2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Improving Visual SLAM in Car-Navigated Urban Environments with Appearance Maps

Alberto Jaenal*, David Zuñiga-Noël*, Ruben Gomez-Ojeda, Javier Gonzalez-Jimenez

Abstract— **This paper describes a method that corrects errors of a VSLAM-estimated trajectory for cars driving in GPS-denied environments, by applying constraints from public databases of geo-tagged images (Google Street View, Mapillary, etc). The method, dubbed Appearance-based Geo-Alignment for Simultaneous Localisation and Mapping (AGA-SLAM), encodes the available image database as an *appearance map*, which represents the space with a compact holistic descriptor for each image plus its associated geo-tag. The VSLAM trajectory is corrected on-line by incorporating constraints from the recognized places along the trajectory into a position-based optimization framework. The paper presents a seamless formulation to combine local and absolute metric observations with associations from Visual Place Recognition. The robustness of the holistic image descriptor to changes due to weather or illumination variations ensures a long-term consistent method to improve car localization. The proposed method has been extensively evaluated on more than 70 sequences from 4 different datasets, proving out its effectiveness and endurance to appearance challenges.**

I. Introduction

Despite great advances in Visual Simultaneous Localization and Mapping (VSLAM) in the last decades, long-term operation in challenging scenarios still remains an open problem [1]. It is well-known that loop closure is one of the keys to achieve a consistent map and precise localization. In the context of car localization and mapping in urban environments, driving in the same place more than once during a tour is not that common, consequently loop closures constraints can not be applied during a large trajectory and the odometry errors grow without bounds. Under the perspective of long-term SLAM, previously built 3D maps can be used as a reference to correct the drift of a real-time computed trajectory, which requires efficient 2D-to-3D matching of local descriptors [2]. This approach, however, presents two serious limitations: first, 3D models scale poorly in large environments [3] and secondly, since the corresponding descriptors are likely to be recorded under very different appearance conditions (e.g. from day/night cycles or cross seasons), feature-based matching is prone to failure.

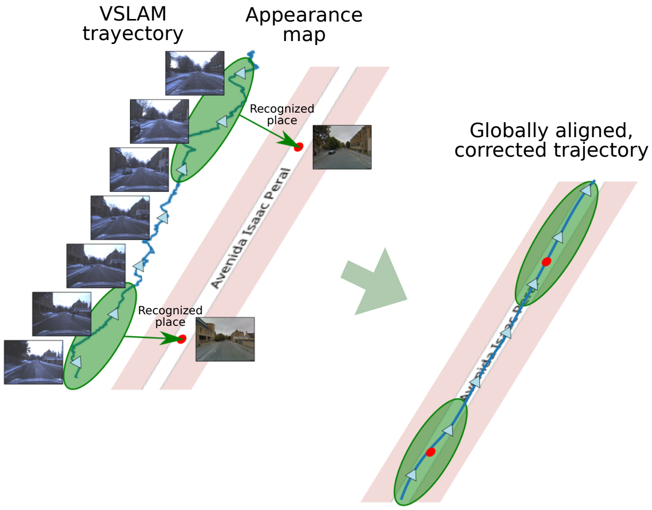

Fig. 1: AGA-SLAM combines geometric information from VSLAM keyframes (blue triangles) and absolute positions from an appearance map (red dots) through holistic VPR techniques (green ellipses), resulting in a corrected trajectory aligned with the map.

Our proposal to correct the trajectory of the vehicle relies on using compact whole-image (holistic) descriptors that can be designed to be robust to lighting changes, as demonstrated in recent publications [4], [5]. Specifically, we propose to enforce the local estimations of any VSLAM solution (ORB_SLAM2 [6], PL-SLAM [7], LSD-SLAM [8], etc) to be compliant with geometric position constraints imposed publicly available 2D databases of geo-tagged images (Google Street View, Mapillary, etc). Such dataset must be previously transformed into a city-scale *appearance map*, composed of holistic image descriptors and the associated geo-tags (i.e. latitude and longitude). Thus, for any keyframe along the trajectory, a Visual Place Recognition (VPR) method [9] can efficiently query the *appearance map*, rising similarity relations between the keyframe and the recognized place positions (see Figure 1).

The combination of the appearance similarity constraints and those from the VSLAM localization output (odometry and loop closure) is not straightforward. We present a seamless formulation that regularizes, through appearance similarity, an optimizable position graph. We report localization improvements of a state-of-the-art VSLAM solution (ORB_SLAM2) in 4 car-mounted public datasets presenting different perceptual challenges.

