Multi-scale Tetrahedral Fusion of a Similarity Reconstruction and Noisy Positional Measurements

Runze Zhang, Tian Fang[1], Siyu Zhu, Long Quan

The Hong Kong University of Science and Technology

Abstract. The fusion of a 3D reconstruction up to a similarity transformation from monocular videos and the metric positional measurements from GPS usually relies on the alignment of the two coordinate systems. When positional measurements provided by a low-cost GPS are corrupted by high-level noises, this approach becomes problematic. In this paper, we introduce a novel framework that uses similarity invariants to form a tetrahedral network of views for the fusion. Such a tetrahedral network decouples the alignment from the fusion to combat the high-level noises. Then, we update the similarity transformation each time a well-conditioned motion of cameras is successfully identified. Moreover, we develop a multi-scale sampling strategy to reduce the computational overload and to adapt the algorithm to different levels of noises. It is important to note that our optimization framework can be applied in both batch and incremental manners. Experiments on simulations and real datasets demonstrate the robustness and the efficiency of our method.

1 Introduction

Monocular SLAM (Simultaneously Localization And Mapping) is only able to reconstruct camera poses and 3D structures, so called *visual measurements*, up to a similarity transformation due to the gauge freedom [11]. Such a similarity reconstruction is not sufficient for the applications on the navigation, osculation avoidance for robots and unmanned aerial vehicles. Moreover, the noise in feature detections, unbalance features [9], local bundle adjustments [4], and biased depth estimators [17] make the visual measurements contain significant drift in both rotation and translation over a long range movement. Drift-free global positional measurements provided by modern global position system (GPS) can be used to address the aforementioned inherent drawback of monocular SLAM. Some works [14, 3] have been done on addressing the scale ambiguity solely. However, the inherent drifting of monocular SLAM makes the error between visual measurements and ground truth no longer follow the normal distribution, which in turn biases the estimation of scale even under maximum likelihood framework. To compensate for the drifting, some other works [6, 15, 8] directly fuse positional measurements with visual measurements by the metric distances between them, which relies on a good initial similarity transformation to resolve the scale ambiguity and requires aligning both set of data in the same coordinate frame. The estimation of such initial transformation of scales can be problematic when the camera moves under

[1] Tian Fang is the corresponding author.

a critical motion such as moving along a straight line, where the rotation around the moving direction is not well constrained. Even worse, low cost GPS sensors on many consumer products give very noisy positional measurement making the estimation of initial similarity transformation less reliable.

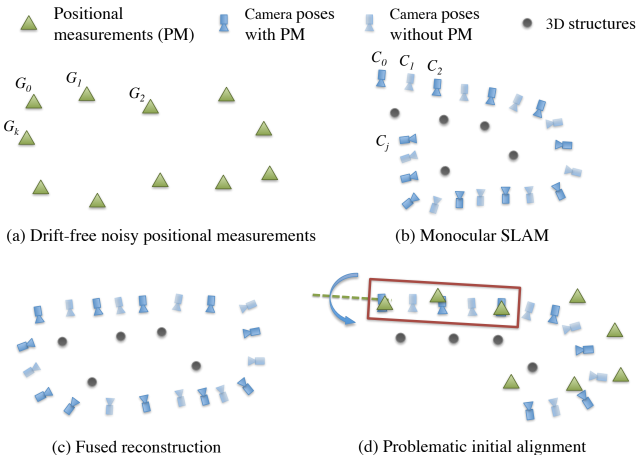
In this paper, we propose a novel multi-scale tetrahedral fusion framework. Based on the invariance of the ratios of two distances under a similarity transformation, we define a ratio constraint over a tetrahedral network defined on the cameras that are associated with positional measurements. This configuration is capable of correcting the drift without the knowledge of the global similarity alignment. The global similarity transformation is in turn estimated later on when there is well-conditioned motion. Moreover, we further propagate such positional measurements to the other cameras via relative pose constraints that retain the local camera motion. Geometric constraints based on reprojection error are involved to ensure the consistency between the reconstructed cameras and 3D structures. Finally, a multi-scale scheme that is adapted to different levels of noise of positional measurements is used to sample the tetrahedral and relative pose constraints. All these constraints are formulated as solving a nonlinear least square optimization. After reviewing the related work in Section 1.1, we first introduce our key idea on tetrahedral network in Section 2.1. Then the formulation is given in Section 2.2 along with a discussion on the details of the optimization in Section 2.4. In Section 3, we finally describe the implementation of our system and present detailed evaluations on our approach.

1.1 Related Work

The work on integrating the positional measurements with a similarity reconstruction can be classified into two categories.

A part of previous research only obtained a metric upgrade by estimating the scale factor between the up-to-scale visual measurement and the metric measurements. Nützi et al. [14] used a spline fitting technique to estimate the scale. Engel et al. [3] proposed a recursive update formula for Maximum likelihood estimation of the scale factor. Their approach takes a metric altitude of a helicopter and the height estimated from a video camera looking downwards to the grounds. In these works, the visual measurement is assumed to be normally distributing around the true estimations. Unfortunately, the assumption is generally not true because the monocular SLAM reconstructs drifted measurements.

