WxBS: Wide Baseline Stereo Generalizations

Dmytro Mishkin[1]
ducha.aiki@gmail.com
Jiri Matas[1]
matas@cmp.felk.cvut.cz
Michal Perdoch[2]
perdoch@vision.ee.ethz.ch
Karel Lenc[3]
karel@robots.ox.ac.uk

[1] Center for Machine Perception
Czech Technical University in Prague
Czech Republic

[2] Computer Vision Laboratory
ETH Zurich, Switzerland

[3] Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

We present a generalization of the wide baseline two view matching problem - WxBS, where x stands for a different subset of "wide baselines" in acquisition conditions such as geometry, illumination, sensor and appearance. We introduce a novel dataset of ground-truthed image pairs which include multiple "wide baselines" and show that state-of-the-art matchers fail on almost all image pairs from the set. A novel matching algorithm for addressing the WxBS problem is introduced and we show experimentally that the WxBS-M matcher dominates the state-of-the-art methods both on the new and existing datasets.

1 Introduction

The Wide Baseline Stereo (WBS) matching problem, first formulated by Pritchett and Zisserman [32], has received significant attention in the last 15 years [26, 40]. Progressively more challenging two- and multi-view problems have been successfully handled [40]; for instance, recent algorithms [30], [28] have successfully matched views of planar objects with out-of-plain orientation differences of up to 160 degrees.

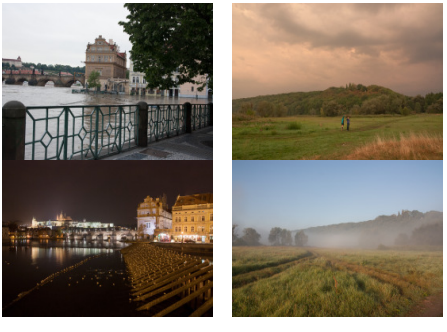


Figure 1: Examples of WxBS problems.

Orientation and viewpoint "baselines" are not the only factors influencing the complexity of establishing geometric correspondence between a pair of images. The standard physical models of image formation and acquisition consider the effects of illumination, the properties of the transparent medium light rays pass through in the scene, surface properties of objects and properties of the imaging sensors.

Pages 12.1-12.12
DOI: https://dx.doi.org/10.5244/C.29.12

In the paper, we consider the generalization of Wide (geometric) Baseline Stereo to WXBS, a two-view image matching problem where two or more of the image formation and acquisition properties significantly change, i.e. they have a *wide baseline*. The "significant change" distinguishes the problem from image registration, where dense correspondence is routinely established between multi-modal images. For registration, various complex transformations have been considered, see Zitová and Flusser [46]. Operationally, the "wide baseline" means "where local, gradient-descent type" methods fail.

The following single wide baseline stereo, or correspondence, problems and their combinations are considered: illumination (WLBS) – difference in position, direction, number, intensity and wavelength of light sources; geometry (WGBS) – difference in camera and object pose, scale and resolution - the "classical" WBS; sensor (WSBS) – change in sensor type: visible, IR, MR; noise, image preprocessing algorithms inside the camera, etc; appearance (WABS) – difference in the object appearance because of time or seasonal changes, occlusions, turbulent air, etc. We denote matching problems, or, equivalently, image pairs, with a significant change in only one of the groups listed as W1BS; if a combination of effects is present, as WXBS. To our knowledge, almost all published image datasets and algorithms are in the W1BS class[26], [30], [42],[3],[17], [18].

We present a new public dataset[1]. with ground truth which combines the above-mentioned challenges and contains both W2BS image pairs including viewpoint and appearance, viewpoint and illumination, viewpoint and sensor, illumination, and appearance change and W3BS – problems where viewpoint, appearance and lighting differ significantly.

We show that state-of-the-art matchers performs poorly on the introduced image pairs, and propose a novel algorithm which significantly outperforms the state-of-the-art without a dramatic loss of speed.

The paper is organised as follows. In Section 2, relevant datasets and matching algorithms are reviewed. The novel WXBS matching algorithm is then introduced in Section 4. The dataset for WXBS problems and the associated evaluation protocol are presented in Section 3. Experimental results are described in Section 5. The paper is concluded in Section 6.

2 Related Work

Viewpoint change. The stereo problem – matching of two images taken from different viewpoints – has always received significant attention of the computer vision community as it is a critical component of the structure from motion task. For images taken concurrently, in both the calibrated and uncalibrated set up, the problem for a narrow baseline is mature [40] and can be now solved in real-time and on a large scale [2].

The standard wide-baseline matching evaluation focuses on the feature detection and description stages[26]. However, the methodology and datasets of [26] are limited to images related by a homography. Attempts have been made to extend the evaluation to 3D scenes [1, 29], but they are significantly less popular. Neither of the above-mentioned protocols evaluates the performance of the matching stage and thus of the full matching pipeline.

