Structure Guided Salient Region Detector

Shufei Fan, Frank Ferrie
Center for Intelligent Machines
McGill University
Montréal H3A2A7, Canada

Abstract

This paper presents a novel method for detection of image interest regions called Structure Guided Salient Regions (SGSR). Following the information theoretic route to saliency detection, we extend Kadir et al.'s *Salient Region* detector by exploiting image structure information. The detected SGSRs are highly distinct and very selective. For planar scenes, their performance on repeatability tests under viewpoint changes is comparable to the state of the art. For 3D scenes, SGSRs are more likely to be repeatably detected under viewpoint change. Their usefulness for wide baseline matching is demonstrated with a real-world example, where their comparative advantages are shown.

1 Introduction

Under wide baseline conditions, two stereo images of the same scene will differ significantly due to occlusion, geometric distortion, and other effects such as illumination change. Local features are the main source of information for establishing image correspondence. Many efforts have been put into finding the local image features that are most likely to be repeatably detected under viewpoint and possibly illumination changes [7, 1, 13, 12, 9, 14, 2, 10, 6].

Many researchers have used corner and edge information to extract features that are likely to be repeatably detected despite viewpoint and scale changes [13, 12, 14, 15]. Some other works find repeatably detectable scene structures based on image intensity, such as the *intensity extrema-based region* detector [13, 12] and the intensity induced *maximally stable extremal region (MSER)* [9]. Line segments are used by Bay et al. [2] to obtain the planar homography, which in turn facilitates epipolar geometry estimation. A convincing attempt to use repetitive patterns as features was made by Chetverikov et al. [3].

Following the seminal work of Lindeberg [7], a family of affine covariant feature detectors was proposed [7, 1, 10]. They typically start by detecting interest points across different scales as candidates. Then each of the candidates is further examined with regard to its scale invariance while simultaneously refining its affine parameters (by affine normalization). The second moment matrix of the intensity gradient is used to find the neighborhood structure of each feature. Combined with the scale selection method, these approaches can find affine covariant interest regions quite accurately. Lowe's *SIFT* detector [8] can repeatably detect features under similarity transforms at their characteristic scales. A different avenue along information theory is explored by Kadir et al. [6] and

BMVC 2008 doi:10.5244/C.22.43

their entropy-based saliency measure is able to select salient elliptical regions (we call them *Salient Regions*) at appropriate scales.

Successful affine covariant detectors can find the same scene structure even though the images undergo scale, viewpoint and illumination changes. According to the findings of a recent benchmark [11], MSER and Hessian-Affine detectors perform consistently better in most of the repeatability and matching score tests. In comparison, *Salient Regions* have poor performance in these respects.

Like many other benchmark efforts, for the convenience of obtaining ground truth, detector comparisons (in [11]) are carried out under simplified circumstances. In the case of comparing repeatability under viewpoint changes, all images are taken from one planar scene from different viewing angles[1]. Affine normalization based detectors need a large neighborhood region to obtain a feature's local structure. If the scene structure of an image feature is indeed locally planar, these methods can detect the same scene structure adapted to different viewpoints with elegant affine warps. Concentrating on local complexity of image patches, the *Salient Regions* detector uses an entropy-based saliency definition. Since its saliency measure is independent of geometric considerations, regions detected by this criterion could be applicable to more general scenes.

We follow the route of information theoretic saliency and propose a different salient region detector called the Structure Guided Salient Region (SGSR) detector. The SGSR detector makes full use of the local intensity structure and intensity probability distribution of regions. It will be shown to be advantageous in two respects: (1) repeatability under viewpoint change using benchmark images provided by Mikolajczyk et al. [11], and (2) real-world application to wide baseline matching.

The outline of this paper is as follows. After reviewing work by Kadir et al. in Section 2, we will describe the proposed method in detail in Section 3. Then, Sections 4 and 5 evaluate its performance in the cases of planar scenes and 3D scenes respectively. Finally, we summarize the conclusions to be drawn in Section 6.

2 Background

Essentially, *Salient Regions* are regions that locally assume maximal signal complexity and at the same time exhibit self-dissimilarity in scale space [5]. The signal complexity is measured by the Shannon entropy (denoted by H) of the local intensity histogram. The self-dissimilarity is approximated by the change of the probability density function (*pdf*) in scale space (denoted by W).

A region's scale saliency Y is defined as the product of the two factors H and W, all of which are functions of scale s and position $\mathbf{x}$. Using $p(d,s,\mathbf{x})$ to describe the region's intensity *pdf* at position $\mathbf{x}$ and scale s, we give the mathematical definitions of Y, H, and W as follows:

$$Y(s,\mathbf{x}) = H(s,\mathbf{x})W(s,\mathbf{x}); \tag{1}$$

$$H(s,\mathbf{x}) = -\sum_{d\in D} p(d,s,\mathbf{x}) log(p(d,s,\mathbf{x})); \tag{2}$$

$$W(s,\mathbf{x}) = \frac{s^2}{2s-1}\sum_{d\in D} |p(d,s,\mathbf{x}) - p(d,s-1,\mathbf{x})|. \tag{3}$$

[1] for example, the *graffiti* image sets used in the benchmark work [11]

In equations (2) and (3), D is the set of possible intensity values.