Another part of research resolved the drifting problem through the fusion given a good initial similarity alignment. Michot et al. [10] augmented the classic bundle adjustment of reprojection error with a penalty term to minimize the error of the difference of the positional measurements and the similarity reconstruction. Lhuillier [8] proposed a constraint optimization framework on the bundle adjustment that guarantees a small change of reprojection error during the fusion. Konolige et al. [6] further marginalized the geometric constraint in the bundle adjustment and optimized a pose graph which constraints only the relative pose between the cameras. However, all these works require an initial similarity alignment between the visual measurement and positional measurements. Such an alignment is estimated with a subset of the measurements at the beginning of the fusion, which could be problematic if such subset of measurements is



(a) Drift-free noisy positional measurements

(b) Monocular SLAM

(c) Fused reconstruction

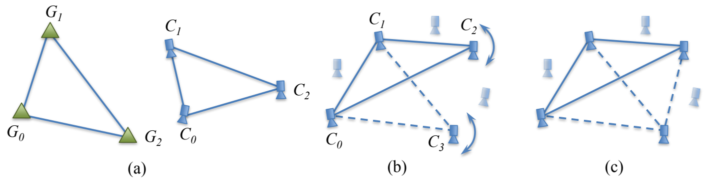
(d) Problematic initial alignment

Fig. 1. The illustration of the fusions. (a) and (b) are the positional measurements (PM) and visual measurements that are under an unknown similarity transformation. (c) is an illustration of a successful fusion which successfully estimate the similarity alignment and correct the drift. (d) illustrates the case that the initial similarity alignment is estimated with a critical motion in the red rectangle. Such alignment is unstable because the rotation (the blue arrow) around the moving direction (the green dashed line) is not well constrained.

in bad condition, such as forming a straight line. Extended Kalman Filter (EKF) [14, 2] has also been used to fuse both the motion and scale simultaneously. In these work, the scale factor is explicitly considered and involved in the state vector of the motion. However, they all required an initial Euclidean registration whose accuracy is very important for the later recursive update.

2 Multi-scale Tetrahedral Fusion

We have two sets of input measurements, a similarity reconstruction generated by monocular SLAM and positional measurements obtained through external sensors such as GPS. The *similarity reconstruction* includes the poses $\mathcal{C} = \{C_j\}$ of the monocular camera at each frame and a set of reconstructed 3D points $\mathcal{P} = \{p_i\}$ as in Figure 1 (b). Each camera is parameterized as $C_j = K_j[R_j|t_j]$. For monocular SLAM, K_j is usually fixed during the capture, so we simply assume K_j is pre-calibrated and drop such terms in the following text. We further denote the extrinsic parameters R_j and t_j as an Euclidean transformation T_j that belongs to $SE(3)$ group. The camera C_j and 3D point p_i are linked by the image feature q_{ij} via the projection function $\mathbb{Q}(C_j, p_i)$ if p_i is visible in C_j. The *positional measurement* $\mathcal{G} = \{G_k\}$ are recorded simultaneously when the camera is moving as in Figure 1 (a). Since the temporal sampling rates of vision and positional measurements are usually different, we explicitly align the index set $\{j\}$ and $\{k\}$ to make sure that k denotes the frames with positional measurements while other frames are generally denoted by index j.



(a) (b) (c)

Fig. 2. The illustration of tetrahedral fusion. (a) A toy example of three GPS and visual measurements; (b) Sparse sampling of triangular constraints is not sufficient to constraint the structure; (c) A tetrahedral structure consisting of four triangular constraints can well preserve the shape.

Since the coordinate frame of the visual measurement generated by monocular S-LAM is unknown, in the alignment-based fusion, a global similarity transformation $S_G = S(G, C, P)$ is first estimated to transform the visual measurement to the coordinate frame of positional measurement as $\hat{\mathcal{C}}, \hat{\mathcal{P}} = S_G(\mathcal{C}, \mathcal{P})$. Then an optimized fused measurement $\mathcal{C}', \mathcal{P}' = \arg\min_{\mathcal{C}', \mathcal{P}'} E_{fusion}(\mathcal{C}, \mathcal{P}, G)$. However, in practice, in the present of noisy positional measurement and the degenerated motions, the estimation of S_G is not always valid and robust. In the following, we introduce a novel framework based on similarity invariants to directly fuse the visual measurement and positional measurement without the knowledge of S_G. The global similarity transformation S_G can then be recovered when there are sufficient measurements and well-conditioned motions. Such decoupling of the alignment from the fusion greatly improves the robustness of the fusion.

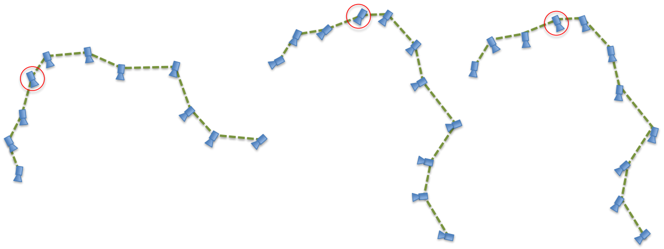
2.1 Overview of Tetrahedral Fusion

To fuse a similarity reconstruction with metric positional measurements without alignment, we must make use of similarity invariant properties that are the ratios of distances and the angles. These invariant properties are completely encoded by the ratios of all edge pairs of a triangle. Let $tri = (G_0, G_1, G_2)$ in Figure 2 (a) be a reference triangle. The ratios between its edges, $\|G_0G_1\|/\|G_1G_2\|$, $\|G_1G_2\|/\|G_2G_0\|$, and $\|G_2G_0\|/\|G_0G_1\|$, remain unchanged under a similarity transformation. Let's further denote another triangle as $tri' = (c_0, c_1, c_2)$ whose vertex c_i corresponds to G_i under an unknown similarity transformation. We introduce a ratio constraint on a pair of triangles:

$$E_{ratio}(c_i, c_j, c_k) = (\|c_i - c_j\| - \frac{\|G_iG_j\|}{\|G_iG_k\|} \cdot \|c_i - c_k\|)^2 \tag{1}$$

For each triangle, three permutations for its vertices give in total three ratio constraints for a triangle.

