Performance of Interest Point Descriptors on Hyperspectral Images

Przemysław Głomb and Michał Cholewa
Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka 5, 44-100 Gliwice, Poland

Keywords: Hyperspectral Images, Interest Point Descriptors, SIFT, SURF, ORB, BRISK.

Abstract: Interest point descriptors (e.g. Scale Invariant Feature Transform, SIFT or Speeded-Up Robust Features, SURF) are often used both for classic image processing tasks (e.g. mosaic generation) or higher level machine learning tasks (e.g. segmentation or classification). Hyperspectral images are recently gaining popularity as a potent data source for scene analysis, material identification, anomaly detection or process state estimation. The structure of hyperspectral images is much more complex than traditional color or monochrome images, as they comprise of a large number of bands, each corresponding to a narrow range of frequencies. Because of varying image properties across bands, the application of interest point descriptors to them is not straightforward. To the best of our knowledge, there has been, to date, no study of performance of interest point descriptors on hyperspectral images that simultaneously integrate a number of methods and use a dataset with significant geometric transformations. Here, we study four popular methods (SIFT, SURF, BRISK, ORB) applied to complex scene recorded from several viewpoints. We presents experimental results by observing how well the methods estimate the 3D cameras' positions, which we propose as a general performance measure.

1 INTRODUCTION

Computer vision applications such as image registration, stereo matching, object and texture recognition, image retrieval, robot simultaneous location and mapping (SLAM) require working with images of the same scene that were taken at different locations and times. Changes in geometry, scene composition, lighting, sensors used to capture images introduce deformations in pixel structure that require specialized algorithms to process. A successful and established approach for this is to use local features (Mikolajczyk and Schmid, 2005; Moreels and Perona, 2007): locate image points with unique characteristics of local pixel neighbourhood, generate their signature descriptors, then match those descriptors across different images. As descriptors are engineered to be robust to potential degradations (Lowe, 2004; Bay et al., 2008) (affine transformations, change of geometry, noise, light etc.) the result should be a match of local areas in two images with high degree of certainty. This match forms the base of subsequent processing, which could be detecting the presence of certain objects in images or estimating the geometric transformation between them, e.g. a homography.

Hyperspectral images combine the spatial dimension of traditional photography with spectral dimension of spectrographs. The recorded spectral data is the result of an interaction of light with surfaces of objects in the scene, thus can be used to identify the materials much more reliably than RGB information. Because of the physical properties of this recording setting, the images from each band (each being essentially a monochromatic bitmap) differ in properties (Mukherjee et al., 2009): amount and characteristics of noise, level of blur, more or less pronounced features (e.g. edges) at given spatial position.

Since their introduction, the local interest point descriptors have been subject of numerous[1] applications and comparisons studies. Various techniques are used for their computation, including Difference of Gaussians (DoG) filter and image gradients (Lowe, 2004), Haar-like features from integral images (Bay et al., 2008) or local intensity comparisons (Leutenegger et al., 2011; Rublee et al., 2011). High dimensional descriptors have been shown to be more robust to geometric transformations than region corre-

[1] As of 2015, the paper introducing one of the most popular methods (SIFT descriptor) has already over 30K citations (source: Google Scholar).

Głomb, P. and Cholewa, M.
Performance of Interest Point Descriptors on Hyperspectral Images.
DOI: 10.5220/0005785001980203
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 3: VISAPP*, pages 200-205
ISBN: 978-989-758-175-5

