Appearance-based Localization across Seasons in a Metric Map

Chris Beall, Frank Dellaert

***Abstract*— In this paper we address the problem of appearance-based long-term outdoor localization across seasons. This is a difficult task due to the changing appearance of visual landmarks across seasons and time of day. Our approach operates based on the premise that combining visual landmarks observed at different times of the year into a single metric map will yield better localization results than a map created from a single sequence alone. We integrate stereo imagery collected at two different times of the year into a unified 3D map, and use this as the basis for localization. A landmark visibility prediction framework is utilized to efficiently retrieve a small subset of landmarks and their feature descriptors from a database of millions of landmarks. The proposed approach is experimentally validated on a challenging sequence collected a year earlier.**

I. Introduction

Vision-based localization systems have received much attention in the past few years. Localization using vision alone is an attractive prospect considering its very low cost compared to other sensor modalities. GPS is useful in many applications, but it is well known that GPS performance is degraded in urban settings due to buildings obstructing the sky. We are particularly interested in the scenario of localizing a moving vehicle, where a coarse localization estimate is available as a prior, either from GPS or from a localization estimate in the immediate past.

While quite a number of visual localization systems have been demonstrated, few have been shown to work robustly in the face of changing scene appearance caused by differences in lighting, seasonal variation, foliage changes, weather, etc. Representing each place as a different experience in a topologically connected map appears to be a particularly promising approach [1], but this sort of technique makes exact localization difficult as the query images are localized in several distinct visual odometry tracks.

In this paper we show that localization across long periods of time (and seasons) within a unified metric map is a feasible approach. We take the view that by combining data from several stereo image sequences into a single map it sufficiently spans the space of possible appearances to enable localization for a wide range of scenarios. This approach clearly presents a number of significant challenges. First, the sequences to be combined into the map must be registered very accurately to ensure the resulting map is geometrically consistent. Since the map contains millions of landmarks, the second challenge is how to decide which landmarks to choose when attempting to localize a query frame.

To the best of our knowledge, this is the first work which explicitly joins data from two sequences into a single metric map as shown in Fig. 1, which is then used for localization.

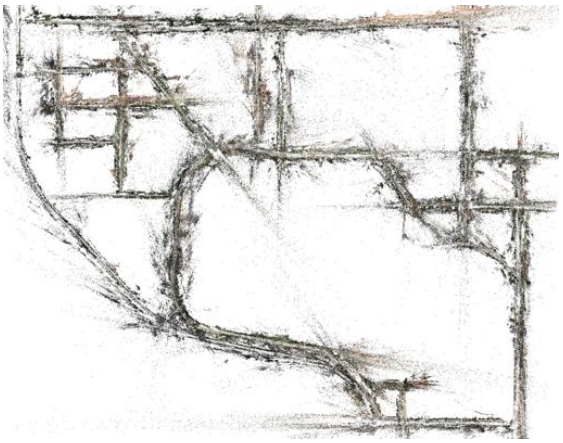


Fig. 1: Color point cloud representing the landmark database of the Georgia Tech campus.

In contrast, previous work which made use of data from different times was topological in nature. The contributions of this paper are:

- Vision-only localization in a large-scale metric map created from data collected during different times.
- Landmark visibility prediction in the context of real-time vehicle localization.

The remainder of this paper is organized as follows. We first discuss related work in section II, followed by a detailed discussion of the 3D map building and localization in section III. Section IV has the results.

II. Related Work

In recent years, many vision-based localization algorithms have been proposed. The work most relevant in terms of its application is that of Churchill et. al. [1]. Visual odometry trajectories, termed experiences, are stored each time the vehicle visits a new place and is unable to relocalize itself within already existing experiences. The system keeps collecting new experiences until they become fully adequate for localization. One disadvantage of this work is that these experiences are only topologically linked, and exact metric pose recovery presents a challenge.

Another interesting approach is that of Lategahn et al., who used a pre-computed 3D map, comprising 3D landmarks and their descriptors, to localize a stereo camera without GPS [2]. Given the previous known pose, all landmarks observed by the nearest camera pose used to build the map are used for descriptor matching. Lategahn et al. took a similar approach in [3], with the notable differences being that a monocular

camera is used during localization, and the resulting pose is refined in a filter together with IMU measurements.

