Place Recognition with WxBS Retrieval

Dmytro Mishkin Michal Perdoch
Jiri Matas
Center for Machine Perception
Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Cybernetics
ducha.aiki@gmail.com {perdom1,matas}@cmp.felk.cvut.cz

Abstract

We present a novel visual place recognition method designed for operation in challenging conditions such as encountered in day to night or winter to summer matching. The proposed WxBS Retrieval method is novel in enriching a bag of words approach with the use of multiple detectors, descriptors with suitable visual vocabularies, view synthesis, and adaptive thresholding to compensate for large variations in contrast and richness of features in different conditions.

The performance of the method evaluated on the public Visual Place Recognition in Changing Environments (VPRiCE) dataset was achieved with precision 0.689 and recall 0.798 and F1-score 0.740. The precision and F1 score are best results so far reported for VPRiCE dataset. Experiments show that the combination of retrieval and matching algorithms with detectors and descriptors insensitive to gradient reversal and contrast lead to both high accuracy and scalability.

1. Introduction

Visual place recognition is not only an interesting problem in its own right, e.g. in the form of localizing historical photographs, but also an enabling technology facilitating applications in areas like autonomous navigation and augmented reality.

The problem in commonly formalized as identification of reference images depicting the same scene as the query which is followed by viewpoint estimation. The time for preprocessing the potentially large corpus of reference images is considered to be unlimited. The query is either a single image, which is most common, or an unorganized set of images or a sequence. Similarly, the reference "map" data may be either images or sequences.

In certain scenarios an approximate location is assumed known from a GPS, GSM or inertial sensors, limiting the number of candidate reference images. Due to inaccuracies, the visual search might still involve matching against tens of thousands of images. We therefore restrict our attention to place recognition method that are able to handle large numbers of reference images - only the earliest place recognition approaches were based on pairwise wide-baseline matching of the query and the database images [19]

Figure 1: Challenges of the Visual Place Recognition in Changing Environments Dataset. Query examples (left), database images highest ranked by the proposed WxBS retrieval (right).

Fast approximate nearest neighbor search techniques and distinctive descriptors [13] enabled localization within thousands of images. Advances in specific image retrieval based on local features, bag of words and fast spatial verification [18, 21] allowed scaling image-based localization to

much larger datasets. Precision of camera localization has also improved significantly, benefiting for instance from 3D structure from motion models and from the use of 2D to 3D matching [20]. In favorable environments and conditions, the basic place recognition is a technical rather than a research problem.

However, new challenges surfaced in the retrieval of images from millions of street level images. The problems of co-occurring features, and confusing features were discussed by Chum *et al*. [6] and Knopp *et al*. [10]. Place recognition in urban "canyons" with many repetitive structures was addressed by Torii *et al*. [23]. Place recognition in mountain environments that are often without dominant landmarks was investigated by Baboud *et al*. [4], and Baatz *et al*. [3]. Location recognition in challenging outdoor conditions such as day-to-night or including seasonal changes is another problem that has received little attention.

In this paper, we present a novel visual place recognition method called *WxBS Retrieval* designed for operation in conditions that at the same time differ significantly in properties like illumination (day, night), the sensor (visible, infrared) , viewpoint, appearance (winter, summer), time of acquisition (historical, current) or the medium (clear, hazy, smoky) *i.e*. exhibit "x wide-baselines" - viewpoint, temporal, appearance, etc. WxBS Retrieval draws heavily on the WxBS-M [15] two-view matching algorithm which proposed a set of features, descriptors, view synthesis steps, and a matching strategies that performed well on WxBS problems. The WxBS-M matcher reflects the progress in recent local feature detectors [16, 24] that present that feature and descriptors that handle some the challenging conditions. We demonstrate that reusing components validated in WxBS within a bag-of-words image retrieval system produces a robust place recognition system.

The WxBS Retrieval is novel in enriching a BoW approach with the use of multiple detectors, HalfRootSIFT [5] and RootSIFT [2] descriptors with suitable visual vocabularies, view synthesis and adaptive thresholding to compensate for large variations in contrast and richness of features in different conditions.

