# Vision-based Vehicle Localization using a Visual Street Map with Embedded SURF Scale

David Wong[1], Daisuke Deguchi[2], Ichiro Ide[1], and Hiroshi Murase[1]

[1] Graduate School of Information Science, Nagoya University, Japan
{davidw, ide, murase}@murase.m.is.nagoya-u.ac.jp
[2] Information and Communications Headquarters, Nagoya University, Japan
ddeguchi@nagoya-u.jp

**Abstract.** Accurate vehicle positioning is important not only for in-car navigation systems but is also a requirement for emerging autonomous driving methods. Consumer level GPS are inaccurate in a number of driving environments such as in tunnels or areas where tall buildings cause satellite shadowing. Current vision-based methods typically rely on the integration of multiple sensors or fundamental matrix calculation which can be unstable when the baseline is small.
In this paper we present a novel visual localization method which uses a visual street map and extracted SURF image features. By monitoring the di erence in scale of features matched between input images and the visual street map within a Dynamic Time Warping framework, stable localization in the direction of motion is achieved without calculation of the fundamental or essential matrices.
We present the system performance in real tra c environments. By comparing localization results with a high accuracy GPS ground truth, we demonstrate that accurate vehicle positioning is achieved.

**Keywords**: Ego-localization, Monocular Vision, Dynamic Time Warping, SURF, Vehicle Navigation

## 1 Introduction

Vehicle ego-localization is an essential component of in-car navigation systems and a necessary step for many of the emerging driver assistance and obstacle avoidance methods. Standard GPS systems can be sensitive to the occlusions common in city driving situations, and rarely manage 5 m accuracy even in ideal environments. For tasks such as lane recognition and obstacle avoidance, higher precision in all environments is required.

For unrestrained motion in unfamiliar environments, Simultaneous Localization and Mapping (SLAM) [1] is an active area of research. Camera-based methods are popular [2], [3], [4]. For automotive navigation, the availability of *a-priori* information and the applicability of known constraints, such as a xed ground plane, allow for simpler localization without the need for simultaneous map construction and loop closure detection. Therefore there are an increasing number of methods that propose the use of cameras with a pre-constructed image database for vehicle positioning [5], [6], [7], [8], [9]. This configuration still has many challenges, including robust localization when lateral translation occurs (for example when a lane change takes place) and computational issues with the calculation of geometry such as the fundamental matrix between views.

In this paper we propose a method for ego-localization that makes use of the scale of Speeded Up Robust Features (SURF) [10] to match images, and show how the use of feature scale improves image match accuracy. A query image is localized by using the known position information of the closest match within a database, or image street map. Unlike other image feature-based localization techniques, no essential or fundamental matrix calculation is required, yet the advantages of feature-based methods including robustness to occlusions and lateral motion are retained. Our method consists of three main components:

1. A visual street map with embedded SURF and accurate position information for every image, constructed from high accuracy sensors including GPS, IMU and odometry
2. A weighted feature matching method which applies the constraints of typical road scenes to the matching of SURF points
3. A localization algorithm that monitors the scale difference between SURF features in the query and street map images within a Dynamic Time Warping (DTW) [11] algorithm to achieve stable localization

We demonstrate the performance of our system in a typical urban traffic environment, and show that using feature scale changes is a simple yet robust way to find the closest street map image and therefore localize the current image. We show how our method is capable of localization even when the traversed lane is different from the lane used for image street map construction.

This paper is organized as follows: In Sect. 2 we give a brief overview of related research. We describe the proposed method in more detail in Sect. 3 and experimental results are presented in Sect. 4. We discuss the results in Sect. 5 before concluding in Sect. 6.

