Abstract

We propose a generalization of the wide baseline two view matching problem - WXBS, where X stands for a different subset of multiple "wide baselines" such as geometric, illumination, sensor and appearance. We introduce a public dataset with a ground truth which combines different types of nuisance factors that influence image matching 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 new and existing datasets.

1. Introduction

The Wide Baseline Stereo (WBS) matching problem, first formulated by Pritchett and Zisserman [22], has received significant attention in the last 15 years [16, 24]. Progressively more challenging two- and multi-view problems have been successfully handled [24] and recent algorithms [20], [17] have shown impressive performance, e.g. matching views of planar objects with orientation difference of up to 160 degrees.

Besides the orientation and viewpoint baseline, other factors influence the complexity of establishing geometric correspondence between a pair of images. The standard physical models of image formation and acquisition consider, beside geometry, the effects of illumination, the properties of the transparent medium light rays pass through in the scene, the surface properties of objects and the properties of the imaging sensors.

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 multimodal images and various complex transformations have been considered, see Žitová and Flusser [27]. 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[16], [20], [25],[3],[10].

We present a new public dataset with a 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 matching 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 3. The dataset for WxBS problems and the associated evaluation protocol are presented in Section 4. Experimental Results are described in Section 5. The paper is concluded in Section 6.

2. Related Work

2.1. 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[24] and can be now solved in real-time and on a large scale[2].

For wide-baseline matching, the standard evaluation protocol focuses on the feature detection and description stages[16]. However, the methodology and datasets of [16] are limited to images related by a homography. Attempts have been made to extend the evaluation to 3D scenes[11][19], 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 which reported good performance and whose binaries are freely available. The ASIFT method[20] method synthetically transforms images in order to improve the range of affine transformations of the DoG detector. This idea have been further extended in MODS[17] which incorporates multiple detectors and adopts an iterative approach that attempts to minimize the matching time. Both algorithms are able to match images with extreme viewpoint changes. Mishkin et al.[17] introduced an extreme-viewpoint dataset that is used to test the ability of the newly proposed WxBS matcher to handle viewpoint changes.

2.2. Multimodal Image Analysis

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[7], 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 divided to area-based and feature-based methods. As we are interested in extending the challenges into multiple-baseline variations, area-based methods are omitted as they lack scale invariance[7].

Feature-based approaches[25] and[7] identify the main issues of existing algorithms in the context of multimodal matching as the selection of the response threshold, i.e. the minimal image contrast which triggers the detector. In[25], the Difference of Gaussian (DoG) response, used in a DoG detector[14], is normalised by local average image intensity in cases when the image contrast is low. Ghassabi et al.[7] 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 [5], 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[14] 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[8]. We refer to this method as inverted-SIFT. The second method[5], denoted as half-SIFT, limits local image gradients directions to $(0, \pi)$ by merging opposite gradient directions in orientation estimation. 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 to 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 as the dominant feature orientation differs
from SIFT’s dominant orientation.

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

2.3. 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 [11], 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. [26] 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. A data presented in [11] are used in Section 5.

Hauagge et al. [10] 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 proposed in [10] is able to match images of architectural objects taken many years apart and even sketches to photos. The dataset introduced in the paper 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 Shadow-Draw is presented in [13]. 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 [23] 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. Matching algorithm for wide multiple baseline stereo

In this section, we propose a WxBS-M matcher designed to perform well on WxBS problems. It includes algorithmic parts of methods that have been shown effective on certain classes of two-view image matching problems and also introduces novel components. The core of the WxBS-M matcher is similar to the MODS [17] view synthesis framework. Its overall structure is shown in Algorithm 1 and each step discussed in the following paragraphs.

