frac{h\mathbf{K}^{-1}\mathbf{b}}{\mathbf{n}^\top \mathbf{K}^{-1}\mathbf{b}}, \tag{2}$$

Similarly, the object height can also be obtained using the estimated ground plane and the 2D bounding box height.

Given 2D object tracks, one may estimate best-fit 3D bounding boxes. The object pitch and roll are determined by the ground plane (see Fig. 2). For a vehicle, the initial yaw angle is assumed to be its direction of motion and a prior is imposed on the ratio of its length and width. Given an initial position from (2), a 3D bounding box can be computed by minimizing the difference between its reprojection and the tracked 2D bounding box.

We defer a detailed description of object localization to future work, while noting two points. First, an accurate ground plane is clearly the key to accurate monocular localization, regardless of the actual localization framework. Second, incorporating cues from detection bounding boxes into the ground plane estimation constitutes an elegant feedback mechanism between SFM and object localization.

Data Fusion with Kalman Filter To combine estimates from various methods, a natural framework is a Kalman filter:

$$\begin{aligned} \mathbf{x}^k &= \mathbf{A}\mathbf{x}^{k-1} + \mathbf{w}^{k-1}, \quad & p(\mathbf{w}) \sim N(0, \mathbf{Q}), \\ \mathbf{z}^k &= \mathbf{H}\mathbf{x}^k + \mathbf{v}^{k-1}, & p(\mathbf{v}) \sim N(0, \mathbf{U}), \end{aligned} \tag{3}$$

In our application, the state variable in (3) is the ground plane, thus, $\mathbf{x} = (\mathbf{n}^\top, h)^\top$. Since $\|\mathbf{n}\| = 1$, n_2 is determined by n_1 and n_3 and our observation is $\mathbf{z} = (n_1, n_3, h)^\top$. Thus, our state transition matrix and the observation model are given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix}^\top, \quad \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

Suppose methods $i = 1, \cdots, m$ are used to estimate the ground plane, with observation covariances $\mathbf{U}_j$. Then, the fusion equations at time instant k are

$$\mathbf{U}^k = (\sum_{i=1}^{m} (\mathbf{U}_i^k)^{-1})^{-1}, \;\; \mathbf{z}^k = \mathbf{U}^k \sum_{i=1}^{m} (\mathbf{U}_i^k)^{-1} \mathbf{z}_i^k. \tag{5}$$

Meaningful estimation of $\mathbf{U}^k$ at every frame, with the correctly proportional $\mathbf{U}_i^k$ for each cue, is essential for principled cue combination. Traditionally, fixed covariances are used to combine cues, which does not account for per-frame variation in their effectiveness across a video sequence. In contrast, in the following sections, we propose a data-driven mechanism to learn models to adapt per-frame covariances for each cue, based on error distributions of the underlying variables.

4. Cues for Ground Plane Estimation

We propose using multiple methods like triangulation of sparse feature matches, dense stereo between successive frames and object detection bounding boxes to estimate the ground plane. The cues provided by these methods are combined in a principled framework that accounts for their per-frame relative effectiveness. In this section, we describe the cues and the next section describes their combination.

Plane-Guided Dense Stereo We assume that a region of interest (ROI) in the foreground (middle fifth of the lower third of the image) corresponds to a planar ground. For a hypothesized value of $\{h, \mathbf{n}\}$ and relative camera pose $\{\mathbf{R}, \mathbf{t}\}$ between frames k and $k+1$, a per-pixel mapping can be computed using the homography matrix

$$\mathbf{G} = \mathbf{R} + \frac{1}{h}\mathbf{t}\mathbf{n}^\top. \tag{6}$$

Note that $\mathbf{t}$ differs from the true translation $\mathbf{t}^*$ by an unknown scale drift factor, encoded in the h we wish to estimate. Pixels in frame $k+1$ are mapped to frame k (subpixel accuracy is important for good performance) and the sum of absolute differences (SAD) is computed over bilinearly interpolated image intensities. With $\rho = 1.5$, a Nelder-Mead simplex routine is used to estimate the $\{h, \mathbf{n}\}$ that minimize:

$$\min_{h,\mathbf{n}} (1 - \rho^{-\text{SAD}}). \tag{7}$$

Note that the optimization only involves h, n_1 and n_3, since $\|\mathbf{n}\| = 1$. Enforcing the norm constraint has marginal effect, since the calibration pitch is a good initialization and the cost function usually has a clear local minimum in its vicinity. The optimization requires about 10 ms per frame. The $\{h, \mathbf{n}\}$ that minimizes (7) is the estimated ground plane from stereo cue.

