Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Predicting the Long-term Robustness of Visual Features

Benjamin Metka, Annika Besetzny, Mathias Franzius, Ute Bauer-Wersing

2015

Preprint:

This is an accepted article published in International Conference on Advanced Robotics. The final authenticated version is available online at: https://doi.org/[DOI not available]

Predicting the Long-term Robustness of Visual Features

Benjamin Metka*, Annika Besetzny*, Ute Bauer-Wersing* and Mathias Franzius†
*Frankfurt University of Applied Sciences
Frankfurt am Main, Germany
Email: bmetka@fb2.fh-frankfurt.de
†Honda Research Institute Europe GmbH
Offenbach, Germany

***Abstract*—Many vision based localization methods extract local visual features to build a sparse map of the environment and estimate the position of the camera from feature correspondences. However, the majority of features is typically only detectable for short time-frames so that most information in the map becomes obsolete over longer periods of time. Long-term localization is therefore a challenging problem especially in outdoor scenarios where the appearance of the environment can change drastically due to different day times, weather conditions or seasonal effects. We propose to learn a model of stable and unstable feature characteristics from texture and color information around detected interest points that allows to predict the robustness of visual features. The model can be incorporated into the conventional feature extraction and matching process to reject potentially unstable features during the mapping phase. The application of the additional filtering step yields more compact maps and therefore reduces the probability of false positive matches, which can cause complete failure of a localization system. The model is trained with recordings of a train journey on the same track across seasons which facilitates the identification of stable and unstable features. Experiments on data of the same domain demonstrate the generalization capabilities of the learned characteristics.**

I. Introduction

The ability to build a map of the environment and to determine the own location is a prerequisite for autonomously acting mobile robots. While localization and mapping can be performed with different kinds of sensors, vision based approaches are especially appealing because of the low cost, weight and the high availability of cameras. Research in the field is steadily progressing and recent work has shown impressive results in mapping large scale environments (e.g. [1]–[3]). However, visual localization and mapping in long-term outdoor scenarios remains a challenging problem since the appearance of a scene can change dramatically due to different lighting conditions, cast shadows, dynamic elements or seasonal effects.

Many approaches use local visual features to perform localization by establishing correspondences between the features in the map and the current view. Distinctive interest points are extracted from an image and a descriptor is computed from the surrounding image patch which enables the re-identification of the feature. Correspondences are established by performing a nearest neighbor search in descriptor space.

Due to dynamic objects and the changing appearance of the environment in long-term outdoor scenarios visual features can usually only be recognized for limited periods of time and the number of true positive matches might decrease drastically even after a few hours [4]. Therefore, most information in the initial map is likely to be only valid for short time-frames and the probability of false positive matches increases over time. In the case of place recognition for loop closure detection even one wrongly established correspondence can cause complete failure of the localization system. Geometric constraints can be used to reject false matches within a RANSAC [5] based scheme but the number of iterations needed to find a hypothesis grows exponentially with the number of outliers and the outlier ratio must not exceed 50% [6].

We propose to learn a discriminative model for long-term robustness prediction of local visual features. The model can be incorporated as an intermediate filtering stage in the standard feature processing pipeline to reject potentially unstable features during the mapping phase. The approach is motivated by the fact that a scene usually contains objects which are likely persistent over time like tree chunks, rocks or houses while others can be expected to be non-stationary or change their appearance over time e.g. cars, people or vegetation. We assume that image statistics around visual features, corresponding to some part of a physical object, can be used to discriminate stable and unstable features.

Robustness characteristics of visual features are learned from texture and color statistics around their corresponding interest points. Data is gathered from recordings of a train journey on the same track over the seasons which allows the easy identification of stable and unstable features. A support vector machine is trained with these statistics with the aim of predicting the long-term robustness of features in unknown scenes of the same domain. Application of the proposed approach results in a smaller set of features with a higher percentage of robust ones, hence it reduces processing time, memory consumption and improves long-term robustness because of fewer false positive matches.