Milford and Wyeth [4] introduced SeqSLAM. Rather than matching local features between images, sequences of images are compared to establish a loop closure. Image similarity is established using sum of absolute differences. Consequently, no lighting/season invariant descriptors are needed. The method works on sequences with drastically different appearance. The method makes assumptions about relatively constant velocity and direction of travel. A related approach is that of Maddern et al. [5], in a system called CAT-SLAM. Sequential appearance based SLAM is enhanced with metric pose filtering to improve the performance.

Valgren et al. [6] also explored an appearance-based approach across scenes with stark appearance changes, using SIFT/SURF descriptor matching, and tuning the parameters for optimal results.

Deciding which map features to match against is a major challenge, and this is especially true in the case of Structure from Motion (SfM), where unordered datasets with mostly unknown location priors are the norm. Li et al. [7] addressed this difficulty by matching 3D points to image features, rather than the more conventional 2D to 3D matching. Points with higher degree are prioritized. This was further improved upon with bi-directional matching in [8]. A similar approach is taken by Sattler et al. [9], where 2D-3D matching is sped up by indexing all image features into a vocabulary tree that was constructed using the 3D model, and the size of each word cluster is used as a proxy for estimated matching speed. Feature matching is prioritized according to cluster sizes. In [10] this approach is further refined with an active correspondence search in both directions.

Another interesting line of attack is reasoning about descriptor occurrence. One such approach is taken in [11], [12] where robust localization is achieved by computing landmark observation likelihoods based on the number of times a landmark was observed across training runs.

It is standard practice to employ a RANSAC [13] framework to achieve robust matching in the presence of outliers. When inlier ratios become very low RANSAC can take many iterations to find a good model. Chum et al. introduced PROSAC [14] , which progressively increases the sample size. This approach assumes that matches can be prioritized, and in the usual case the descriptor distance is suitable. In [15], [16] feature weighting is integrated into the geometric verification procedure (as opposed to post-processing step).

A different approach to solving the data association problem is taken in [17]. The authors proposed a framework for predicting the visibility of landmarks in the scene. Given a new query image with a pose prior, the landmarks which were previously observed by nearby cameras are probabilistically weighted according to a distance metric which is learned in an offline step. The distance metric takes into account camera rotation and translation. This makes it easy to ignore landmarks which were observed by a camera facing in the opposite direction, even though they are very close to the query camera prior. In this paper we are also interested in localizing a query image given a pose prior, and we adopt this same visibility approach for efficiently retrieving likely visible landmarks from our map.

III. Map Building

In this section we describe how we build a map (3D landmark database) which is used for localization. The main steps consist of applying stereo visual odometry to an image sequence, loop closing within and between data sequences, and large scale bundle adjustment. Each of these will be discussed in detail, but first some notation: We define X^s as the set of camera poses $\{x_i^s\}$ for data sequence s. L^s is the set of landmarks $\{l_j^s\}$ observed in sequence s. $\theta^s \hat{=} \{X^s, L^s\}$ is the set of all variables, which together with a camera-landmark visibility table makes up the map M.

A. Stereo Visual Odometry

We run a conventional stereo visual odometry (VO) algorithm to recover the camera trajectory. For each rectified stereo image pair, SIFT features are extracted and matched across the pair. Matches are only retained if they are mutually optimal according to the ratio test [18], and fall within tight threshold of the epipolar line, which is a horizontal scan-line for rectified images. Points with zero disparity are discarded, and 3D points $(X, Y, Z)^T$ are then triangulated. Features are then matched temporally to form a set of putative matches, and a three point algorithm [19] is employed in a RANSAC [13] framework to recover the relative pose.

Features which are successfully tracked for at least two consecutive frames, called feature tracklets, are recorded along with their feature descriptors. As these feature tracklets are geometrically consistent across at least two frames they will be accepted for inclusion in the map. The resulting camera trajectory, together with the accepted landmarks will be optimized later as described in the following sections.

