Do you see the Bakery? Leveraging Geo-Referenced Texts for Global Localization in Public Maps

Noha Radwan Gian Diego Tipaldi Luciano Spinello Wolfram Burgard

***Abstract*— Text is one of the richest sources of information in an urban environment. Although textual information is heavily relied on by humans for a majority of the daily tasks, its usage has not been completely exploited in the field of robotics. In this work, we propose a localization approach utilizing textual features in urban environments. Starting at an unknown location, equipped with an RGB-camera and a compass, our approach uses off-the-shelf text extraction methods to identify text labels in the vicinity. We then apply a probabilistic localization approach with specific sensor models to integrate multiple observations. An extensive evaluation with real-world data gathered in different cities reveals an improvement over GPS-based localization when using our method.**

I. Introduction

Localization is one of the fundamental problems in the area of mobile robotics. The accurate knowledge of the robot position enables a variety of tasks including navigation, transportation, as well as search and rescue. Additionally, the exact information about the position of a user gives the opportunity to offer so-called location-based services with plenty of uses in social networking, health, guidance, entertainment and many others. In outdoor settings, GPS is a popular solution to estimate the position of the robot or the user. Although GPS can theoretically reach an accuracy of a few meters, it cannot always be achieved in practice; due to GPS outages, e.g. inside or near buildings.

Recent advances in the field of computer vision led to a surge in the number of vision-based techniques for localization [11, 18, 23]. The availability of large scale, public, and continually updated comprehensive maps, such as Google Maps and OpenStreetMap, spurred research into utilizing them for robot navigation and localization [1, 22]. In the classical approach, localization is performed after a previous visit of the environment during which a map has been built. The advantage of leveraging and processing publicly available maps lies in the ability to localize without an initial mapping step. The majority of currently available methods mostly focus on only one kind of information provided by those maps, namely geo-tagged street-level imagery.

In this paper, we propose an approach that uses a standard RGB camera to localize on publicly available online maps *without* any use of street-level imagery. The idea is to exploit the rich textual meta-data content of maps, such as the annotations of local shops and businesses as high-level information. Our approach moves away from visual-based feature matching to use mid-level representations for estimating the current geo-location of an image. Specifically, we concentrate on extracting text "in the wild" from images that are cross-referenced from the available annotated map. This enables a new localization form that has global-scale breadth, low bandwidth requirements (no images are transferred in the network) and, lifelong capabilities (users and companies continually update their maps).

Fig. 1. Localization using texts from scene images: Exploiting textual information from surrounding shops enables us to correctly estimate the position of the camera (green star). The images shown were captured by rotating in place. Red rectangles highlight the output of the text extraction phase. Our approach generates pose estimates by matching these labels with a map of geo-referenced texts.

Our procedure is split into three main stages (see Figure 1). First, we extract text from the captured scene images. The extracted texts are then used to identify landmarks in the vicinity of the camera. Finally, we employ a particle filter with a dedicated sensor model to obtain accurate location estimates. We present extensive experiments in three cities in Germany, Switzerland, and England, and quantify the accuracy of the proposed method through a comparison with ground-truth and GPS. The results demonstrate that our technique localizes successfully with a 40% improvement over GPS-based localization.

II. Related Work

A large variety of work has gone into utilizing visual information for localization and navigation tasks. Several

All authors are with the University of Freiburg, Institute of Computer Science, 79110, Germany. This work has been partially supported by the European Commission under the grant number FP7-610603-EUROPA2.

approaches aim at solving the Simultaneous Localization and Mapping (SLAM) problem using vision (refer to the work of Fuentes-Pacheco et al. for a comprehensive overview [5]). Cummins and Newman apply a probabilistic approach based on an approximated Bayes network with the aim of large-scale place recognition [4]. They build a topological map of the environment using features extracted from images to form a visual vocabulary. Konolige and Agrawal [7] formalize the SLAM problem as a non-linear least squares optimization problem. The nodes of the graph represent places, which enforces constraints for loop closures. Lothe et al. [9] use bundle adjustment with camera information. They rely on 3D city models and road homography to reduce error accumulation for SLAM in dense urban environments.