Triangulated 3D Points Next, we consider matched sparse SIFT [14] descriptors between frames k and $k+1$, computed within the above region of interest (we find SIFT a better choice than ORB for the low-textured road and real-time performance is attainable for SIFT in the small ROI). To fit a plane through the triangulated 3D points, one option is to estimate $\{h, \mathbf{n}\}$ using a 3-point RANSAC for plane-fitting. However, in our experiments, better results are obtained using the method of [9], by assuming the camera pitch to be fixed from calibration. For every triangulated 3D point, the height h is computed using (1). The height difference Δh_{ij} is computed for every 3D point i with respect to every other point j. The estimated ground plane height is the height of the point i corresponding to the maximal score q, where

$$q = \max_i \left\{ \sum_{j \neq i} \exp\left(-\mu \Delta h_{ij}^2\right) \right\}, \text{ with } \mu = 50. \tag{8}$$

Note: Prior works like [20, 21] decompose the homography $\mathbf{G}$ between frames to yield the camera height [6]. However, in

practice, the decomposition is very sensitive to noise, which is a severe problem since the homography is computed using noisy feature matches from the low-textured road. Further, the fact that road regions may be mapped by a homography is already exploited by our plane-guided dense stereo.

Object Detection Cues We can also use object detection bounding boxes as cues when they are available, for instance, within the object localization application. The ground plane pitch angle θ can be estimated from this cue. Recall that $n_3 = \sin\theta$, for the ground normal $\mathbf{n} = (n_1, n_2, n_3)^\top$.

From (2), given the 2D bounding box, we can compute the 3D height h_b of an object through the ground plane. Given a prior height $\bar{h}_b$ of the object, n_3 is obtained by solving:

$$\min_{n_3} (h_b - \bar{h}_b)^2. \tag{9}$$

The ground height h used in (2) is set to the calibration value to avoid incorporating SFM scale drift and n_1 is set to 0 since it has negligible effect on object height.

Note: Object bounding box cues provide us unique long distance information, unlike dense stereo and 3D points cues that only focus on an ROI close to our vehicle. An inaccurate pitch angle can lead to large vertical errors for far objects. Thus, the 3D localization accuracy of far objects is significantly improved by incorporating this cue, as shown in Sec. 6.4.

5. Data-Driven Cue Combination

We now propose a principled approach to combine the above cues while reflecting the per-frame relative accuracy of each. Naturally, the combination should be influenced by both the visual input at a particular frame and prior knowledge. We achieve this by learning models from training data to relate the observation covariance for each cue to error behaviors of its underlying variables. During testing, our learned models adapt each cue's observation covariance on a per-frame basis.

5.1. Training

For the dense stereo and 3D points cues, we use the KITTI visual odometry dataset for training, consisting of $F = 23201$ frames. Sequences 0 to 8 of the KITTI tracking dataset are used to train the object detection cue. To determine the ground truth h and $\mathbf{n}$, we label regions of the image close to the camera that are road and fit a plane to the associated 3D points from the provided Velodyne data. No labelled road regions are available or used during testing.

Each method i described in Sec. 4 has a scoring function f_i that can be evaluated for various positions of the ground plane variables $\boldsymbol{\pi} = \{h, \mathbf{n}\}$. The functions f_i for stereo, 3D points and object cues are given by (7), (8) and (9), respectively. Then, Algorithm 1 is a general description of the training.

Intuitively, the parameters $\mathbf{a}_i^k$ of model $\mathcal{A}_i^k$ reflect belief in the effectiveness of cue i at frame k. Quantizing the parameters $\mathbf{a}_i^k$ from F training frames into L bins allows estimating

Algorithm 1 Data-Driven Training for Cue Combination

for Training frames $k = 1 : F$ **do**

- For various values of $\boldsymbol{\pi} = \{h, \mathbf{n}\}$, fit a model $\mathcal{A}_i^k$ to observations $(\boldsymbol{\pi}, f_i(\boldsymbol{\pi}))$. Parameters $\mathbf{a}_i^k$ of model $\mathcal{A}_i^k$ reflect belief in accuracy of cue i at frame k. (For instance, when $\mathcal{A}$ is a Gaussian, $\mathbf{a}$ can be its variance.)
- Compute error $e_i^k = |\arg\min_{\boldsymbol{\pi}} f_i(\boldsymbol{\pi}) - \boldsymbol{\pi}^{*k}|$, where the ground truth ground plane in frame k is $\boldsymbol{\pi}^{*k}$.