In the online localization phase, the local features are extracted and assigned to the closest visual words. Then, a shortlist of most similar images is retrieved using the TF-IDF [21] scoring and spatial verification [18]. Using shortlist of neighboring query images location hypotheses are formed via correspondence between a short sequence of query and database images. Finally, the best location hypothesis is verified by WxBS-M matching algorithm.

In the rest of the paper is structured as follows. First, the WxBS-M algorithm is briefly introduced . Next, each of the steps of the proposed WxBS retrieval algorithm for place recognition is presented in detail. Finally, we evaluate the performance of the underlying WxBS algorithm, the retrieval part of the system and overall performance of the WxBS retrieval on the VPRiCE dataset.

2. WxBS-M Matching Algorithm

The proposed system is derived from the WxBS-M [16] two view matching algorithm intended for challenging environmental changes. The algorithm 1 is presented in detail in [15]. For convenience, we shortly explain its most important parts.

The WxBS-M is an iterative algorithm for matching of two images. In each step a specific combination of artificial view synthesis (step 1) and detectors are run on both images (step 2) to extract affine covariant local features. Next, the HalfRootSIFT and RootSIFT descriptors are computed (step 3) and back-projected to the original images. A set of new tentative correspondences is computed using a variant of the nearest neighbor SIFT ratio test [13] called first geometrically inconsistent nearest neighbor test. This helps to deal with duplicate tentative matches generated because of view synthesis (step 5). All tentative correspondences found so far are then verified by a DEGENSAC [7] algorithm, a variant of RANSAC that simultaneously searches for the most consistent model of epipolar geometry and/or dominant plane homography. Finally, correspondences consistent with epipolar geometry are verified by requiring geometric consistency of the affine frames (step 7). All steps of the algorithm are repeated until a preset number of consistent matches is found or until the last iterations is reached, *i.e*. when finding the relation of the two images with further synthesis steps and different features becomes very unlikely.

Algorithm 1 WXBS-M – a matcher for wide multiple baseline stereo

Input: I_1, I_2 – two images; θ_m – minimum required number of matches; $S_{\max}$ – maximum number of iterations.
Output: Fundamental or homography matrix F or H; a list of corresponding local features.

while $(N_{\text{matches}} < \theta_m)$ and (Iter $< S_{\max}$) **do**
  for I_1 and I_2 separately **do**
    1 Generate synthetic views according to the scale-tilt-rotation-detector setup for the Iter.
    2 Detect local features using adaptive threshold.
    3 Extract rotation invariant descriptors with:
      3a RootSIFT and **3b HalfRootSIFT**
    4 Reproject local features to I_1.
  end for
  5 Generate tentative correspondences with 1st geom. inconsistent rule for RootSIFT and HalfRootSIFT
  6 Geometric verification of *all* TC with modified DEGENSAC estimating F or H.
  7 Check geom. consistency of the LAFs with est. F.
end while

3. WxBS Retrieval for Place Recognition

The high level overview of two phases of the WxBS retrieval algorithm is shown in Algorithm 2 and Algorithm 3. In the following sections, we present in detail the most important parts of the WxBS retrieval system which are in bold font in 2 and 3.

Algorithm 2 WxBS retrieval, offline "mapping" phase

1. Extraction of local features from the database images for all iterations of the WxBS-M algorithm.
2. Quantization to a BoW vocabulary.
3. Inverted file formation.

Algorithm 3 WxBS retrieval, online "localization" phase

1. Extraction of local features from all iterations of the WxBs-M algorithm on the query image.
2. Quantization and inverted file traversal with TF-IDF scoring.
3. Approximate location retrieval – fast spatial verification of the TF-IDF shortlist, re-ranking based on the number of geometrically consistent correspondences.
4. Location hypotheses generation – the top ranked images in the shortlist form "seed" hypotheses – short temporal sequences from shortlists of neighboring query results.
5. WxBS-M based verification and **best location selection** by picking the most consistent "seed" hypothesis.

3.1. Extraction of Local Features

The local feature extraction step follows closely the WxBS-M matcher [15]. The Hessian-Affine and MSER detectors are employed as they have been shown to provide a solid base for solving hard matching problems. The local features are detected on a set of affine-warped views generated from (and including) the original image. The process can be viewed as an extension of the isotropic scale pyramid to an anisotropic pyramid, where image is scaled only along one axis.