Approaches to solve the localization problem can be divided into two groups: topological approaches, and metric approaches. Topological approaches aim at obtaining an estimate of the current position with respect to some known structures in the environment, while metric localization methods typically estimate the position of the robot with respect to geographic coordinates. Brubaker et al. [2] present an approach to topologically localize a robot using a camera and OpenStreetMap data. In their method, they extract a graph from the map information, with nodes representing streets and edges representing intersections. They apply a probabilistic mixture of Gaussian model to estimate the pose and orientation of the robot. Other examples of topological localization include the works of Crandall et al. [3], and Hayes and Efros [6], in which they crawl the image database of Flickr to build their own geo-tagged image database. Both approaches used a clustering approach to extract features from the database images, that are later used to quantify good matches in relation to a query image.

Metric localization approaches can be further split into two subcategories: direct image matching-based techniques, and retrieval-based techniques (refer to Sattler et al. for a comparison [17]). An example of a direct image matching-based technique is the work of Sattler et al. [16], which relies on a preexisting 3D model of the environment as well as a direct matching framework based on image correspondences and a visual vocabulary tree. Similarly, Qu et al. use geo-referenced traffic signs and local bundle adjustment to localize a moving vehicle [14]. They employ a traffic sign detector once the estimated pose is close to a traffic sign in the map, thus providing ground control points to reduce motion drift. Torii et al. present an image retrieval-based technique for localization using interpolation [20]. Initially, they build a database of geo-tagged images and then compute features for each query image. Afterwards, they apply a regression-based approach to search for nearby images using a linear combination of the extracted features for pose estimation. More localization techniques using publicly-available maps are emerging over time. Both, Majdik et al. [10] and Agarwal et al. [1] apply an image-retrieval-based technique for localization using a database of images collected from Google Street View data. Majdik et al. present a solution to the aerial localization problem by generating virtual views from the Street View images and use a histogram-voting scheme to select the best image correspondences to the query image [10]. Agarwal et al., on the other hand, use panorama images from Google Street View as database images [1]. Similar to other image retrieval-based approaches, they obtain the closest matching panorama images through feature correspondence between the query images and the database images. They formulate the problem as a non-linear least squares estimation to compute the rigid body transformation between the Street View panorama and the query image.

The approach proposed in this paper lies in between topological and metric localization approaches, as we estimate our position relative to surrounding textual landmarks. At the same time, we extract features from the images in the form of textual labels. We retrieve the best matching landmarks from a database of geo-tagged textual information. Taking advantage of the text that is abundant in urban environments renders our approach robust to environmental changes, e.g., changes in daylight, scenery changes, etc. It also makes it easy to use with any publicly-available map. Human-readable text has also been exploited in the context of computer vision and robotics. Both, Tsai et al. [21] and Schroth et al. [19], extract text and visual features from query images. They perform feature matching to return the best corresponding images from a database such that both the query image and the retrieved images contain the same textual information. Posner et al. use extracted text from natural scene images to return images that are semantically relevant to a query [13]. They build a generative model to create connections between extracted text and locations in a map. To the best of our knowledge, we are the first to exploit textual information in natural scenes for localization purposes.

III. Text-Based Localization

In this work, we consider the following problem: given that we are standing at a certain position, equipped with an RGB-camera and a compass, can we accurately localize ourselves using surrounding textual information? The answer is yes, given a map of the environment, and at least two text-containing images. Our approach works by extracting textual features from the images and associating them to landmarks in the environment. To obtain a robust estimate of our pose, we adapt Monte Carlo methods accounting for the employed text extraction approach. In the remainder of this paper, we first describe the map representation, followed by the Monte Carlo methods employed for pose estimation. Finally, we outline the text spotting and data association approach used.

A. Map & State Representation

We represent the environment by a set of landmarks, each of which corresponds to a text that could belong to a shop, restaurant, street name, etc. The only assumption that we make is that the text is static, i.e., it is not scrolling over a display. Text signs which are not present in the map, e.g. “Stop”, are not considered a part of our environment model, and hence are not counted as landmarks for pose estimation. We assume that for each landmark l_i in the map, we have the

following set of features: (a) the name, which is the text that appears on the sign, (b) the geo-location (l_{i_x}, l_{i_y}), which is the coordinates of the sign, (c) the orientation (l_{i_o}), which is the orientation angle of the sign (where 0 degrees is north), and (d) size (l_{i_s}), from which we compute the maximum distance of observing the sign. In principle, any publicly available map can be used for the described representation, as the extra features required can be easily inferred from the map structure itself. Landmark orientation can be computed from the street orientation, as text is placed either parallel or orthogonal to the road. The map information provides knowledge about the orientation of the streets with respect to north, which can be directly generalized to all landmarks within that street. A consequence of localizing in an urban environment is that it is unlikely to be able to observe a sign from a shop that is two streets or more away from our location due to occlusions. Accordingly, we estimate the size of the landmark by the width of the nearest street.