B. Closing the Loop

Loop closures are needed to correct for drift in the VO trajectory, as well as to precisely align multiple passes along the same street. Appearance based loop closure detection as in [20] is a popular approach. However, since the data used to build the map has synchronized GPS, we use this to find loop closure candidates. We are not concerned about real-time performance while constructing the map. In a brute force fashion, we find the nearest neighbor camera poses and attempt feature matching and geometric verification as in Sec. III-A. Loop closure landmark observations are recorded to be incorporated into the map (Sec. III-C).

Loop closure detection is also performed between data sequences to provide constraints to align datasets with respect to each other.

C. Map Optimization

Bundle adjustment, or smoothing and mapping (SAM), has been applied to create highly accurate, city-scale reconstructions from large photo-collections[21], [22]. We apply this technique to optimize several data sequences together into a geometrically consistent map.

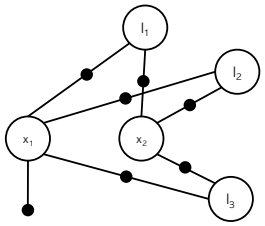


Fig. 2: Factor graph comprising two camera poses, three landmarks, and a GPS prior on camera pose x_1.

The optimization problem at hand is easily represented by a factor graph. A factor graph is a bipartite graph comprising two types of nodes: state variables and factors. Here, the unknown camera poses $X = \{x_i | \, i \in 1...M\}$ and landmarks $L = \{l_j | \, j \in 1...N\}$ make up the set of state variables. The landmark measurements $Z = \{z_k | \, k \in 1...K\}$ as observed by the cameras correspond to factors. An example of a factor graph is shown in Fig. 2.

We minimize the non-linear cost function

$$\sum_{k=1}^{K} \|h_k(x_{i_k}, l_{j_k}) - z_k\|^2_{\Sigma_k} \tag{1}$$

in a least-squares sense, where $h_k(\cdot)$ is the measurement function of landmark l_j from camera x_i, and the notation $\|.\|^2_\Sigma$ represents the squared Mahalanobis distance with covariance Σ. We assume that we have normally distributed Gaussian measurement noise.

For more details on the SAM optimization process, we refer the interested reader to [23].

D. Localization

Given a set of measurements Z_i and the map M, we are interested in efficiently recovering the most likely pose Θ: $P(\Theta|Z_i, M)$. In the case of vehicle localization we also assume that we have a pose prior that comes from the previous pose estimate or GPS. In light of M having many millions of landmarks, it is important to only retrieve landmarks which are likely to be visible in the current stereo frame. We use the visibility prediction framework introduced in [17] to achieve this. The key idea here is that stereo frames which were taken at camera poses X which were nearby the current pose, and also facing in roughly the same direction, are likely to have observed a similar set of landmarks L_v.

The landmark visibility distance metric used in this paper combines Euclidean distance and rotation between the query pose and map poses X. To find the set L_v we compute the distance between the query pose and all poses X, and then collect all of the landmarks observed by the n nearest poses. One important advantage of this approach is that map landmarks observed from a map-building sequence X^s where the vehicle was traveling in the opposite direction along the same road will not be considered visible, which is in accordance with the limits of rotation invariance of the SIFT descriptor.

Date	Frames	VO Fr.	Resolution	Length	Label
Sep 11, 2012	25462	20372	1380×480	10.5km	F
Apr 2, 2013	23090	14053	1384×680	11.38km	K
Aug 1, 2013	21690	15219	1384×680	13.21km	L

TABLE I: Three datasets that were used for the experiments.

Fig. 3: GPS-INS trajectory superimposed on Google Earth imagery. Severe GPS drift due to multi-path issues can be observed to the east of the stadium.

Given L_v, the standard approach is followed to compute a pose estimate: Detect features in the current stereo pair, match and verify with RANSAC.

In practice, some steps can be taken to further speed up the algorithm described above. Computing the visibility distance metric with respect to all poses X can be costly for large M. Instead, we make use of a quad-tree to pre-prune the set of poses, and only compute the visibility for poses that fall within a bounding box of the query pose.