It is theoretically sufficient to exhaustively enumerate all combination of three positional measurements to constraint the fusion. However, such an exhaustive enumeration generates a large number of constraints that overwhelm the computation. A sparse sampling of triangles is thus preferred. Unfortunately, arbitrary sampling does not guarantee a stable fusion. As illustrated in Figure 2 (b), even if the ratio constraints of such



(a) SLAM before fusion

(b) What it should be after fusion

(c) The rotation of cameras are not corrected if only center constraints are performed

Fig. 3. The illustration of the result without rotation constraints. (a) Original SLAM before fusion; (b) Since our fusion result is up to a similarity transformation. The rotation of each camera should remain consistent with original motion. (c) The rotation of each camera cannot be well constrained if only the ratio constraint is applied.

two sampled triangles are met with respect to the referenced triangles, such triangles can still rotate arbitrarily around the edge C_0C_1. Hence, to ensure the stabilities of the structure, we introduce tetrahedral constraints, which ensure that all the triangles in four randomly selected vertices must be sampled at the same time as in Figure 2 (c).

Because the ratio constraints in Equation 1 only maintain the corresponding distance ratio among the position of cameras, the rotation of each camera is not well constrained and is free up to an arbitrary rotation as shown in Figure 3. We further introduce a term $E_{tertrarot}$ as Equation 2 to constrain the rotation of cameras with respect to the edges of the sampled tetrahedron:

$$E_{tetrarot}(D_t) = \sum_{i \in D_t} \sum_{j \neq i, j \in D_t} \left(\frac{(R'_i(c'_j - c'_i))^T (R_i(c_j - c_i))}{||R'_i(c'_j - c'_i)|| \cdot ||R_i(c_j - c_i)||} - 1\right)^2 \tag{2}$$

Actually, the $R_i(c_j - c_i)$ is the projection of translation vector between camera i and j on camera i. And $(R'_i(c'_j - c'_i))^T (R_i(c_j - c_i)) / ||R'_i(c'_j - c'_i)|| \cdot ||R_i(c_j - c_i)||$ is the cosine of angle between such projection before and after fusion. This constraint tries to maintain the consistency between rotation of each camera and the translation between cameras.

Given the tetrahedral constraints, the ratio constraint has a family of trivial solutions that are defined up to a similarity transformation of $\mathcal{G}$. We further introduce two other sets of constraints. One is the relative pose constraint [7], which enforces consistency of the local motion and rotation of the cameras. The other one is the geometric constraints [18], which enforces consistency between the camera poses and 3D points.

2.2 Formulation

Formally, given a reconstruction $\mathcal{C}$ and $\mathcal{P}$ defined up to an arbitrary similarity transformation, and the positional measurement $\mathcal{G}$ in a global geographical reference frame, in the tetrahedral fusion, we are looking for an updated camera poses $\mathcal{C}'$ and 3D points $\mathcal{P}'$, which are still up to a similarity transformation, but associated with $\mathcal{G}$ by minimizing the following energy function.

$$\mathcal{C}', \mathcal{P}' = \arg \min_{\mathcal{C}', \mathcal{P}'} (E_{tetra} + \alpha \cdot E_{pose} + \beta \cdot E_{bundle}), \tag{3}$$

where E_{tetra}, E_{pose}, and E_{bundle} are the energies for tetrahedral constraint, relative pose constraint, and geometric constraint respectively. The tetrahedral constraint is the core of our algorithm, which sets up ratio constraints over a tetrahedral network on the positional measurements. The relative pose constraint embeds the camera poses that do not have corresponding positional measurements into the fusion and makes sure the upgraded camera poses follow the original motion. The last geometric constraint is a classical term that ensures the consistency between camera poses and reconstructed 3D structures. We now present these three terms in details.

Tetrahedral constraint. The tetrahedral constraint is defined as two parts:

$$E_{tetra} = E_{ratiogps} + E_{rot} \tag{4}$$

The first term is to constrain the ratio relationship between GPS and SLAM as Equation 5.

$$E_{ratiogps} = \sum_{\{D_t\}} w_t \sum_{i,j,k \in P^3_{(D_t)}} \frac{1}{12} \cdot E_{ratio}(c'_i, c'_j, c'_k), \tag{5}$$

where $P^3_{(D_t)}$ is all the permutations of three cameras in tetrahedron D_t; and the w_t is the weight of the tetrahedral constraint, normalized by the number of tetrahedrons on its sampling level as Section 2.3. Because the tetrahedral constraints are relative to both GPS and SLAM data, tetrahedrons $\{D_t\}$ are only sampled in the frames where both GPS and SLAM data are available.

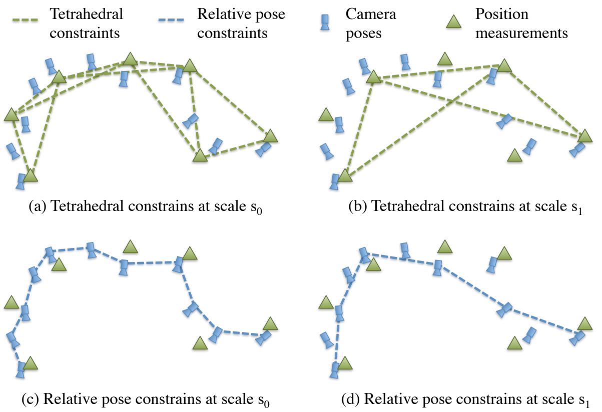
The second part is a tetrahedral constraint for rotation defined in the tetrahedrons D'_t as Equation 6 to maintain the relative rotation among cameras and tetrahedrons.

$$E_{rot} = \sum_{\{D'_t\}} w'_t E_{tetrarot}(D'_t) \tag{6}$$

Rotation constraints involve only SLAM data, so tetrahedrons $\mathcal{D}' = \{D'_t\}$ are sampled among all the frames.