The view synthesis setup adopted from [15] is the one suggested for matching images with high illumination changes. Experiments with the WxBS dataset showed that in most natural scenes with highly textured objects like trees, leaves etc., if MSER detector fails without the view synthesis, it is highly likely to fail with view synthesis as well. Thus, from 3rd iteration, only Hessian-Affine detector is used. The detector and synthesis configurations used are shown in Table 1. The local features from each view are then reprojected to the original image, forming a single array. All used detectors estimate local shape of the feature up to an unknown orientation. To fix the orientation, the gravity vector assumption [17] commonly used in retrieval and visual localization is used instead of the dominant orientation estimation for the retrieval part.

Table 1: Detector and view synthesis configurations of WxBS-M as applied in the location hypothesis verification and propagation step. Each configuration defines a combination of detector and view synthesis parameters.

Iter.	Detector(s) and view synthesis setup
1	MSER,$\{S\} = \{1; 0.25; 0.125\}$, $\{t\} = \{1\}$, $\Delta\phi = 360°/t$
2	MSER,$\{S\} = \{1; 0.25; 0.125\}$, $\{t\} = \{1; 3; 6; 9\}$, HessAff, $\{S\} = \{1\}$, $\{t\} = \{1\}$, $\Delta\phi = 360°/t$
3	HessAff, $\{S\} = \{1\}$, $\{t\} = \{1; 2; 4; 6; 8\}$, $\Delta\phi = 360°/t$
4	HessAff, $\{S\} = \{1\}$, $\{t\} = \{1; 2; 4; 6; 8\}$, $\Delta\phi = 120°/t$

It is important to note that view synthesis improves performance even for pairs of images with no or negligible difference in viewpoint. Many of the VPRiCE live-to-memory pairs have transformations near to identity and yet many of those have been solved only in the 2nd or 3rd view synthesis iteration. The view synthesis can be viewed as a method to increase the density of detected features which makes the matching process more robust to large changes of various image formation factors.

Adaptive thresholding. One of the main problems in matching of day to night, infrared or multimodal images is the low number of detected features. In recent work of Stylianou *et al*. [22], it has been confirmed that the main source of failures in day-night matching is low number of stable features. The problem is acute in dark low contrast like in infrared or badly illuminated images. To improve the performance, under low contrast conditions, the following enhancements of the baseline detectors were performed.

First, all detector local extrema are considered without thresholding the value and sorted according to the magnitude of the response. If the number of detected features with response magnitude $\geq \Theta$ is greater than given R_{min}, the output is the same as for the baseline detector, else the top R_{min} features are used to populate the list.

Finally, to compensate for the decreasing image area in view synthesis, the threshold is adjusted as $R_{\text{curr}} = R_{\text{min}} \cdot S/t$, where S is the scale factor and t is the simulated tilt of the image (c.f. [16] for the details). The R_{min} thresholds were set experimentally so that the average number of detected Hessian-Affine points and MSER regions on the various types of images was R_{HA}=2000$and R_{\text{MSER}}$=500 respectively. This approach gives better matching performance on low-contrast images than IIDOG [15]

For WxBS retrieval, all features are generated at once both in offline and online phase of the algorithm and stored for later BoW quantization and further location hypotheses verification – unlike in the WxBS algorithm where matching proceeds in iterations until enough inliers to either ho-

mography or fundamental matrix are found, or all iterations are exhausted.

Feature description. For the feature description, RootSIFT [2] has been chosen, a modified version of SIFT [13] descriptor which outperformed SIFT both in the WxBS experiments [15] and image retrieval [2]. To facilitate matching of multimodal images where gradient orientations are preserved at discontinuities up to a reversal we have chosen HalfRootSIFT [5] over InvertedRootSIFT [9] following conclusive results of experiments in [15]. Furthermore in the case of fixed dominant orientation, Inverted-RootSIFT and HalfRootSIFT are equally computationally expensive, Inverted-RootSIFT produces a low number of matches complementary to RootSIFT and cannot handle partial gradient reversal.