end for

- Quantize model parameters $\mathbf{a}_i^k$, for $k = 1, \cdots, F$, into L bins centered at $\mathbf{c}_i^1, \cdots, \mathbf{c}_i^L$.
- Histogram the errors e_i^k according to quantized $\mathbf{c}_i^l$. Let v_i^l be the bin variances of e_i^k, for $l = 1, \cdots, L$.
- Fit a model $\mathcal{C}_i$ to observations $(\mathbf{c}_i^l, v_i^l)$.

the variance of observation error at bin centers $\mathbf{c}_i^l$. The model $\mathcal{C}_i$ then relates these variances, v_i^l, to the cue's accuracy (represented by quantized parameters $\mathbf{c}_i^l$). Thus, at test time, for every frame, we can estimate the accuracy of each cue i based purely on visual data (that is, by computing $\mathbf{a}_i$) and use the model $\mathcal{C}_i$ to determine its observation variance.

Now we describe the specifics for training the models $\mathcal{A}$ and $\mathcal{C}$ for each of dense stereo, 3D points and object cues. We will use the notation that $i \in \{s, p, d\}$, denoting the dense stereo, 3D points and object detection methods, respectively.

5.1.1 Dense Stereo

The error behavior of dense stereo between two consecutive frames is characterized by variation in SAD scores between road regions related by the homography (6), as we independently vary each variable h, n_1 and n_3. The variance of this distribution of SAD scores represents the error behavior of the stereo cue with respect to its variables. Recall that the scoring function for stereo, f_s, is given by (7). We assume that state variables are uncorrelated. Thus, we will learn three independent models corresponding to h, n_1 and n_3.

Learning the model $\mathcal{A}_s$ For a training image k, let $\{\widehat{h}^k, \widehat{\mathbf{n}}^k\}$ be the ground plane estimated by the dense stereo method, by optimizing f_s in (7). We first fix $n_1 = \widehat{n}_1^k$ and $n_3 = \widehat{n}_3^k$ and for 50 uniform samples of h in the range $[0.5\widehat{h}^k,\ 1.5\widehat{h}^k]$, construct homography mappings from frame k to $k+1$, according to (6) (note that $\mathbf{R}$ and $\mathbf{t}$ are already estimated by monocular SFM, up to scale). For each homography mapping, we compute the SAD score $f_s(h)$ using (7). A univariate Gaussian is now fit to the distribution of $f_s(h)$. Its variance, $a_{s,h}^k$, captures the sharpness of the SAD distribution, which reflects belief in accuracy of height h estimated from the dense stereo method at frame k. A similar procedure yields variances a_{s,n_1}^k and a_{s,n_3}^k corresponding to orientation variables. Example fits are shown in Fig. 3. Referring to Algorithm 1 above, $a_{s,h}^k$, a_{s,n_1}^k, a_{s,n_3}^k are precisely the parameters

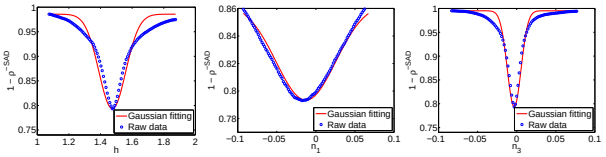


Figure 3. Examples of 1D Gaussian fits to estimate parameters $\mathbf{a}_s^k$ for h, n_1 and n_3 of the dense stereo method respectively.

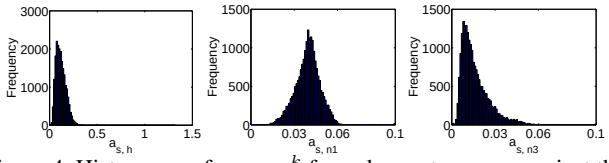


Figure 4. Histograms of errors e_s^k from dense stereo cue against the quantized accuracy parameters $\mathbf{a}_s$ of model $\mathcal{A}_s$, for h, n_1 and n_3.

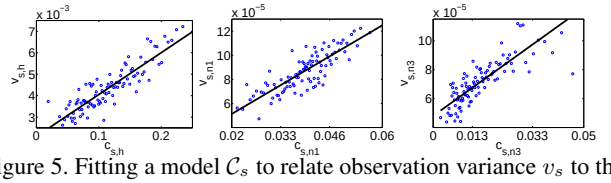


Figure 5. Fitting a model $\mathcal{C}_s$ to relate observation variance v_s to the belief in accuracy $\mathbf{c}_s$ of dense stereo, for h, n_1 and n_3.