Relative pose constraint The relative pose constraint [7] penalizes large changes in the relative pose transformation between two connected cameras, C'_i and C'_j. Let the original relative transformation from C_i to C_j be $\varDelta T_{ij} = T_i^{-1} T_j$. The relative pose constraint is defined as:

$$E_{pose} = \sum_{\{i,j\}} \| \log_{SE(3)} (T'_i \cdot \varDelta T_{ij} \cdot T'^{-1}_j) \|^2_{\Sigma_{ij}}, \tag{7}$$



(a) Tetrahedral constrains at scale s_0

(b) Tetrahedral constrains at scale s_1

(c) Relative pose constrains at scale s_0

(d) Relative pose constrains at scale s_1

Fig. 4. The illustration of multi-scale constraints. For clarity, in (a) and (b), tetrahedral constraints are simplified and illustrated as triangular constraints. Please note that the camera poses and positional measurements are roughly aligned in the illustrations for easy understanding, but in the fusion we do not assume any pre-alignment.

where $\log_{SE(3)}(\cdot)$ measures the relative pose error in the tangent space of $SE(3)$ group and Σ_{ij} is the precision matrix of the Mahalanobis distance $\|\cdot\|_{\Sigma_{ij}}$ for 2-tuple camera pose C_i and C_j. We set Σ_{ij} as:

$$\Sigma_{ij} = w_{ij} \begin{bmatrix} \sigma_{trans}^2 I_{3\times3} & 0 \\ 0 & \sigma_{rot}^2 I_{3\times3} \end{bmatrix} \tag{8}$$

Geometric constraint The geometric constraint mimics the bundle adjustment that minimizes the reprojection error in a maximum likelihood estimation manner. It is defined to be

$$E_{bundle} = \sum_{\{q_{ij}\}} \|q_{ij} - \mathbb{Q}(C_j, p_i)\|_{\Sigma}^2, \tag{9}$$

where Σ is precision matrix for reprojection errors. Conventionally, this term is simply set to identity because the covariance of the reprojection error is hardly known beforehand.

2.3 Multi-scale Constraints

The tetrahedral constraints and relative pose constraints are defined on 4-tuple and 2-tuple relationship respectively. The sampling of such tuples significantly affects the performance of the optimization of Equation 3. A multi-scale scheme to sample such tuples to fuse different level of details of information is therefore necessary.

The tetrahedral tuples are sampled on the positional measurements $\mathcal{G}$ as in Figure 4 (a) and (b). For each scale $s_l = \lfloor n/2^{L-l} \rfloor$ where $L = \lceil log_2 n \rceil, l = 0, \ldots, l_{max}$, we sample the tetrahedron as $(i, i+s_l, i+2s_l, i+3s_l)$ for $i = 0, s_l, 2s_l \ldots \lfloor n/s_l \rfloor s_l$.

The sampled tetrahedral tuples for $E_{ratiogps}$ and E_{rot} in Equation 4 are slightly different. For the first part, the tuples are sampled on overlapped cameras in GPS and SLAM and l_{max} is set as $L-2$ to constrain ratio on GPS from the smallest scale to the largest. For the second parts, they are sampled on all the cameras in SLAM.

The relative pose constraints are defined on the camera poses $\mathcal{C}$. For each scale $s_l = 2^l$ where $l = 0, 1, \ldots, L_r$, we create the relative pose tuple as $(i, i + s_l)$ for $i = 0, s_l, 2s_l \ldots \lfloor n/s_l \rfloor s_l$ as in Figure 4 (c) and (d).

Instead of setting all the weights w_t, w'_t and w_{ij} of the constraints in every scale uniformly, which makes the fine-scale constraints contribute more than the coarse-scale ones do, we set the sum of the weights of each scale identical to each other.

2.4 Optimization

While the general non-linear least squares problem [13] and its concrete application in bundle adjustment [18] have been well studied in the past decades. We still need carefully decide the weight of the energy terms in Equation 3, because such terms penalize different objectives are in different sensor systems that cannot be combined directly . In the following, we discuss the strategies of setting the weights. Then we briefly describe how to extend our method to incremental optimization, which is more useful for real-time applications.

Weight selection We set the β to be E_{tetra}/E_{bundle} with the initial errors empirically according to the extensive evaluation in [8]. However, the similar strategy does not work for setting α, because the initial error of E_{pose} is 0. Instead, we broke α as $\alpha_{rigid}\alpha_{norm}$. α_{norm} is a normalization factor that ensure the sum of all w_{ij} is the same as the sum of all w_t. α_{rigid} is to control how rigid the original local motion should be. For very noisy positional measurements, α_{rigid} should be increased to avoid overfit to $\mathcal{G}$. In our experiment, α_{rigid} is fixed to 0.1.

Incremental optimization For real-time applications, it is unaffordable to setup the tetrahedral network and optimize all the variables globally whenever new measurements arrive. Our framework can be easily modified to support incremental optimization similar to local bundle adjustment [4]. Let's call the last n frames that are involved in the incremental optimization as active frames. Only the constraints in Equation 3 that overlap the active frames are kept. Moreover, the parameters that lay in the non-active frames are fixed during the optimization.