3.2. Approximate Location Retrieval

The initial approximate location is estimated by retrieving one or a sequence of images in the large database. A standard bag of words (BoW) specific image retrieval pipeline with spatial verification is used. First a vocabulary is trained on local features extracted from a set of representative images. In our challenging setup, both features sensitive and insensitive to gradient reversal are used and clustered separately using approximate k-Means. A resulting visual vocabulary is then used in an approximate nearest neighbor (ANN) search, and all local features in the map images are assigned to the closest visual word. The output of this offline phase is an inverted file with a list of occurrences of each visual word in each of the "map" images. A set of labels of features in each image with their geometry data is also stored separately for spatial verification.

In the localization, online phase, the local features are extracted for each unseen – query – image, and assigned using ANN search to the closest visual word. Then, the inverted file is sought and collisions of query labels and database images are scored using TF-IDF weights [21]. A shortlist of locations is formed from the top ranking images. Then, a spatial verification is used to fit an affine transformation between the local features in the query and each of the short-listed images as in [18]. Images are re-ranked based on the number of the matching visual words consistent with the affine transformation. For the details of the BoW retrieval system *c.f*. [18].

3.3. Location Hypotheses Formation

The image retrieval phase provides fast localization in a large database of images. The retrieval performance depends on the specificity or distinctiveness of the scene. Naturally, not all scenes in real world scenario satisfy this requirement. Fortunately, an autonomous system is usually collecting a stream of images at a fixed rate, or a set of key frames is produced by a SLAM system while building a local 3D model of the scene. The same assumption also holds for the VPRiCE "map" images, they are obtained in a systematic way and consecutive images correspond to nearby locations.

We exploit these assumptions and instead of locating a single image, we search for a short sequence-to-sequence correspondences, denoted in the following as *seeds*. A *seed* is a mapping of sequence of n query images to n images in the map. The length of the seed n on query side defines the latency of the localization system. The seed on the database side is formed as sequence of consecutive matching images in the top ranks of retrieved shortlists. A simple dynamic programming algorithm is used to obtain (see Alg. 4), and score (Alg. 5) possible hypotheses that correspond to stationary pose, motion in forward or opposite direction as recorded in the map at potentially different speed (see example of used hypotheses in Figure 3)

The scoring of the *seed* also takes into account whether the neighboring images in the "map" database and in the "live" stream belong to a sequence (predicate SEQ or not NOTSEQ. For the map part, this was verified and stored by WxBS-M in the offline phase by pairwise WxBS-M verification of neighboring images. For the live stream, this requires an additional WxBS-M verification with the previous image.

3.4. Best Location Hypothesis Selection

The image retrieval uses BoW representation to efficiently find a shortlist of matching candidates. The matches verified in spatial verification are based only on co-occurring quantized labels of visual words and are thus affected by quantization noise. Additionally, the affine transformation used in the spatial verification might not be appropriate for significant changes of viewpoint. The significantly smaller number of images in seed hypotheses (usually 3-10) allows verification by a more expensive process of iterative WxBS-M pairwise matching [15], Algorithm 1. The view synthesis and local feature generation were already performed for all "map" images and query image, and thus only steps $5 \dots 7$ of Alg. 1 are performed. Note that more reliable descriptor matching and geometric verification by DEGENSAC is carried out.

The high level location verification algorithm is outlined in Alg. 7, if keeps track if the last seed hypothesis verification succeeded for one of the motion models. The WxBS-M matching algorithm is called in Alg. 6. When a motion model still holds, it is sufficient to match any of the pairs in seed mapping. When the previous location was not available, all elements with the best scored seed are verified.

The criterion for the best match is the number of non-duplicate tentative correspondences which are consistent with estimated homography by LO-RANSAC [12]. The lo-

Figure 2: VPRiCE database. Examples of images from the same location.