As a reference, we adopted two recent algorithms with good reported performance and freely available binaries. The ASIFT method [30] method synthetically transforms images in order to enlarge the matching range of the DoG detector. This idea has been further extended in MODS [27] which incorporates multiple detectors and adopts an iterative approach that

[1] Available at http://cmp.felk.cvut.cz/wbs/index.html

attempts to minimize the matching time. Both algorithms are able to match images with extreme effects induced by viewpoint change. Mishkin *et al*. [27] introduced the extreme-viewpoint dataset that is used to test the ability of the newly proposed WxBS matcher to handle viewpoint changes.

Multimodal image analysis is needed for the alignment of images acquired by different sensors. Most commonly, the problem is encountered in remote sensing and in medical imaging. For instance, in [14], *red-free* and *fluorescein angiographic* images are matched. Similarly, for different modes of magnetic resonance imaging, modality of the captured data depends on the magnetic properties of the scanned chemical compound. In remote sensing, multimodal matching involves *e.g*. registering visual spectrum images against near infrared images (NIR) or Long-Wave infrared (LWIR).

Multimodal registration methods are usually classified as area-based or feature-based methods. Since we are interested in extending the challenges into multiple-baseline variations, area-based methods are omitted as they lack scale invariance [14].

Feature-based approaches [42] and [14] identify the main issues of existing algorithms in the context of multimodal matching as the selection of the the response threshold, i.e. the minimal image contrast which triggers the detector. In [42], the Difference of Gaussian (DoG) [22] response is normalised by local average image intensity in cases when the image contrast is low. Ghassabi *et al*. [14] present a variant of the DoG detector which sets a local response threshold for each image cell on the basis of the image entropy. In [9], it is argued that Harris detector is more suitable for this task as the information along boundaries is preserved in cases of different image modalities.

The main issue of the widely used SIFT descriptor [22] in the context of multimodal images is the lack of invariance to gradient reversal. Two approaches to address this issue have been proposed in the literature. The first generates a second SIFT descriptor of the feature for a gradient reversed image by SIFT vector reordering [15]. We refer to this method as inverted-SIFT. The second method [9], denoted as half-SIFT, limits local image gradients directions to $\langle 0, \pi \rangle$ by treating opposite gradient directions as identical. Unlike the inverted-SIFT, this method allows matching of images that are only partially inverted (per patch), *i.e.*. some gradient directions stay the same while other are reversed. The downside is the reduction of the descriptor discriminability.

The computation of inverted-SIFT has a negligible computational cost, as it can be generated from SIFT descriptors by rearranging the data in the gradient histogram. The only associated computational cost is in the matching since twice as many features are matched in the second image. For the half-SIFT method, the feature patch and its descriptor has to be extracted since the dominant feature orientation differs from SIFT's dominant orientation.

An example of a multimodal image registration dataset is presented in [3]. This dataset consist of 100 pairs of vertically aligned images from a camera and a LWIR thermal sensor. Viewpoint changes between related image pairs are negligible.

Change in object illumination and appearance. Techniques similar to those developed for multimodal image matching can be used for matching of images of differently illuminated objects. In [20], the authors employ half-SIFT and further modify SIFT descriptor in such a way that it collects only gradients located on edges. Yang *et al*. [43] use the Difference of Gaussian features and SIFT to estimate the transformation between the images. If no matches are found, an identity transformation is assumed. From a single local match, multiscale features together with local image statistics are used in an iterative procedure called Dual-Bootstrap to enlarge the region of good alignment.

Hauagge *et al*. [17] argue that local symmetries survive significant illumination changes

and developed a higher-level feature detector for matching of urban scenes where symmetries are abundant. They also assume that the vertical direction is aligned with one of the edges of the image. The method [17] is able to match images of architectural objects taken many years apart and even to match sketches to photos. Their dataset contains 46 pairs of images.

Matching of images depicting very different appearance of the same object arise in computer vision applications. A system for guided drawing of free-form objects called ShadowDraw is presented in [21]. It can be seen as a large-scale image retrieval system which interactively tries to look for images based on sketches given by a user. In the object classification field, the multiple-appearance problem has been investigated in [36] who train a data-driven visual similarity measure in order to match images to sketches or paintings. Those two approaches use global image description rather than local image feature matching.

3 Datasets

Datasets used in experiments are listed in Table 1. When evaluating detectors (Section 5) and the proposed matching algorithm (Section 4) all dataset images are used. However, descriptor evaluation is performed only on a subset of the most challenging pairs (i.e. only pairs 1-6 from OxfordAffine).