IV. Experimental Results

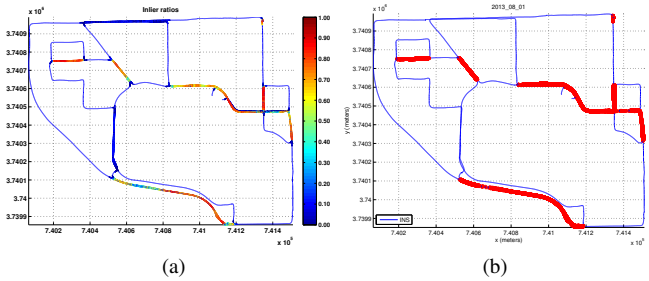
To validate our approach we have built a map using two data sequences collected on our campus. One sequence was collected in April, and the other in August of 2013, called sequences K and L. Sequence F is not included in the map, and is used for localization testing only. A listing of all the data sequences used in this paper is shown in table I. Images were collected using two Point Grey Flea 3 GigE cameras, along with a third color camera for visualization purposes. The cameras were triggered through hardware synchronization at 10Hz.

GPS-INS data was collected using a 3DM-GX3-45 GPS-Aided Inertial Navigation System at up to 100Hz. This data was interpolated and synchronized to camera timestamps. The GPS-INS solution occasionally drifts quite noticeably, particularly when driving next to large buildings which hinder a clear view of the sky in all directions. An example is shown in Fig. 3. Visual Odometry is run on each of the sequences, and feature tracklets, as well as their associated descriptors, are saved for the loop closure step.

A. Closing the Loop

As described in Sec. III-B, loop closure detection is performed within each sequence, as well as between the

Fig. 4: Successful registration and pose recovery on challenging imagery between frames from sequences K (top) and L (bottom). There are notable differences in lighting, foliage, as well as vehicular occlusions. Putative matches are shown in blue, and accepted inlier matches are shown in green.



(a) (b)

Fig. 5: Loop closure results for sequence L. a: Inlier ratios b) Accepted loop closures exceeding the inlier threshold and minimum inlier count are shown in red.

two sequences making up the 3D map. Fig. 4 shows a successful loop closure. The goal is to detect as many loop closures as possible as this promises the most accurate map registration possible. Missed loop closures lead to poor map alignment, while false loop closures present difficulties during optimization. Through empirical experimentation we find that a RANSAC inlier ratio of 0.5, and a minimum inlier count of 10 yield satisfactory results. Fig. 5 shows loop closure results for sequence L.

Fig. 6 shows the loop closure result between sequences K and L. As expected, there are no loop closures where the two trajectories do not overlap, but loop closures are also missed in some places, likely due to vastly different appearance, or due to the RANSAC inlier ratio not meeting the required threshold.

B. Map Optimization

Each sequence is optimized individually before all data are combined into a single map. Camera poses X^s are initialized from GPS, and landmarks L^s are initialized from stereo triangulation. We additionally add weak GPS priors to camera poses so the map remains in true alignment with the

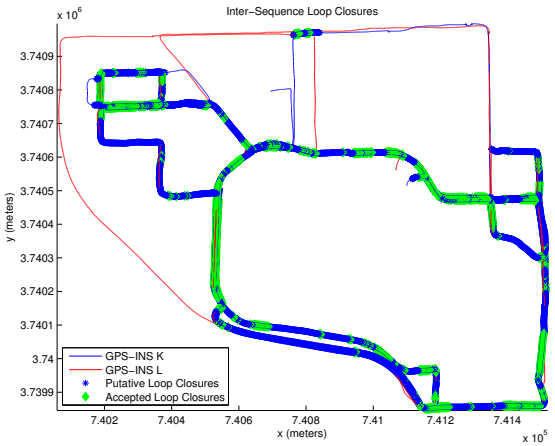
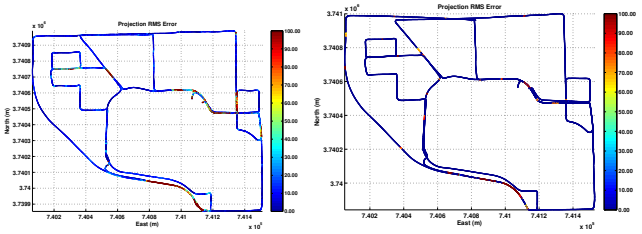


Fig. 6: Loop closures between sequences K & L. Poses where loop closure is possible are shown in blue, and where loop closure was successful is shown in green.