## 2 Related Work

For automotive ego-localization, there are many visual methods which perform vehicle positioning by using a pre-constructed database [5], [8] or image databases such as Google Street View [9]. These systems perform complete localization relative to database images using structure from motion techniques. Such methods allow high accuracy, for example, up to 10 cm precision when combined with an IMU [5], but are also computationally intensive and therefore may barely run in real-time (Lategahn et al. [5] quote 3–10 Hz for their vision + IMU method). They also usually employ supporting sensors in the localization stage — either an IMU [5], or odometry information [9]. A simpler approach is to localize against the closest database image, of which location is known, using DTW [11] (or Dynamic Programming [12]) to remove temporal differences between query and database image streams [6], [13], or by using a low bit-rate image sequence instead of single images [14], which improves stability in varying lighting and weather conditions. The image similarity measure used for matching between sample images and those in the database can be based on average image intensity difference [14], or a kind of template matching [6], [13]. Lane changes and occlusions are not well handled by such methods because they cause a sustained difference in appearance of an appreciable portion of the image. Feature point-based methods are more robust to such changes. Kyutoku et al. [7] matched SIFT [15] features between images to calculate the position of the epipole as a DTW cost measure for comparing image capture positions. The epipole moves away from the vanishing point as the image capture positions become similar. While effective, this technique requires the calculation of the fundamental matrix so can be unstable when the baseline between the query and database images is small.

Vehicle ego-localization is similar to the localization component of the SLAM problem for robotic navigation [1]. There are a number of successful SLAM implementations using a single camera which typically employ structure from motion techniques to determine camera pose [2], [3], [4], [16]. SLAM methods do not easily scale to the large environments found in automotive environments; however the SLAM loop closure problem, where a robot must recognize when it has entered a previously mapped area, is similar to the map relative localization step of automotive ego-localization. State-of-the-art SLAM loop closure methods often use Bag of Features [4], [17], which are excellent at recognizing visually similar areas for loop closure but do not provide a solution for exact localization. They also require the construction of feature vocabularies, which can make scaling to very large environments challenging.

None of the methods mentioned above make use of the scale property of image features to determine image similarity. We show that by using scale differences between matched features as a cost measure for DTW, we can accurately match query images to an image map in an automotive setting. Our method does not require the calculation of feature vocabularies or reconstruction of scene geometry, and since it is feature-based, it continues to work well when partial occlusion or lane changes occur.

## 3 Proposed Ego-localization Method

This section describes a method of ego-localization by comparing images captured from a vehicle-mounted camera and a pre-constructed visual street map. The street map is constructed using data captured from a vehicle equipped with cameras and accurate positioning hardware. Images captured in the localization step are compared to the street map images using DTW to compensate for speed differences in the two image streams. The process is described in more detail below. Sect. 3.1 describes the concept behind SURF scale matching, and Sect. 3.2 details the visual street map construction step. Sect. 3.3 describes the localiza-

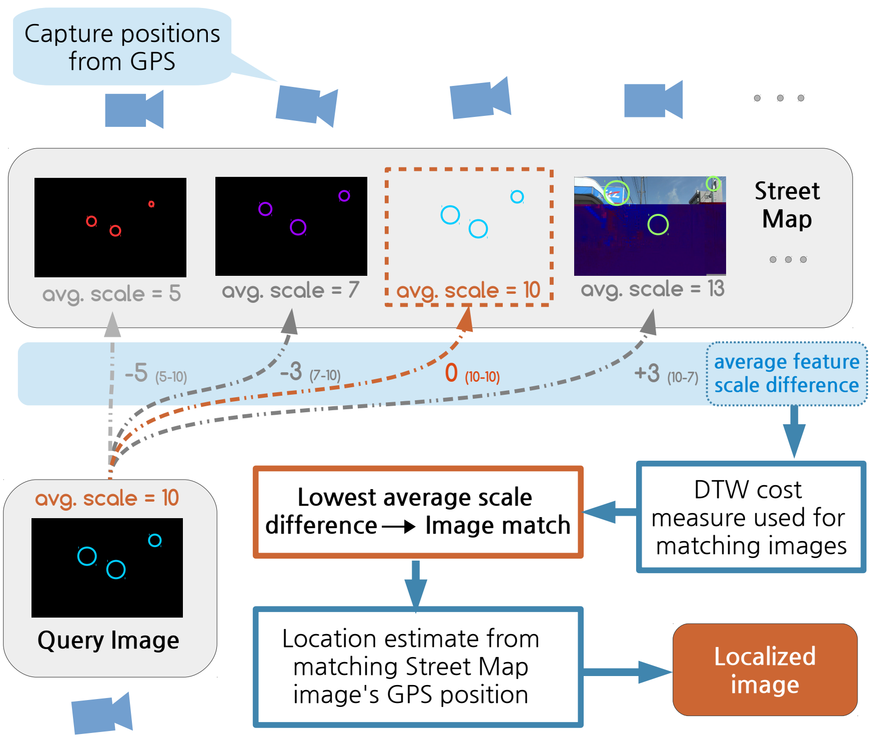


Fig. 1. A ow chart outlining the proposed method. The two main stages are the visual street map construction, and ego-localization by DTW matching the query images from the vehicle to be localized to the street map.

tion of an input query image. An overview of the proposed system is presented in Fig. 1.

