Keypoint Descriptors for Matching Across Multiple Image Modalities and Non-linear Intensity Variations

Avi Kelman, Michal Sofka, and Charles V. Stewart
Rensselaer Polytechnic Institute Department of Computer Science
Troy, New York 12180–3590 *
{*kelmaa,sofka,stewart*}*@cs.rpi.edu*
http://www.vision.cs.rpi.edu/keypoints/

Abstract

In this paper, we investigate the effect of substantial inter-image intensity changes and changes in modality on the performance of keypoint detection, description, and matching algorithms in the context of image registration. In doing so, we modify widely-used keypoint descriptors such as SIFT and shape contexts, attempting to capture the insight that some structural information is indeed preserved between images despite dramatic appearance changes. These extensions include (a) pairing opposite-direction gradients in the formation of orientation histograms and (b) focusing on edge structures only. We also compare the stability of MSER, Laplacian-of-Gaussian, and Harris corner keypoint location detection and the impact of detection errors on matching results. Our experiments on multimodal image pairs and on image pairs with significant intensity differences show that indexing based on our modified descriptors produces more correct matches on difficult pairs than current techniques at the cost of a small decrease in performance on easier pairs. This extends the applicability of image registration algorithms such as the Dual-Bootstrap which rely on correctly matching only a small number of keypoints.

1. Introduction

Keypoint detection, description, and matching techniques have received considerable attention in recent years [14, 15, 16, 19, 24], and the results have been used extensively for image registration and object recognition [2, 11, 20, 24, 21, 25]. These techniques work by detecting keypoints at distinctive image locations, extracting summary descriptions of the image region surrounding the keypoints, and matching these descriptions in order to match the keypoints. The design and experimental evaluation of keypoints and their descriptors has focused on affine changes in both image position and image intensity. This paper considers the effect of more dramatic changes in intensity, including changes in image modality, on the detection, description, and matching of keypoints. Although different modality images potentially measure different phenomena, we assume, as is required for any keypoint method, that correlating features can be detected in the two images. Examples of the types of images for which we would like to extract and match keypoints are shown in Figs. 1 and 2.

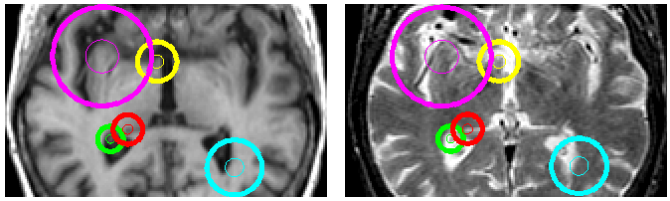


Figure 1. Example of a multimodal pair with a few correct keypoint matches superimposed.

In theory, the assumptions underlying the design of keypoint detection and, especially, description algorithms are violated by these types of image pairs. In practice, however, keypoint algorithms can still be effective. There are several reasons for this. First, many image structures, especially boundaries, tend to be preserved between images under different illuminations and modalities. Texture tends to be more susceptible to change. This suggests that many of the same keypoints, at least those that do not depend on texture, will tend to reappear in the different images. Second, the comparison of keypoint descriptors is a relative measure, and therefore even when a descriptor has changed between images, changes in other descriptors may be even greater, causing the correct match to still be found. Third, given the recent success of the Dual-Bootstrap registration algorithm [25], which is capable of successfully aligning a pair of images starting from just one correct keypoint match, the demands on keypoint matching are not as great as when registration depends entirely on keypoint matching [2]. Thus,

*This article was supported by the DOD and the Medical University of South Carolina under DOD Grant No. W81XWH-05-1-0378. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Department of Defense or the Medical University of South Carolina. This work was also supported by Lockheed Martin and the National Geospatial-Intelligence Agency through CenSSIS, the Center for Subsurface Sensing and Imaging Systems, under the Engineering Research Centers Program of the National Science Foundation (Award Number EEC-9986821).

we will be satisfied with only a small number of correct keypoint matches, provided there is a mechanism for ranking the matches that places these near the top.