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

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

```plaintext
while ($N_{\text{matches}} < \theta_m$) and (Iter $< S_{\text{max}}$) 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 tent. corresp. based on the first geom.
  6 Filter duplicates
  7 Geometric verification of all TC with modified DEGENSAC estimating $F$ or $H$.
  8 Check geom. consistency of the LAFs with est. $F$. end while
```

**View synthesis.** This step is adopted from the MODS framework [17], but the view synthesis setup for the Hessian-Affine feature detector is different. We have observed that in most natural scenes with highly textured objects as trees, leaves etc., if MSER fails without the view synthesis, it is highly likely to fail with the view synthesis as well. Views for Hessian-Affine detector are therefore generated starting from the 2nd iteration. It is important to note that view synthesis improves performance even when images have no or negligible difference in viewpoint. Most image pairs from the GDB-ICP, SymBench and MMS datasets (see Tables 5, 6) have transformation near to identity and yet many of the problems have been solved only in the 3rd or 4th 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 in other image formation factors.

**Feature detection with an adaptive threshold.** One of the main problems in matching of day-night and infrared images is the low number of detected features. The problem is acute in dark low contrast images as in WxBS and MMS [3] datasets. We first considered the illumination invariant Difference of Gaussians approach (IiDoG) of [25] since it was claimed to perform well on problems involving underexposed images. The method is easily generalised to other interest points detectors. However, we found it problematic.
Symmetrical error check. The widely used LO-RANSAC [12] and DEGENSAC [6] minimise the Sampson error [9] which we have experimentally observed to be prone to failure when view synthesis is used [28]. Near-degenerate homography and fundamental matrix are often output due to multiple near duplicate matches that affect the numerical stability of the model estimation.

The use of the standard degeneracy test [9] for homography does not solve the problem either. The test verifies if rank\( \mathbf{H} \) = 3 which is in practice approximated by testing the following condition

\[
\left| \frac{\det(\mathbf{H})}{h_{33}^2} \right| > 0.1. \tag{1}
\]

Figure 3 illustrates an example of such failure, when a number of similar scene parts are matched to multiple observations of the same part with small localization noise introduced in the view synthesis. The use of symmetric re-projection errors \( \epsilon_{\text{SymH}} \) (see 7) or epipolar (see 8) avoids the near-degenerate cases, however the experiments has shown that it leads to less accurate models for correspondences with noisy positional information in case of homography estimation – see Table 1.

\[
\epsilon_{\text{SymH}}(\mathbf{H}, \mathbf{u}, \mathbf{v}) = d(\mathbf{u}, \mathbf{Hv})^2 + d(\mathbf{v}, \mathbf{H}^{-1}\mathbf{u})^2, \tag{2}
\]

\[
\epsilon_{\text{SymEG}}(\mathbf{F}, \mathbf{u}, \mathbf{v}) = (\mathbf{v}^T \mathbf{Fu})^2 \times \frac{1}{(\mathbf{Fu})_1^2 + (\mathbf{Fu})_2^2 + \frac{1}{(\mathbf{F}^T \mathbf{v})_1^2 + (\mathbf{F}^T \mathbf{v})_2^2}} \tag{3}
\]

The problem can be solved by performing check of the symmetric error (see Eq. 7/8) with loosen (by times of 3 or 4) threshold for the inliers of the model. It is not necessary that all inliers pass this test - a few over the minimal number necessary to estimate the model suffice to check that model is not degenerate.

Local affine frame consistency check. Since the WxBS-M matches are not restricted to planar scenes, the epipolar
Figure 3: An example of a degenerate solution undetected by check (6), \( \left| \det(H)/h_{13} \right| = 79.8 \). Many of the inlier correspondences are mapped to a single feature.

Table 1: Symmetric epipolar error on ground truth points on Oxford dataset [16] for RANSAC minimizing different errors

| RANSAC function | mean error [px] | max. error [px] |
|-----------------|----------------|-----------------|
| Sampson H       | 1.3            | 2.8             |
| Symmetric H     | 1.5            | 3.1             |
| Sampson F       | 2.6            | 11.4            |
| Symmetric F     | 1.3            | 4.6             |

geometry constraint is used for verification of the tentative correspondences. On top of that, WxBS-M can also use homography to find (cutting, virtual) dominant plane. The problem is that the epipolar geometry constraint is less restrictive than the homography. In some cases, the wrong correspondences as well as the correct ones are consistent with some fundamental matrix (see Figure 12, left). Moisan and Stival proposed an ORSA [18] method, used e.g. in ASIFT [20] matching that exploits an a-contrario statistic-based approach to detect incorrect epipolar geometry. In our experiments [28], this approach was overperformed by the following method.