Most of the published datasets (with exception of the LostInPast dataset [17]) include only a single nuisance factor per image pair. This is suitable for evaluation of the robustness to a particular nuisance factor but fails to predict performance in more complex environments. One of the motivations of the proposed WxBS datasets is to address this issue.

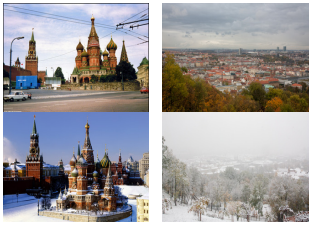
Table 1: Datasets used for evaluation

Short name	Proposed by	#images	Type
GDB	Kelman *et al.* [20], 2007	22 pairs	WLBS, WSBS
SymB	Hauagge and Snavely [13], 2012	46 pairs	WABS, WLBS
MMS	Aguilera *et al.* [1], 2012	100 pairs	WSBS
EVD	Mishkin *et al.* [27], 2013	15 pairs	WGBS
OxAff	Mikolajczyk *et al.* [26], [24], 2013	8 sextuplets	WGBS
EF	Zitnick and Ramnath *et al.* [35],2011	8 sextuplets	WGBS,WLBS
Amos	Jacobs *et al.* [15],2007	$>$ 100K	WLBS,WABS
VPRiCE	VPRICE Challenge 2015 [28]	3K pairs	WGABS, WGLBS,WGSBS,
Past	Fernando *et al.* [10], 2014	502 images	WGABS
WxBS	here	37 pairs	WABS,WGABS,WGLBS, WGSBS,WLABS,WGALBS

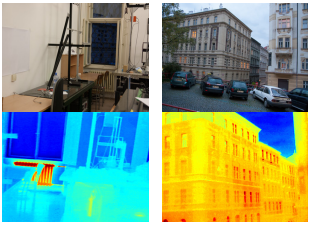
WxBS dataset and evaluation protocol. A set of 37 image pairs has been collected from Flickr and other sources. The dataset is divided into 6 categories based on the combinations of nuisance factor present, see Table 2. For every image, a set of approximately 20 ground-truth correspondences has been annotated. Selected examples are presented in Figure 2. The resolution of the majority of the images is 800×600 with the exception of LWIR images from the WGSBS dataset which were captured by a thermal camera with a resolution of 250×250 pixels. The selected image pairs contain both urban and natural scenes.

Table 2: The WxBS datasets categories

Short name	Nuisance	#images	Avg. # GT Corr.
MAP2PH	appearance (map to photo)	6 pairs	homography provided
WGABS	viewpoint, appearance	5 pairs	22 per img.
WGLBS	viewpoint, lighting	9 pairs	21 per img.
WGSBS	viewpoint, modality	5 pairs	18 per img.
WLABS	lighting, appearance	4 pairs	25 per img.
WGALBS	viewpoint, appearance, lighting	8 pairs	17 per img.

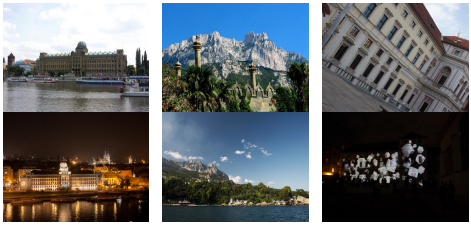


a) WGABS (5 pairs)



b) WGSBS (5 pairs)

c) WLABS (4 pairs)



d) WGLBS (9 pairs)

e) WGALBS (8 pairs)

Figure 2: Examples of image pairs from the WXBS dataset.

Ground truth and the evaluation protocol. In the image registration tasks, it is often sufficient to define ground truth as a homography between an image pair. However, the WxBS dataset contains significant viewpoint changes. In the case of a non-planar scene a homography can, at best, cover the dominant plane.

We assume that an ideal algorithm matches the majority of the scene content, thus our ground truth is a set of manually selected correspondences which evenly cover the part of the scene visible in both images. The average number of correspondences per image pair is shown in Table 2.

The evaluation protocol for the WxBS dataset. For each image pair indexed with $i \in \mathbb{Z}$ we have manually annotated a set of correspondences $(\mathbf{u},\mathbf{v}) \in C_i$ where $\mathbf{u}$ and $\mathbf{v}$ are coordinates of manually picked positions in the 1st and the 2nd image respectively. For epipolar geometry we use *symmetric epipolar distance* $e(\mathbf{F}_i,\mathbf{u},\mathbf{v})$ and for homography he *symmetric reprojection error* $e(\mathbf{H}_i,\mathbf{u},\mathbf{v})$[16]. The choice is based on preliminary experiments where Sampson error was prone to quasi-degenerate solutions in presence of "one-to-many" correspondences.