* These authors contributed equally.

This work has been funded by the Government of Spain in part under project FPU17/04512, and in part under project WISER (DPI2017-84827-R), financed by the European Regional Development's funds (FEDER). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal used for this research.

The authors are with Machine Perception and Intelligent Robotics (MAPIR) Group, Department of Systems Engineering, University of Malaga, 29071 Malaga, Spain {alberto.jaenal, dzuniga, rubengooj, javiergonzalez}@uma.es

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

The reader may be wondering why not use GPS data to correct the visual trajectory, for example, integrating both estimations with a bayesian filter (e.g. Kalman filter). The main reason is that the precision of GPS receivers can drop substantially (or even fail) in urban, densely built-up environments due to the so-called *street canyon* effect [10]. Additionally, such integration is not immediate [11], involving additional problems, such as keeping accurate time synchronization for long periods. Then, having a solution that can work in GPS-denied areas is of clear interest.

The novelty of this work is supported by the following contributions:

- A new vision-based position localization framework robust to appearance changes, which enhances existing stereo-based VSLAM systems operating in city-scale environments
- A regularization-based position fusion between local estimates from VSLAM and absolute constraints from the appearance map through topological relations based on visual similarity (described in Section III-E.3)
- As a side effect, our system can geo-tag each keyframe, which allows its further incorporation into the appearance map
- To the best of our knowledge, holistic appearance constraints are used to correct metric estimations for the first time

An running example of AGA-SLAM is shown in `https://youtu.be/mEDW_dB-EK4`.

II. Related Work

Adding absolute constraints to improve the performance and robustness of Visual Odometry (VO) and VSLAM techniques has been explored from different perspectives. These can be classified in three categories: relying on additional sensors, taking a pure visual approach or a combination of both.

In the first category lays the work of Rehder *et. al.*, who use sparse GPS measurements in [12] to improve the robustness of VO in low-textured environments. The relative observations (odometry) are merged with global constraints (GPS) into a graph for optimization. In the context of car navigation, the authors of [13] exploit the chain structure of the graph of constraints to adjust the computation time depending on the available resources. Recently, Qin *et. al.* [11] developed a general graph-based framework aiming to fuse odometry estimates with global constraints. The framework is general in the sense that, apart from GPS, other sensors providing partial global localization measurements (barometers, magnetometers, etc) can be easily integrated.

Pure vision-based approaches typically match local features against 3D models to estimate geometric constraints in the form of relative poses (known as visual localization). This is the case of [14], where the trajectory of a camera is estimated within a known scene, interleaving feature tracking and spatio-temporal coherent matching in Structure-from-Motion models. Due to poor scalability of 3D model-based visual localization, several distributed solutions have been proposed [15], [16], where lightweight visual trackers are executed on client devices and the more expensive computations are left to a dedicated server. The server, in general, provides global constraints for the client-side bundle adjustment. Rather than using pre-built 3D maps, some works use constraints obtained from feature-based visual localization on 2D geo-tagged image databases. The authors of [17] address global aerial localization in urban environments fusing local estimates from feature tracking and global constraints from air-ground matches with database images, refined with 3D cadastral models. Agarwal *et. al.* [18] propose a two-step approach to visual localization in urban environments. The first performs feature tracking for a short image sequence, and the second estimates the relative pose between the reconstruction and the Google Street View panoramas.

The performance of visual techniques can worsen due to perceptual problems or large models, therefore some works combine them with external sensors to improve it. Surber *et. al.* [19] use a Kalman filter framework to explicitly estimate the transformation between the local and the global frame, i.e. the so-called baseframe transformation. The framework improves visual-inertial odometry using GPS priors to globally localize on pre-built 3D maps. Instead of relying on filtering, the authors of [20] combine inertial measurements with visual localization in a graph-based formulation to estimate the baseframe transformation. The inertial measurements are exploited to make 2 DoF of the rotational part of the baseframe transformation observable, enhancing this way the optimization convergence. Lynen *et. al.* [2] describe a global localization system that works on large-scale, compressed 3D models, providing absolute constraints to trajectories estimated through visual-inertial SLAM. The system is decoupled in a server-client architecture and efficiently localizes in city-scale scenarios, exploiting the environment appearance and geometry. In [21], the authors substitute feature-based visual localization with a Deep Neural Network, which provides accurate global localization from VO and additional measurements (form GPS and IMU). The resulting global trajectory is further refined in a pose graph optimization framework in a sliding-window to provide smoother results.