The goal of this paper is to investigate keypoint detection, description, and matching for image pairs involving substantial changes in illumination and differences in modalities. We use a suite of representative image pairs with known inter-image transformations as the basis for our investigation. We focus on just three of the top detection techniques, (a) the Laplacian-of-Gaussian [11], (b) Harris corners [7], and (c) maximally-stable extremal regions (MSERs) [12], investigating the repeatability of their locations and orientations. Among descriptors we focus on variations of the SIFT [11] and shape-context descriptors [1], both of which emphasize the distribution of points and gradients. The primary differences between these, once gradient information is added to shape-contexts, are the spatial organization of the bins and the choice of points — all points in a region or just the edge points. To keep our analysis simple, we use the square grid of the SIFT descriptor, but we do examine the choice of points. Moreover, we do not conflate questions of affine invariance [4, 9, 13, 15] with our primary investigation, choosing our data set to avoid substantial viewpoint effects. Thus, we focus on addressing the primary issue: how to best capture in a descriptor the information preserved between images. In doing so we compare the original SIFT descriptor to modified descriptors involving different ways to employ gradient and edge information.

2. Detection Techniques

We investigate three representative detectors: (1) The Laplacian-of-Gaussian (LoG) detector [11] finds peak Laplacian responses across both spatial and scale dimensions in a Gaussian scale-space image representation. (2) Harris corners [13] are complementary features to blobs, and are detected by finding maxima of the Harris cornerness measure [7]. Similar detectors have proven useful in the medical imaging literature [8]. (3) Maximally-stable extremal regions (MSERs) [12], which are found as image areas that are stable with respect to the change of intensity thresholds. In the experiments, we used our own implementation of LoG and publicly-available MSER [12] and Harris corner [23] executables. The LoG implementation uses filtering techniques similar to [3]. We choose these three detectors (a) to represent effective local and region-based techniques and (b) because of their relative efficiency (e.g. over entropy-based methods such as [10]) both in practice and in running a large suite of experiments.

3. Descriptors

The SIFT descriptor [11] is computed by partitioning the image region surrounding each detected keypoint into a 4×4 grid of subregions, and computing an orientation histogram of 8 bins in each subregion. The grid is square, with the x-axis oriented along the keypoint gradient direction, and the width of the grid being approximately 12-times the detected scale of the keypoint. Within each subregion, the gradient orientation of each pixel is entered into the orientation histogram, with weighted vote proportional to the gradient magnitude. A normalized 128-component vector is formed by concatenating the 16 region containers. Keypoints are matched between images or between an image and a keypoint database by minimizing the distance between descriptors. A ratio test, comparing the distances between the best and the second best match for a given keypoint, is used as a measure of match quality. Originally, all keypoints with a ratio below 0.8 were considered strong candidates for being correct [11]. In the Dual-Bootstrap [25], keypoints are instead ranked-ordered by this distinctiveness measure.

While the SIFT descriptor is invariant to linear changes in intensity, the image pairs we consider here involve nonlinear changes. The questions we address are (1) how badly does this affect performance, and (2) can anything be done to improve performance. Under the latter category, we consider two alternatives, moving toward a structural view of the keypoint neighborhood:

Gradient Mirroring (GM): The first alternative simply associates anti-parallel gradient directions and therefore considers gradient directions in the interval $[0, \pi)$ instead of $[0, 2\pi)$. In effect this makes the descriptor invariant to contrast reversals. While this makes sense for many multimodal image pairs (Fig. 2), there is an associated loss of information — the descriptor is now length 64. Similar steps have been used in object detection applications [5], where the goal is determining if an image (or image region) of an object is an instance of the class.

Edge Precursors (EP): The second alternative adopts the shape-context [1, 17, 18, 22] idea of using only detected points. Features are found by (1) computing the gradient outer product matrix over a small neighborhood at each pixel, (2) computing the trace of this matrix at each pixel, and (3) selecting pixels that are local maxima of this trace along the dominant direction of their matrix. This is similar to the edge computation described in the original Harris corner detector paper [7]. The resulting points may be viewed as edge precursors. This realizes the intuition that the information preserved under modality and strong illumination changes is primarily along the boundaries. In computing the actual descriptor, the 4×4 SIFT grid is used with gradient directions in the interval $[0, \pi)$, as above. Thus, the primary difference with gradient mirroring is that a selected subset of pixel locations is used to