The recall on ground truth correspondences C_i of image pair i and for geometry model $\mathbf{M}_i$ is computed as a function of a threshold θ

$$\mathrm{r}_{i,\mathbf{M}_i}(\theta) = \frac{|\{(\mathbf{u},\mathbf{v}) : (\mathbf{u},\mathbf{v}) \in C_i, e(\mathbf{M}_i,\mathbf{u},\mathbf{v}) < \theta\}|}{|C_i|} \tag{1}$$

using appropriate error functions $e(\mathbf{F}_i,\mathbf{u},\mathbf{v})$ or $e(\mathbf{H}_i,\mathbf{u},\mathbf{v})$. For all pairs of each category W we define an overall recall per category as:

$$\mathrm{r}_W(\theta) = \frac{\sum_{i \in W} \mathrm{r}_{i,M_i}(\theta)}{|W|}, \tag{2}$$

the fraction of the ground truth annotated correspondences which are consistent with the output of wide baseline stereo algorithm for a given threshold in a nuisance category.

4 Algorithm for Wide Multiple Baseline Stereo Matching

Algorithm 1 MODS-WxBS – 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 rSIFT** and **3b hrSIFT**
    4 Reproject local features to I_1.
  end for
  5 Generate tent. corresp. based on the first geom. inconsistent rule for rSIFT and hrSIFT separately using kD-tree
  6 Filter duplicates
  7 Geometric verification of *all* TC with modified DEGENSAC [10] estimating F or H.
  8 Check geom. consistency of the LAFs with est. F or H.
end while

This section introduces WxBS-MODS (WxBS-M), a variant of MODS [27, 28], a matcher designed for WxBS problems. Its structure is presented in Algorithm 1. The basic idea is to repeat the matching steps – synthesize artificial views of given images, detect and describe local features, match and geometrically verify them – until a reliable geometrical model is found.

MODS was chosen as the core of the proposed WxBS algorithms since its showed state-of-the-art performance on two-view matching problems with extreme change in geometry [28].

The differences between MODS and WxBS-M are: (1) the detection threshold for all detectors is set adaptively, as detailed below, (2) the use of multiple descriptors – RootSIFT [6] (rSIFT) and HalfRootSIFT [20] (hrSIFT, gradient orientation $\in [0;\pi)$), and (3) descriptors from different detectors as well as for different descriptors are placed in separate kD-trees.

Descriptor selection for WxBS-M. We considered the following descriptors for the WxBS-M matcher: SIFT [22], rSIFT [6], hrSIFT (gradients in interval $[0;\pi)$) [20], InvSIFT (SIFT with reordered histogram bins as for inverted image) [15], LIOP[42], AKAZE [5], MROGH [11], FREAK [4], ORB [33], SymFeat [17], SSIM [35] (implementation [8]), DAISY [39] and – as a baseline – L_2-normalized raw grayscale pixel intensities.

The evaluation protocol was as follows. The dataset for descriptors evaluation consisted of 40 image pairs from datasets listed in Table 1 divided into 5 parts according to the nuisance factors: geometry, appearance, illumination, sensor, map versus photo (as showed in Figure 3).

For each reference image of the pair, local affine-covariant features were detected by Hessian-Affine, MSER and FOCI. Multiple detectors were used in order to minimize the selection bias towards a specific detector.

The affine-covariant regions were assinged dominant orientation and then reprojected to the second image by the ground truth homography. All the images are either without significant relative depth or taken from virtually identical viewpoints, so homography is the appropriate two-view relationship. Regions not visible in the second image were discarded.The geometric repeatability of such regions is by construction 100% and the maximum possible recall is 1.

Color-to-grayscale image transformation have been done via channel averaging, which gives best matching performance [19].Then affine regions were normalized to patch size 41x41 (scale $\sigma = 3\sqrt{3}$) and described with given descriptors. An affine-normalization procedure is performed even for the fast binary descriptors, which is rarely used because of the significant additional processing time. However, the goal of the experiment is to explore

descriptor performance in challenging conditions, not their speed. The procedure helps – the typical threshold of the Hamming distance for binary descriptors on unnormalized patch is around 60-80, while on affine normalized patches similar performance is obtained with a threshold around 10-30. All descriptors benefit from the affine normalization procedure, e.g. the graffiti 1-6 pair from the OxAff dataset could be matched with FREAK descriptor only when using a normalized patch.

Floating point descriptors were compared using L_2 distance, binary using Hamming distance. The Recall-Precision curves are shown in Figure 3. The second-nearest distance ratio is used to parameterize the curve for floating point descriptors, the Hamming distance for binary ones.

The results shows that gradient-histogram based SIFT and its variants including DAISY are the best performing descriptors by a big margin in the presence of any (geometric, illumination, etc) nuisance factors despite the fact that some of the competitors – LIOP, MROGH – have been specifically designed to deal with illumination changes. The second best descriptor is, surprisingly, the patch with contrast-L_2-normalized pixels beating all other descriptors. It ha as huge memory footprint – 1681 floats, but the affine-photo-L_2-normed grayscale pixel intensities are a strong descriptor baseline.