(a) (b)

Fig. 7: Sequence L. a) Per-camera RMS errors before optimization b) RMS errors after optimization

streets. The Huber cost function is used to achieve robustness against possible outliers. RMS projection errors per camera, before and after optimization, are shown in Fig. 7.

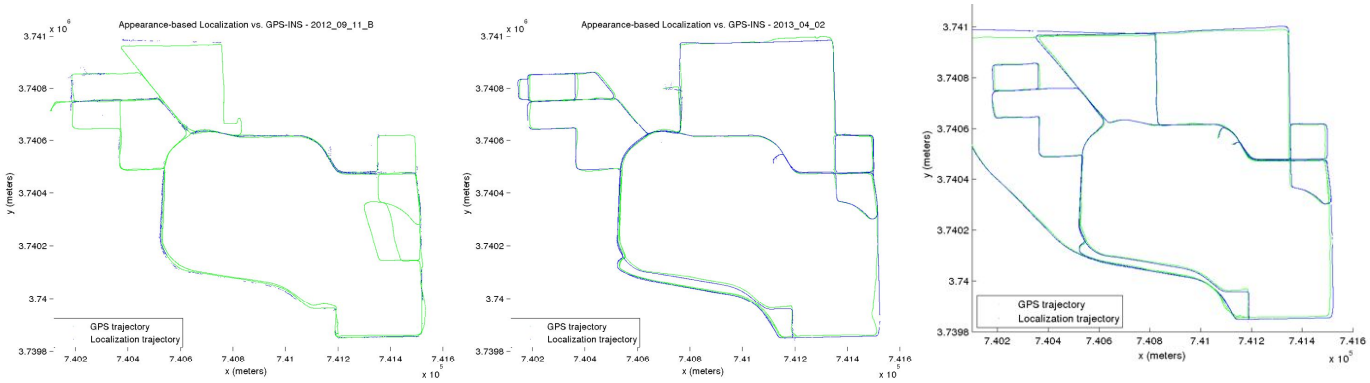
Finally, the two optimized sequences are combined, and landmarks which were observed in both sequences are represented as a single landmark. The final optimized camera trajectories are shown in Fig. 8.

To fully appreciate the structure of the 3D map, Fig. 9 shows a top-down view of all contained landmarks, with landmarks observed in sequences K and L shown in blue and green, respectively. Fig. 1 shows the color point cloud. The complete map, inclusive of feature descriptors has a size of approximately 1.4GB on disk.

C. Localization

We have conducted localization experiments for each of the three sequences, shown in Fig. 10. It is expected that sequences K & L will perform very well, as these contributed to the map. Sequence F, however, is a lot more challenging, since this sequence was taken in the previous year, and scene appearance was drastically different in many places across campus.

Fig. 11 shows a visualization of the smallest visibility distance for each query pose. The smaller the distance, the more likely the camera is to have observed the same



(a) Sequence F (b) Sequence K (c) Sequence L

Fig. 10: Localization results with KL map. Estimated poses are shown blue, GPS-INS priors are shown in green.

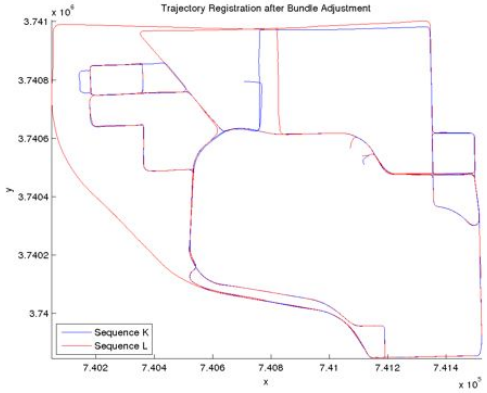


Fig. 8: Optimized camera trajectories after full bundle adjustment of over 12 million factors and over 2.2 million variables.