For localizing, we rely on a number of observations using an RGB camera and a compass, such that each observation is associated with an image of the observed landmark, the orientation angle β with which the landmark is observed, and the estimated maximum distance d. We do not assume any prior knowledge of our location. Moreover, as we focus on one-shot global localization, we do not share information between the different poses, rather try to estimate a location for each pose separately using the observations at that time step. Sharing information between the different poses for tracking purposes will be addressed in future work. The goal of each time step is to estimate the pose in geo-coordinates.

B. Pose Estimation

At the heart of our system is the pose estimation phase. We assume for the time being that we already extracted text from the scene images, and each observation is associated with a, possibly empty, set of landmarks. The goal of this phase is to obtain a probabilistic estimate of our location. More formally, we wish to estimate the probability $p(x \mid z_{1:n}, m)$ of being at location x, given the observations $z_1, \ldots, z_n$, and the map m. First, we make the frequently made assumption, that the individual measurements are independent given x which in turn leads to

$$p(x \mid z_{1:n}, m) \quad = \quad \eta \prod_{i=1}^{n} p(x \mid z_i, m). \tag{1}$$

To calculate $p(x \mid z, m)$ we integrate over all different landmark associations a, that are obtained from the data association phase described in Section IV:

$$\begin{aligned} p(x \mid z, m) &= \sum_a p(x, a \mid z, m) \\ &= \sum_a p(x \mid a, z, m) \cdot p(a \mid z, m). \end{aligned} \tag{2}$$

Since the belief computed by Equation (2) is multi-modal with the number of modes growing combinatorially with the possible data associations, we approximate it with a weighted sample set. To sample from it we resort to the importance sampling principle and choose as proposal distribution

$$\begin{aligned} \pi(x) &= p(x \mid z_i, m) \\ &= \frac{p(z_i \mid x, m) \cdot p(x)}{p(z_i)}, \end{aligned} \tag{3}$$

where we chose the measurement z_i uniformly at random. According to the importance sampling principle, we compute for each sample its importance weight

$$\begin{aligned} w(x) &= \frac{p(x \mid z_{1:n}, m)}{p(x \mid z_i, m)} \\ &= p(z_{1:n})^{-1} p(z_i) \prod_{l \neq i} p(z_l \mid x, m) \\ &\propto \prod_{l \neq i} p(z_l \mid x, m) \end{aligned} \tag{4}$$

We model the individual likelihood with a mixture distribution over the latent data association variables. In this work, we assume the set of associations returned from the text spotting phase to be equally likely. This results in:

$$\begin{aligned} p(z_i | x, m) &= \sum_{a_i} \frac{1}{|a_i|} p(z_i | a_i, x, m) \\ &= \sum_{a_i} \frac{1}{|a_i|} \mathcal{U}(d_i, 0, \hat{d}_i) \mathcal{N}(\beta_i, \hat{\beta}_i, \sigma^2) \end{aligned} \tag{5}$$

where $|a_i|$ denotes the number of data association from text spotting, β_i and d_i are respectively the angle and distance measurements, and $\hat{\beta}_i$ and $\hat{d}_i$ are the predicted values.

To be robust to outliers and false measurements from the data association phase, we apply a robust method to compute the particle weights, inspired by the trimmed estimator approach [15]. A trimmed estimator excludes extreme values while computing the desired statistics. Extreme values can be either the lowest/highest 5th percentile or the n-th maximum/minimum points. In our work, we discard the lowest percentile of likelihood values to compute the weights.