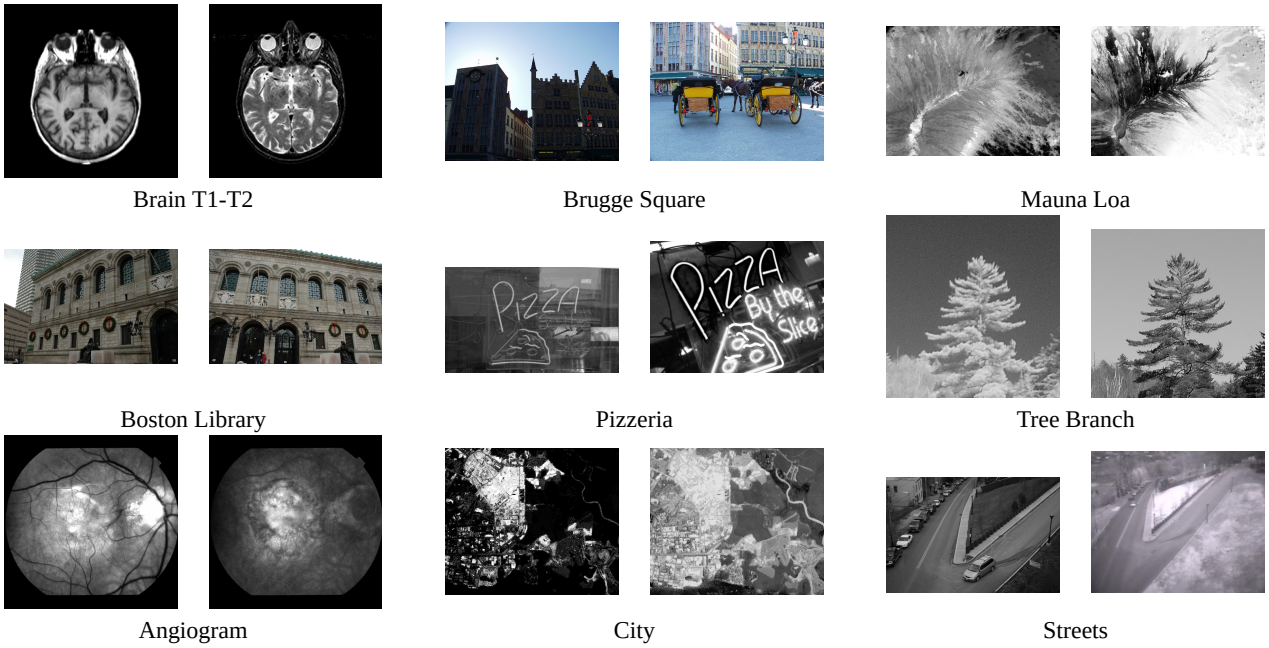


Figure 2. Several example pairs from our 21 image-pair dataset. These include T1 and T2 phases of MRI (Brain T1-T2), different camera exposures (Brugge Square), thermal and short wave IR (Mauna Loa), a more straight-forward pair (Boston Library), a neon sign during the day and in the evening (Pizzeria), a visible spectrum and IR (Tree Branch) pair, different phases of retinal imaging modalities (Angiogram), multispectral and visible spectrum images of a city (City), and a digital camera and a web cam with removed IR filter (Streets). See `http://www.vision.cs.rpi.edu/keypoints/` for the complete dataset.

compute the descriptor.

We have tried two other techniques: (1) ignoring gradient magnitudes when using the edge-based technique, similar to earlier versions of shape-contexts [1], and (2) taking the structure preservation idea further by using the output of the Canny edge detector. In both cases the results are nowhere near competitive with the techniques examined carefully here, so we do not discuss them.

Overall, we now have two alternatives for the descriptor, SIFT-GM and SIFT-GMEP, to use each with the scale-space LoG detector, the MSER detector, and the multiscale Harris corner detector. Together with the original SIFT this gives nine detection/descriptor combination methods to test.