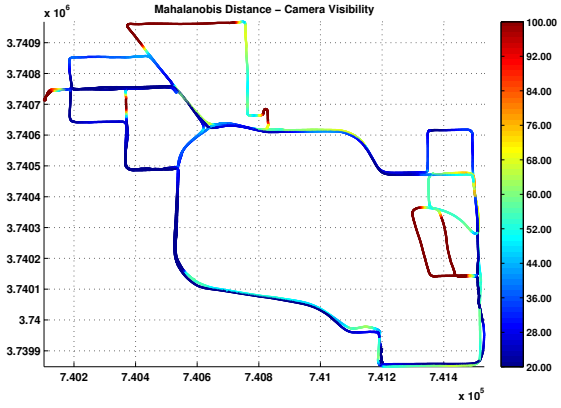


Fig. 11: Camera visibility scores (Mahalanobis distance) per GPS query pose for sequence F with respect to the full database. Lower (blue) is better. Streets which were not covered by the database, or which were traveled in the opposite direction have a large distance (red).

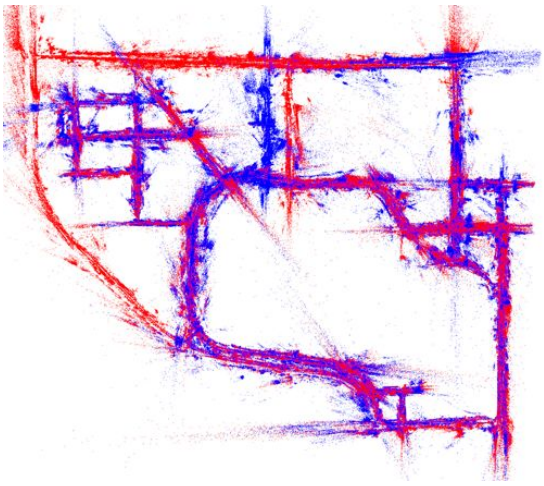
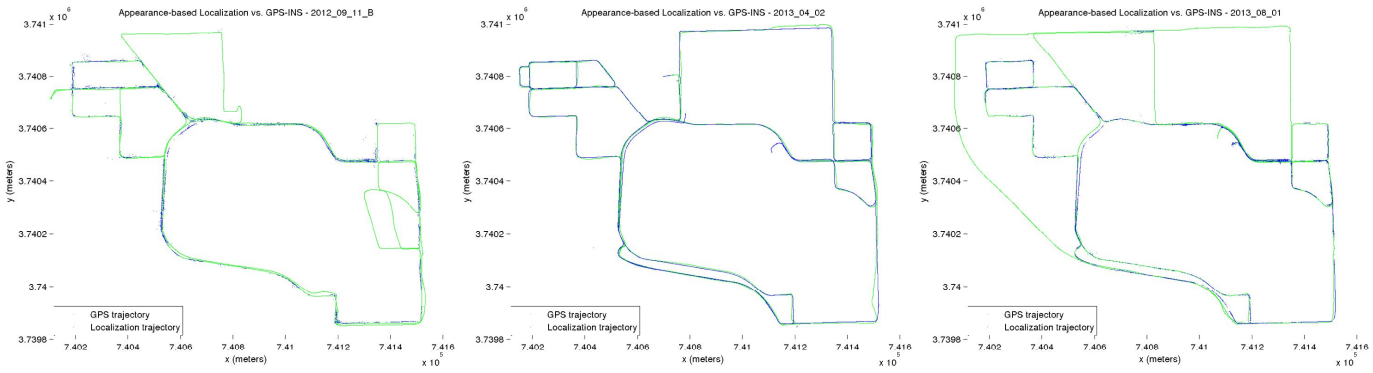


Fig. 9: Point cloud of tracked landmarks. Points shown in blue and red are from sequences K and L, respectively.

landmarks. For example, note that to the east there is a street block which was not covered in the map, and therefore has a very large visibility distance (deep red).