In this work, we rely on appearance maps to provide global position constraints from a pure visual approach. Specifically, the VSLAM trajectory is globally constrained by finding matches between keyframes and mapped places through appearance-based VPR. To the best of our knowledge, we use for the first time holistic appearance-based topological constraints to perform metric corrections. We choose holistic image descriptors as the core of the appearance-based VPR as they are robust to strong perceptual changes (illumination, weather, view-point) [4], [9]. We focus on car-navigated urban environments, which allows simplifications of appearance-based VPR tasks and to use high quality, publicly available geo-tagged image databases such as Google Street View for the map generation.

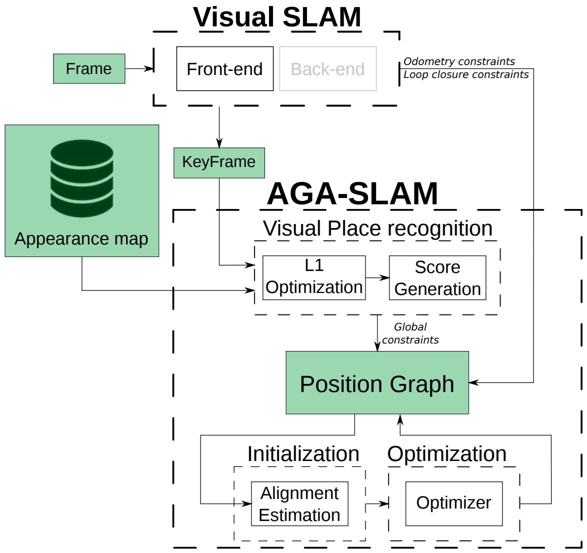

Fig. 2: Block diagram representation of AGA-SLAM. Note that the VSLAM back-end module is substituted by the proposed framework.

III. System Overview

In this Section, we describe the main building blocks of AGA-SLAM, depicted in Figure 2. Our method attaches to an existing keyframe-based VSLAM system and consists of three main modules: visual place recognition, alignment estimation and position information fusion.

A. Assumptions

The proposed framework is designed to operate attached only to metric VSLAM systems (stereo or visual-inertial, namely). Besides, we assume car-mounted cameras facing forwards or backwards and travelling through urban environments. Therefore, we assume that that any street can be observed in two different directions (even single-way ones).

B. Local Position Estimates

The local position estimation module is not a contribution of this paper, and we rely on existing VSLAM algorithms. Nevertheless, minor modifications might be needed in order to provide the information as required by our framework.

Generally speaking, the local position ${}^{\boldsymbol{L}}\boldsymbol{t_i} \in \mathbb{R}^3$ and covisibility factors[1] $c_{ij} \in \mathbb{R}^+$ with $j \in \{i-n_c, \dots, i-1\}$ are required, for each keyframe K_i. The parameter n_c defines the size of the covisibility window and, therefore, the density of the final graph (Section III-E).

In the event of loop closure between keyframes K_i and K_l, the relative translation ${}^{\boldsymbol{i}}\boldsymbol{t_l} \in \mathbb{R}^3$ (as estimated by SLAM) and the covisibility factor c_{il} between them are also required.

C. Visual Place Recognition

The visual place recognition module compares the appearance of a new keyframe K_i with a sparse appearance map through holistic image descriptors in order to relate the local VSLAM frame to the global map frame.

1) Appearance map: The sparse appearance map M consists of n_M descriptor-pose pairs

$$M = \Big\{ \big(\boldsymbol{d_j^M}, {}^M T_j\big),\ j = 1, \dots, n_M \Big\}, \tag{1}$$

where $\boldsymbol{d_j^M} \in \mathbb{R}^{n_d}$ refers to the appearance-based image descriptors and ${}^M T_j \in \mathbb{R}^3 \times SO(2)$ to the absolute pose[2] (in global map coordinates) of the j-th entry of M, respectively. In terms of VPR, we stack the map descriptors to construct the image descriptor database $\boldsymbol{D_M} \in \mathbb{R}^{n_M \times n_d}$.

2) Appearance-based VPR: The execution time of the place recognition module is critical for real-time performance, specially when working with city-scale maps. For this reason, we assume a sparse map, so that the appearance of an image query can be explained by few dictionary elements.