4. Experiments

Our experiments evaluate keypoint detection and matching for multimodal image pairs and for image pairs involving strong illumination changes. In this evaluation it is important to keep two goals in mind. The first is maximizing the overall number of correct keypoint matches. This is consistent with object detection [5], object recognition [6, 18, 24], and keypoint-based registration [2]. The second goal is ensuring that at least a few keypoints are matched correctly and that these matches occur near the top of the rank ordering. This enables the success of registration algorithms such as the Dual-Bootstrap [25] that are capable of aligning images starting from just one keypoint match, but consider several different initial matches.

4.1. Data Set

We collected an experimental dataset of image pairs (Fig. 2), including 8 pairs taken by different sensors, 5 medical image pairs, and 9 visible spectrum pairs, of which 3 have extreme illumination change. Image sizes range from 352×233 to 2532×2483. This dataset represents a sampling of the endless variation in possible appearance and modality changes, and therefore our results must be viewed as suggestive of the effectiveness of keypoint detection and matching in any context.

In order to generate our results, a verified transformation was obtained for each image pair by running the publicly-available Dual-Bootstrap executable [25] and supplying manual initial estimates for pairs that failed to initialize automatically. The transformation was manually verified as being correct by careful examination of the final alignment result. Most of the pairs chosen are quite difficult, but a few are relatively easy. These were included to measure the potential loss in performance on more straightforward pairs when attempting to improve performance on challenging ones.

4.2. Detection Repeatability

In evaluating detection repeatability, each tested detector is applied to each image separately, and then the verified transformation is applied to the moving image to map its keypoints into the fixed image. We can then find the clos-

est fixed image keypoint to each mapped keypoint. If the location is (generously) within 6 pixels, the orientation difference is within 10 degrees, and the ratio between scales is within $0.67 - 1.5$, we consider the match to be "correct". This notion of correctness is generally sufficient for the Dual-Bootstrap to succeed, although it is perhaps a little too generous for the keypoints-only registration technique of [2]. We use this definition of correctness throughout the experiments.

The ideal analysis would be an ROC curve for each of the detectors, summarized across all image pairs. We found this meaningless to generate because there is an extremely-wide fluctuation in the performance across different images. Therefore, we set an operating point of the LoG and Harris detectors by limiting the total number of points and of the MSER detector by choosing the default parameters. Raw numbers and percentages of repeated detections on a per-image-pair based are summarized in Table 1. Observe that the overall percentage of correct matches varies from as low as 0.5% to as high as 48.5%. On one-third of the image pairs all detectors had less than 10% repeatability. Next, note that the LoG consistently had the most matches and MSERs the fewest, but when we compare repeatability percentages, the results are mixed. Finally, which detector does better overall depends on which image pairs are considered.

Technique	Min / Out of	Avg / Out of	Max / Out of
LoG	20 / 2302	279 / 1891	793 / 2027
Harris	10 / 2000	219 / 1452	771 / 2000
MSER	7 / 375	131 / 1391	699 / 1961

Table 1. Summary of keypoint detection results. Minimum, average, and maximum numbers of correct matches out of how many were possible for all pairs in the dataset. MSERs produce the least number of correct matches overall.

4.3. Descriptor Experiments

In evaluating the descriptors, for each image pair we compute keypoints and their descriptors in each image and then match them between the two images. For each keypoint, the top two matches are found and the descriptor distance ratio between these two is computed as the "distinctiveness" measure of the best match. The set of best matches for all keypoints is rank-ordered by distinctiveness. The best 100 are then evaluated to determine which are correct. The number correct among these is the starting point for comparison across methods. We assess the performance of the two variations of the descriptor using LoG keypoints, MSER keypoints, and Harris corners and compare it to the original SIFT.

The results are evaluated in several ways. First, the raw numbers are presented using a bar chart in Fig. 3. This shows several things: (a) that no one method is better than any other on all pairs, and (b) that the number of correct matches varies dramatically across the data set, and (c) that there are some image pairs for which certain combinations produce no correct matches.