Note that most of the descriptors gain significantly from photometric normalization, cf. the first two columns of Figure 3. The published implementations are clearly sensitive to contrast variations.

Most of descriptors, despite their different underlying assumptions and algorithmic structure, successfully match almost the same patches (see third column in Figure 3) – and the most complementary descriptor to the leading rSIFT is its gradient-reversal-insensitive version hrSIFT.

The most difficult nuisance factor for tested descriptors is modality – the best recall for images acquired in different ways (infrared, visible and map) is 0.06..0.12 – even for the "ideal" detector. Descriptors are much more robust to geometry and illumination differences, yet not able to successfully match even half of detected regions.

The results confirming the domination of SIFT-based methods are in agreement with [37] and [12] despite the fact that they adopted a rather different evaluation methodology. However, we could not confirm clear superiority of the SSIM over SymFeat descriptors, which could be explained by the fact that the SSIM descriptor was designed for use only with the SSIM detector.

Adaptive threshold of the detector response. One of the main problems in matching of day to night and infrared images is the low number of detected features. The problem is acute in dark low contrast images in the WGSBS and MMS [3] datasets. A possible approach addressing the problem is iiDoG [27] where the difference of Gaussians is normalized by sum of Gaussians. It works well, but cannot be easily applied for other types of detectors, i.e. MSER.

Instead, we propose to use the following adaptive thresholding for all feature detectors. First, all local extrema of the response function are detected, i.e. no thresholding takes place. Next, the detected features are sorted according to the response magnitude. If the number of detected features with response magnitude $\geq \Theta$ is greater than a given threshold $R_{\min}$, these are output and the algorithm terminates (this is the standard approach). If there is not enough features above the threshold, take top $R_{\min}$ features for output.

