16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Integrating Visual and Geometric Consistency for Pose Estimation

Huiqin Chen, Emanuel Aldea, Sylvie Le Hégarat-Mascle
SATIE - CNRS UMR 8029, Paris-Sud University, Paris-Saclay University
Building 660, rue Noetzlin, Gif-sur-Yvette 91190, France
{huiqin.chen, emanuel.aldea, sylvie.le-hegarat}@u-psud.fr

Abstract

In this work, we tackle the problem of estimating the relative pose between two cameras in urban environments in the presence of additional information provided by low quality localization and orientation sensors. An M-estimator based approach provides an elegant solution for the fusion between inertial and vision data, but it is sensitive to the prior importance of the visual matches between the two views. In addition to using cues extracted from local visual similarity, we propose to rely at the same time on learned associations provided by the global geometrical coherence. A conservative weighting scheme for combining the two types of cues has been proposed and validated successfully on an urban dataset.

1 Introduction

This paper aims at exploiting and improving the fusion of pose priors from sensors (GPS, accelerometers, gyroscopes, magnetometers) and image features for the estimation of relative camera pose. The estimation of relative camera pose is a fundamental step for many computer vision tasks, such as structure from motion(SFM), SLAM or object tracking. Traditional pose estimation relies on image features: given a pair of images (I_a, I_b), the standard procedure may be described as follows:

- Extract SIFT keypoints for (I_a, I_b) and assign a built-in descriptor to each keypoint;
- Construct keypoint correspondences from I_a to I_b by matching them using SIFT descriptor distance minimization with a rejection threshold based on the ratio of the two smallest distances;
- Calculate the relative pose (R, t) by decomposing the fundamental or essential matrix after having computed it using a robust estimation RANSAC on correspondences, or estimate (R, t) directly using an iterative estimation method such as an M-estimator.

However, this traditional strategy only based on image feature is limited by inaccurate matching due to the presence of repetitive features and occlusions, which often occur in urban scenes. Nowadays, sensors available in GPS receivers or IMUs are widely integrated with camera devices in smart phones. These sensors provide a readily available camera pose information, albeit often very imprecise. Despite its uncertainty, the sensor prior can play a crucial role in improving pose estimation based on image feature.

There exist some fundamental approaches for integrating the pose prior with image information. The authors of [1] fuse the pose prior provided by IMU during the matching step in a process which can be considered as filtering for the relative pose estimation. They constrain the search area for correspondences around the epipolar line defined by sensor data in order to guide the matching. As mentioned in [2], this method is very sensitive to sensor noise.

For many works on visual inertial SLAM using *temporal* sequences, a common fusion method is to constrain separately the image based estimation and pose prior as prediction step and correction step for Kalman filtering [11, 3]. Provided that video sequences cover in detail the area of interest, even pure visual SLAM may provide accurate results for registering cameras in urban scenes [10]. However, in our work we study the more constrained scenario in which a single pair of images is available, along with a low quality pose prior provided by low cost GPS and inertial sensors. Indeed, in metropolitan areas video-recording (especially using UAVs) is highly regulated, and even if videos from ground-level dynamic cameras are available, they have often low quality and are heavily occluded.

In the last years, minimizing a loss function including both image features and pose prior by non linear optimization has been shown to be more accurate than Kalman filtering in visual-inertial SLAM [4]. To the best of our knowledge, [2] is the first work who extends a similar idea for the relative camera pose estimation for a single pair of views. Instead of using RANSAC, they propose an algorithm called SOREPP which relies on the fusion of putative correspondences and of noisy pose priors from sensors. In order to be robust to outliers, they use the following M-estimator:

$$\hat{s} = \arg\min_{s} \left\{ c \left(\sum_{k \in \Omega} w(k)(1 - g(k, s)) \right) + \lambda(s)^2 \right\}, \tag{1}$$

where c is a weighting parameter, Ω is the set of putative correspondences and $w(\cdot)$ their weights, $g(k, s)$ is a Gaussian score evaluated for the correspondence k with respect to the relative pose s, and the regularization term $\lambda(s)$ is a distance measure between s and the sensor prior.