Image pair name	SIFT, LoG	SIFT, corners	SIFT, MSER	SIFT-GM, LoG	SIFT-GM, corners	SIFT-GM, MSER	SIFT-GMEP, LoG	SIFT-GMEP, corners	SIFT-GMEP, MSER
Angiogram	4	1	1	8	4	1	27	10	45
Bay	1	1	1	2	1	10	1	29	1
Boston	1	1	1	1	1	1	1	1	1
Boston Library	1	1	1	1	1	1	1	1	1
Brain T1-T2	76	–	24	3	2	1	2	5	8
Brain T1-PD	1	6	2	2	3	8	1	11	5
Brain T2-PD	1	1	1	2	2	1	1	3	1
Brugge Square	1	1	1	1	1	1	1	12	1
Brugge Tower	1	1	1	1	1	2	1	1	1
Capital Region	52	–	2	–	–	4	–	–	–
City	1	1	6	1	1	1	1	4	1
Day Night	6	5	1	5	–	1	7	5	7
EO-IR-1	90	–	1	–	58	13	28	55	74
EO-IR-2	2	31	11	4	2	1	3	15	9
Grand Canyon 2	1	17	1	1	1	1	1	1	1
Mauna Loa	1	1	2	1	1	1	1	2	6
MR-CT	90	1	7	36	6	17	17	3	3
Pizzeria	1	4	2	1	1	2	1	3	7
Satellite	1	1	1	1	1	1	1	2	1
Streets	1	56	–	2	7	12	1	2	2
Tree Branch	–	–	–	2	1	1	1	3	2
White Tower	1	1	1	57	3	1	1	87	1

Table 2. The position of the first correct keypoint in the rank-ordering for each detector-descriptor and each image pair. For many pairs the first available keypoint match is the correct one and the position improved dramatically for difficult pairs (e.g. Brain T1-T2, EO-IR-1). SIFT-GM computed at MSERs gives at least one correct match for all pairs.

Our second evaluation is motivated by the observations that the Dual-Bootstrap algorithm needs only one match to start its registration process (it tests multiple such initializations), and the algorithm produces an accurate final transformation for more than 80% of the correct initial matches. Thus, for the different detector-descriptor combinations we evaluate the position of the first correct match in the rank-ordering for each detector-descriptor combination. It can be seen from Table 2 that this is the top-ranked match for many pairs. On the other hand, for several difficult image pairs (e.g. Brain T1-T2, EO-IR-1) the position of the correct match improved dramatically for all descriptor modifications. SIFT-GM computed at MSERs gives at least one correct match for all pairs and the position of the first correct match found is close to the beginning. Therefore it performs best in this ranking.