II. Related Work

Appearance changes of the environment induce high visual diversity into images of the same place visited at different times. This poses a severe challenge for vision based localization and mapping methods and different approaches towards long-term autonomy have been proposed recently.

The special case of appearance changes caused by different illumination conditions is addressed by several authors. In [7]

978-1-4673-7509-2/15/$31.00 2015 IEEE

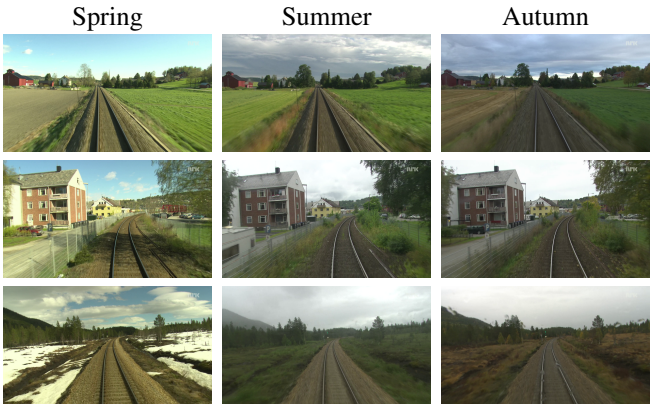


Fig. 1. Varying appearance of the same location over time. Images are from the Nordlandsbanen Railway data set which consists of recordings of the same track in every season featuring seasonal, weather and lightning effects in man-made and natural environments. Note that season winter was not considered in this work since the drastic changes in appearance do not allow a reasonable amount of feature matches. Content licensed under Creative Commons, Source: NRKbeta.no https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/.

the effects of shadows are mitigated by transforming images to a shadow invariant representation where pixel values are a function of the underlying material property. Mapping and localization is then performed in parallel with standard gray-scale and illumination invariant images.

Work on learning descriptors which are more robust with respect to illumination changes than hand designed descriptors has been done by [8] and [9]. In [8] features are tracked over a sequence of a time-lapse video containing dynamic lighting conditions. Matching and non-matching pairs of image patches are discriminated by a contrastive cost function. Genetic optimization is used in [9] to obtain a descriptor from a pool of elementary descriptor building blocks. Although all authors demonstrate superior performance with respect to standard feature matching, illumination invariance addresses only a part of possible appearance changes.

Another approach of Churchill et. al [10] is to build and maintain dynamic maps of the environment where the diversity in the appearance of a place is captured by different visual experiences which are linked topologically. Every time the system fails to localize in the current map a new experience is created. The authors demonstrate localization and mapping in an outdoor environment at different daytimes and changing weather conditions over the course of three month.

In [11] and [12] a certain track is traversed several times under different conditions while keeping track of feature occurrences per place. The statistics collected during the training runs allow to model the probability of feature visibility per place.

Milford et al. [13] demonstrated localization along one dimensional routes across difficult conditions causing severe changes in appearance. The approach, called *SeqSLAM*, matches sequences of images rather than finding a single global best match. Matching is performed directly on the down-sampled, patch-normalized images. The holistic image matching over sequences restricts this approach to one dimensional traversals along a defined route without deviations in lateral position.

In [14] the authors propose to predict the appearance changes of a place to perform global image matching between different points in time. A dictionary of corresponding visual words from aligned image streams is constructed which allows to translate images of a place from one distinct condition to another. Instead of directly matching images of places from different conditions matching is performed with the translated version of the image. The authors demonstrate that holistic place recognition (*SeqSLAM*) benefits from appearance change prediction.

We propose a generic approach of learning a model for long-term robustness prediction of visual features which does not depend on a specific interest point detector/descriptor combination as long as a suitable feature representation is chosen. The aim is to learn a model of the characteristics of robust features which allows generalization to different scenes of the same domain.