$\mathbf{a}_s^k$ that indicate accuracy of the stereo cue at frame k.

Learning the model $\mathcal{C}_s$ For frame k, let $e_{s,h}^k = |\widehat{h}^k - h_k^*|$ be the error in ground height, relative to ground truth. We quantize the parameters $a_{s,h}^k$ into $L = 100$ bins and consider the resulting histogram of $e_{s,h}^k$. The bin centers $c_{s,h}^l$ are positioned to match the density of $a_{s,h}^k$ (that is, we distribute F/L errors $e_{s,h}^k$ within each bin). A similar process is repeated for n_1 and n_3. The histograms for the KITTI dataset are shown in Fig. 4. We have now obtained the $\mathbf{c}_s^l$ of Algorithm 1.

Next, we compute the variance $v_{s,h}^l$ of the errors within each bin l, for $l = 1, \cdots, L$. This indicates the observation error variance. We now fit a curve to the distribution of $v_{s,h}$ versus $c_{s,h}$, which provides a model to relate observation variance in h to the effectiveness of dense stereo. The result for the KITTI dataset is shown in Fig. 5, where each data point represents a pair of observation error covariance $v_{s,h}^l$ and parameter $c_{s,h}^l$. Empirically, we observe that a straight line suffices to produce a good fit. A similar process is repeated for n_1 and n_3. Thus, we have obtained models $\mathcal{C}_s$ (one each for h, n_1 and n_3) for the stereo method.

5.1.2 3D Points

Similar to dense stereo, the objective of training is again to find a model that relates the observation covariance of the 3D points method to the error behavior of its underlying variables. Recall that the scoring function f_p is given by (8).

Learning the model $\mathcal{A}_p$ We observe that the score q returned by f_p is directly an indicator of belief in accuracy of the ground plane estimated using the 3D points cue. Thus, for

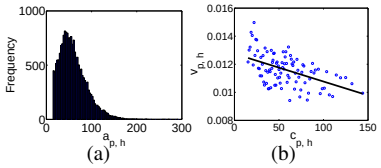


Figure 6. (a) Histogram of height error $e_{p,h}$ and cue accuracy $a_{p,h}$. (b) Relating observation variance $v_{p,h}$ to expected accuracy $c_{p,h}$.

Algorithm 1, we may directly obtain the parameters $a_p^k = q^k$, where q^k is the optimal value of f_p at frame k, without explicitly learning a model $\mathcal{A}_p$.

Learning the model $\mathcal{C}_p$ The remaining procedure mirrors that for the stereo cue. Let $\widehat{h}_p^k$ be ground height estimated at frame k using 3D points, that is, the optimum for (8). The error $e_{p,h}^k$ is computed with respect to ground truth. The above $a_{p,h}^k$ are quantized into $L = 100$ bins centered at $c_{p,h}^l$ and a histogram of observation errors $e_{p,h}^k$ is constructed. A model $\mathcal{C}_p$ may now be fit to relate the observation variances $v_{p,h}^l$ at each bin to the corresponding accuracy parameter $c_{p,h}^l$. As shown in Fig. 6, a straight line fit is again reasonable.

5.1.3 Object Detection

We assume that the detector provides several candidate bounding boxes and their respective scores (that is, bounding boxes before the nonmaximal suppression step of traditional detectors). A bounding box is represented by $\mathbf{b} = (x, y, w, h_b)^\top$, where x, y is its 2D position and w, h_b are its width and height. The error behavior of detection is quantified by the variation of detection scores α with respect to bounding box $\mathbf{b}$.

Learning the model $\mathcal{A}_d$ Our model $\mathcal{A}_d^k$ is a mixture of Gaussians. At each frame, we estimate 4×4 full rank covariance matrices $\mathbf{\Sigma}_m$ centered at $\boldsymbol{\mu}_m$, as:

$$\min_{A_m, \boldsymbol{\mu_m}, \mathbf{\Sigma}_m} \sum_{n=1}^{N} \left(\sum_{m=1}^{M} A_m e^{-\frac{1}{2}\boldsymbol{\epsilon}_{mn}\mathbf{\Sigma}_m^{-1}\boldsymbol{\epsilon}_{mn}} - \alpha_n \right)^2, \quad (10)$$

where $\boldsymbol{\epsilon}_{mn} = \mathbf{b}_n - \boldsymbol{\mu}_m$, M is number of objects and N is the number of candidate bounding boxes (the dependence on k has been suppressed for convenience). Example fitting results are shown Fig. 7. It is evident that the variation of noisy detector scores is well-captured by the model $\mathcal{A}_d^k$.