The third method of evaluation, presented in Fig. 5, at-

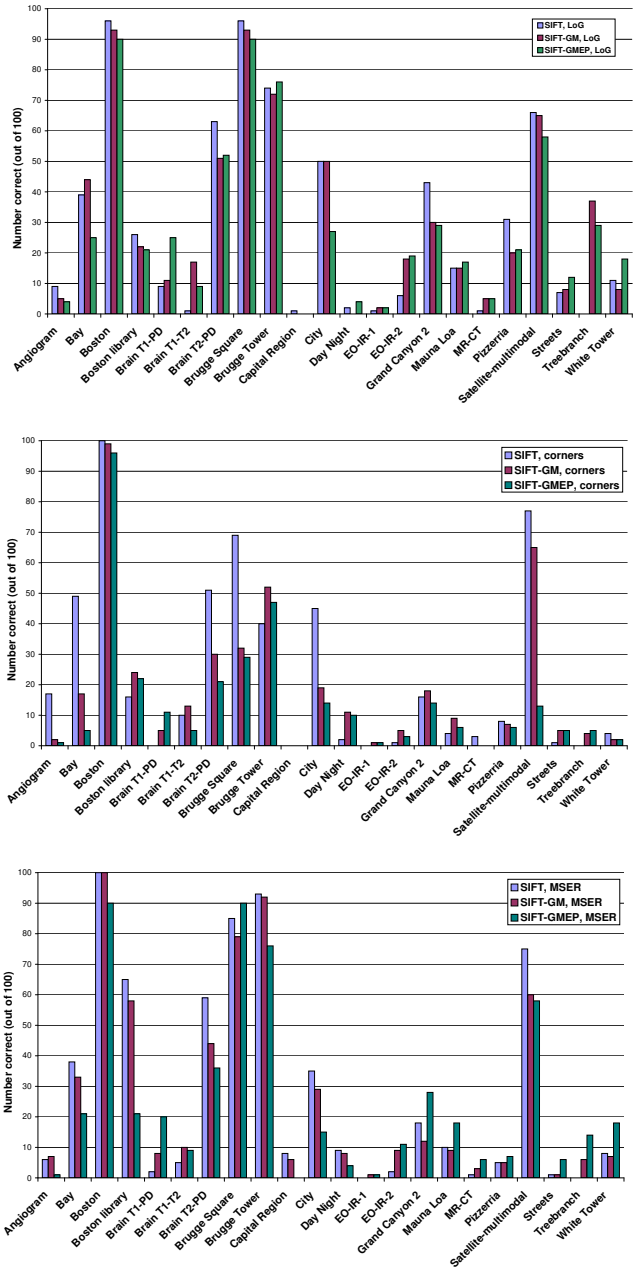


Figure 3. Bar charts showing the number of correct keypoints among the top 100 for each of the methods. For ease of reading, this is split across descriptors computed for LoG keypoints, descriptors computed for corners, and descriptors computed for MSERs.

tempts to increase understanding of these results. The vertical axis in the figure represents the number of correct keypoint matches, k, in the top 100, while the horizontal axis represents the number of image pairs, p. A curve is plotted for each detector-descriptor combination. A point (k, p) on the curve means that p image pairs had at least k correct matches. Thus, if the curve for one combination is consistently higher than the curve for another, it means there are more pairs with that number of correct matches. Particular attention should be paid to *larger values of* p.

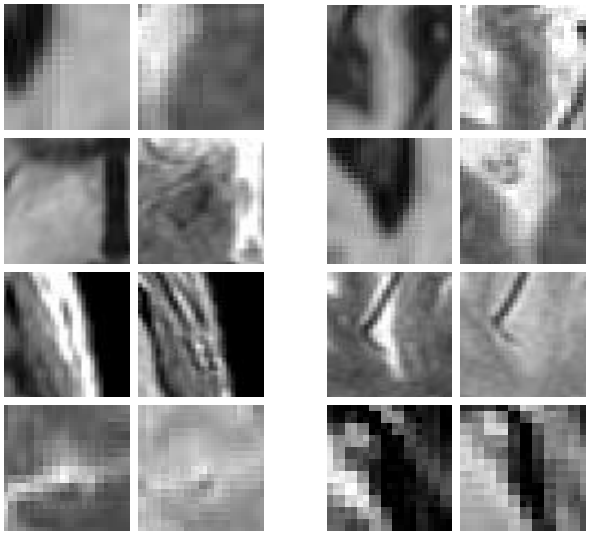


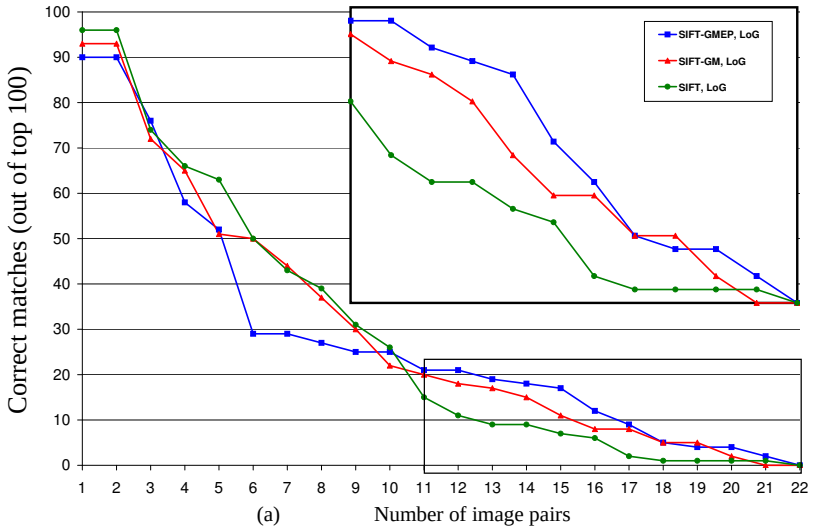
Figure 4. Example of correctly matched keypoint regions using SIFT-GM evaluated at LoG keypoints. Notice the dramatic non-linear changes between images within the pairs. The regions were resized for display.