lation (Mikolajczyk and Schmid, 2005). Their performance is consistent across different experimental settings (Moreels and Perona, 2007). At the same time, while there have been extensions to multispectral images (see e.g. (Brown and Susstrunk, 2011)), the studies performed on hyperspectral images have been comparatively limited. Hyperspectral images substantially differ from RGB or monochrome images because of, among else, much larger data size, varying performance of sensor of different frequencies, complex statistical relationships between recorded spectra, variation in data resulting from push-broom recording scheme, and varying noise across frequency spectrum (Mukherjee et al., 2009; Vakalopoulou and Karantzalos, 2014). While they can be reduced to monochrome images (that could be a simple input to classical interest point algorithms), that conversion is not trivial and could lose structural information (Dorado-Munoz et al., 2012). Most approaches in hyperspectral domain are based on the SIFT algorithm: as classification support (Xu et al., 2008), algorithm extension (Mukherjee et al., 2009; Dorado-Munoz et al., 2012), aligning image strips for change detection (Ringaby et al., 2010) or optimizing parameters for hyperspectral image matching (Sima and Buckley, 2013). A different approach is taken in (Vakalopoulou and Karantzalos, 2014), where SIFT and SURF are combined in working with spectral bands groups.

We identify two important practical shortcomings of current studies. One is the lack of including a significant geometric deformations in the test data set–currently used images differ by time of acquisition and selected affine parameters only (translation (Mukherjee et al., 2009; Ringaby et al., 2010; Dorado-Munoz et al., 2012; Vakalopoulou and Karantzalos, 2014) and scale (Dorado-Munoz et al., 2012)), obtained by down-looking satellite or plane-mounted camera. The only exception is (Sima and Buckley, 2013), where tripod-acquired geological data show some geometric deformations. The second problem is the lack of comparing side-by-side the performance of different methods. The focus is commonly on only one method, even at verification stage (Xu et al., 2008; Mukherjee et al., 2009; Ringaby et al., 2010; Dorado-Munoz et al., 2012; Sima and Buckley, 2013). The one exception is (Vakalopoulou and Karantzalos, 2014), where SIFT and SURF are compared.

Our focus in this paper is the investigation of performance of interest point descriptors on hyperspectral images of a 3D scene. We make two novel contributions: first, we compare four separate descriptor algorithms: SIFT (Lowe, 2004), SURF (Bay et al., 2008), ORB (Rublee et al., 2011) and BRISK (Leutenegger et al., 2011); second, we use a specially prepared dataset of scene of mixed natural and man-made objects, imagined from different view points. Our experimental setting is as follows: we use the interest point algorithms to detect and match points in two images, then evaluate them based on quality of estimation of relative 3D camera positions.

This paper is organized as follows: next section presents the experimental setting. The results are presented in the third section, and the last section presents discussion and concluding remarks.

2 METHODS

Data Set. To compare the descriptors, we use a specially prepared data set that allows to test image processing methods on images with significant geometric deformations, resulting from hyperspectral imagining a 3D scene with total viewpoint change of about 45^o[2]. To our best knowledge, this is the first dataset of such kind.

We use a scene (cf. Figure 1) containing both natural and artificial fruits of several categories. This produces images rich in structure in both visual and NIR spectral ranges (in the former, color based edges are the strongest, in the latter, neighborhoods of materials of different types). The scene also contains checkerboard-type markers for calibration and ground-truth estimation, and Munsell grey panel for light calibration. The scene is lighted with multi point halogen light, supported by UV lamp (Omnilux CFL UV 25W with color temperature 6000 UV K). Images are recorded with Surface Optics SOC-710VP 375-1045 nm camera from five points. The angle steps are at $\approx 11^o$ intervals, this choice is based on analysis of (Moreels and Perona, 2007), where it has been observed that viewpoint change of more than at 30^o drastically reduces the feature matching effectiveness.

Descriptors. For comparison, we select four descriptors: SIFT (Lowe, 2004) and SURF (Bay et al., 2008) because of their popularity and reported good performance; and ORB (Rublee et al., 2011) and BRISK (Leutenegger et al., 2011), proposed as alternative descriptors with good time efficiency. We used the implementations available in the OpenCV library (Bradski, 2000). For matching, we use ratio filtering (Lowe, 2004): we only consider points for which the ratio of distance to first and to second nearest neighbour is lower than $r_0 = 0.8$, thus excluding points that could be well matched to several locations.