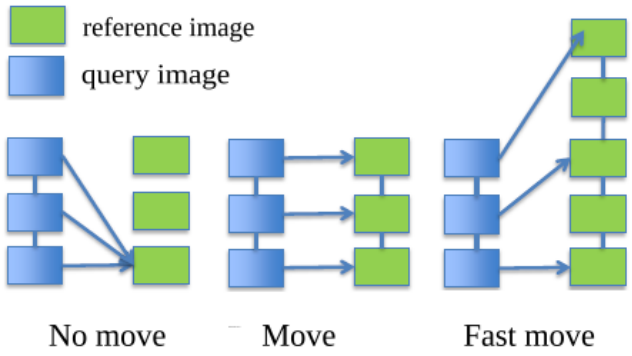

Figure 3: Examples of used adjacency models

cal affine frame consistency check (LAF-check) is applied for elimination of the incorrect correspondences. We use coordinates of the closest and furthest ellipse points from the ellipse center of both matched local affine frames to check whether the whole local feature is consistent with estimated geometry model.

Algorithm 4 GENERATESEEDHYPOTHESES

Input: n_{prev} motion model from previous image or $\varnothing$.
Output: set of seed motion models S_m

$N := ((0,1,2,...),(0,0,1)...)$ – seed "motion" models
if $n_{prev} \neq \varnothing$ **then** $S_m := N\,[\,n_{prev}\,]$
else $S_m := N$

4. Experimental Evaluation

In the experimental validation, two principal components of the system are evaluated: the initial location hypotheses formation using image retrieval and the WxBS retrieval algorithm with verification of location hypotheses.

4.1. VPRiCE dataset

For the location retrieval, the dataset from Visual Place Recognition in Changing Environments (VPRiCE) 2015 challenge [1] comes at hand with a wide range of realistic outdoor sequences. To address specifically each of the

Algorithm 5 SCOREHYPOTHESES

Input: s_m - seed motion model, T_i - shortlists of query images, l_{idx} live image index, m_{idx} memory image index.
Output: sc - seed hypothesis score, s - seed hypothesis mapping live indices to memory indices.

$m := s_m + m_{\text{idx}}$ – compute abs. indices in memory
$l := s_m + l_{\text{idx}}$ – compute abs. indices in live
$sc := \sum m_i \in T_i - \sum m_i \notin T_i$
sc += $\sum$ SEQ$(m_i, m_{i+1}) - \sum$ NOTSEQ(m_i, m_{i+1})
sc += $\sum$ SEQ$(l_i, l_{i+1}) - \sum$ NOTSEQ(l_i, l_{i+1})
$s = l \leftrightarrow m$

Algorithm 6 VERIFYHYPOTHESES

Input: s - seed hypothesis mapping
Output: M - result of verification

function VERIFYHYPOTHESES(S)
  if SeedProp **then** $M := \bigcup_{i=1}^{n}$ WXBSMATCH(l_i,m_i)
  else $M := \bigcap_{i=1}^{n}$ WXBSMATCH(l_i,m_i) **return** M
end function

Algorithm 7 Location verification

Input: Retrieval shortlists T_i for query images I_i, $i \in \{0,1,..\}$,s_{prev} - previous verified seed hypothesis mapping.
Output: s - seed hypothesis mapping

for each $l_{idx} \in$ **Live** sequence **do**
  1. S_m := GENERATESEEDHYPOTHESES(n_{prev})
  2. SC := SCOREHYPOTHESES(S_m,T_i,l_{idx}, m_{idx})
  3. M := VERIFYHYPOTHESES(S), S: $SC(S) > sc_{min}$
  4. if M **then** Store s, n_{prev} := MOTIONMODEL(s), best matching motion model
  else
  5. if $n_{prev} \neq \varnothing$ **then**
  $n_{prev} := \varnothing$; GoTo 1.
  else Store $l_{idx} \leftrightarrow -1$, $n_{prev} := \varnothing$
end for

challenges of changing environments we also use the WxBS dataset proposed in [16].

The VPRiCE 2015 challenge aims at focusing efforts of the visual place recognition community. It consists of two parts *memory* and *live* with 7778 images of outdoor environments under various viewing conditions. The *memory* part of the dataset consists of the images the robot observed during its first visit of the environment. It is the reference, the images from the *live* part of the dataset have to be matched against. The footage has been recorded from trains, cars, buses, bikes, or pedestrians and represents *memory* - offline map and *live* - online localization of an autonomous vehicle or robot. The order of appearance of places is not the same in the *live* and *memory* part of the dataset (the robot takes different routes through the environment). Examples of matching locations in the VPRiCE datasets are shown in Figure 2.