We formulate the VPR of the current keyframe descriptor $\boldsymbol{d_{K_i}} \in \mathbb{R}^{n_d}$ and the sparse image descriptor database $\boldsymbol{D_M}$ as a noise-aware ℓ_1-minimization problem. The sparse optimization problem regards noise bases within the reconstruction error, expressed as

$$\hat{\boldsymbol{x}}^* = \min_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{x}}\|_1 \quad \text{subject to } \boldsymbol{d_{K_i}} = \hat{\boldsymbol{D}}_{\boldsymbol{M}} \hat{\boldsymbol{x}}, \tag{2}$$

where $\hat{\boldsymbol{D}}_{\boldsymbol{M}} = \begin{bmatrix} I_{n_d} & D_M \end{bmatrix}$ is the noise-aware image descriptor database and $\hat{\boldsymbol{x}} \in \mathbb{R}^{n_M + n_d}$ is the noise-aware sparse solution. Subsequently, the relaxed, unconstrained minimization problem associated to (2) is solved by means of the homotopy algorithm associated to the Basis Pursuit Denoising (BPDN) problem. For further information about its computation, we refer the interested reader to [22].

We compute an approximated solution with the homotopy solver in few iterations and later we obtain the most contributing basis of the dictionary as $j^* = \operatorname{argmax} \frac{\hat{\boldsymbol{x}}^*}{\|\hat{\boldsymbol{x}}^*\|}$.

3) Similarity score: Finally, the module provides an appearance-based similarity score $s_{ij^*} \in \mathbb{R}^+$ between the current keyframe descriptor $\boldsymbol{d_{K_i}}$ and the matched map descriptor $\boldsymbol{d_{j^*}^M}$ for the regularization of the graph-based optimization (Section III-E). Similarly to [23], we use a normalized similarity score, although based on the cosine similarity S_C, as

$$s_{ij^*} = \mu_{n_s} \cdot \frac{S_C(\boldsymbol{d_{K_i}}, \boldsymbol{d_{j^*}^M})}{S_C(\boldsymbol{d_{K_i}}, \boldsymbol{d_{K_{i-1}}})}, \tag{3}$$

where μ_{n_s} is a score obtained from the previous n_s keyframes that weights s_{ij^*} according to the sequential appearance evolution.

D. Initialization

The initialization module aims to provide an initial estimate of the rigid body baseframe transformation ${}^M\tilde{T}_L \in$ SE(3) that aligns the local frame to the frame defined by the appearance map. After initialization, the baseframe transformation is subsequently refined (Section III-E).

[1] In this paper, we use the term covisibility as defined in [6].

[2] We assume that the movement of a car in urban environments is modelled with a 3D position plus heading direction.

1) Prior conditions: In order to avoid unsuccessful or degenerated solutions, we impose some initialization conditions. First, the number of matched keyframes has to exceed a threshold τ_k (50 in our experiments). Secondly, keeping the proportion between the singular values of the trajectory in the range $[0.1, 10]$, we ensure enough curvature for a consistent initialization. This way, we avoid initializing with purely straight trajectories as they present ambiguous rotations along the direction of motion.

2) Alignment estimation: The alignment is performed through Random Sample Consensus (RANSAC) [24] on all candidates. The aim of the alignment is to find the 6DoF rigid body transformation between the local trajectory and the appearance map positions that minimizes the distance error metric:

$$e_{RANSAC} = \sum_{(i,j)\in C_{PR}} \left\| {}^{\boldsymbol{M}}\boldsymbol{t_j} - {}^{M}\tilde{T}_L\, {}^{\boldsymbol{L}}\boldsymbol{t_i} \right\|, \tag{4}$$

where C_{PR} is the set of indices between the i-th keyframe and the matched j-th entry of the appearance map M. In each RANSAC iteration, the transformation ${}^{M}\tilde{T}_L$ is estimated using Umeyama's closed-form solution [25] with 3 pairs from C_{PR}, sampled according to their similarity scores. Despite metric VSLAM do not suffer from strong scale drift, we empirically found that using the 7DoF closed-form solution and then filtering out candidate solutions whose scale factor is not close enough to the identity ($[0.8, 1.25]$ in our experiments) produces more consistent initializations.

E. Position Information Fusion

The information fusion between local estimates from the VSLAM system and additional, global constraints from the appearance map is carried out by optimizing a position graph.

1) Graph Structure: In our formulation, we consider three different types of observations (see Figure 3):

- *Odometry*: providing sequential relative translations between keyframes (from VSLAM). For each new keyframe, we consider the n_c previous keyframes with the highest covisibility factors. The parameter n_c controls the density of the graph, and setting it to 1 yields the essential graph [6].
- *Loop closure*: providing relative translations between distant keyframes in time (from VSLAM). These constraints are sparse, less frequent than the odometry ones.
- *Place recognition*: providing global position constraints for keyframes (from appearance map). These constraints are also sparse, but more frequent than loop closure and one-to-many.