3 Implementation and Experiments

We implemented a standard visual SLAM system to generate the visual measurement from a video taken by a monocular camera. We first detect and track features with Harris corners and KLT tracker. Only the tracks spanning more than five frames are kept for camera pose estimation. Visual keyframes are inserted whenever less than 70% of tracks are kept from last keyframes. To initialize the reconstruction, a sliding window

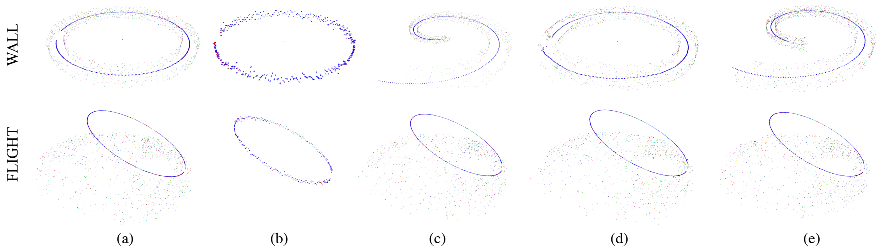


Fig. 5. Simulation datasets and results. Top row: WALL; bottom row: FLIGHT. (a) Ground-truth; (b) Perturbed GPS; (c) Visual SLAM; (d) Results by our method; (e) Results by IBA [8].

containing the last three consecutive keyframes is used to scan the tracked frames until a 5-point triplet reconstruction [12] is succeeded with a sufficient baseline and enough inlier tracks. Then the camera poses of consecutive frames are resectioned using 3-point pose estimation algorithm. A new 3D point is triangulated and verified whenever the baseline among its visible cameras is large enough. Local bundle adjustment [4] is used to improve the local consistency of the estimated 3D points and cameras. The implementation of our tetrahedral fusion according to Equation 3 is quite straight forward. Ceres Solver [1] is used to solve the non-linear least square optimization. All our implementation is written in C++ without any GPU optimization. The experiments are carried out on a PC equipped with Intel Core i7-930 CPU and 16GB ram. All parameters are set to the default value introduced in the paper, while the weight in Equation 3 are set by the strategy described in Section 2.4. The fusion process now runs on our test platform at about 9 fps.

3.1 Simulation Experiments

Because of the lack of ground-truth for the experiment data, we generate two simulation datasets, named as WALL and FLIGHT. WALL is to simulate a common forward moving motion for humans and ground vehicles. It is generated by constructing two concentric circular walls and making the camera move along the corridor between the two walls. The camera is kept looking forward during the whole simulation. FLIGHT is to simulate a typical flight of an UAV that takes video using a camera looking downwards vertically at the ground. The simulated camera takes off and lands, moving in circle and shooting at a flat ground. Random 3D points are sampled on the synthetic scenes and projected back to the moving cameras to construct the simulated feature tracks for SLAM. Each projection in feature tracks is perturbed by a random noise at 0.5 pixels. The predefined camera moving trajectories are considered as the ground truth GPS measurement $\{G_k^*\}$. The ground truth are shown in Figure 5 (a). Random perturbation σ_{gps} is added to $\{G_k^*\}$ to generate the perturbed GPS measurements $\{G_k\}$ as shown in Figure 5 (b). The reconstructed trajectories by SLAM as shown in Figure 5 (c) are used as the visual measurement $\{C_j\}$. In the following, we first evaluate our fusion result via visual inspection and then carefully study how the fusion works with respect to different parameter settings and noise levels comparing with the state-

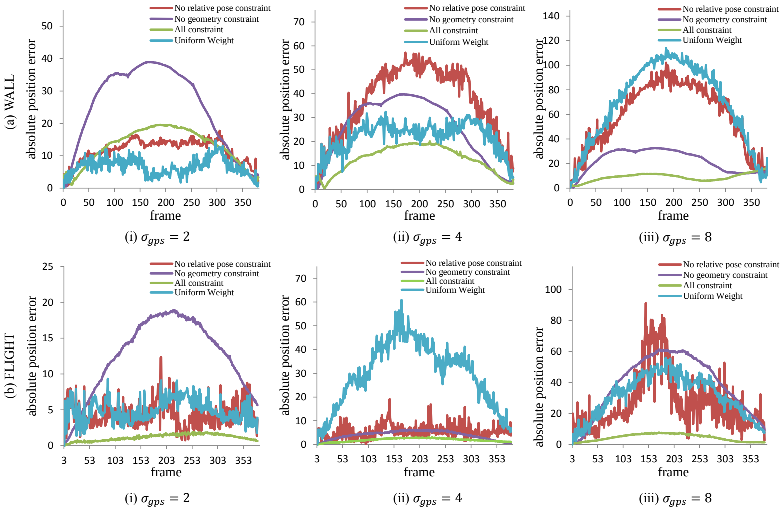
Table 1. Mean (m), standard deviation (σ) and maximum value (∞) of absolute position error of camera locations with respect to perturbed GPS (gps) and ground-truth GPS (gt). SLAM is the result of visual SLAM. Tetra is our method. IBA is method in [8].

		m^{gps}	σ^{gps}	∞^{gps}	m^{gt}	σ^{gt}	∞^{gt}
WALL	SLAM	123.5	67.39	212.1	123.5	67.42	212.0
	Tetra	**12.33**	**5.603**	**19.53**	**12.53**	**5.475**	**20.17**
	IBA	103.0	56.08	176.2	103.0	56.05	176.2
FLIGHT	SLAM	6.317	1.909	12.76	6.317	2.797	10.40
	Tetra	3.386	**1.833**	9.885	**1.481**	**0.7583**	**3.051**
	IBA	**3.299**	1.900	**9.771**	1.608	1.011	3.261

of-the-art positional fusion IBA [8]. The following experiments are performed by batch version of our method and IBA [8].

Qualitative evaluation The visual SLAM result of dataset WALL, as shown in Figure 5 (c), suffers from serious drift, because the forward moving motion gives very narrow baseline between consecutive frames and makes the reconstruction with large bias. However, after the tetrahedral fusion with the noisy GPS data as shown in Figure 5 (b), we get a visually plausible trajectory as shown in Figure 5 (d) that almost close the loop perfectly. In contrast, as shown in the top row of Figure 5 (e), IBA [8] cannot deal with drift because the initial similarity alignment is biased due to the drifted visual measurement and noisy positional measurements. In the dataset FLIGHT, Figure 5 (d) and (e) do not show too much difference visually between our method and IBA, since the result of visual SLAM has very small error.