### 3.1 Concept: image matching using SURF scale

Scale invariant features such as SURF are commonly used for their robustness to changes in lighting and view orientation. One of the properties of SURF keypoints is their size, or scale. The method proposed in this paper is based around the use of the scale of these features for image matching and therefore localization. If two images have the same viewing direction, their corresponding SURF feature points will have a similar scale when the capture positions were spatially close. As the distance between the images increases, the di erence in corresponding feature scales also increase. The proposed method makes use of this change to match images between the query image and the street map by averaging the scale change of the matched features. The street map image with the smallest average scale change from the query is selected as a match, therefore

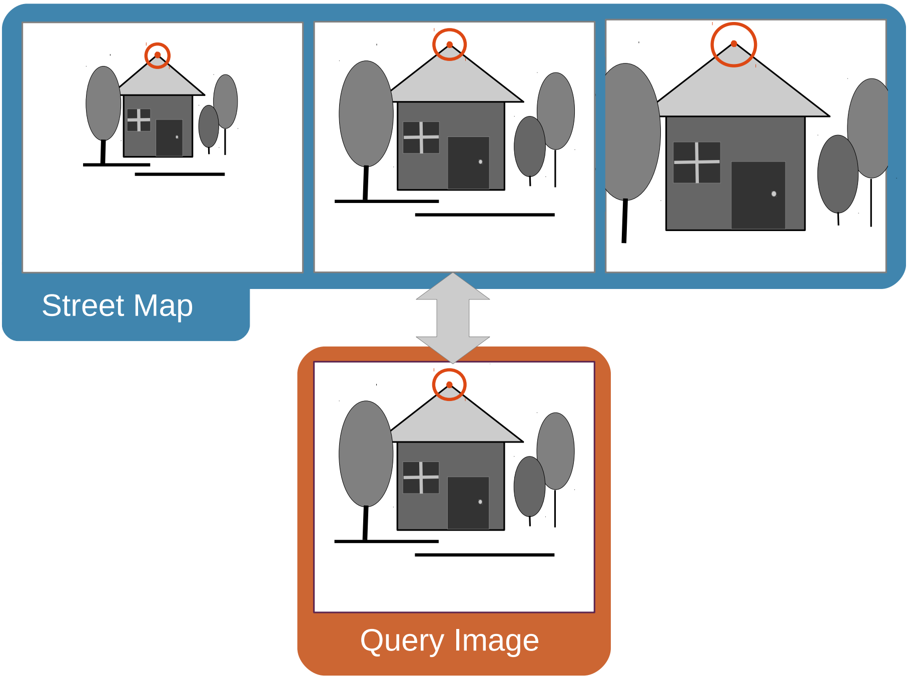


Fig. 2. A demonstration of scale change for image matching. The street map images are in the top row and the query image below. A sample feature matched across all views shows how the scale change is used to nd the closest match.

localizing the query image in the direction of motion. In the context of DTW, features extracted from street map images behind the query image's location will have a smaller scale than the corresponding features of the query image; conversely, corresponding features from street map images from in front of the query image will have larger scale. Sequentially running through the street map images and nding where the feature scale changes are minimized leads to the spatially closest street map image for each query image, and because we know the location of each street map image in world co-ordinates, we can assign this position to the query image for localization. Only the scale change between the matched features is used, so there is no need to calculate fundamental matrices. This allows matching even when the baseline between query and street map images is small. When lateral motion occurs, the average scale change remains constant so the method is also robust to changes in lane and road position when localizing in the direction of motion. Fig. 2 shows the concept of using feature scale change for image matching.