Recall that the scoring function f_d of (9) estimates n_3. Thus, only the entries of $\mathbf{\Sigma}_m$ corresponding to y and h_b are significant for our application. Let σ_y and σ_{h_b} be the corresponding diagonal entries of the $\mathbf{\Sigma}_m$ closest to the tracking 2D box. We combine them into a single parameter, $a_d^k = \frac{\sigma_y \sigma_{h_b}}{\sigma_y + \sigma_{h_b}}$, which reflects our belief in the accuracy of this cue.

Learning the model $\mathcal{C}_d$ The remaining procedure is similar to that for the stereo and 3D points cues. The accuracy parameters a_d^k are quantized and relaed to the corresponding variances of observation errors, given by the f_d of (9). The fitted linear model $\mathcal{C}_d$ that relates observation variance of the detection cue to its expected accuracy is shown in Fig. 8.

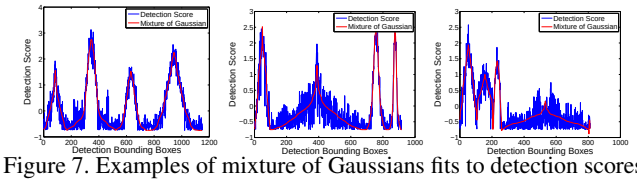


Figure 7. Examples of mixture of Gaussians fits to detection scores. Note that our fitting (red) closely reflects the variation in noisy detection scores (blue). Each peak corresponds to an object.

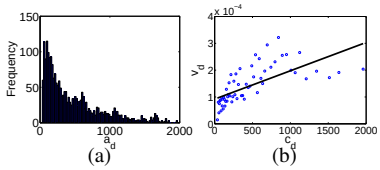


Figure 8. (a) Histogram of n_3 error and cue accuracy $a_{d,h}$. (b) Relating observation variance $v_{d,h}$ to expected accuracy $c_{d,h}$.

5.2. Testing

During testing, at every frame k, we fit a model A_i^k corresponding to each cue $i \in \{s, p, d\}$ and determine its parameters $\mathbf{a}_i^k$ that convey expected accuracy. Next, we use the models $\mathcal{C}_i$ to determine the observation variances.

Dense Stereo The observation $\mathbf{z}_s^k = (n_1^k, n_3^k, h^k)^\top$ at frame k is obtained by minimizing f_s, given by (7). We fit 1D Gaussians to the homography-mapped SAD scores to get the values of $a_{s,h}^k$, a_{s,n_1}^k and a_{s,n_3}^k. Using the models $\mathcal{C}_s$ estimated in Fig. 5, we predict the corresponding variances v_s^k. The observation covariance for the dense stereo method is now available as $\mathbf{U}_1^k = \text{diag}(v_{s,n_1}^k, v_{s,n_3}^k, v_{s,h}^k)$.

3D Points At frame k, the observation $\mathbf{z}_p^k$ is the estimated ground height h obtained from f_p, given by (8). The value of q^k obtained from (8) directly gives us the expected accuracy parameter a_p^k. The corresponding variance $v_{p,h}^k$ is estimated from the model $\mathcal{C}_p$ of Fig. 6. The observation covariance for this cue is now available as $\mathbf{U}_p^k = v_{p,h}^k$.

Object Detection At frame k, the observation $\mathbf{z}_d^{k,m}$ is the ground pitch angle n_3 obtained by minimizing f_d, given by (9), for each object $m = 1, \cdots, M$. For each object m, we obtain the parameters $a_d^{k,m}$ after solving (10). Using the model $\mathcal{C}_d$ of Fig. 8, we predict the corresponding error variances $v_d^{k,m}$. The observation covariances for this method are now given by $\mathbf{U}_d^{k,m} = v_d^{k,m}$.

Fusion Finally, the adaptive covariance for frame k, $\mathbf{U}^k$, is computed by combining $\mathbf{U}_s^k$, $\mathbf{U}_p^k$ and the $\mathbf{U}_d^{k,m}$ from each object m. Then, our adaptive ground plane estimate $\mathbf{z}^k$ is computed by combining $\mathbf{z}_s^k$, $\mathbf{z}_p^k$ and $\mathbf{z}_d^{k,m}$, using (5).