2) State Vector and Error Function: The state vector, defined by AGA-SLAM, comprises the 3D position of each keyframe with a global rotation

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x_1} & \cdots & \boldsymbol{x_i} & \cdots & \boldsymbol{q} \end{bmatrix}^\top, \tag{5}$$

where $\boldsymbol{x_i} \in \mathbb{R}^3$ denotes the position vector for the keyframe K_i and $\boldsymbol{q}$ represents the rotation of the baseframe transformation, parametrized as a unit quaternion.

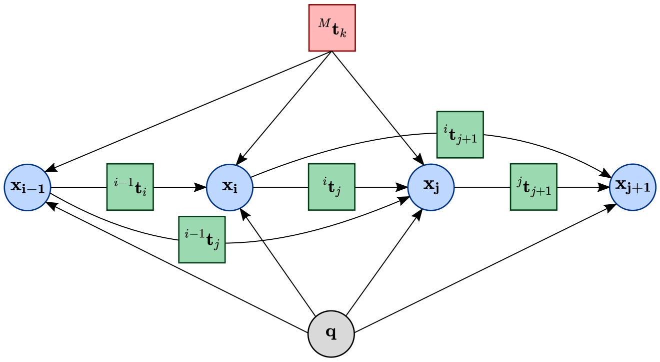

Fig. 3: Representation of the graph structure to be optimized. The reference positions from the appearance map are represented with ${}^{\boldsymbol{M}}\boldsymbol{t_k}$, and the local VSLAM estimates (odometry and loop closure) are represented by ${}^{\boldsymbol{i}}\boldsymbol{t_j}$. The state vector is represented as variable nodes in the graph.

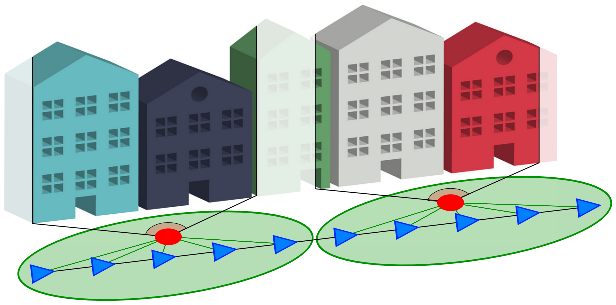

Fig. 4: The sparsity of the appearance map and the geometric covisibility between keyframes foster that multiple covisible keyframes (blue triangles within green ellipses) will match to the same place (red).

The error function is defined to be:

$$F(\boldsymbol{x}) = \sum_{(i,j)\in C_{PR}} f_{ij}(\boldsymbol{x_j}, \boldsymbol{q}) + \sum_{(i,j)\in C_{Odo}} \lambda_1\, g_{ij}(\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{q}) + \sum_{(i,j)\in C_{LC}} \lambda_2\, g_{ij}(\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{q}), \tag{6}$$

where

$$\boldsymbol{f_{ij}}(\boldsymbol{x_j}, \boldsymbol{q}) = s_{ij} \left\| {}^{M}\tilde{T}_L^{-1}\, {}^{\boldsymbol{M}}\boldsymbol{t_i} - \boldsymbol{q} \otimes \boldsymbol{x_j} \right\|^2 \tag{7}$$

is the distance between the i-th map position and the j-th keyframe, weighted by the similarity score s_{ij}, and

$$\boldsymbol{g_{ij}}(\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{q}) = c_{ij} \left\| {}^{\boldsymbol{i}}\boldsymbol{t_j} - \boldsymbol{q} \otimes (\boldsymbol{x_i} - \boldsymbol{x_j}) \right\|^2 \tag{8}$$

is the distance between the observed and the current translation vector form the i-th to the j-th keyframe, weighted by the covisibility factor c_{ij}. Note that through the similarity score s_{ij}, (7) can be seen as a regularization term of the global function (6).

C_{Odo} and C_{LC} represent the odometry and loop closure connections between keyframes, respectively. The $\otimes$ operator refers to the unit quaternion rotation of 3D points (or vectors). The two regularization terms λ_1 and λ_2 in (6) control the relative rigidity of the odometry chain and the relative importance of loop closure, respectively. The relation of these parameters with appearance is discussed in Section III-E.3.