IV. Text Spotting & Data Association

To recognize texts in natural scene images, we employ the method from Neumann and Matas, which falls into the category of approaches that use region groupings [12]. In their work, the authors train a sequential classifier for character detections to select extremal regions from the component tree of the image. They further use a number of heuristic functions to effectively prune the selected regions. This allows for a fast exhaustive search of the state space of character sequences before grouping the regions into high level text blocks. We adopted this method in our work due to its robustness, and relied on an open-source implementation by the authors. Note that our approach is independent of the particular text extraction method used.

The text extraction method provides a list of the different detected words, each with an associated confidence score. We perform two-phase post-processing on the extracted text.

In the first stage, we filter the extracted words based on both confidence scores and the word structure. More precisely, we discard words with confidence scores lower than 50%, single letter words and words with multiple consecutive character occurrences, e.g., "gaummmm". The goal of the second stage is to fix any substitution errors (e.g., the letter "l" and the number "1"). For this purpose, we use the GNU Aspell[1] spell checker with a custom dictionary that contains only the words that occur in our map. We check each extracted word against the dictionary to replace it by the closest matching word, if needed. If the extracted word matches an existing dictionary word then it does not get replaced. However, in the case where it does not match, it gets replaced with the closest matching word, only if the number of edits is less than half of the length of the word. An example case is the detected word: "volksl". Using the custom dictionary, the word: "oska" has a smaller edit distance than the word: "volksbank", which is the correct spelling in this case. However, since the number of edits is more than half the length of the word, we do not make the correction, and retain the word as it is. This adds flexibility for landmark association, by not committing to a correction when we are uncertain.

After the post-processing stage, we use the extracted text to assign a set of landmarks for each image. The size of the landmark set varies depending on the quality of the extracted text. The closer the extracted text is to landmarks in our world representation, the smaller the size of the landmark set. We measure closeness of text to a landmark using a probability mass function based on the Levenshtein distance [8] between the text and the landmark. The probability of observing a landmark l_j, with extracted text t_i is approximated by a Gaussian distribution on the Levenshtein distance score s_{ij}. This probability is conditioned on the observation angle β_i to ensure text readability, i.e.,

$$p(t_i \mid l_j) = \begin{cases} \mathcal{N}(s_{ij}, 0, 1) & \text{if visible}(l_j, \beta_i) \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

We compute the above probability for all landmarks in the map. To avoid exponential blow-up with big maps, we return the top n landmarks, where the probability $p(t_i \mid l_j)$ is higher than some threshold. Assigning multiple landmarks to a single observation results in having multiple hypotheses for a single pose, which is in turn handled by the pose estimation step (Section III-B). In the event that the extracted text does not occur in the map, or does not match any of the landmarks, then this observation is discarded.

V. EXPERIMENTS

We evaluated our method on three different datasets. For all datasets, we analyzed the performance of our method by evaluating both the whole pipeline, and the pose-estimation using ground-truth text labels. Evaluating using ground-truth text labels serves to have a baseline of the best achievable performance with perfect text recognition. We collected the first dataset in Freiburg using a Google Tango tablet, while

[1]K. Atkinson. GNU Aspell, 2003. http://aspell.net

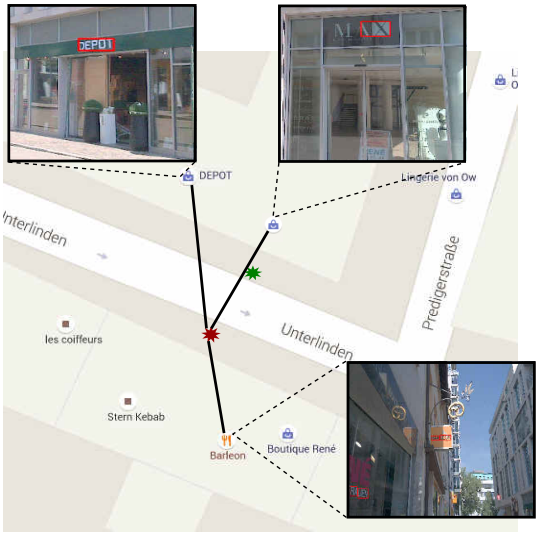

Fig. 2. Example pose from the Freiburg dataset. The green star represents the ground-truth position, the blue star shows the estimated pose from our approach. Lines connect the pose with the observed landmarks. Red rectangles in the images show the output of the text-spotting phase.