Table 1. Localization Results

| Method | Avg. error (m) | Max. error (m) |
|---|---|---|
| Proposed method (with SURF scale, one camera) | 1.96 | 8.10 |
| Proposed method (with SURF scale, two cameras) | 1.56 | 6.00 |
| Comparative method (without SURF scale, one camera) | 8.45 | 16.12 |

matches was typically relatively small so a $w_s$ value of approximately ten times $w_d$ and $w_r$ (which were approximately equal) was found to be effective. This configuration still prioritized the *SSD* of feature descriptors for determining the best feature match.

Localization relative to the street map was performed using one camera and repeated with two cameras. For a comparative method that does not make use of feature scales, the inverse of the number of matched features was used for the DTW matching cost [18]. In the comparative method, the cost measure in (2) was replaced with the following:

$$g(t;\tau) = 1/N_{t,\tau}. \tag{3}$$

If enough features are extracted and the same feature match filtering described by (1) is employed, the comparative method offers a reasonably effective way of identifying the general street map area of the current query image. However, the results we present below show how additionally monitoring the scale of the matched features allows a much more refined comparison of the input and street map images.

The localization accuracy for each method was evaluated by using the MMS localization data. The ground truth localization information associated with the query images was used to calculate the actual closest visual street map image, creating an image match ground truth. The match result of the query image relative to the street map was compared to the image match ground truth. The image matching results of the three methods are presented in Fig. 3.

The results show that the use of the scale of matched features gives a more robust distance measure between the query and street map images, even when the query images are captured in a different lane from the street map images. Comparison of the one and two camera results shows that wider field of view provided by two cameras increases the image matching performance of the method considerably. Fig. 4 shows a comparison of the street map images selected as matches and sample query images, for both the proposed method and comparison method. The image matching performance of the proposed method is consistently good, with the GPS ground truth showing that the system finds the correct closest street map image for 70 % of the time, and is always within 3 frames of the correct match (when both cameras are used).

The cameras capture images at distance rather than time intervals, so there is a high accuracy penalty for each incorrectly matched image frame. An incorrect match results in a localization error of approximately 2 m or a multiple of 2 m because of the fixed frame capture separation of the MMS system. Despite of this, the average localization error of our system was less than 2 m, which is within one street map image interval, even when a single camera was used. The average localization accuracy results are presented in Table 1.

## 5 Discussion

Even though the street view map was constructed using a different lane from the query image views, successful localization was performed, illustrating the robustness of using feature scale for image matching when lateral change in viewpoint occurs. Because it is a feature-based localization method, it also demonstrated robust matching in the presence of occlusion. An example of successful matching in an occluded scene is shown in Fig. 5. Unlike most feature-based localization methods, no calculation of image geometry is required, so the method is simple. It also demonstrated good recovery from incorrect matches. In the dataset used for the experiments, the vehicle never came into a situation where localization was not possible within the spatial constraints applied by the DTW method. In the case where this could happen though, if the vehicle became lost a regressive image matching method from a wider selection of the street map images may need to be performed.

There were a number of issues specific to the image capture method which limited the accuracy of the proposed method in our experiments. The 2 m capture interval of the camera meant that the metric localization error of individual image matches could only be evaluated in multiples of approximately two meters. The accuracy of the system is highly dependent on the frame rate of capture and also visual street map image interval, so a higher frame rate would provide far superior results and more effective analysis of accuracy. The localization accuracy could also be potentially improved by applying a motion model and interpolating the query image position between the two closest matched street map images.

The use of both forward facing cameras improved the results, because of the wider field of view they enabled. The two cameras were not used as a stereo pair so a similar result could be achieved using a monocular camera and a lens providing a wide field of view.

The experiments in this method used the same vehicle and cameras for the street map construction and localization stages. This is quite a favorable configuration, so future work will include testing with images from a range of cameras and lens types as well as determining how differing camera heights and environmental conditions affect the stability of the system.


Fig. 4. Sample images showing DTW matching results. The central column is the query image, and the one on left the corresponding matched street map image using the proposed method. The column on the right shows the matched street map images using the comparative method. The numbers in the top right of the images are the sequence frame numbers, showing how DTW matching absorbs di erences in vehicle speeds between the visual street map and query image sequences. Note that although results from using both cameras are displayed, only the left-hand camera image is shown for clarity.