Thus, we have described a ground plane estimation method that uses models learned from training data to adapt the relative importance of each cue – stereo, 3D points and detection bounding boxes – on a per-frame basis.

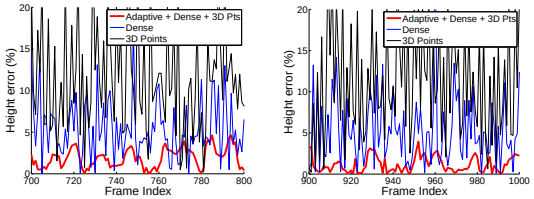


Figure 9. Height error relative to ground truth over (left) Seq 2 and (right) Seq 5. The effectiveness of our data fusion is shown by less spikiness in the filter output and a far lower error.

6. Experiments

We present extensive evaluation on the KITTI dataset [8], which consists of nearly 50 km of driving in various conditions. Our experiments are performed on an Intel i7 laptop. The SFM modules occupy three CPU threads and the ground plane estimation occupies two threads. 3D object localization is demonstrated using object detection and tracked bounding boxes computed offline using [1, 18].

6.1. Accuracy of Ground Plane Estimation

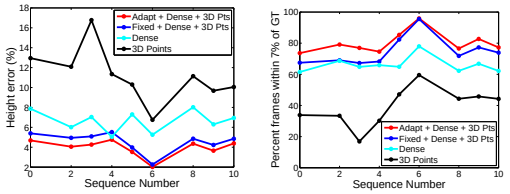
In consideration of real-time performance, only the dense stereo and 3D points cues are used for monocular SFM. Detection bounding box cues are used for the object localization application where they are available.

Fig. 9 shows examples of error in ground plane height relative to ground truth using 3D points and stereo cues individually, as well as the output of our combination. Note that while individual methods are very noisy, our cue combination allows a much more accurate estimation than either.

Next, we demonstrate the advantage of cue combination using the data-driven framework of Sec. 5 that uses adaptive covariances, as opposed to a traditional Kalman filter with fixed covariances. For this experiment, the fixed covariance for the Kalman filter is determined by the error variances of each variable over the entire training set (we verify by cross-validation that this is a good choice).

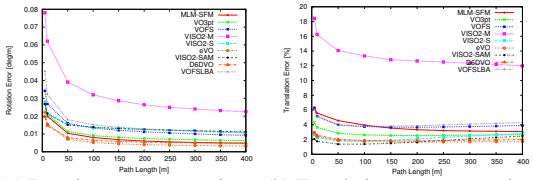
In Fig. 10, using only sparse feature matches causes clearly poor performance (black curve). The dense stereo performs better (cyan curve). Including the additional dense stereo cue within a Kalman filter with fixed covariances leads to an improvement (blue curve). However, using the training mechanism of Sec. 5 to adjust per-frame observation covariances in accordance with the relative confidence of each cue leads to a further reduction in error by nearly 1% (red curve). Fig. 10(b) shows that we achieve the correct scale at a rate of 75 – 100% across all sequences, far higher than the other methods.

In particular, compare our output (red curves) to that of only 3D points (black curves). This represents the improvement by this paper over prior works like [9, 20, 21] that use only sparse feature matches from the road surface.



(a) Height error (b) Success rate

Figure 10. Error and robustness of our ground plane estimation. (a) Average error in ground plane estimation across Seq 0-10. (b) Percent number of frames where height error is less than 7%. Note that the error in our method is far lower and the robustness far higher than achievable by either method on its own.



(a) Rotation error comparisons (b) Translation error comparisons

Figure 11. Comparison with other SFM systems. Our system is labeled MLM-SFM, shown in solid red. Note that we nearly match even state-of-the-art stereo for rotation and compete well against stereo in translation as well. Note the wide improvement over other monocular systems. See KITTI webpage for complete details.

	VISO2 (Stereo)	VISO2 (Mono)	Ours (Mono)
Rot (deg/m)	1.36e-02	3.41e-02	1.00e-02
Trans (%)	1.54	11.25	3.37

Table 1. Accuracy and robustness of monocular SFM that uses our cue-combined scale correction. The errors are averages for Seq 0-10 in KITTI, computed over 50 trials to demonstrate robustness. Note that our rotation error is lower than stereo. Translation error is comparable to stereo and much lower than other monocular systems.