Quantitative evalution Here we quantitatively evaluate our method using the absolute camera position error with respect to the ground truth GPS. Since our fusion framework yields only an up-to-scale reconstruction, an optimal similarity transform is computed to align the SLAM result and GPS before the computation of the absolute error. First, Table 1 illustrates the absolute position error of our method compared with original visual SLAM and IBA when the perturbation added to the ground-truth GPS is $\sigma_{gps} = 4$, where the size of the simulation scene is roughly 100. Then to evaluate the effectiveness of the energy terms in Equation 3, we carry out the fusion without the relative pose constraint E_{pose} and geometry constraint E_{bundle}. Moreover, we study the performance of our framework with uniform weights for each piece of constraint in Equation 3. In these experiments, we vary the perturbation of GPS to $\sigma_{gpsz} = 2, 4, 8$. The results are plotted and listed in Figure 6 and Table 2. We can easily find that our fusion method with adaptive weighting gives best results in terms of absolute position error in most cases. However, an exception is the case when the noise level is low. Such result is reasonable, because given the GPS measurements with very little noise, any good algorithm should rely on GPS measurement directly. Therefore, our fusion without relative pose gives better results than the fusion with all constrains. The uniform weighting also performs better because the uniform weighting strategy essentially weights less the relative pose



(i) $\sigma_{gps} = 2$ (ii) $\sigma_{gps} = 4$ (iii) $\sigma_{gps} = 8$

(i) $\sigma_{gps} = 2$ (ii) $\sigma_{gps} = 4$ (iii) $\sigma_{gps} = 8$

Fig. 6. Absolute position error of each frame with respect to the ground-truth GPS for the fusion using different configurations of energy terms and weights.

Table 2. Mean (m), standard deviation (σ) and maximum value (∞) of absolute position error with respect to the ground-truth GPS (gt) on simulation datasets by the fusion using different configurations of energy terms and weights.

		$\sigma_{gps} = 2$			$\sigma_{gps} = 4$			$\sigma_{gps} = 8$		
		m^{gt}	σ^{gt}	∞^{gt}	m^{gt}	σ^{gt}	∞^{gt}	m^{gt}	σ^{gt}	∞^{gt}
WALL	No relative pose	11.63	3.518	17.7	36.63	14.77	57.17	58.84	26.28	102.1
	No geometry	24.98	12.47	38.95	25.54	12.71	39.70	22.32	8.927	32.64
	All constraint	12.83	5.581	19.53	**12.63**	**5.600**	**20.04**	**8.979**	**3.169**	**16.78**
	Uniform Weight	**6.105**	**2.700**	**14.71**	21.43	6.407	32.34	68.40	32.26	114.0
FLIGHT	No relative pose	3.810	1.853	12.37	4.878	2.904	19.09	26.96	18.64	91.22
	No geometry	12.06	5.175	18.90	3.447	1.807	6.245	38.16	18.48	61.34
	All constraint	**0.614**	**0.696**	**1.908**	**1.481**	**0.7583**	**3.051**	**3.971**	**2.297**	**7.750**
	Uniform Weight	4.381	1.626	9.326	29.32	13.50	60.85	30.90	13.53	56.05

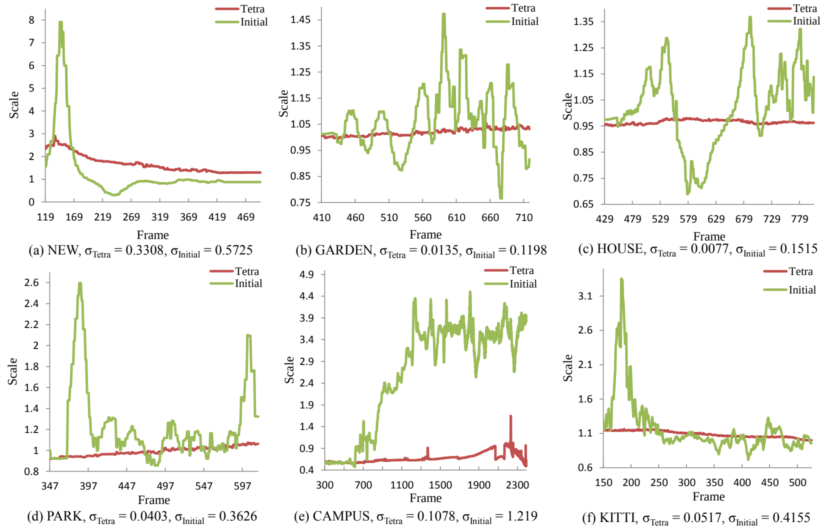
term since the number of the relative pose constraints is far less than the tetrahedral constraints.

3.2 Real-video Experiments

In this section, we test our fusion with six real videos divided in two groups. The first group includes five real videos with noisy or incomplete GPS data. The first video is part of part 1 in New College Dataset[16], which is called "NEW" in the following. The next three video "GARDEN", "HOUSE", "PARK" are taken by a monocular camera mounted on an unmanned aerial vehicle, the GPS measurement is output by the on-board flight controller. These videos are taken at 10Hz with the resolution 686×452 pixels, while the positional measurements are recorded at about 3Hz. The last video "CAMPUS" is captured on a ground vehicle with the resolution 640×480 pixels at 15Hz, while the G-

Table 3. Statistics on the running time of batch fusion for the real video datasets.