4.2. Image Retrieval Experiment

The following experiment was performed to tune the parameters and performance of the image retrieval system. We have considered images from both the *memory* and *live* parts of the Visual Place Recognition in Changing Environments (VPRiCE) dataset to capture the variability of the challenging environmental changes. In practice, it is also expected that a robot will receive a representative set of images covering targeted environments and operational conditions. The WxBS detection employed Hessian Affine [17] and MSER detectors [14]. RootSIFT (RS) and HalfRootSIFT (HRS) descriptors were computed with up-right assumption. All local features were used to create two visual vocabularies one for RS and one for HRS descriptors. The RS vocabulary (1M visual words) was further split per feature type to light, dark, saddle Hessian points and MSER+ and MSER- proportionally to their average occurrence in the images and each part clustered separately. Light and dark blobs, and MSER+/- were merged together to allow clustering of features with gradient reversals for HRS representation. This split significantly speeds up quantization and has negligible effect on retrieval performance. Finally, an inverted file was built and additional geometry data of each feature were stored for spatial verification. During evaluation, each query image was indexed into the visual word vocabularies and a shortlist of thousand most similar images formed using TF-IDF scoring and inverted file. Each image in shortlist was verified using spatial verification, by finding the best affine transformation between the query and database image. The affine transformations were constrained to those preserving up-right orientation. The number of colliding labels, consistent with affine transformation (inliers) between query and each image in the shortlist, was used to get the final ranking of the shortlisted images.

A ground truth similar to Oxford buildings protocol [18] was created for tuning of the image retrieval system. We have manually labeled three sets of *memory* images for each of the challenging queries from *live* part of VPRiCE dataset: *Good* - correct, closest location image, *Ok* - images nearby correct location with substantial scene overlap, and *Junk* - images from the correct location with minimal (horizon) or low overlap. The 52 query images were selected to proportionally cover the *live* part of the dataset and were further split to groups to see performance on different setups (Car, Train, Campus, Campus IR, Bike).

For the evaluation, a mean average precision measure (mAP) among all the queries was used. The average precision was computed as the area under the precision-recall curve, considering the ranks of *Good* and *Ok* images as positive examples, ignoring *Junk* images and counting all other as negative examples. Results with three different setups: using only HRS labels, only RS labels and aggregating matches from both HRS and RS are shown in Table 2.

Table 2: Image retrieval scores (mAP) on selected sequence from the VPRiCE dataset for three different descriptor choices – RootSIFT (RS), HalfRootSIFT (HRS), both. The mAP is computed using Oxford Buildings style ground truth.

Sequence	HRS	RS	HRS+RS
Bike	0.002	0.002	0.003
Campus-Day	0.947	0.906	0.935
Campus-IR	0.428	0.564	0.600
Car	0.498	0.486	0.478
Train	0.309	0.319	0.390
Total mAP	0.440	0.463	0.504

Table 3: Location recognition results according to the VPRICE protocol for different stages of the proposed method – reference round truth, +/-1 frame error tolerance.

Method	Precision	Recall	F1
BoW HalfRootSIFT	0.530	0.890	0.665
BoW Half&RootSIFT	0.538	1.000	0.700
BoW Half&RootSIFT + MODS + adj. model	0.821	0.825	0.823
Competitors			
MAPIR (CNN) [8]	0.747	0.836	0.789
Bonn (CNN) [11]	0.726	0.758	0.741

4.3. Evaluation on VPRiCE dataset

The main goal of the proposed system is to accurately recognize location of an image or a short sequence of images. To measure the overall performance we follow the

Table 4: Location recognition results according to the VPRICE protocol (with sequenced partial ground truth) for different stages of the proposed method per sequences. $\pm$ 1 stands for taking match as correct, if predicted number differs from ground truth by one