The affine covariant detectors provide matches of full local affine frames i.e. three points-to-three points matches. The computation of features and descriptors already assumes that the scene is at least locally planar. We exploit these two assumptions to verify if all three points of each tentative correspondences are consistent with model. The consistency is verified by using the \( \epsilon_{\text{SymH}} \) or \( \epsilon_{\text{SymEG}} \) error depending on the estimated model.

The check is performed for the inliers of the RANSAC model. Tentative correspondences which do not pass the check are deleted from the inlier list. This helps to clean-up incorrect tentative correspondences which are accidentally consistent with good model or eventually reject all correspondences if the model is not valid. In the later case, the WxBS-M algorithm performs another step of the view synthesis to add more correspondences. As shown in Table 2, this step significantly improves the quality of the estimated model, although it does not solve the problem in all cases. The experiments has shown that it is better to threshold the sum of errors of the three LAF points instead of having a threshold for each of the points in the consistency check due to noise in estimation of the local affine frames [28].

Figure 4: Correspondences obtained without (left) and with (right) the LAF check. Incorrect correspondences are in red.

Table 2: Symmetric epipolar error on manually selected not-all-in-plane ground-truth points on the EVD [17] dataset

|                | mean error [px] | max. error [px] |
|----------------|----------------|-----------------|
| No check       | 25.7           | 708.2           |
| LAF-check      | 9.3            | 277.4           |

4. WxBS dataset and evaluation protocol

A set of 31 image pairs has been collected from Flickr and other sources. The dataset is divided into 5 categories based on the combinations of nuisance classes summarized in Table 3. For every image, a set of approximately 20 ground-truth correspondences has been annotated. Selected examples are presented in Figure 5. The resolution of the majority of the images is 800 \( \times \) 600 with the exception of LWIR images from the WGSBS dataset which were captured by a thermal camera with a resolution of 250 \( \times \) 250 pixels. The selected image pairs contain both urban and natural sceneries. The dataset and the ground truth correspondences will be made public.

4.1. 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 presented dataset contains significant viewpoint changes. In case of a non-planar scene - a homography can, at best, cover a dominant plane. The tested state-of-the-art algorithms use different geometry models and therefore the ground truth has to be applicable for both epipolar geometry and homography.

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 3.

Table 3: The WxBS datasets categories

| Short name | Nuisance               | #images | Avg. #GT Corr. |
|------------|------------------------|---------|----------------|
| 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.    |
| WGLBS      | viewpoint, appearance, lighting | 8 pairs | 17 per img. |

The evaluation protocol. For each image pair indexed with \(i \in \mathbb{Z}\) we have manually annotated a set of correspondences \((u_i, v_i) \in C_i\) where \(u\) and \(v\) are positions in the 1st and the 2nd image respectively. For epipolar geometry we use the symmetric epipolar distance (8) and the symmetric reprojection error (7) for homography (9). Recall on ground truth correspondences \(C_i\) of image pair \(i\) and for geometry model \(M_i\) is computed as a function of threshold \(\theta\)

\[
\tau_{i,M_i}(\theta) = \frac{|\{(u_i, v_i) : (u_i, v_i) \in C_i, e(M_i, u, v) < \theta\}|}{|C_i|}
\]

using appropriate error functions (7) or (8). For all pairs of each category \(W\) we define an overall recall per category as:

\[
\tau_W(\theta) = \frac{\sum_{i \in W} \tau_{i,M_i}(\theta)}{|W|}
\]

This measure is as the fraction of the confirmed annotated correspondences for a given threshold in a nuisance category. Plotting the recall measure against the threshold shows the distribution of geometric errors of the ground truth correspondences. The thresholds for the different geometric models are not comparable as each of them has a different geometric meaning.