III. Learning Robustness Characteristics of Visual Features

In this section we describe the process of training data generation and the learning approach for robustness prediction of local visual features. The proposed method is based on the assumption that texture and color statistics around visual features can be used to train a discriminative model that enables the prediction of a features' long-term robustness. Stable and unstable features are identified in images of the same location at different points in time and under varying environmental conditions. A support vector machine is then trained with labeled feature vectors computed from the visual features' surrounding image patches.

A. Data Set

The data set we use in this work consists of high definition video material from a TV documentary recorded on a train journey on the same 729 kilometer railway track once in every season (see Fig. 1 for example images). It features seasonal effects like snow covered ground and color changing foliage as well as different weather and lighting conditions in natural and man-made environments. GPS-readings were recorded in conjunction with the video and used to time-synchronize the recordings of the different seasons. After synchronization the position of the train in one frame from one video corresponds to the same frame in the other videos. Thus frame-accurate ground truth information is available within the accuracy of the GPS-localization.

We extracted frames at a fixed interval throughout the whole videos of every season to generate a manageable amount of image data. The resulting frames contain a broad variety of locations each with a season specific appearance. To further refine the GPS-based alignment the frames from one season are defined as reference and the best matching frame from the other seasons is found within a sequence around the extracted key-frames. Matching is performed by computing the normalized cross-correlation between the down-sampled and patch-normalized images from the reference frame and the ones from the sequence. Frames extracted from parts of the video when the train passed a tunnel, stopped at signal lights or had the windshield wiper turned on were sorted out since they would distort the results.

The final training data set contains images of 164 locations in seasons spring, summer and autumn. Recordings from the winter season were not considered for the training process nor