Method	Prec	Rec	Prec ±1	Rec ±1
Train1				
BoW (HRS)	0.477	0.477	0.594	0.594
BoW (HRS, RS)	0.562	0.562	0.682	0.682
BoW, WxBS-M, seed	0.983	0.983	0.984	0.983
Train2-TN				
BoW (HRS)	0.000	1.000	0.000	1.000
BoW (HRS, RS)	0.000	1.000	0.000	1.000
BoW, WxBS-M, seed	0.802	1.000	0.802	1.000
Train3				
BoW (HRS)	0.508	0.508	0.636	0.636
BoW (HRS, RS)	0.564	0.564	0.688	0.688
BoW, WxBS-M, seed	0.712	0.712	0.740	0.740
Train4				
BoW (HRS)	0.512	0.512	0.652	0.652
BoW (HRS, RS)	0.616	0.616	0.744	0.744
BoW, WxBS-M, seed	0.731	0.731	0.915	0.915
Campus-Day				
BoW (HRS)	0.235	0.235	0.560	0.560
BoW (HRS, RS)	0.240	0.240	0.600	0.600
BoW, WxBS-M, seed	0.790	0.790	0.995	0.995
Campus-IR				
BoW (HRS)	0.120	0.120	0.390	0.390
BoW (HRS, RS)	0.145	0.145	0.420	0.420
BoW, WxBS-M, seed	0.140	0.140	0.330	0.330
Car				
BoW (HRS)	0.083	0.085	0.166	0.170
BoW (HRS, RS)	0.092	0.094	0.180	0.185
BoW, WxBS-M, seed	0.201	0.207	0.414	0.426
Bike				
BoW (HRS)	0.017	0.017	0.045	0.045
BoW (HRS, RS)	0.026	0.026	0.062	0.062
BoW, WxBS-M, seed	0.065	0.065	0.072	0.072
TOTAL, Sequence GT				
BoW (HRS)	0.343	0.324	0.458	0.435
BoW (HRS, RS)	0.362	0.343	0.476	0.451
BoW, WxBS-M, seeds	0.623	0.657	0.710	0.761
TOTAL, full reference GT				
BoW (HRS)	0.380	0.853	0.530	0.890
BoW (HRS, RS)	0.403	1.000	0.538	1.000
BoW, WxBS-M, adj.model	0.689	0.798	0.821	0.825
BoW (HRS, RS)	0.403	1.000	0.538	1.000
BoW, WxBS-M, adj.model	0.689	0.798	0.821	0.825

VPRiCE evaluation protocol. The goal is to output for each query image in the *live* part of the dataset the index of the closest location from the *memory* images. Some locations might not be present in the memory and the system should report -1. The performance is assessed in terms of precision, recall and F1 score. The precision was computed as the number of correct answers (positive or negative) out of all queries, recall then as the number of correct positive answers. The overall official results of our method are presented in Table 3. We have evaluated separately the initial location hypotheses generation (labeled *BoW (HRS) and BoW (HRS, RS)*) by simple considering the top ranking image from the retrieval shortlists. The BoW (HRS) method uses only HalfRootSIFT vocabulary, while the BoW (HRS, RS) uses both HalfRootSIFT and RootSIFT features. Finally, the proposed seeds hypotheses and verification method is denoted by *BoW + WxBS-M + seeds* takes the shortlists of the retrieval BoW (HRS,RS) and verifies them with the location verification algorithm (Alg. 7).

To provide deeper insight into performance of the algorithms, we split the *live* dataset into continuous sequences and labeled them by acquisition setups. The official sequenced ground truth was not available at the time of submission, thus we have manually selected the ground truth response for each of the subset *live* images. The setups represent different challenges of the VPRiCE dataset. The results are summarized in Table 4. The WxBS retrieval system reported ground truth location with precision 0.623 and recall 0.657. It is clear that some of the sequences (Train, Campus-Day) are almost solved by the algorithm, while other (Car, Bike) are still too challenging. The seeds based hypothesis verification improves most of the sequences.