Finally, the optimal solution $\boldsymbol{x}^*$ is given by the state vector that minimizes the error function in (6):

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} F(\boldsymbol{x}). \tag{9}$$

3) Appearance regularization: The fusion of metric information (VSLAM keyframes and map positions) topologically related (through VPR) is a challenging issue, since the appearance space is not directly relatable with metric observations. In order to perform this fusion, we optimize the graph regarding the relative rigidity of the nodes (determined by λ_1 and λ_2) and the regularization term (7), determined by the similarity score s_{ij}. The score is the result of a VPR-based optimization on the assumption of a sparse appearance map (Section III-C.2), which spawns a sparse topology having two main implications (see Figure 4):

- First, the appearance overlap between elements of the map is limited. One place is sufficiently described by one element of the database, and this element will be sufficiently distinct from the remain ones.
- Secondly, highly covisible keyframes from VSLAM will match to the same place, so the map elements will have one-to-many connections with keyframes.

IV. Experimental Validation

We extensively evaluated the proposed framework on various urban stereo datasets and compared it with ORB_SLAM2 [6], an open-source state-of-the-art VSLAM solution. Altogether, we evaluated our approach in more than 70 sequences from 4 different datasets, in terms of local accuracy, time efficiency and geo-tagging precision.

We chose ORB_SLAM2 as the underlying VSLAM system for a fair comparison and used 4096-dimensional NetVLAD descriptors [9] (each one requiring ~16 KB) to capture the holistic appearance for the VPR. We carried out the experimental evaluation with an Intel Core i7-6700K desktop computer with 16-GB RAM and a Titan X Pascal GPU.

As stated earlier, GPS receivers suffer in urban environments and thus can lead to inaccuracies when using it as a ground-truth. We tried our best to identify and remove inaccurate GPS measurements for the datasets considered in the experimental evaluation.

A. Datasets

The appearance map was built taking images from Google Street View[3], providing finely geo-tagged images captured under homogeneous appearance conditions. We extracted NetVLAD descriptors from images oriented along the two main street directions [26], covering the city surroundings of the evaluation sequences for each dataset[4] (see Table I).

TABLE I: Employed Google Street View databases

City	Area	Images	Appearance map required space
Oxford	11.2 km^2	9820	76.71 MB
Seognam	7.1 km^2	11082	86.57 MB
Malaga	11.1 km^2	37828	295.53 MB

TABLE II: Evaluation of translational RMSE (m)

Dataset		ORB_SLAM2	AGA-SLAM
Oxford RobotCar	No LC	8.77	**7.58**
Complex Urban	LC	16.53	**14.31**
	No LC	17.05	**15.39**
Malaga Urban	LC	**8.83**	10.96
	No LC	11.01	**10.88**
New Malaga Stereo	LC	**9.65**	12.79
	No LC	39.74	**15.24**

Oxford RobotCar Dataset [27]: collects stereo at 16 Hz sequences grabbed on a car on Oxford from May 2014 to December 2015 without loop closures, providing strong appearance seasonal challenges.

Complex Urban Dataset [28]: provides stereo images of the downtown of Seongnam at 10 Hz, consisting in large avenues and tall buildings. The resemblance of the whole downtown conforms an environment prone to perceptual aliasing.

Malaga Urban Dataset [29]: presents over 35 km of stereo images collected at 20 Hz on December 2009 over different areas of the city of Malaga. The dataset provides loop closures and a temporal shift for almost a decade with respect to the Google Street View database (grabbed approximately on 2018).

New Malaga stereo Dataset: data collected on our own at 20 Hz in the same areas than the Malaga Urban Dataset, gathered at January 2019 under different weather conditions.

B. Local position estimation

In order to test the accuracy of each method on the evaluation sequences, we measured the average absolute translation RMSE [6] of both approaches over 5 executions. Table II depicts the results, distinguishing between sequences containing loop closures and not.

The results show that our system improves the performance of ORB_SLAM2 in sequences where no loop closures are present. In such cases, the performance of ORB_SLAM2 depends solely on its VO module, while AGA-SLAM manages to find matches with the appearance map, thus limiting the drift accumulation. Also, our approach is not able to initialize when VO drifts excessively, leading to inconsistent and intractable trajectories.

In sequences containing loop closures, the performance of AGA-SLAM drops compared to ORB_SLAM2 in some sequences. In those cases, the holistic appearance regularization shows a deterioration of the estimations since the

[3] Provided and copyrighted by Google: www.google.com/maps

[4] Experimental validation was not possible in KITTI Vision Benchmark due to the lack of Google Street View information in Karlsruhe.