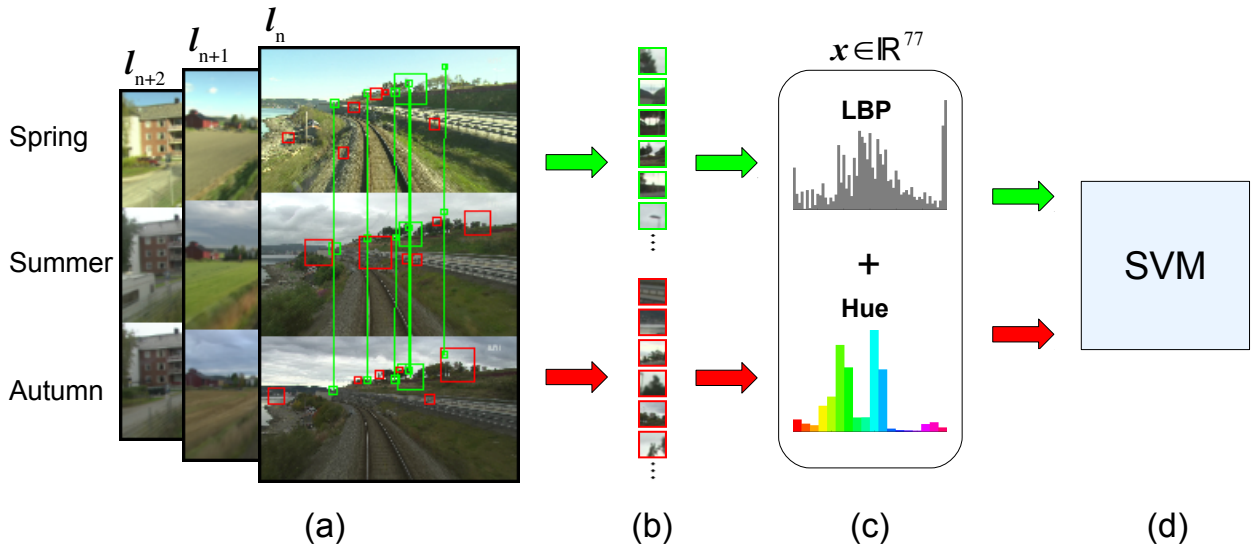


Fig. 2. **Illustration of the training process.** (a) Stable and unstable features are identified in the seasonal image sequence of all locations $l_{1\ldots164}$ and selected for training. Features which can be tracked across the seasonal images of the current location l_n, depicted with green squares, are selected as stable samples. Features which were extracted in one season but could not be matched in any of the other seasons are defined as unstable. Since the number of occurrences of unstable features is many times higher only a subset is randomly chosen to obtain balanced classes. The selected unstable samples are represented by red squares. The size of the squares is determined by the scale of the corresponding interest point. (b) Patches of the selected stable and unstable features are resized to 48x48 pixel for feature vector computation. (c) A histogram of uniform *Local Binary Patterns* and a hue histogram are computed on the image patch. Concatenation of the histograms yields the 77-dimensional feature vector. (d) Finally a support vector machine is trained with the labeled feature vectors.

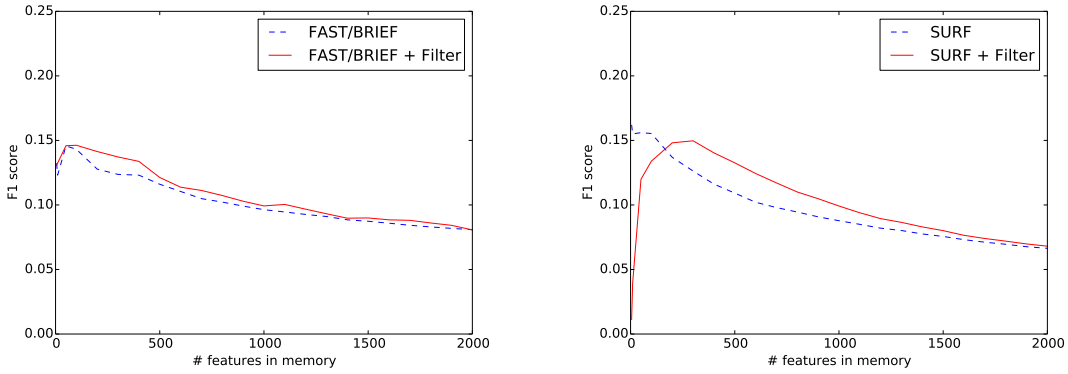
the experiments since the extreme appearance did not allow a reasonable amount of feature matches. The TV logo in the top right corner of the extracted frames is masked out for interest point detection and feature description.

B. Training Data Generation

To generate labeled training data we identify stable and unstable features in the data set by tracking them over the seasonal images of every location. For every of the 164 locations we detect interest points in the corresponding seasonal images and compute their descriptors. Descriptors are then matched between all image combinations by performing a k-nearest neighbor search in descriptor space with $k = 2$ and the application of the distance ratio test with a ratio of $r = 0.75$. We consider two features as equal if their descriptors match and the Euclidean distance in image space is smaller than a threshold. We set this threshold to 40 pixel since the viewpoint of the camera is not exactly the same over the different recordings. Features which were successfully matched over all seasonal images of one location are labeled as stable and added to the training set. Features which could not be matched to any of the other seasonal images are labeled as unstable. Since the amount of unstable features is usually many times larger than the stable ones only a subset is chosen randomly so that the classes are balanced. The result of the selection process is illustrated in Fig. 2a, 2b.
We use the implementations from the *openCV*-library [15] of the *FAST* interest point detector [16] and the binary *BRIEF* descriptor [17] as well as *SURF* [18] for interest point detection and description. The orientation of the features is not encoded since the target application is localization of vehicles moving on the ground plane. The threshold for the detector response has been set to a value so that roughly 1000 interest points were detected per image. From the total number of features found in the 164 seasonal images the selection process yields a balanced set of 10000 and 14000 labeled features when using *FAST/BRIEF* and *SURF*, respectively. The higher number of selected *SURF* features can be explained by the fact that interest point detection is performed at multiple scales while *FAST* interest points are extracted at a fixed scale. Furthermore parts of the image corresponding to physically nearby regions are frequently affected by motionblur which is disadvantageous for the *FAST* detector since it responds to corner like image structures.