[2]The dataset will be made available on-line, link removed for anonymization purposes.

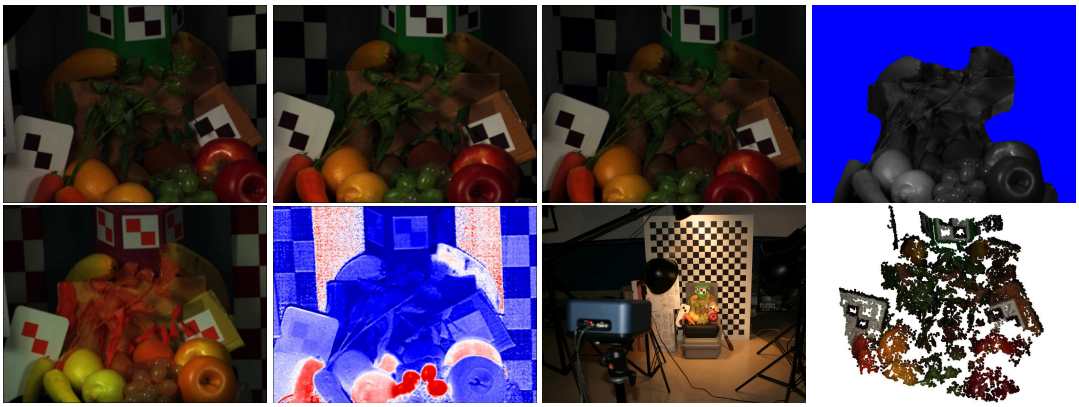

Figure 1: Renderings of the scene used in the experiment. Top row: color RGB renderings from three separate angles and the mask used for removing calibration markers. Bottom row: false color NIR (Near Infrared), 1000nm bands, acquisition setting (one of camera locations) and an example 3D point cloud computed from matched keypoints. Note the easy separation of artificial (plastic) and natural (food) objects on the 1000nm band image, which is much more difficult on visual band range.

Experiment. While there exist a number of different methods that are used for evaluating descriptor performance (see e.g.(Mikolajczyk and Schmid, 2005; Moreels and Perona, 2007)) we argue that in most practical applications, one is interested either in extracting 3D scene parameters, or in object recognition. We focus on the former problem, as it usually has more general (less application specific) formulation. Also, the quality of 3D estimation verifies descriptors' sensitivity to actual scene information over acquisition conditions. The 3D scene estimation from uncalibrated images is recently becoming a popular application of image processing methods. It involves recovering sparse or dense scene 3D point cloud and camera parameters, from a sequence of images in process called Structure From Motion (SFM). For this task, technique of Bundle Adjustment (see e.g. (Wu et al., 2011)) is often selected.

We propose a simple evaluation scheme, which measures the relative error of estimating 3D camera position. We argue that while individual measures (e.g. repeatability, stability) are important for qualitative assessment of algorithms performance, the final application result is very often of key importance. In case of 3D reconstruction, the quality can be measured on how well the scene parameters are estimated, e.g. with reference to ground truth data. Camera positions are a good benchmark because their precise estimation leads to high quality of the scene point cloud, and at the same time the position error in 3D is simple and intuitive. As the scene parameters are estimated up to a translation and scale factor, the error of camera estimation may be misleading in some cases[3], we propose a performance measure based on error of relative camera position estimation:

$$\varepsilon = \frac{1}{n_p} \sum_{(i,j),i\neq j} \|\vec{d}_{ij} - \vec{d}'_{ij}\|_2 \qquad (1)$$

where $\vec{d}_{ij} = \vec{c}_j - \vec{c}_i$ is the true, and $\vec{d}'_{ij} = \vec{c}'_j - \vec{c}'_i$ estimated distance between positions of cameras i and j; n_p is the number of camera pairs combinations.