The performance of SOREPP depends on the sensor precision as well as on the quality of correspondences in the set Ω and the values of $w(\cdot)$. If the GPS/IMU uncertainty is too high, the regularization term value $\lambda(s)$ decreases and impacts less the optimization. The presence of a significant ratio of outliers in Ω makes it difficult for the M-estimator to find good solutions, except with very precise pose priors. In order to guide the M-estimator, SOREPP estimates a correspondence

weight in the form of a rough approximation of the prior probability of being an inlier [2] (see Section 2.1). Although it has the advantage of being simple, the proposed weighting tends to be unreliable in scenes with repetitive patterns such as urban contexts.

Our work aims to further improve the estimation of weights for correspondences, and therefore to improve the performance of pose estimation based on SOREPP. To this aim, we propose that weights be also based on a global geometry consistency estimated by a deep network recently introduced in [5]. Then, we explore how to combine this additional criterion with the weighting strategy based on local classical features in SOREPP. Our main contributions consist in:

- applying jointly a neural network with a pose prior for robust relative camera pose estimation;
- developing an effective correspondence weighting strategy by exploiting SIFT features as well as geometric consistency.

2 Proposed estimation

In the following, we discuss the weighting strategy based on local visual features from [2], as well as on global geometry consistency as introduced recently in [5]. The last part of this section highlights how these two weighting strategies guided by fundamentally different objectives may be used jointly in order to obtain robust correspondence weights.

2.1 Weights from classical features

A set of putative correspondences Ω is constructed using the classical SIFT nearest neighbor strategy [6] with a standard ratio threshold of 0.75. SOREPP defines independently the weight $w(k)$ for each correspondence k based on the SIFT ratio test [6] :

$$\forall k \in \Omega, w(k) = 1 - \frac{d_{1NN}(k)}{d_{2NN}(k)}. \quad (2)$$

The underlying assumption is that more distinctive the first nearest neighbor, more likely the matching corresponds to an inlier. However, this cue which is based only on the descriptor appearance, is not robust in presence of repetitive features and occlusion. Furthermore, while decreasing the number of outliers, the non-adaptive ratio threshold reduces as well the number of inliers, an outcome which is undesirable when inliers are scarce. Next we will discuss the weighting strategy based on global geometry consistency which avoids relying on a fixed ratio threshold.

2.2 Weights from the geometry network

Given all coordinates of putative correspondences in $\Omega = \{k_1, k_2, ..., k_i, ..., k_N\}$, with $k_i = (x_{ai}, y_{ai}, x_{bi}, y_{bi}) \in \Omega$, for an image pair (I_a, I_b), which is constructed by classical SIFT nearest neighbor strategy without ratio threshold, a weakly supervised deep network is trained in [5] to estimate a global geometrically consistent weight $w_k \in [0, 1]$ for each correspondence:

$$\{w(k_1), w(k_2), ..., w(k_i), ..., w(k_N)\} = f_\Phi(\Omega), \quad (3)$$

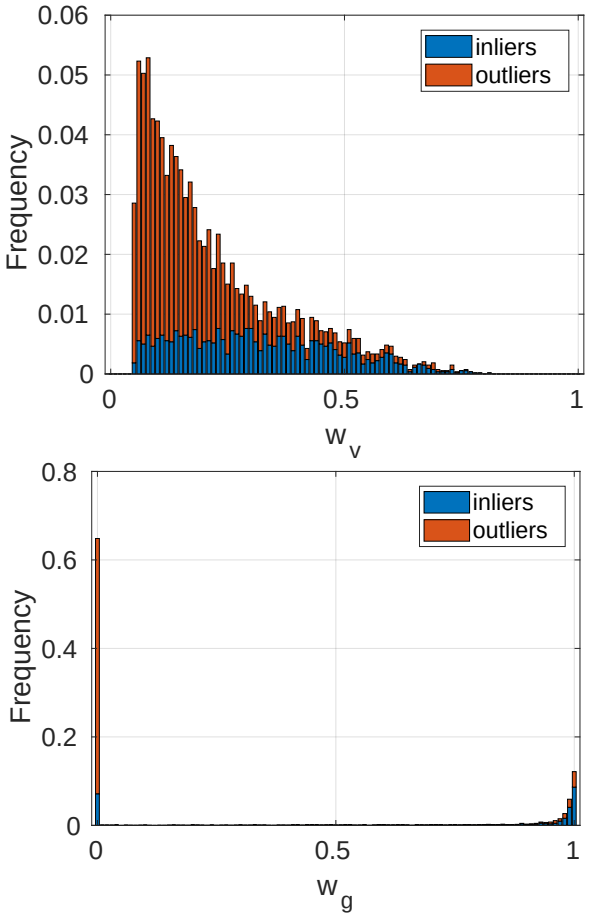


Figure 1: Histogram of inliers and outliers for w_v (above) and w_g (below). The neural network provides a much crisper output compared to a traditional appearance based evaluation.

where $f_\Phi(\cdot)$ is the regression of a deep network with parameters Φ. If $w(k_i) = 0$, the correspondence k_i is likely to be an outlier and classified as such, following the precautionary principle.

The core idea is that inliers are geometrically structured, conversely to outliers. Unlike the weight based on ratio test in Section 2.1 that depends only on the considered correspondence, when considering geometric consistency all correspondences contribute to the estimation of $w(k_i)$. Because this globally geometric weight is independent of appearance, it can potentially avoid the appearance ambiguity for repetitive features.

2.3 Exploiting both cues

It is useful to exploit jointly the local appearance cue and global geometric consistency cue because of their complementary information and their relative balance in performance and precision. In the following part, different combinations of these cues are investigated and compared. For the sake of simplicity, w_v denotes the weight from classical visual features given by Eq. (2), w_g the weight from the geometry network given by Eq. (3) and $w^{(i)}$ the combined weight to be investigated for a specific correspondence k.

Linear operation. The simplest fusion approach

is to consider the average of w_g and w_v:

$$w^{(1)} = \frac{w_g + w_v}{2}, \tag{4}$$

Averaging can be seen as a weight filtering which allows to keep considering correspondences having possible a very low weight w_v or w_g, while reinforcing correspondences with both high weights. In our case, this averaging approach is taken as the baseline.

Regression. Instead of applying a simple averaging operation, it is possible to learn automatically the fusion weight by logistic regression :

$$w^{(2)} = g(w_g, w_v), \tag{5}$$

where $g(\cdot)$ represents the bivariate logistic regression function.

In order to train the regression model, we choose image pairs in the dataset of [5] and match corners using the classical nearest neighbor SIFT strategy. The values w_v are computed during the matching step, whereas values w_g are obtained by taking the matched correspondence coordinates for each image pair as the input of neural network model in [5] with the trained model. By choosing a fixed Sampson distance threshold of matching points to the epipolar line, each correspondence is labeled as 1 for inliers or 0 for outliers. As expected, in our setting the resulting weight is close to 1 when w_v is larger than 0.8 irrespective of the value of w_g. This seems to be reasonable since for few outliers the ratio test is higher than 0.8 according to Fig. 1, upper histogram. Compared to averaging, regression tends overall to increase the combined weight, which might increase more the influence of potential inliers but also that of some outliers.

Conservative weighting. The previous strategies introduce more or less some confusion between inliers and outliers in terms of weighting for the M-estimator based optimization. In order to be stricter, we propose to adopt a conservative weighting based on a pessimistic function such as $\min(w_g, w_v)$. One significant consequence is that, in contrast to previous strategies, the min operation discards entirely the correspondences with $w_g = 0$, which exhibit a high ratio of outliers.

Since w_v seems to be more unreliable according to Fig. 1, the symmetric weight from visual cues denoted by w_{vs} (computed by interchanging for the association the source and destination images in the pair) is also taken into account in order to constrain more the influence of visual part, and w_v is substituted by $min\{w_v, w_{vs}\}$. However, the geometry network exhibits already a conservative behaviour in the way it outputs the weights (see Fig. 1, lower histogram), thus the symmetric weight from the network denoted by w_{gs} is used conversely in order to allow more prospective points identified by geometric coherence: $\max\{w_g, w_{gs}\}$. Finally, by taking into account the particular behaviour of the two weighting algorithms, the fusion weight is computed by the following formula:

$$w^{(3)} = \min\left\{\max\left\{w_g, w_{gs}\right\}, \min\left\{w_v, w_{vs}\right\}\right\}, \tag{6}$$

The result in Eq. (6) is also due to the fact that the loss function used in the neural network encourages a crisp decision regarding the nature of each input observation, a behavior which is clearly visible as well in Fig. 1 when comparing the geometry weighting with the histogram of w_v values.