5. Results

The primary evaluation criterion was the ability of matcher algorithm to find sufficiently a correct geometric transformation in a reasonable time. An accurate geometry transformation can be found in a consecutive step. In order to show the influence of different improvements (see Section 3), results are reported for various WxBS matcher
configurations (see Table 4). We have also computed separately results for the homography (denoted $WXBS-Mn.H$) and epipolar geometry model (denoted $WXBS-Mn.F$).

The performance of the proposed $WXBS-M$ matcher was compared with state-of-the-art matchers ASIFT [20], Dual Bootstrap (DBstrap) [26] and MODS [17] using the evaluation protocol outlined in the previous section.

The matchers were divided into two categories based on the geometric model. Dual bootstrap and MODS estimate homography transformation, while ASIFT and the proposed $WXBS-M$ matcher estimates epipolar geometry of the scene. For a homography, an image pair is considered solved when at least four ground truth correspondences have the symmetric reprojection error $e_{SYM.H} < 10$. For epipolar geometry, the number of confirmed correspondences must be at least ten, with error $e_{SYM.EG} < 10$.

### 5.1. Comparison on $W1BS$ datasets

#### Table 4: Different configurations of the $WXBS$ matcher

| Configuration | Thresholding | Descriptor             |
|---------------|--------------|------------------------|
| $WXBS-M1$     | Adaptive     | SIFT + HalfSIFT        |
| $WXBS-M2$     | Adaptive     | SIFT                   |
| $WXBS-M3$     | Adaptive     | SIFT + InvSIFT         |
| $WXBS-M4$     | Adaptive     | SIFT + InvSIFT + HalfSIFT |
| $WXBS-M5$     | Fixed        | SIFT + HalfSIFT        |

In this experiment, the performance of the proposed $WXBS-M$ matcher was compared on four publicly available $W1BS$ datasets: Kelman [11], SymBench [10], EVD [17] and MMS [3]. The transformations between the image pairs in these datasets are homographies. Table 6 shows that the state-of-the-art $W1BS$ matchers perform well only for some particular problems while the proposed $WXBS$ matcher significantly outperforms them in all cases in terms of number solved image pairs. All computations have been performed on Intel i7 3.9GHz server (8 cores) with 16GB RAM.

#### Table 5: $W1BS$ datasets used for evaluation

| Short name | Proposed by | #images | Type     |
|------------|-------------|---------|----------|
| GDB-ICP    | Kelman et al. [11], 2007 | 22 pairs | W1BS, W2BS |
| SymBench   | Hauagge and Snavely [10], 2012 | 46 pairs | W1BS, W2BS |
| MMS        | Aguilera et al. [3], 2012 | 100 pairs | W1BS |
| EVD        | Mishkin et al. [14], 2013 | 15 pairs | W1BS |

#### 5.2. $WXBS$ dataset

The previous experiment was conducted on mostly planar scenes and exercising only some groups of image nuisances. In this experiment, we compare the matchers on the above discussed $WXBS$ dataset.

The results are summarized in Table 7. Note that the state-of-the-art matchers were not able to match almost any image pair which combines more nuisance factors. The proposed $WXBS-M$ matcher shows much better performance, but still is not able to solve even half of the new dataset pairs.

In Figure 7 we show the recall on ground truth correspondences against the geometric error. This measure does not indicate number of solved pairs but illustrates overall geometric precision of estimated geometry models. It clearly shows that for some dataset categories, none of the algorithms is able to produce valid results. Still, the $WXBS-M$ matcher outperforms other algorithms in the number of confirmed correspondences.

### 6. Conclusions

We have introduced a new problem – the wide multiple baseline stereo ($WXBS$) which considers matching of images which simultaneously differ in more than one image acquisition factors such as viewpoint, illumination, sensor or where object appearance changes significantly, e.g. over time.