We estimate the true positions–ground truth for the experiments–using semi-automatic point selection (user assisted with automatic refinement) from calibration markers, inspired and in part similar to earlier work on descriptor comparison (Mikolajczyk and Schmid, 2005; Moreels and Perona, 2007).

For given range of bands from hyperspectral images and parametrized descriptor algorithm, we proceed as follows:

1. Locate interest points and compute descriptors separately in each image. Only the scene data is used; calibration markers are masked out (see Figure 1).
2. For each pair of images, descriptor matches are computed using Euclidean or Hamming norm (depending on descriptor algorithm) and filtered using neigborhood ratio $r_0 = 0.8$.
3. Matched interest points are input to the SFM algorithm, where 3D scene parameters are estimated. Relative camera positions are measured, and compared with estimated ground truth.

We use VisualSFM (Wu, 2011) software for 3D scene estimation, as it was easy to integrate with experiment suite and was found to perform well in the initial tests. The experiments are performed separately for different groups of bands (UV, visual, near infrared range).

[3]Initial experiments suggested that distortion in relative position estimation is a better quality predictor of the final point cloud than absolute position.

Table 1: Performance of interest point descriptors, as measured with relative camera position estimation. Columns denote results for given spectral range, rows descriptor algorithms with set parameters (see text). Errors are given as percentage of a camera model estimated (related to number of cameras identified) and mean of their relative errors of cameras' distances measurement (brackets). '–' denotes that camera model estimation did not converge, in most cases because lack of enough stable feature correspondences.

	VIS[a]	VIS-B[a]	VIS-G[a]	VIS-R[a]	NIR[b]	NIR-1[b]	NIR-2[b]	NIR-3[b]
SIFT-A	100% (1.15)	–	100% (0.39)	100% (0.49)	–	100% (0.37)	–	10% (2.32)
SIFT-B	100% (1.93)	60% (1.91)	100% (0.45)	60% (2.33)	10% (0.47)	100% (1.22)	10% (0.71)	100% (0.66)
SIFT-C	60% (2.00)	–	100% (1.04)	100% (3.95)	–	100% (0.55)	–	10% (2.13)
SURF-100	100% (1.50)	10% (0.47)	100% (0.71)	100% (0.58)	10% (1.00)	100% (2.97)	–	100% (0.33)
SURF-300	30% (0.77)	–	100% (0.48)	60% (7.80)	–	30% (2.33)	–	10% (2.35)
SURF-500	30% (2.60)	–	60% (1.90)	100% (6.24)	–	30% (0.63)	–	–
BRISK-5	10% (2.54)	10% (2.37)	10% (0.91)	10% (1.16)	–	10% (2.22)	–	–
ORB	10% (2.41)	–	10% (2.36)	10% (2.36)	–	–	–	–

[a] Visual spectral ranges: all $400-700$nm, bands $6-63$; blue $400-500$nm, bands $6-24$; green $500-590$nm, bands $25-42$; red $590-700$nm, bands $64-82$.

[b] Near infrared ranges: all $700-1050$nm, bands $64-127$; sub-range 1 is $700-800$nm, bands $64-82$; sub-range 2 is $800-900$nm, bands $82-100$; sub-range 3 is $900-150$nm, bands $101-127$.

3 RESULTS

Parameters. Interest point descriptor algorithms typically have a number of parameters, e.g. related to sensitivity of the detector or type of description generated. For SIFT method, we use the following sets of parameters: original values as recommended in (Lowe, 2004) (SIFT-A); modification of $\sigma = 1.0, contrastThreshold = 0.02, edgeThreshold = 10$ based on results of (Sima and Buckley, 2013) (SIFT-B); and modification of $\sigma = 1.0$ as proposed in (Vakalopoulou and Karantzalos, 2014) as more sensitive version (SIFT-C). For SURF method, we observe that the main parameter influencing the result is the Hessian threshold h; implementation (Bradski, 2000) uses default $h = 100$ and recommends $h \in (300, 500)$, following that we define three sets of parameters denoted as SURF-100, -300, -500. For BRISK method, we use the threshold parameter t; proposed value of $t = 30$ has been observed to generate small number of keypoints, so we introduce two other values $t = 15$ and $t = 5$ to increase sensitivity, final sets are denoted as BRISK-30, -15 and -5. For ORB method, we use original recommended parameters (ORB-A), version with FAST detector instead of Harris (ORB-F), and version with increased *patchSize* and *edgeThreshold* to $p_s = 51$ (ORB-ps), to improve detection on noisy bands.