3 Experiments

In the experimental part, we evaluate the performance of the weighting strategies introduced in the previous section, and we compare them to SOREPP algorithm [2] and to the geometric network [5] on a dataset collected in a challenging urban environment. We start by presenting the dataset, then we introduce the implementation details followed by the evaluation metric. The quantitative details are then discussed.

3.1 Dataset

We collected an urban scene dataset containing 32 images acquired with a smart phone in front of a major railway station, some of the images being taken at ground level and some from the upper floors of a building. As in [2], the *GeoCam* application was used to record in each image header the approximate pose provided by the embedded sensors. The ground truth was constructed by feeding all the images into VisualSFM [7]. The relative poses were varied gradually in order to ensure a high-quality VisualSFM estimation, and also to provide image pairs with varying degrees of difficulty for our pose estimation problem. The dataset is provided to the academic community[1].

3.2 Implementation

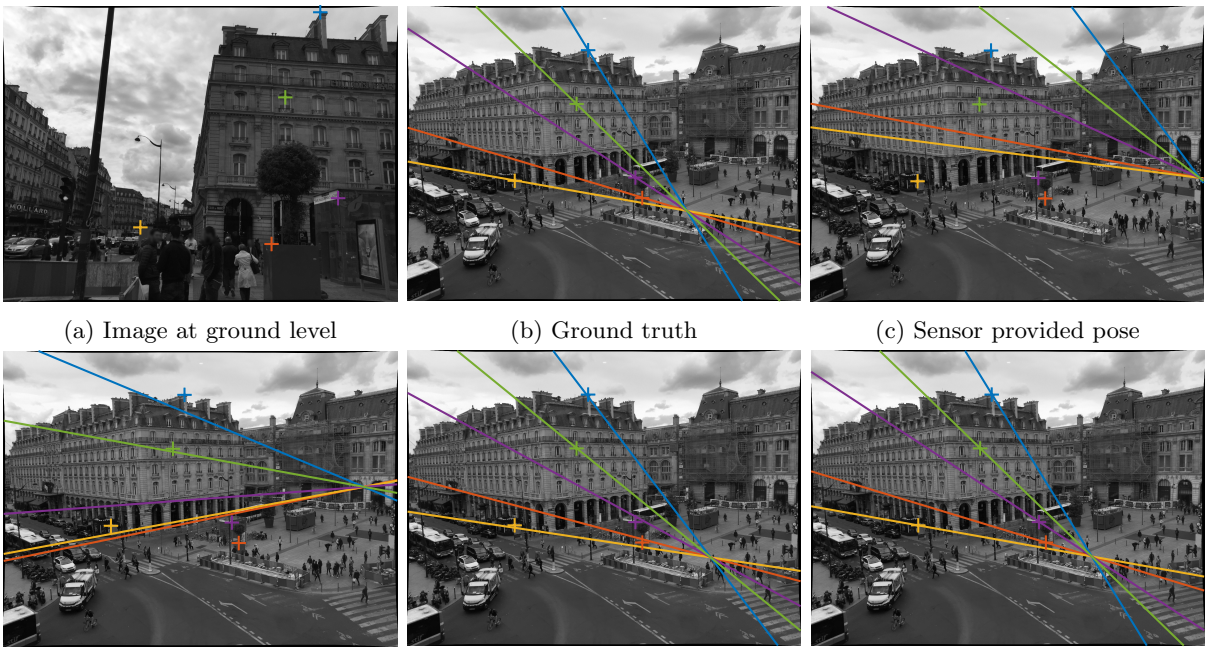
Given an image pair from the dataset, we compute two correspondence sets Ω_1 and Ω_2 using the SIFT nearest neighbor strategy with different ratio test thresholds. A standard ratio test threshold 0.75 is chosen for Ω_1. For Ω_2, we choose a very high value 0.95 which almost amounts to cancelling the ratio test, but it actually still helps eliminating a fair number of correspondences. We compute the weight w_v as in Eq. (3) for each correspondence in Ω_1 and Ω_2 during the matching step. Furthermore, we take all normalized matching point coordinates in Ω_2 as the input of the geometric network in [5] and we get a geometric weight w_g for each correspondence in Ω_2. We use Ω_2^* to denote the subset of Ω_2 where w_g is greater than 0 and each correspondence in Ω_2 is normalized. We applied the model trained by the authors of [5] on outdoor scenes. At the same time, we also compute three versions of fusion weight $w^{(1)}$, $w^{(2)}$ and $w^{(3)}$ for Ω_2 separately with Eq. (4), (5) and (6).