	# of iterations	# of visual frames	# of GPS frames	# of tracks	# of projections	Total time
NEW	163	2552	324	74717	1087035	21348.6s
GARDEN	158	881	288	47088	454655	172.5s
HOUSE	154	826	322	52904	453929	132.8s
PARK	186	632	247	30597	379133	250.9s
CAMPUS	167	2395	2395	149725	1011778	15996.9s
KITTI	171	666	564	162509	990748	13657.3s



(a) NEW, $\sigma_{\text{Tetra}} = 0.3308$, $\sigma_{\text{Initial}} = 0.5725$

(b) GARDEN, $\sigma_{\text{Tetra}} = 0.0135$, $\sigma_{\text{Initial}} = 0.1198$

(c) HOUSE, $\sigma_{\text{Tetra}} = 0.0077$, $\sigma_{\text{Initial}} = 0.1515$

(d) PARK, $\sigma_{\text{Tetra}} = 0.0403$, $\sigma_{\text{Initial}} = 0.3626$

(e) CAMPUS, $\sigma_{\text{Tetra}} = 0.1078$, $\sigma_{\text{Initial}} = 1.219$

(f) KITTI, $\sigma_{\text{Tetra}} = 0.0517$, $\sigma_{\text{Initial}} = 0.4155$

Fig. 7. The scale estimation on the real video datasets. Green line: the initial scale estimated by aligning the SLAM measurements of the latest 20 frames with the corresponding GPS. Red line: the scale estimated by aligning our fused measurements with the corresponding GPS measurements. $\sigma_{initial}$ and σ_{tetra} are the standard deviation of the initial scales and the scale estimated by our fusion respectively.

PS measurements are recorded at about 10Hz. The second group contains one real video “2011_09_26_drive_0117” in the raw data in KITTI vision benchmark[5], which is one of the longest videos of raw data. In the following, this video is called “KITTI”. The GPS in this data is enough accurate and regarded as ground-truth on the benchmark. We regard the original GPS as ground-truth in our experiment and add Gaussian noise to the original GPS data to generate noised GPS. The scale of datasets in our experiment and running time of batch fusion is listed in Table 3.

Group without ground-truth Since we do not have the access to the specification of the GPS sensors, we have little knowledge on the accuracy of the GPS measurement. To quantize the magnitude of the noise of GPS, we compute the standard deviation of the magnitude of angular acceleration, listed as σ_a at the top row of Table 4. Figure 8 (i-i) shows the noisy GPS measurement of NEW, GARDEN and CAMPUS visually .

Table 4. Mean (m) and standard deviation (σ) of absolute position error of camera locations with respect to GPS measurement (gps) for real data. m^{2d} is the mean value of the ratio between the reprojection error after fusion and the reprojection error of SLAM [8]. Tetra is our method. IBA is method in [8].

	NEW, $\sigma_a = 15.66°$			GARDEN, $\sigma_a = 25.80°$			HOUSE, $\sigma_a = 24.44°$		
	m^{gps}	σ^{gps}	m^{2d}	m^{gps}	σ^{gps}	m^{2d}	m^{gps}	σ^{gps}	m^{2d}
SLAM	6.12037	3.60924	(1)	2.833	1.000	(1)	4.040	1.365	(1)
Tetra	**0.03089**	**0.5483**	**1.004**	0.115	**0.566**	**1.624**	**0.2857**	**0.6410**	**3.713**
IBA	2.38272	1.56625	24.82	**0.0729**	0.5662	1.64673	1.500	0.7680	4.47081
	PARK, $\sigma_a = 24.68°$			CAMPUS, $\sigma_a = 31.04°$					
	m^{gps}	σ^{gps}	m^{2d}	m^{gps}	σ^{gps}	m^{2d}			
SLAM	3.170	2.223	(1)	157.1	80.97	(1)			
Tetra	**0.2713**	**0.6276**	**9.752**	**2.520**	2.698	**1.582**			
IBA	0.4737	0.6376	26.2039	7.319	**1.360**	2.8881			

Table 5. Mean (m^{gt}) and standard deviation (σ^{gt}) of absolute position error of camera locations with respect to original GPS measurement for “KITTI”.m^{gps} and σ^{gps} is with respect to noised GPS measurement.m^{2d} is the mean value of the ratio between the reprojection error after fusion and the reprojection error of SLAM [8]. Tetra is our method. IBA is method in [8].

	m^{gt}	σ^{gt}	∞^{gt}	m^{gps}	σ^{gps}	∞^{gps}	m^{2d}
SLAM	5.456	2.174	8.731	5.863	2.151	9.436	(1)
Tetra	**0.5621**	**0.6313**	**1.928**	**0.9096**	**0.7564**	**3.335**	**0.9969**
IBA	0.6365	0.8466	2.495	0.9273	0.8843	3.417	7.340

Though the σ_a of NEW is not large, Figure 8 shows that its GPS data is incomplete, which corresponds to the situation where GPS information cannot be obtained in urban valley environment. With such noisy GPS measurements, we estimate the scale factor between GPS and visual measurements by finding the best similarity transformation to align the latest 20 GPS measurements with the corresponding SLAM measurements. The green plots in Figure 7 show that the estimated scale is very noisy, which makes the fusion with positional measurement not stable using alignment-based method. Even worse, the SLAM measurements contain very large drifting as shown in Figure 8 (c.i) and Figure 7 (e). In contrast, even without the initial alignment, our method successful fuses the GPS and SLAM measurements, which in turn makes the estimation of scales very robust as shown in the red plots of Figure 7.