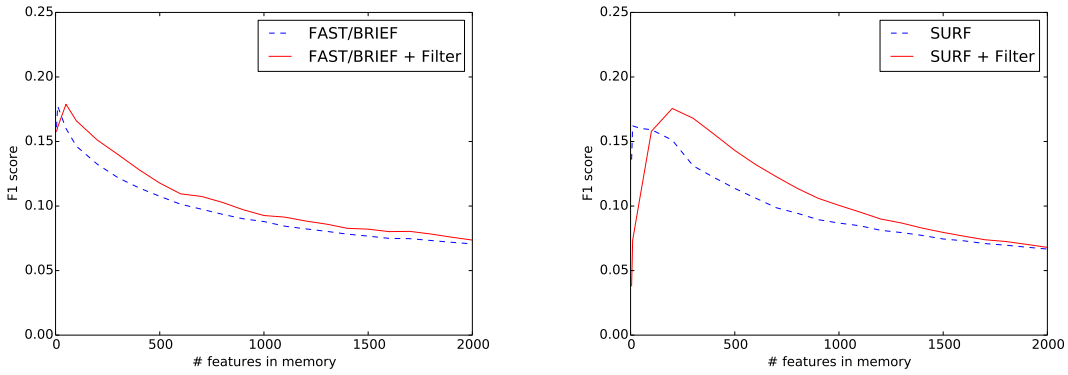
C. Training Process

Feature vectors for the support vector machine training are constructed from the pixel values around the interest points of the stable and unstable features which have been selected for the training set. Features are a combination of low level texture and color information. Texture information is obtained by computing a histogram of the uniform *Local Binary Patterns* [19]. Color information is encoded in an 18 bin hue histogram. The sample points within the patch region around the interest point are weighted by a two dimensional Gaussian window for histogram computation. Concatenation of the histograms yields the 77-dimensional feature vector.
The size of the patch used for the computation of the feature vector is chosen to be equal to the size of the descriptor window which is determined by the scale of the interest point. Since the *FAST* interest points are not localized in scale the image patch is fixed to 48x48 pixel. When using *SURF* the descriptor window varies with scale. Instead of scaling the LBP-operator we chose to resize the patch to 48x48 pixel with bilinear interpolation for feature computation.
A support vector machine with RBF-kernel is trained with



(a) (b)

Fig. 3. **Matching from summer to spring**. (a) Results from the experiment with *FAST/BRIEF*. Application of the filter based on robustness prediction yields an average improvement of 4.74%. (b) Results from the experiment with *SURF*. In the low memory where false classifications have a stronger impact the F1-score is worse when the robustness filter is applied. In the subsequent memory regions performance with the robust filter leads to a substantial performance gain so that the overall improvement results to 7.72%.



(a) (b)

Fig. 4. **Matching from autumn to spring**. (a) Results from the experiment with *FAST/BRIEF*. Application of the robustness filter results in an average performance gain of 8.88%. (b) Results from the experiment with *SURF*. As in the previous experiment the F1-score obtained with the robustness filter is worse when using up to 150 features for memory. For larger memory sizes feature matching benefits from the application of the filter so that the average performance is increased by 11.79%.

feature vectors from the balanced stable and unstable visual features resulting from the selection process. Parameters of the support vector machine are determined by a grid search with five-fold cross validation. Single steps of the whole process are illustrated in Fig. 2.

IV. Experiments

The aim of the robustness prediction is to filter out potentially unstable visual features during the mapping phase. This results in more compact maps and a reduced probability of false positive matches compared to conventional feature processing.