As the SFM algorithm is probabilistic in nature, it was run $n = 10$ times for each case and the best model was selected–this correspond to typical usage scenario.

Test Data. The resulting hyperspectral image set comprises of five images of dimensions $696 \times 520 \times 128$, the last dimension corresponds to spectral range $375-1045$nm, with average spectral resolution close to 5nm and 12 bits of precision. Spectral data at each pixel is normalized to enhance contrast. For input to descriptor algorithms, individual bands are aggregated into groups: (partial) ultraviolet range, $375-400$nm, bands $1-6$ and four visual and four near infrared ranges (see Table 1 notes). The objective is to reduce noise persistent in individual bands, while retaining ability to observe performance on different parts of spectral range.

Camera Position Estimation Errors. The results are presented in Table 1. As it can be expected, estimation of 3D scene parameters from medium-to-low resolution of hyperspectral camera is a challenging task for all tested algorithms. In many cases (denoted '-') the number of matched points was not enough for parameter estimation. In other cases, only some of cameras were properly identified–to signify this, a percentage score was added to the error to show how many relative camera positions could be processed. Acceptability criteria could be formulated as score of 100% with error close to zero. We note that SIFT and SURF outperform the BRISK and ORB methods, with SIFT marginally ahead of SURF in the visual ranges. Post experiment analysis suggests that performance of ORB suffered because of low number of detected points, while for BRISK it was high number of similar matches.

Sensitivity to Spectral Range. Due to varying degree of blur, noise, and changing scene spectral characteristics with frequency, the responses of each feature algorithm have a lot of variance. In particular the UV range, even with additional light, was too noisy to produce estimates for any methods. Difficult range was also whole and middle of near infrared

(NIR and NIR-2), which have a low contrast; however, when separated the NIR-1 and NIR-3 ranges are reliable for estimation. This is expected as scene contains both natural and man-made objects, thus their materials' reflectances differ in narrow sub-ranges of the NIR range. Visual range performs best, especially the VIS-G region, where contrast, sharpness and signal to noise ratio is at maximum. It is important to note that the performance in different spectral ranges is dependent on scene composition, but in general similar performance can be observed in visual and selected bands of the NIR range.

Sensitivity to Parameters. Standard SIFT-A parameters perform well, as expected, for visual range. Increasing sensitivity improves estimation in the far NIR-3 range, however it must be done with caution, as can be observed on results of SIFT-B and -C. Both have been proposed as more sensitive and hyperspectral-tuned versions, yet the C loses performance in the visual range. For SURF, increased sensitivity translates in general to better performance. Similarly for BRISK, where only the -5 parameter set was considered, as remaining produced too many convergence errors. For ORB, modifications of recommended parameters (ORB-F, ORB-ps) did not produce improvement in performance.

4 DISCUSSION

Individual Method Performance. While the exact numbers allow to create ranking list of methods, the large variability suggests less strict categorization. It seems that performance of SIFT and SURF is comparable, and better than BRISK and ORB. For the latter two methods, the performance could be perhaps improved by exhaustive parameter optimization combined with image preprocessing. This may be important in practical applications, as ORB and especially BRISK are much faster than remaining methods, which can be of use as hyperspectral cameras produce naturally high volume of data. Increased sensitivity through parameter settings does seem to be a good strategy; this is similar as reported in (Sima and Buckley, 2013).