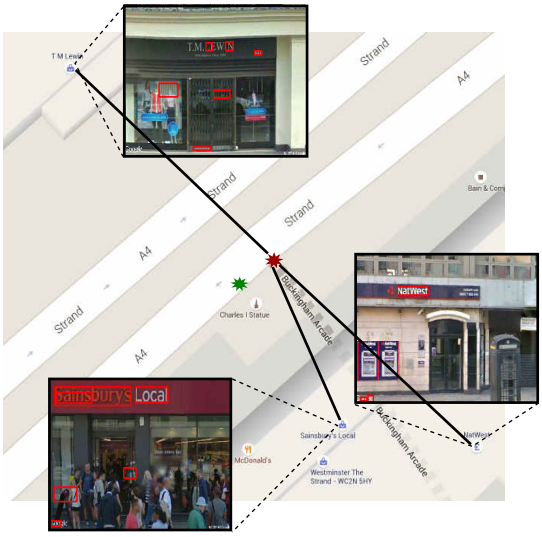

Fig. 3. Example pose from the London dataset. The green star represents the ground-truth position, the blue star shows the estimated pose from our approach. Lines connect the pose with the observed landmarks. Red rectangles in the images show the output of the text-spotting phase.

we obtained the two remaining datasets for London and Zurich using Google Street View.

For each pose, we collect a minimum of two observations. At each time step, the observations were captured by standing in a certain pose, and rotating in place. For both the Freiburg and London dataset, we were able to obtain on average 3 observations per pose. However, in Zurich, we were only able to obtain an average of 2 observations per pose. This is a consequence of the difficulty of obtaining Street View images from Zurich. The restricted availability of Street View images in down-town areas, and the presence of motion blur in some images increased the difficulty of collecting more observations per pose, which in turn affects the quality of the

localization results. The quality of all the results is expected to improve by sharing information between the different poses, e.g. in tracking. However, we intend to investigate this in future work.

In order to quantify the localization results, we report the mean location of the data association mode with the highest average weights. Furthermore, text-spotting localization failures are defined as cases where the text-spotting method fails to extract any text for 50% or more of the captured images; which causes our algorithm to output "not enough data, unknown location".

Figure 2 and Figure 3 show examples of applying our approach in Freiburg and London respectively. In both figures, the ground-truth position is shown in green, and the estimated location is shown in red. For both poses, we have an error of approximately 8 m. Notice in Figure 3, in the *T.M. Lewin* image, despite the bad performance of the text-spotting, we are still able to achieve 100% data association accuracy for that pose. We are able to achieve this result as we select the best N matches for each image; for the *T.M. Lewin* image we select multiple shops as the text is not distinctive enough. We rely on the pose estimation phase to handle the multiple hypotheses; which in this case entails that particles with wrong data associations receive low weights, and in turn are replaced in the resampling step.

A. Freiburg Dataset

During the evaluation phase, we manually added a few annotations to tag some more shops in Google Maps which were not annotated, or changing the position of a label to match changes due to construction sites. We collected a total of 60 poses from different locations in the city of Freiburg. Additionally, our map consists of approximately 180 landmarks. Due to the unavailability of Google Street View in Freiburg, we used the odometry obtained from Google's Tango tablet as ground-truth. Furthermore, we collected GPS coordinates for each pose, in order to compare the localization performance of our method with standard GPS obtained from a mobile device.

Figure 4 displays the cumulative error plot of the presented approach in comparison to GPS. Our method has a mean localization error of 10.7 m (7.0 m with manual text labeling), versus a mean of 27.0 m from GPS obtained poses. In 85% of the cases, our approach performs better than GPS localization alone. Furthermore, over 80% of our localization results are at a distance between 0 and 20 m from the ground-truth. On the other hand, 80% of the GPS localization results are at a distance between 0 and 40 m from the ground-truth position. On this dataset, we have 18.3% text-spotting localization failure, which justifies the performance gap between the whole pipeline, and using ground-truth text labels. Another source of error comes from the data association phase, when the text does not match the text in the map.

B. London Dataset

Google's Street View maps are used for data collection and as a source of ground-truth for the London dataset. We collected approximately 300 poses from different locations in London, and have a map of around 1,000 landmarks from different shops, restaurants and signs. The poses were collected from different districts in the city ranging between urban, rural, and motorway regions.