We evaluate our filtering approach and the conventional feature processing on a separate test data set which is created in the same way as described in section III-B but with an offset in time. This ensures that the training and test set do not contain images of the same location. We perform cross-season feature matching on sequences of 60 extracted frames and store the number of true positive and false positive matches. Features from summer and autumn are matched against features from spring which is defined as the reference season. Two features from the reference and query season are defined as equal, and counted as true positive, if their descriptors match and the Euclidean distance in image space is smaller than 40 pixel. If the descriptor of a feature was matched to the descriptor of a feature in the other season but the Euclidean distance is greater than 40 pixel it is considered a false positive match.

The cross-season matching along the sequence is performed several times with an increasing size of memory. The number of extracted features was set to values in the range between 5 and 2000 for the reference and query frames. With the conventional method the appropriate number of features according to the memory limit is obtained by adjusting the threshold of the interest point detector until the limit is reached. In case there were more features extracted, the ones with lower response are sorted out. The robustness filter is only applied for feature extraction from the reference frames. In this case the selection process is different. In the first step a number of interest points is detected that is equal to 1.5 times of the final memory sizes. For every feature we compute the confidence with the trained support vector machine and sort them by their confidence value. Then the top features are selected according

to the memory size.
True positive and false positive rates are averaged over one run through the image sequence and memory step. Performance of the conventional feature matching and our robustness filtering approach is compared in terms of the F1-score as the harmonic mean of precision and recall.

A. Results

Results of the comparison between conventional feature matching and our robustness filtering approach are illustrated in Fig. 3 and Fig. 4 showing the F1-scores from experiments matching from summer to spring and from autumn to spring, respectively. As expected curves from all experiments decrease with growing memory size since the probability of false positive matches is getting larger. It can be observed that the application of the robustness filter during feature extraction from images of the reference season generally increases feature matching performance compared to conventional feature processing. Applying the proposed additional robustness filter results in an average performance gain between 4.74%-11.79% in terms of the F1-score.
In the experiments with *SURF* the performance of our robustness filtering approach is worse than the conventional method in low memory regions using up to about 150 features. The model apparently rejected robust features as unstable which has a negative impact on performance especially when only few features are memorized. The same effect can be observed in both experiments since the robustness filter is applied for the reference season spring. In the subsequent memory regions performance with the robust filter leads to a substantial performance gain so that the overall average improvement results to 7.72% and 11.79%.
In the experiments with *FAST/BRIEF* the performance gain is smaller but more steady regarding the low memory regions. The average performance improvement is equal to 4.74% and 8.88% when matching from summer to spring and matching from autumn to spring respectively.
Surprisingly the performance of both variants is better when matching features from autumn to spring than matching features from summer to spring even if the temporal distance is shorter. We assume that this effect results from the fact that the vegetation might be more similar.

V. Conclusion

In this paper we have presented a method to learn a model of the robustness characteristics of visual features which was trained with stable and unstable features identified in images of the same location across seasons. The model can be incorporated into the standard feature processing pipeline as an intermediate filtering stage to predict the long-term robustness of the extracted features and reject the potentially unstable ones.
Experiments on a separate test set have shown the capability of the model to generalize to unseen features of the same domain. Integration of the model for robustness prediction during feature extraction from the reference images resulted in an increased F1-score between 4.74%-11.79% compared to conventional feature processing.
While we expected a higher gain in performance it has to be noted that learning a model to distinguish between stable and unstable features is a hard problem because of the high diversity of visual features and an overlap between the classes. In real world scenarios overlapping classes can not be avoided since even the most stable features might be occluded, overgrown by vegetation or disappear due to cast shadows. Therefore a misclassification can always occur although a features' characteristics conform to the learned model.
In general the results were quite surprising since we expected to find robust features on man-made objects. Instead the most stable features are found at tree tops of conifers with high contrast against the skyline (see Fig. 5). This can be explained by the nature of the data set which contains a lot more natural than man-made urban scenes and the selection bias of the interest point detectors.
Since the approach is simple and generally applicable it would be interesting to evaluate it in different kinds of environments. An interesting data source would be time-lapse recordings which capture the appearance changes of a location over time with a fixed camera because they allow the easy identification of stable and unstable features. However, recordings from several different places would be required to capture general feature characteristics and to learn a useful model.
An extension to further improve long-term robustness would be the integration of lighting invariant descriptors e.g. [8], [9] which could be easily incorporated into the proposed approach.