6.2. Benchmark Monocular SFM on KITTI

We now show the impact of our ground plane estimation for the monocular SFM. The SFM evaluation sequences in KITTI are 11-21, for which ground truth is not public. Our system's performance is accessible at the KITTI evaluation website, under the name **MLM-SFM**[1]. Comparison with other systems is given by Fig. 11. Note that all the other systems, except VISO2-M [9], are stereo, yet we achieve close to the best rotation accuracy. Our translation error is lower than several stereo systems and far lower than VISO2-M.

Another important benefit of our scale correction is enhanced robustness. As demonstration, we run 50 trials of our system on Seq 0-10, as well as other stereo and monocular systems VISO2-S and VISO2-M [9]. Errors relative to ground truth are computed using the metrics in [8] and the average errors are summarized in Table 1. Note our vast performance improvement over VISO2-M, a rotation error better than VISO2-S and translation error comparable to stereo.

Fig. 1 shows recovered camera paths from our monocular system, overlaid with ground truth. Compare the accurate scale maintained by our system (a), as opposed to a method that uses only 3D points (b). This again shows the effectiveness of our data-driven cue combination, which leads to monocular SFM performance close to stereo (c).

[1] www.cvlibs.net/datasets/kitti/eval_odometry.php

6.3. Monocular SFM on Hague Dataset

We show additional results on the publicly available Hague dataset [4]. It consists of three sequences of varying lengths, from 600 m to 5 km. It is challenging due to low resolution images, as well as several obstacles due to crowded scenes and moving vehicles close to the camera. Accurate scale drift correction allows us to successfully complete such sequences, in contrast to prior monocular SFM systems. In the absence of ground truth, Table 2 reports figures for loop closure or end-point error relative to map information.

Seq	Frames	Length (m)	End-point error (%)
1	2500	609.34	5.37
2	3000	834.39	1.99
3	19000	5045.45	4.85

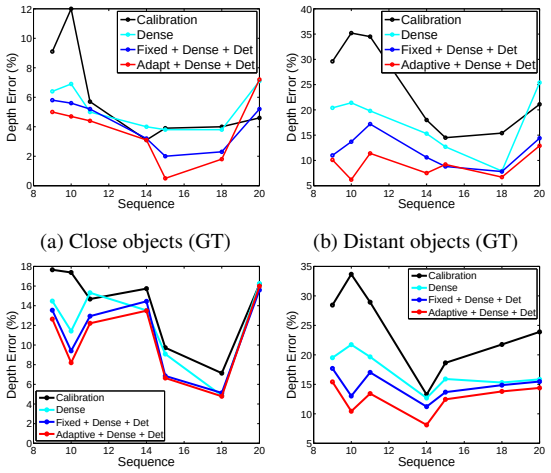
Table 2. End-point errors for sequences in The Hague dataset.

6.4. Accuracy of 3D Object Localization

Now we demonstrate the benefit of the adaptive ground plane estimation of Sec. 5 for 3D object localization. KITTI does not provide a localization benchmark, so we instead use the tracking training dataset to evaluate against ground truth. We use Seq 1-8 for training and Seq 9-20 for testing. The metric we use for evaluation is percentage error in object position. For illustration, we consider only the vehicle objects and divide them into "close" and "distant", where distant objects are farther than 10m. We discard any objects that are not on the road. Candidate bounding boxes for training the object detection cue are obtained from [7].

Fig. 12 compares object localization using a ground plane from our data-driven cue combination (red curve), as opposed to one estimated using fixed covariances (blue), or one that is fixed from calibration (black). The top row uses ground truth object tracks, while the bottom row uses tracks from the state-of-the-art tracker of [18]. For each case, observe the significant improvement in localization using our cue combination. Also, from Figs. 12(b),(d), observe the significant reduction in localization error by incorporating the detection cue for ground plane estimation for distant objects.

Finally, we compare our localization results to those of Choi and Savarese [1]. We use the object tracking output of [1] provided on a few KITTI raw sequences and show in Table 3 that using our adaptive ground plane estimation yields a lower error. Note that the ground plane estimation of [1] suffers due to sole reliance on salient features on the road surface. This demonstrates how our framework for cue-combined ground plane estimation may complement existing localization systems to significantly enhance their accuracy.



(a) Close objects (GT) (b) Distant objects (GT)

(c) Close objects (DP) (d) Distant objects (DP)

Figure 12. Comparison of 3D object localization errors for calibrated ground, stereo cue only, fixed covariance fusion and adaptive covariance fusion of stereo and detection cues. (Top row) Using object tracks from ground truth (Bottom row) Using object tracks from [18]. Rrrors reduce significantly for adaptive cue fusion, especially for distant object where detection cue is more useful.