According to the various inputs and algorithms, we classify the different estimation methods as shown in Tab 1.

3.3 Evaluation metric

Given the estimated rotation matrix and translation vector (R_{es}, t_{es}) and the ground truth (R_{gd}, t_{gd}) for relative camera poses, we compute the rotation error δR and translation error δt separately for the evaluation.

[1]The data used in this work may be found at: `http://hebergement.u-psud.fr/emi/S2UCRE/ChenMVA19.zip`



(a) Image at ground level (b) Ground truth (c) Sensor provided pose

(d) GEO-RANSAC [5] (e) SIFT-SOREPP [2] (f) Ours-CNSV

Figure 2: Illustration of pose estimation results between a ground level view (Fig. 2a) and an overview camera. The ground truth (Fig. 2b) is computed using SfM on a larger set of images. The quality may be assessed visually based on the position of the manually defined control points with respect to their epilines, and based on the location of the epipole.

Table 1: Estimation method

Name	Algorithm	Input
Sensor	-	-
SIFT-RANSAC	RANSAC	Ω_1
GEO-RANSAC [5]	RANSAC	Ω_2^*
SIFT-SOREPP [2]	SOREPP	Ω_1, w_v, sensor
GEO-SOREPP	SOREPP	Ω_2 , w_g, sensor
Ours-LO	SOREPP	Ω_2, $w^{(1)}$, sensor
Ours-REG	SOREPP	Ω_2, $w^{(2)}$, sensor
Ours-CNSV	SOREPP	Ω_2 , $w^{(3)}$, sensor

For δR, we firstly compute the relative rotation matrix ΔR between R_{es} and R_{gd}, with $\Delta R = R_{es}^T \cdot R_{gd}$. ΔR can be represented by a rotation of angle ϕ around a vector v. Smaller is ϕ, closer R_{es} is to R_{gd} . Thus the extent of the rotation error δR can be approximated by ϕ which is computed as (see page 584 in [8]):

$$\delta R \approx \phi = \arccos\left(\frac{\mathrm{Tr}(\Delta R) - 1}{2}\right), \tag{7}$$

with $Tr(\cdot)$ the trace of a matrix.

The translation error δt is computed as the angle between t_{es} and t_{gd} :

$$\delta t = \arccos\left(\frac{t_{es} \cdot t_{gd}}{\|t_{es}\|\|t_{gd}\|}\right), \tag{8}$$

In the following, we take the pose estimation error as the maximum value between δR and δt. To evaluate the performance of each method, we compute the cumulative curve of pose estimation error taken, as previously performed in [5] and in [9]. For each method, the corresponding curve presents the percentage of image pairs whose pose is successfully estimated with respect to a given error threshold. As a qualitative result, we illustrate in Fig. 2 a relative pose estimation for an image pair, and the corresponding epipolar lines and epipole locations for the different methods considered.

3.4 Quantitative evaluation

The evaluation results are presented in Fig. 3. When comparing the weighting strategies and the other methods, the Ours-CNSV approach which uses the conservative fusion rule shows the best performance. The important role of the noisy pose priors provided by sensors is confirmed by the SOREPP based estimation (SIFT-SOREPP) which improves over the pure vision based algorithms, namely classical SIFT-RANSAC and the geometry based network (GEO-RANSAC). The linear weighting and the bivariate regression weights are less effective in supporting the M-estimator, especially in the high precision range (error smaller than 5 degrees) which is desirable for the accurate localization of elements of interest in the camera fields of view. Our experiments confirm systematically that relying on the more permissive Ω_2 threshold *for visual filtering* is more effective than using the stricter Ω_1, because the *global* geometric consistency provided by the neural network identifies more reliably the inliers than the *local* fixed ratio test threshold used classically for SIFT matching.