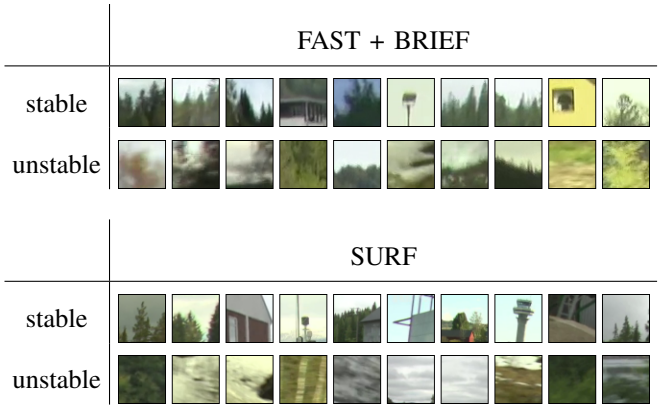


Fig. 5. **Most stable and unstable features**. Features from all seasonal images of the test data set were ranked by the confidence value of the robustness prediction to determine the 10 most stable and unstable features. Features with high confidence are found on man-made objects like lanterns or buildings which is what we expected. Surprisingly features on top of conifers against the skyline give the highest confidence values. Low confidence features are mainly found in regions with little texture and contrast.

References

[1] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," in *Robotics: Science and Systems V, University of Washington, Seattle, USA, June 28 - July 1, 2009*, 2009.

[2] G. Sibley, C. Mei, I. D. Reid, and P. M. Newman, "Vast-scale outdoor navigation using adaptive relative bundle adjustment," *I. J. Robotic Res.*, vol. 29, no. 8, pp. 958–980, 2010.

[3] H. Lategahn, M. Schreiber, J. Ziegler, and C. Stiller, "Urban localization with camera and inertial measurement unit," in *2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast City, Australia, June 23-26, 2013*, 2013, pp. 719–724.

[4] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, 2013, pp. 3791–3798.

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

[6] F. Fraundorfer and D. Scaramuzza, "Visual odometry : Part ii: Matching, robustness, optimization, and applications," *Robotics Automation Magazine, IEEE*, vol. 19, no. 2, pp. 78–90, June 2012.

[7] C. McManus, W. Churchill, W. P. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014*, 2014, pp. 901–906.

[8] N. Carlevaris-Bianco and R. M. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Chicago, IL, USA, September 2014, pp. 2769–2776.

[9] H. Lategahn, J. Beck, and C. Stiller, "DIRD is an illumination robust descriptor," in *2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, June 8-11, 2014*, 2014, pp. 756–761.

[10] W. Churchill and P. M. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *IEEE International Conference on Robotics and Automation, ICRA 2012, 14-18 May, 2012, St. Paul, Minnesota, USA*, 2012, pp. 4525–4532.

[11] E. Johns and G. Yang, "Dynamic scene models for incremental, long-term, appearance-based localisation," in *2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, 2013, pp. 2731–2736.

[12] ——, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, 2013, pp. 3212–3218.

[13] M. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation, ICRA 2012, 14-18 May, 2012, St. Paul, Minnesota, USA*, 2012, pp. 1643–1649.

[14] P. Neubert, N. Sünderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *2013 European Conference on Mobile Robots, Barcelona, Catalonia, Spain, September 25-27, 2013*, 2013, pp. 198–203.

[15] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.

[16] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, 2006, pp. 430–443.

[17] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: binary robust independent elementary features," in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, 2010, pp. 778–792.

[18] H. Bay, T. Tuytelaars, and L. J. V. Gool, "SURF: speeded up robust features," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, 2006, pp. 404–417.

[19] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.