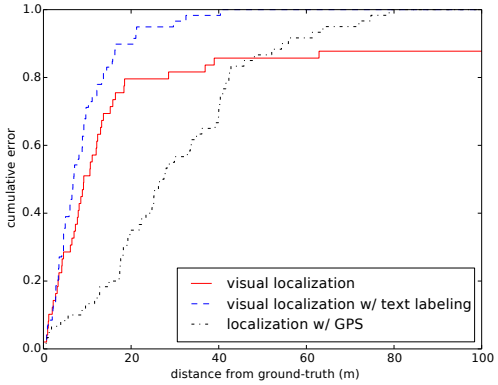

Fig. 4. Freiburg dataset cumulative error plot. The x-axis shows the distance from ground-truth position in meters, and the y-axis shows the percentage of points with distance less than or equal to the x-value. Results show that the visual localization approach has higher percentage of points lying within a low error distance in comparison to localization using GPS.

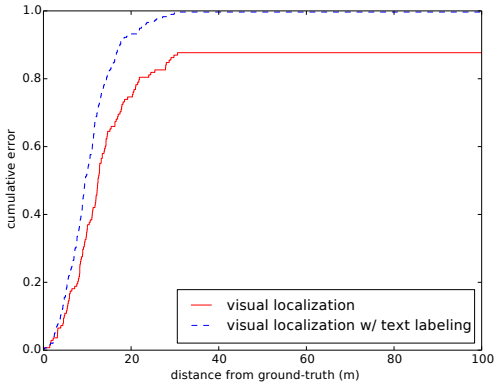

Fig. 5. London dataset error histogram. dataset cumulative error plot. The x-axis shows the distance from ground-truth position in meters, and the y-axis shows the percentage of points with distance less than or equal to the x-value. The plot compares the performance of our full approach (red plot) versus using our approach with manually labeled text (blue plot). We show an average localization error of 12.5 m and a baseline of 9.6 m.

The results of our approach can be seen in Figure 5. For this dataset, as the poses are extracted from Google's Street View, unlike the Freiburg dataset, we do not compare the results with GPS, since we do not have access to raw GPS measurements from Google's Street View maps. Our method has a mean localization error of 12.5 m and an error of 9.6 m with the baseline approach. Furthermore, on this dataset, 80% of the localization poses are at a distance between 0 and 25 m from the ground-truth position. We suffer from 2.33% localization failures due to incorrect data association, and 53.3% text-spotting localization failures.

C. Zurich Dataset

The Zurich dataset contains approximately 300 poses and 900 landmarks obtained using Google's Street View maps.

The poses were collected from central, industrial and rural regions. The histogram error plots using our approach are presented in Figure 6. Similar to the London dataset, we compare the performance of our approach to the baseline with perfect text detection. The results show a mean localization error of 23.2 m for the full approach versus 9.7 m for the approach with ground-truth text labeling. On this dataset, 60% of the localization poses lie within 0 and 25 m distance from the ground-truth position. We suffer from 3.66% localization failures due to incorrect data associations, and 57.6% text-spotting localization failures. This plus the increased mean error are due to the larger difficulty of the text extraction for this dataset because of the motion distortion and the smaller number of observations per pose.

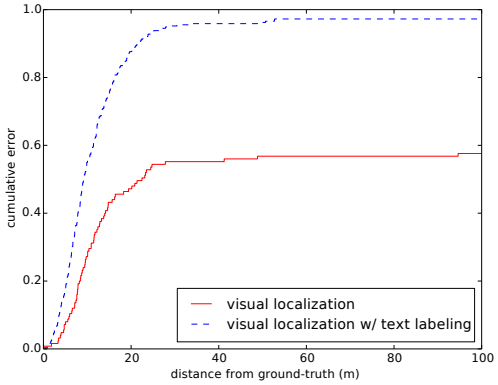

Fig. 6. Zurich dataset error histogram. dataset cumulative error plot. The plot compares using the full approach (red plot) versus the baseline using manual text labeling (blue plot). The x-axis shows the distance from ground-truth position in meters, and the y-axis shows the percentage of points with distance less than or equal to the x-value. On this dataset, our approach results in an average localization error of 23.2 m, and an improved error (average 9.7 m) with manual text labeling.