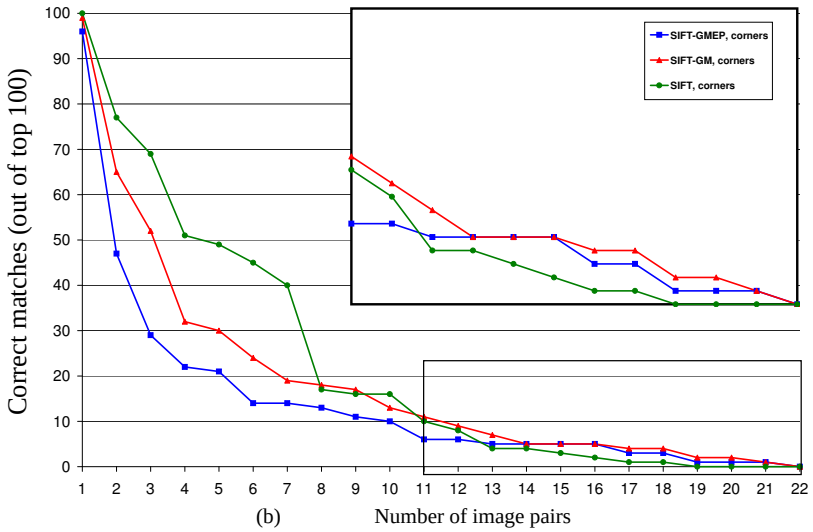
This plot now makes clear the trade-offs introduced by the different descriptors considered. By moving away from standard SIFT, using Gradient Mirroring (GM) and using Edge Precursors (EP), we obtain better performance on difficult pairs (as seen on the right side of the curve) at the expense of reduced matches on easier pairs. When a registration algorithm needs only one or two correct keypoint matches (or even just a few) to initialize the successful alignment of an image pair, this is an acceptable trade-off, especially since the trade-off does not tend to affect the rank-ordered positions of the top matches. Thus, all descriptor modifications proposed here give better results on harder pairs than the standard SIFT. Overall, SIFT-GMEP at LoG locations gives the best results in terms of the number of correct matches even though, as seen in Table 2, SIFT-GM at MSERs is most successful in terms of position of the first correct keypoint. This suggests that success of a detector-descriptor pairing is not dependent solely on the individual success of the detector or the descriptor. Interestingly, MSERs performed the worst in terms of detection repeatability (Table 1). The types of regions it detects, however, are highly descriptive which makes them successful when used in matching. Examples of the descriptor regions for several correct matches are in Fig. 4.

5. Conclusions

Several conclusions may be drawn from our experimental results.

- While keypoint detectors and the SIFT descriptor were designed under the assumption of linear changes in in-





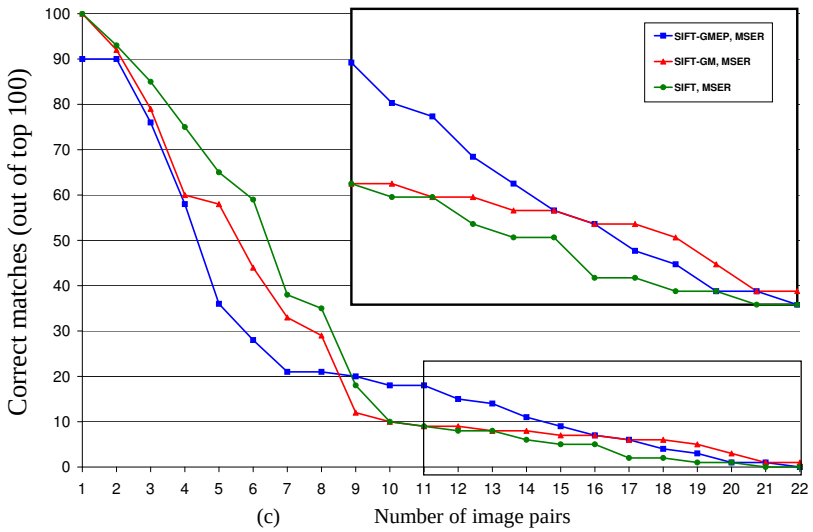


Figure 5. Descriptors evaluated at **(a)** LoG, **(b)** corners, and **(c)** MSERs. Plot of the number of correct matches, *k* (vertical) and the number of image pairs, *p*, (horizontal) having *at least* that number of correct matches. Higher curves are better, especially on the right side of the plot.