Seq	Num of objs	Avg num frms	3D Error (%)			
			MTT[1]	Ours		
				Dense	Fixed	Adapt
0004	4	199	14.4	9.1	7.5	6.1
0056	1	293	13.9	8.2	5.0	5.5
0047	5	235	13.8	11.2	10.8	10.0
Average			14.0	9.4	7.6	7.1

Table 3. Comparison with 3D object localization of [1]. Dense: using dense stereo cue for ground plane estimation. Fixed: combine cues in a traditional Kalman filter. Adapt: the cue combination of Sec. 5. We use the tracked 2D bounding boxes of [1]. Note the improvement by incorporating both dense stereo and detection cues. Also note the advantage of using our adaptive cue combination, that leads to over 6% improvement in location error over the system of [1]

Fig. 1 shows an example from our localization output. Note the accuracy of our 3D bounding boxes (red), even when the 2D tracking-by-detection output (cyan) is not accurate.

7. Conclusion and Future Work

We have demonstrated that accuracte ground plane estimation allows monocular vision-based systems to achieve performance similar to stereo. In particular, we have shown that it is beneficial to include cues such as dense stereo and object bounding boxes for ground estimation, besides the traditional sparse features used in prior works. Further, we proposed a mechanism to combine those cues in a principled framework that reflects their per-frame relative confidences, as well as prior knowledge from training data.

Our robust and accurate scale correction is a significant step in bridging the gap between monocular and stereo SFM. We believe this has great benefits for autonomous driving applications. We demonstrate that the performance of real-time monocular SFM that uses our ground plane estimation is comparable to stereo on real-world driving sequences. Further, our accurate ground plane easily benefits existing 3D localization frameworks, as also demonstrated by our experiments.

In future extension of this work, we will explore a deeper integration of ground plane and SFM cues with object detection, to obtain accurate and real-time 3D multi-target tracking.

Acknowledgments This research was conducted at NEC Labs America during the first author's internship in 2013.

References

[1] W. Choi and S. Savarese. Multi-target tracking in world coordinate with single, minimally calibrated camera. In *ECCV*, 2010.
[2] B. Clipp, J. Lim, J.-M. Frahm, and M. Pollefeys. Parallel, real-time visual SLAM. In *IROS*, pages 3961–3968, 2010.
[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *PAMI*, 29(6):1052–1067, 2007.
[4] G. Dubbelman and F. Groen. Bias reduction for stereo based motion estimation with applications to large scale odometry. In *CVPR*, 2009.
[5] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Robust multiperson tracking from a mobile platform. *PAMI*, 31(10):1831–1846, 2009.
[6] O. D. Faugeras and F. Lustman. Motion and Structure From Motion in a Piecewise Planar Environment. *Pat. Rec. AI*, 2(3):485–508, 1988.
[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.
[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.
[9] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3D reconstruction in real-time. In *IEEE Int. Veh. Symp.*, 2011.
[10] P. Hansen, H. S. Alismail, P. Rander, and B. Browning. Monocular visual odometry for robot localization in LNG pipes. In *ICRA*, 2011.
[11] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007.
[12] G. Klein and D. Murray. Improving the agility of keyframe-based SLAM. In *ECCV*, 2008.
[13] A. Kundu, K. M. Krishna, and C. V. Jawahar. Realtime multibody SLAM with a smoothly moving monocular camera. In *ICCV*, 2011.
[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[15] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *CVPR*, pages 652–659, 2004.
[16] K. Ozden, K. Schindler, and L. Van Gool. Simultaneous segmentation and 3D reconstruction of monocular image sequences. In *ICCV*, 2007.
[17] K. E. Ozden, K. Schindler, and L. V. Gool. Multibody structure-from-motion in practice. *PAMI*, 32(6):1134–1141, 2010.
[18] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.
[19] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *ICCV*, 2009.
[20] D. Scaramuzza and R. Siegwart. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE Trans. Robotics*, 24(5):1015–1026, 2008.
[21] S. Song, M. Chandraker, and C. C. Guest. Parallel, real-time monocular visual odometry. In *ICRA*, pages 4698–4705, 2013.
[22] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Scale drift-aware large scale monocular SLAM. In *Robotics: Science and Systems*, 2010.
[23] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: Understanding multi-object traffic scenes. *PAMI*, 35(4):882–897, 2013.