The Characteristics of Hyperspectral Images. The complex properties of hyperspectral images have large influence on method performance. While per-band strategy can be effective in locating various stable features (e.g. color boundaries in VIS region or material edges in NIR range), the change of image properties makes the performance of methods uneven. As for scene composition, there's a known effect of scene-dependent performance (cf. conclusions in (Moreels and Perona, 2007)); the adaptation of methods to characteristics of hyperspectral images and inclusion of other types of objects will be the subject of further studies.

Relation to other Works. Our results are in general agreement with analysis done for 'regular' images (Mikolajczyk and Schmid, 2005; Moreels and Perona, 2007). In particular, we affirm the performance and robustness of the SIFT descriptor for hyperspectral images. This supports the research done on extending or applying SIFT to hyperspectral images (Xu et al., 2008; Mukherjee et al., 2009; Ringaby et al., 2010; Dorado-Munoz et al., 2012). We note, however, that there's a high chance of attaining similar level of performance with the SURF algorithm. With regards to (Vakalopoulou and Karantzalos, 2014), we don't see their level of error when using standard SIFT or SURF parameters in the visual spectral range. On the contrary, this is the region where both methods perform quite well; this is in fact expected, as bands in visual range are close to grayscale images used to derive original parameters. One possible reason for this difference could be some unique properties of the capturing equipment or imagined scenes used by (Vakalopoulou and Karantzalos, 2014). Our study is less investigative in the interplay of different parameters than (Sima and Buckley, 2013), as we have not one, but four different parameter sets to analyze.

Conclusions. We've presented a performance analysis of several interest point descriptor methods as applied to hyperspectral images in a novel setting. We show the SIFT and SURF both produce quite good results across bands. Application of BRISK and ORB, while faster, did not result in stable 3D estimation from this type of images. For improving performance of descriptors on hyperspectral data, we recommend increasing sensitivity through parameter settings, and location of informative spectral ranges, e.g. with prior knowledge or additional preprocessing.

ACKNOWLEDGEMENTS

This work has been supported by the project 'Representation of dynamic 3D scenes using the Atomic Shapes Network model' financed by National Science Centre, decision DEC-2011/03/D/ST6/03753.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V. (2008). Speeded-up robust features (SURF). *Computer Vision*

and Image Understanding, 110(3):346 – 359. Similarity Matching in Computer Vision and Multimedia.

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Brown, M. and Susstrunk, S. (2011). Multi-spectral SIFT for scene category recognition. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 177–184.

Dorado-Munoz, L., Velez-Reyes, M., Mukherjee, A., and Roysam, B. (2012). A vector SIFT detector for interest point detection in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4521–4533.

Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

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

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630.

Moreels, P. and Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3):263–284.

Mukherjee, A., Velez-Reyes, M., and Roysam, B. (2009). Interest points for hyperspectral image data. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3):748–760.

Ringaby, E., Ahlberg, J., Wadströmer, N., and Forssén, P.-E. (2010). Co-aligning aerial hyperspectral push-broom strips for change detection. In *Proc. SPIE*, volume 7835, pages 78350Y–78350Y–7.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: an efficient alternative to SIFT or SURF. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Sima, A. A. and Buckley, S. J. (2013). Optimizing SIFT for matching of short wave infrared and visible wavelength images. *Remote Sensing*, 5(5):2037–2056.

Vakalopoulou, M. and Karantzalos, K. (2014). Automatic descriptor-based co-registration of frame hyperspectral data. *Remote Sensing*, 6(4):3409–3426.

Wu, C. (2011). VisualSFM: A visual structure from motion system. http://ccwu.me/vsfm/.

Wu, C., Agarwal, S., Curless, B., and Seitz, S. (2011). Multicore bundle adjustment. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3057–3064.

Xu, Y., Hu, K., Tian, Y., and Peng, F. (2008). Classification of hyperspectral imagery using SIFT for spectral matching. In *2008 Congress on Image and Signal Processing, CISP '08.*, volume 2, pages 704–708.