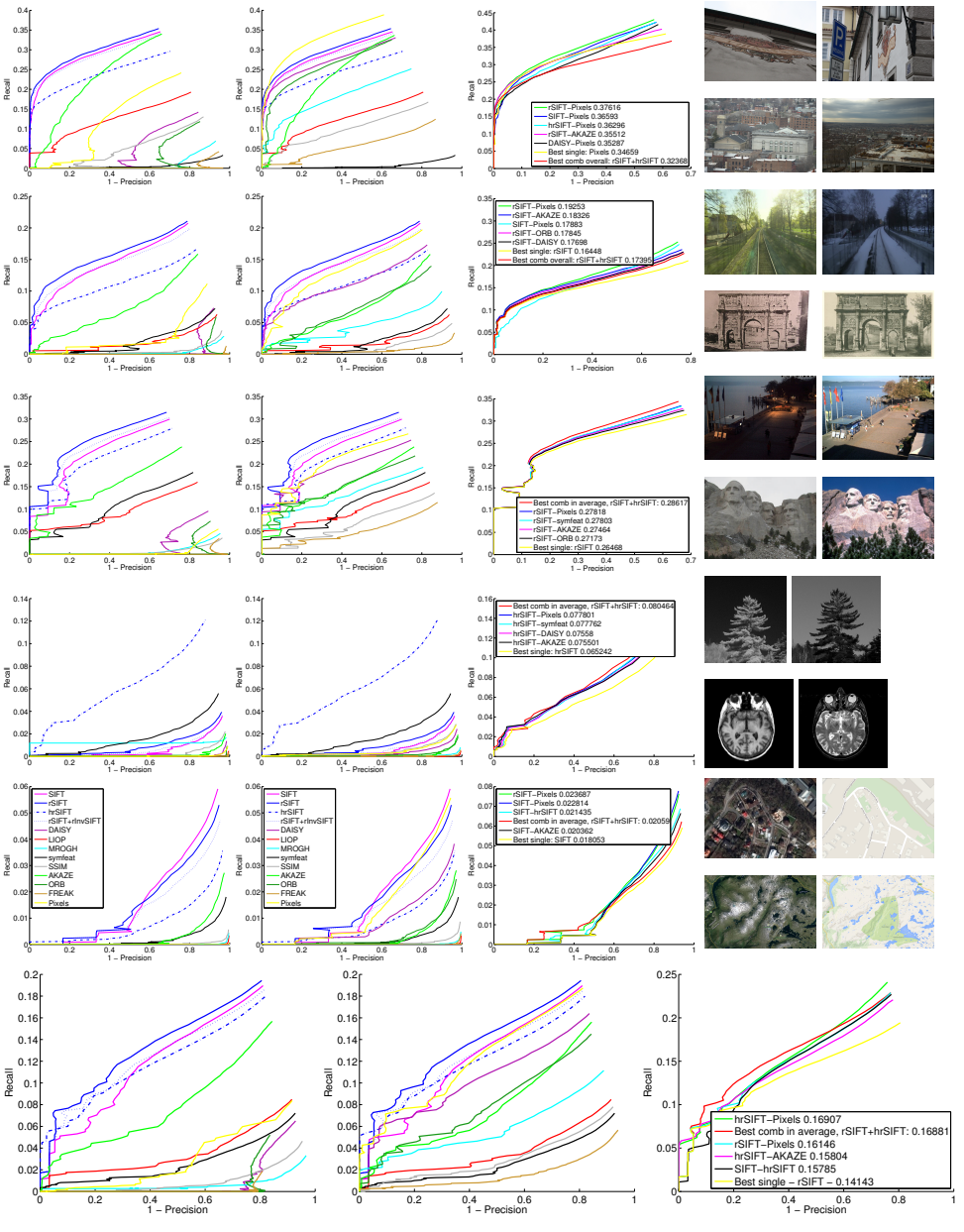


Figure 3: First column: descriptors computed using authors' implementation, second column - descriptors computed on photometrically normalized patches (mean = 0.5, var = 0.2) patches as done in SIFT. Third column: top 5 complementary pairs of descriptors (photometrically normalized). Forth column: examples of images with current nuisance factor, rows: WGBS,WABS,WLBS,WSBS,MAP2PH, ALL. The numbers in legend are mean average precision. Note that axis scales differs in each row, i.e. for different WxBS problems.

Table 3: Matching algorithm comparison. The number of matched image pairs (left) and the average running time (right). The FOCI detector is run through MS Windows simulator wine, the time includes a big overhead.

Alg.	EF		EVD		MMS		WGABS		WGALBS		WGLBS		WGSBS		WLABS		Past		OxAff		SymB		GDB	
	#	time	#	time	#	time	#	time	#	time	#	time	#	time	#	time	#	time	#	time	#	time	#	time
	33	[s]	15	[s]	100	[s]	5	[s]	8	[s]	9	[s]	5	[s]	4	[s]	172	[s]	40	[s]	46	[s]	22	[s]
Threshold adaptation																								
MSER	16	1.4	3	1.4	1	0.3	0	2.0	0	1.3	0	1.3	0	0.8	1	1.2	8	1.3	40	3.5	23	2.4	9	2.4
AdMSER	25	3.4	8	4.0	6	1.0	0	4.0	0	3.2	0	3.3	0	1.4	1	2.6	11	2.9	40	5.7	26	4.6	13	6.9
DoG	29	2.3	0	2.8	10	0.8	0	2.7	0	2.3	0	2.1	0	1.0	1	2.4	13	2.0	38	4.8	29	2.7	12	4.7
iiDoG	29	3.1	0	3.0	11	1.2	0	3.2	0	2.9	0	2.8	0	1.2	1	2.5	13	2.2	38	8.0	29	2.9	12	6.1
AdDoG	29	2.6	0	3.4	11	1.2	0	3.3	0	3.0	0	3.0	0	1.5	1	2.7	13	2.7	38	4.1	30	3.0	12	4.8
HesAf	32	4.6	1	5.2	15	1.2	0	5.5	0	3.8	0	4.2	0	2.0	1	3.6	24	4.0	40	11	35	5.8	17	9.1
AdHesAf	33	5.7	2	7.6	35	2.9	0	7.2	1	6.5	0	6.0	0	3.2	1	4.9	25	5.4	40	10	35	7.2	18	13
Other detectors																								
WαSH	0	1.8	0	5.4	0	0.6	0	2.8	0	2.5	0	1.4	0	1.8	0	1.2	0	1.9	24	4.1	3	2.8	3	6.9
ORB	3	4.1	0	3.6	1	0.8	0	2.8	0	2.7	0	3.6	0	1.6	0	2.8	1	2.3	28	8.7	5	3.0	3	6.1
SURF	27	2.3	0	2.4	7	1.0	0	2.5	0	1.9	0	2.1	0	0.9	1	1.4	10	1.9	38	5.8	31	2.9	15	4.0
AKAZE	28	4.3	0	3.6	10	0.8	1	4.7	0	3.4	0	4.0	0	1.3	1	2.7	25	3.6	38	13	35	5.6	17	6.4
FOCI	29	12	0	39	14	11	1	32	0	29	0	29	0	20	1	29	21	13	38	35	35	27	17	45
SFOP	25	11	0	16	12	4.7	0	12	0	10	0	10	0	9.2	0	7.5	11	12	36	15	24	11	8	17
WADE	16	14	0	20	0	3.4	0	58	0	11	0	14	0	7.9	1	8.3	20	23z	34	60	34	46	13	77
State-of-art matchers																								
ASIFT	23	27	5	12	18	3.2	0	52	0	32	0	35	0	12	1	30	62	32	40	102	27	14	15	41
MODS	33	4.8	15	11	27	11	2	41	2	31	1	46	0	17	1	26	94	27	40	3.4	42	18	18	11
DBstrap	31	26	0	18	79	9.3	0	11	0	13	0	13	0	4.7	0	15	16	28	36	24	38	21	16	17
Proposed matcher																								
WXBS-M	33	4.7	15	14	82	12	3	40	3	63	3	61	0	26	3	28	107	42	40	5.1	43	18	22	12