tensity, they can be effective in matching image pairs taken under substantially different illumination conditions and even changes in modality. Here, effectiveness means that they produce at least a few correct matches whose distinctiveness measure ranks them near the top. This is sufficient for these image pairs to be accurately aligned by algorithms such as the Dual-Bootstrap.

- The effectiveness of the SIFT descriptor in matching of challenging image pairs may be improved by equating anti-parallel gradient directions (SIFT-GM) and focusing the calculation on edge precursors (SIFT-GMEP). This is achieved with a negligible loss in performance for easier image pairs.
- The repeatability of keypoint detection under changes in illumination and modality is disappointingly low, reinforcing a result reported in [25] that corner matching alone is not sufficient for the most difficult image registration problems.
- No one keypoint detector is most effective for all pairs, suggesting that a combination of detectors be used in practice.
- Finally, while keypoint detection, description, and matching on challenging image pairs are usually effective for initializing some registration algorithms, it seems quite unlikely that they are sufficient for recognition algorithms that depend on large numbers of correct keypoint matches.

The final point clearly highlights a challenge for future work on keypoint detection and description.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(4):509–522, June 2002. 2, 3

[2] M. Brown and D. Lowe. Recognising panoramas. In *Proc. ICCV*, 2003. 1, 2, 3, 4

[3] M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. In *Proc. CVPR*, volume 1, pages 510–517, 2005. 2

[4] O. Chum and J. Matas. Geometric hashing with local affine frames. In *Proc. CVPR*, volume 1, pages 879–884, New York, NY, USA, 2006. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 2, pages 886–893, 2005. 2, 3

[6] A. Frome, D. Huber, R. Kolurri, T. Buelow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *Proc. Eigth ECCV*, 2004. 3

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988. 2

[8] T. Hartkens, K. Rohr, and H. S. Stiehl. Evaluation of 3d operators for the detection of anatomical point landmarks in mr and ct images. *Comput. Vis. Image Und.*, 86(2):118–136, May 2002. 2

[9] T. Kadir and M. Brady. Saliency, scale and image description. *Int. J. Comp. Vis.*, 45(2):83–105, 2001. 2

[10] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. Eigth ECCV*, 2004. 2

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vis.*, 60(2):91–110, November 2004. 1, 2

[12] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *IVC*, 22(10):761–767, Sept. 2004. 2

[13] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. J. Comp. Vis.*, 60(1):63–86, 2004. 2

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(10):1615–1630, 2005. 1

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int. J. Comp. Vis.*, 65(1–2):43–72, 2005. 1, 2

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

[17] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(11):1832–1837, 2005. 2

[18] G. Mori and J. Malik. Recognizing objects in adversarial clutter: breaking a visual captcha. In *Proc. CVPR*, volume 1, pages 134–141, 2003. 2, 3

[19] E. N. Mortensen, H. Deng, and L. Shapiro. A sift descriptor with global context. In *Proc. CVPR*, volume 1, pages 184–190, 2005. 1

[20] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Comp. Vis.*, 66(3):231–259, Mar. 2006. 1

[21] C. Steger. Occlusion, clutter, and illumination invariant object recognition. *Int. Arc. Photo. Remote Sensing*, XXXIV, part 3A:345–350, 2002. 1

[22] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *Proc. CVPR*, volume 1, pages 127–133, 2003. 2

[23] Visual Geometry Group. Affine covariant features. `http://www.robots.ox.ac.uk/~vgg/research/affine/`. Last accessed 14 March 2007. 2

[24] Štěpán Obdržálek and J. Matas. Sub-linear indexing for large scale object recognition. In *Proc. British Machine Vision Conf.*, pages 1–10, 2005. 1, 3

[25] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai. Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence. *IEEE Trans. Pattern Anal. Machine Intell.*, 2007. Accepted for publication. 1, 2, 3, 7