For the $WXBS$ problem, a new public dataset with a ground truth for evaluation of the matching algorithms has been introduced. A novel matching algorithm for address-
Table 7: Matching results on the WxBS dataset

|        | WxBS-M1 | WxBS-M2 | WxBS-M3 | WxBS-M4 | WxBS-M5 |
|--------|---------|---------|---------|---------|---------|
| GA     | 32.7    | 32.7    | 34.0    | 32.7    | 35.6    |
| BS     | 10      | 16.4    | 22.8    | 26.6    | 38.2    |
| WxBS   | 18.3    | 18.1    | 14.4    | 15.1    | 20.9    |
| DBstrap|x        | 9.41    | 22      | 26.6    | 46.8    |
| ASIFT  | 0       | 0       | 0       | 0       | 0       |
| MODS   | 1       | 7.9     | 1       | 3.12    | 5.75    |
| WxBS-X | 0       | 0       | 0       | 0       | 0       |
| LA     | 20.9    | 26.6    | 26.6    | 26.6    | 26.6    |
| GS     | 7.9     | 16.4    | 22.8    | 26.6    | 46.8    |
| GL     | 11.1    | 4.52    | 0       | 0       | 0       |
| GLA    | 11.1    | 18.1    | 22      | 26.6    | 46.8    |
| BS     | 0       | 0       | 0       | 0       | 0       |
| WxBS   | 0       | 0       | 0       | 0       | 0       |
| DBstrap|x        | 9.41    | 22      | 26.6    | 46.8    |
| ASIFT  | 0       | 0       | 0       | 0       | 0       |
| MODS   | 1       | 7.9     | 1       | 3.12    | 5.75    |

Figure 8: Average processing time of matching on the WxBS datasets.

References

[1] H. Aans, A. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *IJCV*, 97(1):18–35, 2012.
[2] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building Rome in a day. In *ICCV*, 2009.
[3] C. Aguilería, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, 2012.
[4] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.
[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, T. Werner, and J. Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, 2005.
[7] Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, and E. 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*, (2013:1):1–16, 2013.
[8] 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. In *ICMR*, 2011.
[9] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*, volume 2.
[10] D. Hauagge and N. Snavely. Image matching using local symmetry features. In *CVPR*, 2012.
[11] A. Kelman, M. Sofka, and C. V. Stewart. Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations. In *CVPR*, 2007.
[12] K. Lebeda, J. Matas, and O. Chum. Fixing the locally optimized RANSAC. In R. Bowden, J. Collomosse, and K. Mikolajczyk, editors, *BMVC*, 2012.
[13] Y. J. Lee, C. L. Zitnick, and M. F. Cohen. Shadowdraw: real-time user guidance for freehand drawing. In *ACM Transactions on Graphics (TOG)*, volume 30, page 27. ACM, 2011.
[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.
[16] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, 2005.
[17] D. Mishkin, M. Perdoch, and J. Matas. Two-view matching with view synthesis revisited. *CoRR*, abs/1306.3855, 2013.
[18] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *IJCV*, 57(3):201–218, 2004.
[19] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. In *ICCV*, 2005.
[20] J.-M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.
[21] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, 2009.
[22] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, 1998.
[23] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. In *ACM Transactions on Graphics (TOG)*, volume 30, page 154. ACM, 2011.
[24] T. Tuytelaars and K. Mikolajczyk. Local Invariant Feature Detectors: A Survey. 2008
[25] V. Vonikakis, D. Chrysostomou, R. Kouskouridas, and A. Gasteratos. A biologically inspired scale-space for illumination invariant feature detection. *Measurement Science and Technology*, 24(7):074024, 2013.
[26] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai. Registration of challenging image pairs: Initialization, estimation, and decision. *PAMI*, 29(11):1973–1989, 2007.
[27] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003.
[28] Authors. Tech.Report. Configuring the wide multiple baseline stereo matcher. Supplied as additional material.
Recall on GT pts. 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6