5 Experiments

The performance of the proposed WxBS-M matcher was compared with the state-of-art matchers: ASIFT [30], Dual Bootstrap (DBstrap) [43], and MODS [28] and with a number of algorithms that use only a single feature: MSER [23], DoG [22], Hessian-Affine [25] (implementation [31]), FOCI [45], IIDOG [42], WADE [34], WαSH [41], SURF [7], SFOP [13], AKAZE[5]. For single features algorithms, matching is done as in Algorithm 1 except that no view synthesis is performed; both rSIFT and hrSIFT descriptors are used.

We focus on getting a reliable answer to the "match/non-match" question for challenging image pairs. Therefore performance is measured by the number of successfully matched pairs. Image pairs are considered matched if $\geq$15 correct inliers to a homography are found. Since the Lost-in-the-Past dataset contains 2300 matchable image pairs, which is unfeasible for all-pairs matching, we selected a subset of 172 medium-difficulty image pairs. Other datasets (see Table 1) are used fully.

The results are summarized in Table 3. The classical datasets (OxAff, EF, etc.) consider only geometrical or illumination differences and are no more challenging. The sensor IR vs. visible change (MMS) is the most challenging nuisance factor. Even if it is the only one present, most matchers success rates are below 10%, with the exception of AdHesAf (35%), DBstrap (79%) and WxBS-M (82%).

The state-of-the-art matchers were able to match almost no image pairs with more nuisance factors. The proposed WxBS-M shows much better performance, but still is unable to solve more than half of the new WxBS dataset pairs.

Results in Table 3 confirm that the proposed adaptive thresholding strategy works as well as, or even better, than iiDoG for DoG, but it is 1.5 times faster (versions of detectors with adaptive thresholding are denoted with "Ad" prefix). It also significantly improves results of the MSER and Hessian-Affine, even when the most prominent nuisance is the viewing

geometry (on the EVD dataset).

Comparing single detectors, (adaptive) Hessian-Affine still shows best performance among more recent detectors.

6 Conclusions

We have presented a new vision problem – the wide multiple baseline stereo (WxBS) – which considers matching of images that simultaneously differ in more than one image acquisition factor such as viewpoint, illumination, sensor type or where object appearance changes significantly, e.g. over time. A new dataset with the ground truth for evaluation of matching algorithms has been introduced and made public.

We have extensively tested a large set of popular and recent detectors and descriptors and show than the combination of RootSIFT and HalfRootSIFT as descriptors with MSER and Hessian-Affine detectors works best for many different nuisance factors. We show that simple adaptive thresholding improves Hessian-Affine, DoG, MSER (and possibly other) detectors and allows to use them on infrared and low contrast images.

A novel matching algorithm for addressing the WxBS problem has been introduced. We have shown experimentally that the WxBS-M matcher dominates the state-of-the-art methods both on both the new and existing datasets.

7 Acknowledgements

The authors were supported by The Czech Science Foundation Project GACR P103/12/G084, ERC project VarCity (#273940), and by CTU student grant SGS15/155/OHK3/2T/13.

References

[1] Henrik Aanæs, Anders Lindbjerg Dahl, and Kim Steenstrup Pedersen. Interesting interest points. *IJCV 2012*, pages 18–35, 2012.

[2] S. Agarwal, N. Snavely, I. Simon, S.M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV 2009*, pages 72–79, 2009.

[3] Cristhian Aguilera, Fernando Barrera, Felipe Lumbreras, Angel D Sappa, and Ricardo Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, 2012.

[4] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. FREAK: Fast Retina Keypoint. In *CVPR 2012*, 2012.

[5] P. F. Alcantarilla, J. Nuevo, and A. Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *BMVC 2013*, 2013.

[6] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR 2012*, 2012.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV 2006*, pages 404–417, 2006.

[8] K. Chatfield, J. Philbin, and A. Zisserman. Efficient retrieval of deformable shape classes using local self-similarities. In *Workshop on Non-rigid Shape Analysis and Deformable Image Alignment, ICCV*, 2009.

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

[10] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR 2005*, pages 772–779, 2005.