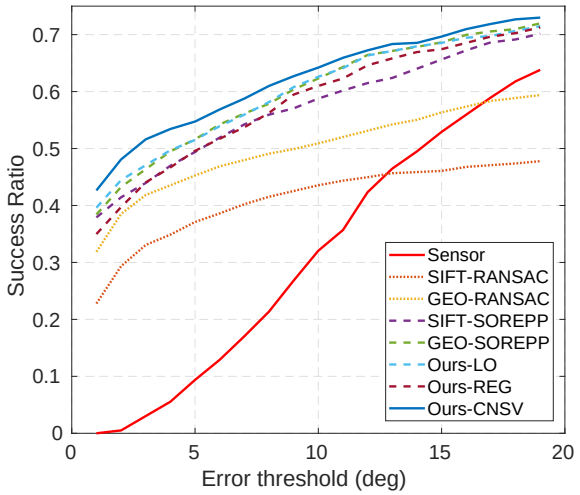


Figure 3: Rotation and translation success ratio of various pose estimation algorithms depending on a given error threshold.

4 Conclusion

For the relative pose estimation problem, we proposed a strategy to rely at the same time on local information provided by visual similarity, and by a global geometrical coherence of the transformation between two views. We combined the two types of cues along with localization and orientation information within an M-estimator, with excellent results on real data acquired in an urban scenario.

In future work, we intend to formalize the usage of prior knowledge in the combination process in order to extend it for other sources of information assessing the reliability of interest point matches, i.e. belonging to the same semantic category or exhibiting local similarity in terms of a learned function.

Acknowledgements

This work was funded by the French National Research Agency grant ANR-16-SEBM-0001 (the "S^{2}UCRE" project) within the French-German funding framework "Safety and Security of Urban Crowded Environments".

References

[1] DR Wong, MP Hayes and al.: "IMU-aided SURF feature matching for relative pose estimation," *Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference of*, pp.1–6,IEEE, 2010.

[2] Y Goldman, E Rivlin,I Shimshoni: "Robust epipolar geometry estimation using noisy pose priors," *Image and Vision Computing*, vol.67, pp.16–28, 2017.

[3] R Brockers, S Susca, D Zhu and al.: "Fully self-contained vision-aided navigation and landing of a micro air vehicle independent from external sensor inputs," *Unmanned Systems Technology XIV*, vol.8387, pp.83870Q, International Society for Optics and Photonics, 2012.

[4] S Leutenegger, S Lynen, M Bosse and al.: "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol.34, n.3, pp.314–334, SAGE Publications Sage UK: London, England, 2015.

[5] KM Yi, E Trulls, and Y Ono , V Lepetit and al.: "Learning to Find Good Correspondences," *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, CONF, 2018.

[6] DG Lowe: "Distinctive image features from scale invariant keypoints," *International journal of computer vision*, vol.60, n.2, pp.91–110, 2004.

[7] C Wu: "VisualSFM: A visual structure from motion system," 2011.

[8] R Hartley, A Zisserman: "Multiple view geometry in computer vision," *Cambridge university press*,2003.

[9] JW Bian, WY Lin, Y Matsushita and al.: "Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence," *Computer Vision and Pattern Recognition, 2017 IEEE Conference on*, pp.2828–2837, IEEE, 2017.

[10] T Pollok, E Monari: "A visual SLAM-based approach for calibration of distributed camera networks," *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS*, pp.429–437, IEEE, 2016.

[11] M Pollefeys, D Nistér, J-M Frahm, A Akbarzadeh, P Mordohai, B Clipp, C Engels,D Gallup, S-J Kim, P Merrell and al.: "Detailed real-time urban 3d reconstruction from video," *International Journal of Computer Vision*, vol.78, n.2-3, pp.143–167, Springer, 2008.