To compare the absolute position errors with IBA, we take all SLAM measurements and corresponding GPS measurements to estimate a robust similarity transformation, so that IBA can successfully fuse the GPS and SLAM measurements. We also compute the ratio between the reprojection error of fusion results and the reprojection error of SLAM for a comparison. The statistic of the comparison listed in Table 4 shows that our fusion method gives the best results in most cases. Although sometimes IBA gives comparable results, it is noted that IBA requires a robust alignment before fusion, while ours can fuse the positional measurements even such alignment is not valid.

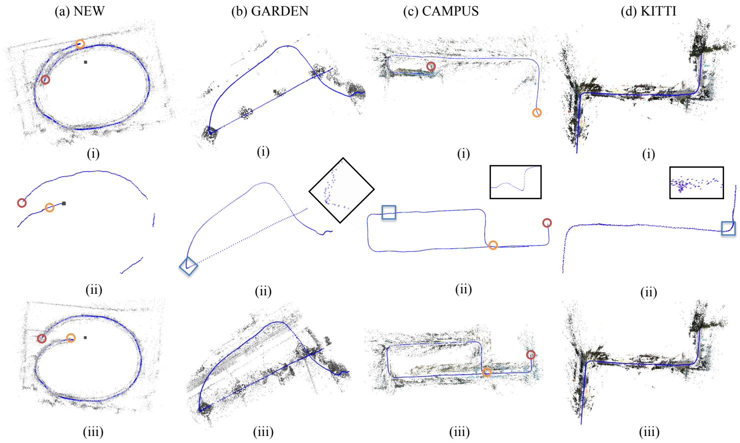


Fig. 8. The results of batch version of our method on NEW, GARDEN, CAMPUS and KITTI. (i) The results of SLAM; (ii) GPS measurements(noised GPS for KITTI), the region in the blue rectangle is zoomed in and shown in the black rectangle; (iii) Our fusion results. The orange and red circles indicate the corresponding visual and GPS measurements to show the large drift in visual SLAM.

Group with ground-truth Since the GPS data of "KITTI" is enough accurate and regarded as ground-truth, we add Gaussian noise with $\sigma = 0.2$ in three directions to the original GPS data and the original GPS data is used as ground-truth to give quantitative evaluation in our experiment. The detail of noised GPS is shown as Figure 8 (d.ii). As shown in Table 5, our method generates better results than the-state-of-art method IBA [8]. We also estimate the scale factor between GPS and visual measurements as above experiment on dataset without ground-truth and Figure 7 (f) shows that the result is robust.

4 Conclusions

In this paper, we propose a multi-scale tetrahedral fusion framework. The key insight of our method is the usage of the ratio of distances that is invariant under similarity transformation, which decouples the task of fusion from the task of similarity alignment. The tetrahedral network ensures a sparse sampling of the ratio constraints, while the multi-scale scheme further adapts the fusion to different level of noisy positional measurements. Our framework is capable of fusing a similarity reconstruction with positional data even when the similarity alignment is not valid. The fused results can help to resolve the scale ambiguity robustly.

Acknowledgement. Real-time videos Garden, House, Park are provided by DJI and CAMPUS is provided by Key Laboratory of Machine Perception (Ministry of Education) in Peking University. This work is supported by RGC-GRF 618711, RGC/NSFC N_HKUST607/11, ITC-PSKL12EG02, and National Basic Research Program of China (2012CB316300).

References

1. Agarwal, S., Mierle, K.: Ceres Solver: Tutorial & Reference. (Google Inc.)
2. Dusha, D., Mejias, L.: Error analysis and attitude observability of a monocular gps/visual odometry integrated navigation filter. The International Journal of Robotics Research **31** (2012) 714–737
3. Engel, J., Sturm, J., Cremers, D.: Camera-based navigation of a low-cost quadrocopter. In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, IEEE (2012) 2815–2821
4. Eudes, A., Lhuillier, M.: Error propagations for local bundle adjustment. In: CVPR. (2009) 2411–2418
5. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
6. Konolige, K., Agrawal, M.: FrameSLAM: From bundle adjustment to real-time visual mapping. Robotics, IEEE Transactions on **24** (2008) 1066–1077
7. Kummerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: g^2o: A general framework for graph optimization. In: ICRA. (2011) 3607–3613
8. Lhuillier, M.: Incremental fusion of structure-from-motion and gps using constrained bundle adjustments. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2489–2495
9. Mei, C., Sibley, G., Cummins, M., Newman, P., Reid, I.: Rslam: A system for large-scale mapping in constant-time using stereo. International Journal of Computer Vision **94** (2011) 198–214
10. Michot, J., Bartoli, A., Gaspard, F.: Bi-objective bundle adjustment with application to multi-sensor slam. 3DPVT **3025** (2010)
11. Morris, D.D.: Gauge Freedoms and Uncertainty Modeling for Three-dimensional Computer Vision. PhD thesis, Carnegie Mellon University (2001)
12. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004) 756–777
13. Nocedal, J., Wright, S.: Numerical Optimization. Springer (2000)
14. Nützi, G., Weiss, S., Scaramuzza, D., Siegwart, R.: Fusion of imu and vision for absolute scale estimation in monocular slam. Journal of intelligent & robotic systems **61** (2011) 287–299
15. Rehder, J., Gupta, K., Nuske, S., Singh, S.: Global pose estimation with limited gps and long range visual odometry. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on, IEEE (2012) 627–633
16. Smith, M., Baldwin, I., Churchill, W., Paul, R., Newman, P.: The new college vision and laser data set. The International Journal of Robotics Research **28** (2009) 595–599
17. Sibley, G., Sukhatme, G., Matthies, L.: The iterated sigma point kalman filter with applications to longrange stereo. In: Proceedings of Robotics: Science and Systems. (2006) 263–270
18. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment - a modern synthesis. In: Proceedings of the International Workshop on Vision Algorithms: Theory and Practice, Springer-Verlag (2000) 298–372