[11] Bin Fan, Fuchao Wu, and Zhanyi Hu. Rotationally invariant descriptors using intensity order pooling. *PAMI 2012*, 34(10):2031–2045, 2012.

[12] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Lost in the past: Recognizing locations over large time lags. *CoRR*, abs/1409.7556, 2014.

[13] W. Förstner, T. Dickscheid, and F. Schindler. Detecting interpretable and accurate scale-invariant keypoints. In *12th IEEE International Conference on Computer Vision (ICCV'09)*, Kyoto, Japan, 2009.

[14] Zeinab Ghassabi, Jamshid Shanbehzadeh, Amin Sedaghat, and Emad Fatemizadeh. An efficient approach for robust multimodal retinal image registration based on ur-sift features and piifd descriptors. *EURASIP Journal on Image and Video Processing*, (1):1–16, 2013.

[15] Jonathon S. Hare, Sina Samangooei, and Paul H. Lewis. Efficient clustering and quantisation of sift features: Exploiting characteristics of the sift descriptor and interest region detectors under image inversion. In *ICMR 2011*, pages 2:1–2:8. ACM, 2011.

[16] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[17] D.C. Hauagge and N. Snavely. Image matching using local symmetry features. In *CVPR 2012*, pages 206–213, 2012.

[18] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent Temporal Variations in Many Outdoor Scenes. In *CVPR 2007*, 2007.

[19] Christopher Kanan and Garrison W. Cottrell. Color-to-grayscale: Does the method matter in image recognition? *PLoS ONE*, 2012.

[20] Avi Kelman, Michal Sofka, and Charles V Stewart. Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations. In *CVPR 2007*, 2007.

[21] Yong Jae Lee, C Lawrence Zitnick, and Michael F Cohen. Shadowdraw: real-time user guidance for freehand drawing. In *ACM Transactions on Graphics (TOG)*, volume 30, page 27. ACM, 2011.

[22] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV 2004*, 60(2): 91–110, 2004.

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

[24] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI 2005*, 27 (10):1615–1630, 2005.

[25] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *IJCV 2004*, 60(1):63–86, 2004.

[26] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *IJCV 2005*, 65(1-2):43–72, 2005.

[27] Dmytro Mishkin, Michal Perdoch, and Jiri Matas. Two-view matching with view synthesis revisited. In *IVCNZ 2013*, pages 436–441, 2013.

[28] Dmytro Mishkin, Michal Perdoch, and Jiri Matas. Mods: Fast and robust method for two-view matching. *CoRR*, abs/1503.02619, 2015.

[29] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. In *CVPR 2005*, volume 1, pages 800–807 Vol. 1, 2005.

[30] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[31] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR 2009*, pages 9–16, 2009.

[32] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV 1998*, pages 754–760, 1998.

[33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV 2011*, pages 2564–2571, 2011.

[34] Samuele Salti, Alessandro Lanza, and Luigi Di Stefano. Keypoints from symmetries by wave propagation. In *CVPR 2013*, pages 2898–2905, 2013.

[35] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR 2007*, 2007.

[36] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Data-driven visual similarity for cross-domain image matching. In *ACM Transactions on Graphics (TOG)*, volume 30, page 154. ACM, 2011.

[37] A. Stylianou, A. Abrams, and R. Pless. Characterizing feature matching performance over long time periods. In *WACV 2015*, pages 892–898, 2015.

[38] Niko Suenderhauf and Arren Glover. The vprice challenge 2015: Visual place recognition in changing environments, 2015. URL https://roboticvision.atlassian.net/wiki/pages/viewpage.action?pageId=14188617.

[39] E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. *PAMI 2010*, 32(5):815–830, 2010.

[40] Tinne Tuytelaars and Krystian Mikolajczyk. *Local Invariant Feature Detectors: A Survey*. Now Publishers Inc., 2008.

[41] C. Varytimidis, K. Rapantzikos, and Y. Avrithis. Wash: Weighted α-shapes for local feature detection. In *ECCV 2012*, 2012.

[42] Vasillios Vonikakis, Dimitrios Chrysostomou, Rigas Kouskouridas, and Antonios Gasteratos. A biologically inspired scale-space for illumination invariant feature detection. *Measurement Science and Technology*, 24(7), 2013.

[43] Gehua Yang, Charles V Stewart, Michal Sofka, and Chia-Ling Tsai. Registration of challenging image pairs: Initialization, estimation, and decision. *PAMI 2007*, 29(11):1973–1989, 2007.

[44] Bin Fan Zhenhua Wang and Fuchao Wu. Local intensity order pattern for feature description. In *ICCV 2011*, pages 603–610, 2011.

[45] C. Lawrence Zitnick and Krishnan Ramnath. Edge foci interest points. In *ICCV 2011*, pages 359–366, 2011.

[46] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003.