We have conducted the same experiments using only sequence K as the basis for the landmark map, and these results are shown in Fig. 12. As expected, the results for sequence K are virtually unchanged, and sequence L has gaps in localization where its trajectory does not overlap with K. Sequence F is relatively similar to the previous result, with the notable difference that localization was somewhat worse in areas where the two sequences K & L had poor loop closures. In other words, these were areas where there might exist alignment problems in the map. This underscores the need for very good registration when combining data from multiple sequences into a single metric map, and this is to be addressed in future work. Table II shows the localization performance of the three sequences with respect to a map constructed from sequence K alone vs. a map constructed from K+L.

The visual odometry component of our system runs faster



(a) Sequence F (b) Sequence K (c) Sequence L

Fig. 12: Localization results with K map. Estimated poses are shown blue, GPS-INS priors are shown in green.

Map	F	K	L	Total
K	5335	22969	10540	38844
K+L	4417	22819	21245	48481

TABLE II: Number of successfully localized frames of sequences F, K, L against maps created from sequence K alone, and from sequences K and L.

than real-time (10Hz). The performance of the localization module varies greatly, depending on the number of landmarks returned from the map, and depending on the inlier ratio. In the successful case it takes about 5-10ms, depending on the sequence (localizing K or L against the map is faster than F). When localization fails it can take up to hundreds of ms, dependent on RANSAC termination thresholds. However, these results were obtained with unoptimized code, and localization of individual image frames is easily parallelizable.

CONCLUSION

In this paper we presented a robust localization system based on a unified metric landmark map created from two stereo sequences collected at different times of the year. Efficient vision-based localization was performed by relying on a visibility prediction framework to retrieve a subset of landmarks which are used for descriptor matching. Experiments on real data showed the effectiveness of the approach. In future work we plan to incorporate more datasets into the map, and extending the visibility prediction framework to handle seasonal appearance explicitly.

REFERENCES

[1] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1645–1661, 2013.

[2] H. Lategahn and C. Stiller, "City gps using stereo vision," *ICVES*, 2012.

[3] H. Lategahn, M. Schreiber, J. Ziegler, and C. Stiller, "Urban localization with camera and inertial measurement unit," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, 2013, pp. 719–724.

[4] M. Milford and G. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, may 2012, pp. 1643 –1649.

[5] W. Maddern, M. Milford, and G. Wyeth, "Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory," *The International Journal of Robotics Research*, vol. 31, no. 4, pp. 429–451, 2012.

[6] C. Valgren and A. J. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149–156, 2010.

[7] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," pp. 791–804, 2010.

[8] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," pp. 15–29, 2012.

[9] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 667–674.

[10] ——, "Improving image-based localization by active correspondence search," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 752–765.

[11] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *Proc. ICRA*, 2013.

[12] ——, "Dynamic scene models for incremental, long-term, appearance-based localisation," in *Proc. ICRA*, 2013.

[13] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[14] O. Chum and J. Matas, "Matching with PROSAC - progressive sample consensus," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[15] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 500–513.

[16] R. Raguram, J. Tighe, and J.-M. Frahm, "Improved geometric verification for large scale landmark image collections." in *BMVC*, 2012, pp. 1–11.

[17] P. F. Alcantarilla, K. Ni, L. M. Bergasa, and F. Dellaert, "Visibility learning for large-scale urban environment," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*. IEEE, 2011. [Online]. Available: http://frank.dellaert.com/pub/Alcantarilla11icra.pdf

[18] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[20] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *Intl. J. of Robotics Research*, vol. 27, no. 6, pp. 647–665, June 2008.

[21] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," in *SIGGRAPH*, 2006, pp. 835–846.

[22] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *Intl. Conf. on Computer Vision (ICCV)*, 2009.

[23] F. Dellaert, "Square Root SAM: Simultaneous location and mapping via square root information smoothing," in *Robotics: Science and Systems (RSS)*, 2005.