### Appendix

7. Configuring the wide multiple baseline stereo matcher

8. Adaptive threshold of the detector response

Three approaches for detector response threshold are compared in this Section. The iiDoG algorithm normalises the feature response by the average image intensity brightness of the point neighbourhood. Adaptive thresholding used in WxBS-M, takes all features with response greater than given threshold. If number features above the threshold is less, than desired, algorithm take top $N$ features. The last tested setup is standard fixed threshold on detector response. Experiments were performed on the selected image pairs, shown in Figure 9 and also with a subset of 73 image pairs from the MMS dataset [3] where precise homography between the images has been established. The results are shown in the Table 8.  $-2k$ detectors use the adaptive threshold method for detection of $N = 2000$ features with the highest feature response. Fixed threshold $\theta$ was set to 5. The number of correct matches is computed using method defined by Mikolajczyk et al. [15].

| Matcher       | DoG   | DoG-2k | iiDoG | iiDoG-2k |
|--------------|-------|--------|-------|----------|
| #M           | #F    | #M    | #F    | #M       | #F     |
| bikes        | 82    | 391   | 392   | 1749     | 157    | 692   | 590   | 1748 |
| notredame16  | 91    | 640   | 104   | 866      | 83     | 562   | 97    | 773  |
| vatican      | 2     | 121   | 18    | 645      | 7      | 143   | 19    | 639  |
| bdom         | 32    | 719   | 125   | 1846     | 109    | 1191  | 124   | 1824 |
| MMS Avg.     | 3     | 123   | 16    | 628      | 8      | 217   | 16    | 628  |

Table 8: Number of correct matches (#M) and the number of detected features in reference image (#F) per variant of DoG detector.

iiDoG outperforms standard DoG in images with big illumination change with fixed threshold. DoG with adaptive threshold shows the best performance among all tested configurations.

9. HalfSIFT vs. InvSIFT

In this part performance of several SIFT descriptor variants is compared. Features detected in the reference image were reprojected into the tested image using the ground truth homography (upright features are used only) in order to weaken the influence of a the feature detector. Features are detected with Hessian Affine feature detector with adaptive threshold (response thr. $\theta = 8$, min. num. features $N = 2000$).

9.1. Descriptor performance

Descriptor performance have been evaluated with protocol defined by Mikolajczyk et al. [15]. SIFT [14] ($S \rightarrow S$), half-SIFT [5]($H \rightarrow H$) and inverted SIFT [8]($S \rightarrow SI$) were compared. It is important to note that in the case of the inverted-SIFT experiment, only descriptors from the tested image are “inverted”. In our notation, this is expressed as $S \rightarrow SI$ - i.e. SIFTs from reference image matched to SIFTs and inverted-SIFTs from the tested image.

As the features are reprojected, the geometric repeatability of the image frames is generally 100%. Therefore recall and precision defined by Mikolajczyk [15] is always equal. That is why only the recall varying the First geometrically inconsistent nearest neighbour (FGINN) ratio is plotted. The results for selected image pairs from Figure 9 are shown in Figure 10.
notredame16  bdom  vatican  bark  graf  wall  bikes

Figure 9: Image pairs used in detector and descriptor experiments. WiBS pairs notredame16, bdom and vatican are from SymBench dataset [10]. The standard testing WgBS image pairs bark, graf, wall and bikes (which contains mainly out-of-focus blur) are from [16] and for the pairs bark, graf and bikes, refined ground truth from [?] is used.

Figure 10: Performance of different SIFT descriptor variants on reprojected image features.

$H \rightarrow H$ configuration showed the lowest performance. Only in cases of multimodal pairs the $H \rightarrow H$ configuration has got comparable performance as standard descriptors. Therefore using only half-SIFT descriptors affects the matcher performance.