**Fig. 5.** Example of successful image matching when occlusions occur in either the query image or street map image.

## 6 Conclusion

We proposed a method for ego-localization using the average SURF scale change across matched features as a cost measure for sequential image matching against a pre-constructed visual street map. The experimental results show that e ective street map image matching can be achieved with an average error of 1.56 m using two cameras. The system performs well even when the query images are captured in a di erent lane from the street map images, and is robust to occlusions in either image streams. There are potential improvements in localization accuracy to be made by using a higher frame rate image capture for the visual street map and localization stages. Interpolating the vehicle position between several of the closest matched street map images rather than taking the position of the single closest matched street map image is another potential extension to the method.

Future work will include the construction of a street map with a higher camera frame rate for greater localization accuracy, and testing in a larger variety of environmental conditions. We also plan to test the method with di erent cameras and lenses, for example a single camera with a wide angle lens con guration for a wide eld of view to replace the two camera experimental setup presented in this paper.

## References


1. Durrant-Whyte, H.F., Bailey, T.: Simultaneous localization and mapping: part I. IEEE Robotics and Automation Magazine **13**(2) (2006)
2. Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P.: Monocular vision based SLAM for mobile robots. In: Proc. 18th International Conference of Pattern Recognition (ICPR2006). Volume 3. (2006) 1027{1031
3. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Trans. Pattern Analysis and Machine Intelligence **29**(6) (2007) 1052{1067
4. Botterill, T., Mills, S., Green, R.D.: Bag-of-words-driven, single-camera simultaneous localization and mapping. Journal of Field Robotics **28**(2) (2011) 204{226
5. Lategahn, H., Schreiber, M., Ziegler, J., Stiller, C.: Urban localization with camera and inertial measurement unit. In: Proc. 2013 IEEE Intelligent Vehicles Symposium (IV2013). (2013) 719{724
6. Uchiyama, H., Deguchi, D., Takahashi, T., Ide, I., Murase, H.: Ego-localization using streetscape image sequences from in-vehicle cameras. In: Proc. 2009 IEEE Intelligent Vehicles Symposium (IV2009). (2009) 185{190
7. Kyutoku, H., Takahashi, T., Mekada, Y., Ide, I., Murase, H.: On-road obstacle detection by comparing present and past in-vehicle camera images. In: Proc. 12th IAPR Conference on Machine Vision Applications (MVA2011). (2011) 357{360
8. Nedevschi, S., Popescu, V., Danescu, R., Marita, T., Oniga, F.: Accurate ego-vehicle global localization at intersections through alignment of visual data with digital map. IEEE Trans. Intelligent Transportation Systems **14**(2) (2013)
9. Badino, H., Huber, D.F., Kanade, T.: Real-time topometric localization. In: Proc. 2012 IEEE International Conference on Robotics and Automation (ICRA2012). (2012) 1635{1642
10. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up Robust Features (SURF). Computer Vision and Image Understanding **110**(3) (2008) 346{359
11. Muller, M.: Dynamic time warping. In: Information Retrieval for Music and Motion. Springer Berlin Heidelberg 69{84
12. Wagner, D.B.: Dynamic programming. The Mathematica Journal **5**(4) (1995) 42{51
13. Sato, J., Takahashi, T., Ide, I., Murase, H.: Change detection in streetscapes from GPS coordinated omni-directional image sequences. In: Proc. 18th International Conference of Pattern Recognition (ICPR2006). (2006) 935{938
14. Milford, M.: Visual route recognition with a handful of bits. In: Proc. 2012 Robotics: Science and Systems. (2012) 297{304
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91{110
16. Se, S., Lowe, D., Little, J.: Vision-based mobile robot localization and mapping using scale-invariant features. In: Proc. 2001 IEEE International Conference on Robotics and Automation (ICRA2001). (2001) 2051{2058
17. Cummins, M., Newman, P.: Appearance-only slam at large scale with fab-map 2.0. The International Journal of Robotics Research **30**(9) (2010) 1{24
18. Kameda, Y., Ohta, Y.: An implementation of pedestrian localization by rst-person view camera in urban area. In: Proc. 13th Meeting on Image Recognition and Understanding (MIRU2010). (2010) 364{369