The preliminary evaluation protocol of the VPRiCE dataset requires algorithm to report the closest position. Some parts of the dataset were sampled quite coarsely, e.g. in the CAR sequence are images 20-30m apart. The subsequent "live" sequences are of course not perfectly synchronized even when they tend to follow the same frequency of sampling. Additional sources of noise as different viewpoint, slightly different trajectory make it a very hard task even for the human observer to pick the closest position. Additionally, if the algorithm sticks to an "earlier" or "later" frame consistently for some time, it is easy to fall in an off-by-one error w.r.t ground truth. Thus we have decided to report modified recall and precision values in the last two columns (labeled Prec ±1 and Rec. ±1 of Table 4, i.e. the result is counted as matching when it is within one frame difference from the ground truth location. The locations within ±1 frame were found with precision 0.821 and recall 0.825, and within ±5 frames – precision 0.923 and recall 0.841.

5. Conclusions

The WxBS retrieval system for outdoor localization based on large scale image retrieval and WxBS matching has been presented. Its performance was evaluated on the public Visual Place Recognition in Changing Environments

(VPRiCE) dataset achieving ground truth location precision 0.689 and recall 0.798. Locations within ± 1 frame were found with precision 0.821 and recall 0.825.

Acknowledgements

The authors were supported by the Czech Science Foundation Project GACR P103/12/G084, the Technology Agency of the Czech Republic research program TE01020415 (V3C – Visual Computing Competence Center) and by the MSMT LL1303 ERC-CZ grant.

References

[1] The vprice challenge 2015 visual place recognition in changing environments. https://roboticvision.atlassian.net/wiki/pages/viewpage.action?pageId=14188617. Accessed: 2015-04-20.

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[3] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *Computer Vision–ECCV 2012*, pages 517–530. Springer, 2012.

[4] L. Baboud, M. Cadik, E. Eisemann, and H.-P. Seidel. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 41–48. IEEE, 2011.

[5] J. Chen, J. Tian, N. Lee, J. Zheng, R. Smith, and A. Laine. A partial intensity invariant feature descriptor for multimodal retinal image registration. *Biomedical Engineering, IEEE Transactions on*, 57(7):1707–1718, 2010.

[6] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3416–3423. IEEE, 2010.

[7] O. Chum, T. Pajdla, and P. Sturm. The geometric error for homographies. *Computer Vision and Image Understanding*, 97(1):86 – 102, 2005.

[8] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez. Training a Convolutional Neural Network for Appearance-Invariant Place Recognition. *ArXiv e-prints*, May 2015.

[9] J. S. Hare, S. Samangooei, and P. H. Lewis. Efficient clustering and quantisation of sift features: Exploiting characteristics of the sift descriptor and interest region detectors under image inversion. ICMR '11, pages 2:1–2:8, New York, NY, USA, 2011. ACM.

[10] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *Computer Vision–ECCV 2010*, pages 748–761. Springer Berlin Heidelberg, 2010.

[11] J. Knopp, J. Sivic, and T. Pajdla. Lazy sequences matching under substantial appearance changes. In *ICRA 2015 Workshop on Visual Place Recognition in Changing Environments*, 2015.

[12] K. Lebeda, J. Matas, and O. Chum. Fixing the locally optimized ransac. In R. Bowden, J. Collomosse, and K. Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference*, 2012.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extrema regions. In *In British Machine Vision Conference*, pages 384–393, 2002.

[15] D. Mishkin, J. Matas, M. Perdoch, and K. Lenc. WxBS: Wide Baseline Stereo Generalizations. *CoRR*, abs/1504.06603, Apr. 2015.

[16] D. Mishkin, M. Perdoch, and J. Matas. Mods: Fast and robust method for two-view matching. *CoRR*, abs/1503.02619, Mar. 2015.

[17] M. Perd'och, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 9–16, June 2009.

[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[19] D. P. Robertson and R. Cipolla. An image-based system for urban navigation. In *BMVC*, pages 1–10, 2004.

[20] 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*, pages 667–674. IEEE, 2011.

[21] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477 vol.2, 2003.

[22] A. Stylianou, A. Abrams, and R. Pless. Characterizing feature matching performance over long time periods. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 892–898, Jan 2015.

[23] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *CVPR*, 2013.

[24] Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit. Tilde: A temporally invariant learned detector. *arXiv preprint arXiv:1411.4568*, 2014.