9.2. Performance of the descriptor combination

Performance of the difference descriptor combinations have been evaluated. In case of $S,H \rightarrow S,H$, tentative correspondences for SIFT and half-SIFT descriptors are generated separately using the FGINN ratio of 0.8 [17] and the accepted matches are joined together. If both SIFT and half-SIFT matches were found for a reference image feature, the SIFT match is preferred. The last examined configuration, $S,H \rightarrow SI,H$, matches SIFT features from reference image to SIFT + inverted-SIFT features from the tested image. The number of accepted matches for each configuration is shown in Table 9. Results are computed for pairs from 9 and also for the subset of the MMS dataset [3] (see 8).

Table 9 shows that SIFT performs the best as single descriptor and SIFT + halfSIFT combination outperforms other configurations including $S,H \rightarrow SI,H$. The last exam-
Table 9: Number of matched reprojected frames using different combinations of descriptors.

Table 10: Consistency of geometric transformation obtained by RANSAC minimizing Sampson error. Average errors for the model inliers.

Table 11: Symmetric transfer errors on ground truth points for geometry model estimated by RANSAC have been calculated.

Table 12: Symmetric transfer errors on ground truth points for RANSAC which minimizes different errors.

10. RANSAC error functions

Testing the WxBS-M matcher on various datasets [3, 10, 11, 17] revealed that both LO-RANSAC [12] and DegenerSac [6] implementations, which minimize the Sampson error [9], often output near-degenerate solution (see Figure 11).

This problem was more common in homography estimation, but happened in epipolar geometry estimation as well. Fundamental matrices were not, strictly speaking, degenerate, but the epipole was placed near incorrect matches. The phenomenon does not occur on a standard (easy) dataset like Oxford [16].

Homography matrix is degenerate when one of its rows or columns is linearly dependent [9], i.e. rank(H) = 2. Direct test for matrix rank is not applicable due to numerical issues and is instead approximated by testing the following condition

\[ \det(H) > 0.1. \]  

This condition holds true for all observed near-degenerate cases, i.e. it is not sufficient for their detection.

Symmetric error [7, 8] unlike Sampson error is big (see Table 10 “multiple-to-one”, and can be used for near-degenerate transformation detection.

\[ e_{\text{SymH}}(H, u, v) = d(u, Hv)^2 + d(v, H^{-1}u)^2, \]  

\[ e_{\text{SymEG}}(F, u, v) = \left( v^T Fu \right)^2 \times \left( \frac{1}{(Fu)^1_1 + (Fu)^1_2} + \frac{1}{(F^Tv)^1_1 + (F^Tv)^1_2} \right) \]

The accuracy of model estimation with RANSAC minimizing Symmetrical error (a) and Sampson error which checks the far-the-best model by symmetric error (b) has been tested on two dataset. Symmetrical transfer (H-case) or epipolar (F-case) errors on ground truth points for geometry model estimated by RANSAC have been calculated.

Comparisons have been performed on three datasets with manually obtained ground truth points by Lebeda [12] – homogr (H), middlebury (F), kusvod2 (F) and on Oxford dataset [16] (H). Ground truth points for the Oxford dataset have been obtained generating 50 random points in reference image. Then those points have been reprojected to the second image using the ground truth homographies.

Geometrical model was estimated using tentative correspondences from MSER and Hessian-Affine local features detected without view synthesis. Error threshold has been set to 2 pixels for all experiments in this and the next Section.

Results are shown in Tables 11-12. Sampson error minimization followed by symmetric error check for the far-the-best model (SymmCheck) has slightly better accuracy of the model than simple symmetric homography error minimization. Symmetrical error performs better in epipolar geometry case.

11. Local affine frame consistency check

Moisan and Stival’s ORSA [18] method, used in ASIFT [20] matching, exploit an a-contrario statistic-based approach to detect incorrect epipolar geometry. Instead of having constant threshold for error, ORSA looks for the matches that have most “diameter”, i.e. the better cover im-
Figure 11: Examples of a near-degenerate geometry models, when many of the inlier correspondences are mapped to a single feature. First row – homographies, second row – epipolar geometry. Image pairs, from left to right: EO-IR-1 [11], pantheon [10], dum [17].