VI. CONCLUSIONS

In this paper, we presented a novel approach to the global localization problem that exploits the abundance of textual information in urban environments. Our method first extracts texts from the natural scene images, associates it to a map consisting of landmarks and corresponding text labels and then estimates the pose of the camera based on the angle and size of the extracted texts. In extensive experiments we evaluate the performance of the suggested approach, and the results demonstrate an accuracy of up to 1 meter, which corresponds to a 40% improvement over GPS poses obtained with a mobile device. Furthermore, unlike feature-based visual localization approaches, our proposed method is robust to scenery and environmental changes. This clearly demonstrates the potential of using text as a source of information in localization applications. Note that the only sensory requirement is a stream of camera images and a map of landmarks with labels that can easily be created. This makes our method easy to deploy and affordable to use. In future work we will consider improving on the text-spotting and the data association approaches used, as they are the current bottleneck of the approach. In addition, we plan to investigate the performance of this approach in a tracking scenario, where the goal is to optimize the whole path error.

REFERENCES

[1] P. Agarwal, W. Burgard, and L. Spinello. Metric localization using google street view. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[2] M. A. Brubaker, A. Geiger, and R. Urtasun. Lost! leveraging the crowd for probabilistic visual self-localization. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2013.

[3] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Int. Conf. on World Wide Web*, pages 761–770, 2009.

[4] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. of Robotics Research (IJRR)*, 30(9):1100–1123, 2011.

[5] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. Visual simultaneous localization and mapping: A survey. *Artificial Intelligence Review*, 43(1):55–81, 2015.

[6] J. Hayes and A. A. Efros. IM2GPS: Estimating geographic information from a single image. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[7] K. Konolige and M. Agrawal. FrameSLAM: from bundle adjustment to realtime visual mapping. *IEEE Transactions on Robotics*, 24(5): 1066–1077, 2008.

[8] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Dokl. Akad. Nauk*, volume 163, pages 845–848, 1965.

[9] P. Lothe, S. Bourgeois, E. Royer, M. Dhome, and S. Naudet-Collette. Real-time vehicle global localisation with a single camera in dense urban areas: Explotitation of coarse 3D city models. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 863–870, 2010.

[10] A. L. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza. MAV urban localization from google street view data. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3979–3986, 2013.

[11] N. Mattern, R. Schubert, and G. Wanielik. High-accurate vehicle localization using digital maps and coherency images. In *Intelligent Vehicles Symposium (IV)*, pages 462–469, 2010.

[12] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3538–3545, 2012.

[13] I. Posner, P. Corke, and P. Newman. Using text-spotting to query the world. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3181–3186, 2010.

[14] X. Qu, B. Soheilian, and N. Paparoditis. Vehicle localization using mono-camera and geo-referenced traffic signs. In *Intelligent Vehicles Symposium (IV)*, pages 605–610, 2015.

[15] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*, volume 589. John Wiley & Sons, 2005.

[16] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *Int. Conf. on Computer Vision*, pages 667–674, 2011.

[17] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In *British Machine Vision Conference (BMVC)*, volume 6, page 7, 2012.

[18] M. Schreiber, F. Pggenhans, and C. Stiller. Detecting symbols on road surface for mapping and localization using OCR. In *Int. Conference on Intelligent Transportation Systems (ITSC)*, pages 597–602, 2014.

[19] G. Schroth, S. Hilsenbeck, R. Huitl, F. Schweiger, and E. Steinbach. Exploiting text-related features for content-based image retrieval. In *Int. Symposium on Multimedia (ISM)*, pages 77–84, 2011.

[20] A. Torii, J. Sivic, and T. Pajdla. Visual localization by linear combination of image descriptors. In *Int. Conf. on Computer Vision Workshops (ICCV Workshops)*, pages 102–109, 2011.

[21] S. S. Tsai, H. Chen, D. M. Chen, and B. Girod. Mobile visual search with word-HOG descriptors. In *Data Compression Conference (DCC)*, pages 343–352, 2015.

[22] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *European Conf. on Computer Vision (ECCV)*, pages 255–268. 2010.

[23] W. Zhang and J. Kosecka. Image based localization in urban environments. In *Int. Symposium on 3D Data Processing, Visualization and Transmission*, pages 33–40, 2006.