TABLE III: Timing results of each module in miliseconds per frame (median time)

			ORB_SLAM2		AGA-SLAM			
Dataset		N	GBA	Total	Descriptor ext.	ℓ_1 opt.	Pos. graph opt.	Total
Oxford RobotCar	No LC	28	-	256.19	5.12	9.16	3.55	255.87
Complex Urban	LC	6	52.97	181.31	15.11	28.63	294.66	475.70
	No LC	7	-	171.59	14.90	27.46	16.01	195.21
Malaga Urban	LC	8	45.23	218.31	8.81	41.97	28.18	272.37
	No LC	13	-	204.84	10.78	49.24	7.12	256.39
New Malaga Stereo	LC	3	57.88	259.93	12.21	57.32	25.18	325.44
	No LC	9	-	241.03	18.73	92.84	33.09	355.24

topological relations provided by VPR are unable to replace the accuracy of Global Bundle Adjustment (GBA).

The statistics also represent the influence of the environment appearance on the performance of out approach, worsening in scenarios with uniform appearance. The appearance ambiguity hinders the place characterization and the initialization, making the graph optimization less accurate. This case is particularly remarkable in the Complex Urban dataset, where the propensity to perceptual aliasing impedes precise corrections. On the other hand, the proposed framework outperforms the VSLAM method in the Oxford RobotCar dataset, demonstrating its robustness under seasonal and weather changes. A feasible explanation is that historical and characteristic buildings of central Oxford contribute to the place characterization and to the accuracy in general.

We compared the time requirements of both approaches, measuring the average processing time per frame over 5 executions for each sequence and summarized them in Table III. The results show a general increase of the time spent by AGA-SLAM with respect to ORB_SLAM2 since every module demands extra time, although similar performance is achieved in a CPU-GPU framework without requiring multi-threading. In the version of ORB_SLAM2 attached to our framework, GBA is the only process disabled, for which we measured its execution time in the original implementation.

As the results show, the execution time of the VPR module (descriptor extraction + ℓ_1 optimization) is constant for each dataset. The descriptor extraction is carried out on a GPU, with constant times for similar image sizes, and the ℓ_1 optimization depends mainly on the size of the database of descriptors.

The position graph is the most time-consuming module, running only in sequences with appearance constraints (VPR, loop closures or both). We observed that the complexity of the graph optimization grows in two different situations. First, in those sequences where VO drifts excessively, the position graph needs higher deformations, which requires more time to be completed. Secondly, loop closures detected after a long time forces our approach to optimize a large portion of the graph at once. This last case is particularly noticeable in Complex Urban dataset.

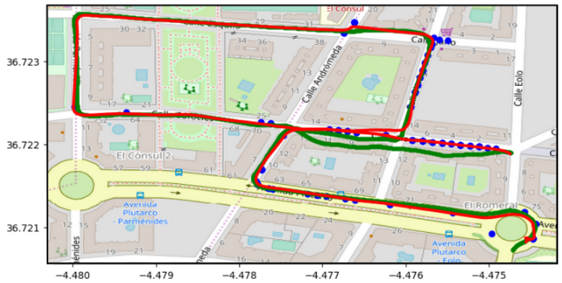

Fig. 5: A geo-tagged, corrected trajectory (red) compared with the GPS ground-truth trajectory (green). AGA-SLAM shows higher drift where there are not enough matches with the appearance map (blue dots).

C. Geo-tagging estimation

Once the VSLAM trajectory is matched and aligned with an appearance map constructed from geo-tagged images, our system can produce geo-tags for each keyframe by expressing the positions $\mathbb{R}^3$ in GPS coordinates.

The amount of correctly recognized places and low VO drift are the main conditions for a successful trajectory alignment with the appearance map. For instance, trajectories with few matches do not surpass the initial conditions imposed in the alignment estimator, as happens in environments presenting perceptual challenges. The achieved geo-tagging precision shows that the assumptions to fuse metric and topological observations are correct, but there is still room for improvements. Presumably, the appearance map sparsity may not be sufficient for the current holistic descriptor. The low accuracy achieved in the Complex Urban dataset may be due to the high perceptual aliasing present in the environment, adding uncertainty to the recognition of places from the appearance map.