Table 12: Symmetric epilolar errors on ground truth points for RANSAC minimizing different errors.

| Function to optimize            | avg. [px] | max [px] |
|--------------------------------|-----------|----------|
| Oxford dataset [16]             |           |          |
| Sampson + SymmCheck             | 2.6       | 11.4     |
| Symmetric                       | 1.3       | 4.6      |
| Lebeda middlebury dataset [12]  |           |          |
| Sampson + SymmCheck             | 5.7       | 43.5     |
| Symmetric                       | 3.4       | 18.9     |
| Lebeda kusvod2 dataset [12]     |           |          |
| Sampson F +                     | 39.7      | 246.2    |
| Symmetric F                     | 38.4      | 299.1    |

We use coordinates of the short and long axis of the corresponding ellipses to check whether whole local feature is consistent with estimated geometry model (LAF-check). This check is performed for the RANSAC output model and regions which do not pass the check are deleted from the inlier list. If number of correspondences after the LAF-check is less than user defined minimum number, matcher goes for the next step of the view synthesis.

Two variants have been considered. The first is to have error threshold for each of three points - center, short and long axis while have minimum possible error and estimate if such inliers could be non-random.

Figure 12: Correspondences obtained without (left) and with (right) the LAF check. Incorrect correspondences are in red.
Table 13: Symmetric epipolar error on manually picked ground truth points

| Dataset              | Method            | avg. [px] | max [px] |
|----------------------|-------------------|-----------|----------|
| Lebeda middlebury    | RANSAC            | 3.4       | 18.9     |
|                      | DegenSAC F        | 2.7       | 10.6     |
|                      | DegenSAC F+LAF    | 5.1       | 31.5     |
|                      | ORSA              | 1.5       | 5.0      |
|                      | ORSA+LAF          |           |          |
| Lebeda kusvod2       | DegenSAC F        | 39.7      | 246.2    |
|                      | DegenSAC F+LAF    | 38.8      | 198      |
|                      | ORSA              | 32.9      | 230.7    |
|                      | ORSA+LAF          | 32.9      | 230.7    |
| EVD                  | DegenSAC F        | 25.7      | 708.2    |
|                      | DegenSAC F+LAF-check | 9.3    | 277.4    |
|                      | ORSA              | 120.0     | 2078.0   |
|                      | ORSA+LAF          | 46.0      | 283.0    |

long ellipse axis correspondences, the second - to have threshold for the sum of the errors - ideal correspondence of the one axis would allow bigger error for another. First variant turned out to be too strictive - it has eliminated all matches in 5 out of 15 images of the EVD dataset and haven’t given significant increasing of the model quality on other images, so we recommend to use threshold on for sum of the errors squared.

ORSA and DegenSAC with and without applying LAF-check procedure have been compared. ORSA implementation taken from the ASIFT code [19]. Experiment setup is similar to the previous Section, but now we have used view synthesis, because EVD dataset isn’t solvable without it. Results are shown in Table 13.

ORSA showed competitive performance on the Lebeda middlebury and kusvod2 datasets but failed on the EVD dataset. In 5 out of 15 cases neither geometric transformation neither most of inliers were correct or “matching is not significant” output got. Inliers of the DegenSAC output were mostly correct small with exception of those, which were shifted along the epipolar lines. Still, consistency between estimated model and manually picked ground truth points is not enough without LAF-check.

LAF-check allow to clean up ORSA output, but 5 images remain unsolved. The only difference than instead of incorrect epipolar geometry, ORSA + LAF-check outputs “no match” answer. That is why we used DegenSAC in WxBBS-M matcher.

Note, that difference in the quality of the geometric model with and without LAF-check is possible only when LAF-check leads to the rejection of the all inliers and next step of the view synthesis. Otherwise all the difference is only in less number of (inconsistent) inliers – see Figure 12.