An example of georeference is depicted in Figure 5 over a map[5], where the final trajectory appears with several matched places. As can be seen in the image, the ground-truth GPS trajectory is inconsistent due to the *street canyon* effect in a portion of the sequence. Despite some outliers, the precise absolute positions of the appearance map lead AGA-SLAM

[5] Extracted from https://www.openstreetmap.org

TABLE IV: Geo-tagging accuracy

Dataset	Mean RMSE (m)	Aligned sequences (%)
Oxford RobotCar	11.18	61.90
Complex Urban	20.29	55.00
Malaga Urban	16.08	74.36
New Malaga Stereo	17.28	100

to produce accurate and visually consistent estimations.

V. Conclusions

This work presents AGA-SLAM, a novel, vision-based framework that extends VO from any stereo-based VSLAM system by including geometric constraints from appearance maps. Hereby, we contribute with a pure visual system robust to appearance changes that provides accurate position estimates and heads towards persistent long-term maps for VSLAM. These claims are supported by an extensive evaluation on 4 public datasets, each one targeting different perceptual challenges. We showed improvements over the state-of-the-art ORB_SLAM2 system in urban scenarios, achieving high performance with city-scale appearance maps built from a public geo-tagged image database (Google Street View). The proposed framework is able to integrate metric estimations topologically connected, under a sparse map assumption.

In future work, the proposed framework might be extended including orientation constraints for full pose corrections, as the works based on 3D models [2]. Our work might also be improved to face a wider set of configurations, such as operating with monocular VSLAM. Finally, the geo-tagging capabilities can be employed to update appearance maps.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec 2016.

[2] S. Lynen, B. Zeisl, D. Aiger, M. Bosse, J. Hesch, M. Pollefeys, R. Siegwart, and T. Sattler, "Large-scale, real-time visual-inertial localization revisited," *arXiv preprint arXiv:1907.00338*, 2019.

[3] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla, "Are large-scale 3d models really necessary for accurate visual localization?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1637–1646.

[4] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-invariant place recognition by discriminatively training a convolutional neural network," *Pattern Recognition Letters*, 2017.

[5] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[6] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[7] R. Gomez-Ojeda, F. Moreno, D. Zuñiga-Noël, D. Scaramuzza, and J. Gonzalez-Jimenez, "PL-SLAM: A stereo SLAM system through the combination of points and line segments," *IEEE Transactions on Robotics*, vol. 35, no. 3, pp. 734–746, June 2019.

[8] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[9] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[10] Y. Cui and S. S. Ge, "Autonomous vehicle positioning with gps in urban canyon environments," *IEEE transactions on robotics and automation*, vol. 19, no. 1, pp. 15–25, 2003.

[11] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," *arXiv preprint arXiv:1901.03642*, 2019.

[12] J. Rehder, K. Gupta, S. Nuske, and S. Singh, "Global pose estimation with limited gps and long range visual odometry," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 627–633.

[13] C. Merfels and C. Stachniss, "Sensor fusion for self-localisation of automated vehicles," *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 85, no. 2, pp. 113–126, 2017.

[14] H. Lim, S. N. Sinha, M. F. Cohen, M. Uyttendaele, and H. J. Kim, "Real-time monocular image-based 6-dof localization," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 476–492, 2015.

[15] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular slam on a mobile phone," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 4, pp. 531–539, 2014.

[16] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *European conference on computer vision*. Springer, 2014, pp. 268–283.

[17] A. L. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza, "Air-ground matching: Appearance-based gps-denied urban localization of micro aerial vehicles," *Journal of Field Robotics*, vol. 32, no. 7, pp. 1015–1039, 2015.

[18] P. Agarwal, W. Burgard, and L. Spinello, "Metric localization using google street view," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 3111–3118.

[19] J. Surber, L. Teixeira, and M. Chli, "Robust visual-inertial localization with weak gps priors for repetitive uav flights," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 6300–6306.

[20] H. Oleynikova, M. Burri, S. Lynen, and R. Siegwart, "Real-time visual-inertial localization for aerial and ground robots," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2015, pp. 3079–3085.

[21] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2616–2625.

[22] Y. Latif, G. Huang, J. Leonard, and J. Neira, "Sparse optimization for robust and efficient loop closing," *Robotics and Autonomous Systems*, vol. 93, pp. 13–26, 2017.

[23] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[24] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[25] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[26] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, vol. 43, no. 6, pp. 32–38, 2010.

[27] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[28] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 642–657, 2019.

[29] J.-L. Blanco-Claraco, F.-Á. Moreno-Dueñas, and J. González-Jiménez, "The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario," *The International Journal of Robotics Research*, no. 2, pp. 207–214.