The *Salient Regions* detector was later generalized to be invariant to affine transforms induced by viewpoint changes [6]. This invariance is achieved by replacing the circular sampling window (parameterized by scale s) with an ellipse, which is summarized by the vector $\{s, r, \theta\}$, where s is the scale, r is the aspect ratio of the major axis versus the minor axis of the ellipse, and θ is the orientation of the major axis. Brute-force searching over the three-parameter space can be very expensive. Therefore, Kadir et al. proposed a seeding and local adaptation approach. They start by finding seed regions conforming to the original saliency criterion using circular sampling windows. The seed regions are then locally adapted by searching for optimal s, r and θ values (equivalent to deforming the seed circles to an ellipse at an optimal scale s), to maximize the regions' saliency measure. This local adaptation method greatly improves efficiency.

However, there are a few drawbacks to the above-mentioned method. First, the circular sampling window used in the seeding procedure may prefer isotropic structure to anisotropic structure. This bias may contribute to low repeatability scores under viewpoint change. Because a change of viewing angle will skew isotropic structures in one image to anisotropic ones in the other, they do not get equal chance of being detected. Second, feature locations detected with circular sampling windows will need additional adjustment to fine-tune the center of the deformed region. This positional refinement was not conducted in the original work. Nevertheless, the authors' innovative attempt at introducing information theory into feature detection is in line with human attention to features. We believe this saliency measure may capture more of the intrinsic structures in the scene and is more likely to be repeatably detected under a wide baseline condition.

3 The Structure Guided Salient Region

Based on the theory of entropy-based saliency for identifying features, we propose a different route to salient region detection using seeding with local structure. To be specific, we propose a two-step detection procedure of seeding and local saliency detection, where our seeding will take into account local intensity structures of the image.

We will first briefly describe our representation of features in Section 3.1. Then, we will present the two steps in Sections 3.2 and 3.3 respectively. Lastly, in Section 3.4 we will introduce our method for robustly estimating the region *pdf*.

3.1 Representation of the Scale And Affine Invariant Features

We describe a scale and affine invariant feature by $f_l = \{\mathbf{x}_l, s_l, t_l, v_l\}$, where $\mathbf{x}_l$ is a 2×1 vector $(x_0, y_0)^T$, signifying the center of the feature region; s_l is a scalar describing the feature's scale; and t_l indicates the shape of the image region covered by this feature. We represent t_l by a normalized 2×2 symmetric matrix $\begin{pmatrix} A & B \\ B & C \end{pmatrix}$, called the structure tensor. Its symmetry and normalization reduces the two-by-two matrix to two degrees of freedom. It is equivalent to representing an elliptical shape by aspect ratio and orientation, but the tensor representation is more convenient to work with in our case. Finally, v_l contains the descriptor values for this feature.

In essence, image feature detection is the estimation of $\{\mathbf{x}_l, s_l, t_l, v_l\}$ for all points of interest. Feature matching is the process of establishing correspondences between

features from two images by examining similarity of the feature descriptor values v_l.

3.2 Seeding Using Local Structure

Since the desired salient features should have a relatively large change of *pdf* over scale, they typically are image blobs that have large intensity variation with respect to their surrounding pixels. We propose to use these blobs as seeds for saliency detection.

Scale invariant blob detection techniques can be used to extract blobs. For example, Lindeberg [7] detected blobs by searching for local extrema of Laplacian-of-Gaussian filtered images in scale space. But this method detects circular-shaped blobs only. For arbitrary blob shapes, one needs an affine-invariant blob detector like the Hessian-Affine detector [11]. But its affine-adaptation will need to compute the structure tensor of a region's neighborhood, which is usually much larger than the region itself. For images of 3D scenes, this large neighborhood is likely to cover surface depth change, in which case the local neighborhoods are no longer covariant to affine transform.

We use blobs detected by MSER [9] as our seeds. Since their detection procedure relies solely on image intensity contrast, those with high intensity variation with respect to their surrounding neighbors are preferred over those with low contrast. We loosen the requirement on neighbor contrast by lowering the minimum margin between inner and outer regions. This will result in a large collection of regions, many of which may be detected due to noise. These noisy regions will be eliminated when their statistical properties are further examined, as will be described in detail in the next section.

One interesting property of these seeds is that their shape is readily obtained by analyzing the region boundary. The region is enclosed by an ellipse, represented by the seed's location $\mathbf{x}_l = (x_0, y_0)^T$, scale s_l, and tensor $t_l = \begin{pmatrix} A & B \\ B & C \end{pmatrix}$. The ellipse is defined by the quadratic equation:

$$(\mathbf{x} - \mathbf{x}_l)^T \begin{pmatrix} A & B \\ B & C \end{pmatrix} (\mathbf{x} - \mathbf{x}_l) = s_l^2. \tag{4}$$

3.3 Local Salient Region Adaptation

Now that we have obtained the initial set of feature seeds $\mathscr{F} = \{f_1, ...f_N\}$, where $f_l = \{\mathbf{x}_l, s_l, t_l\}$, $l \in 1, ...N$, we will examine their saliency as defined in Equation (1). We will also locally adapt the seeds to choose the position and scale for which they achieve optimal saliency. Since the region boundary already gives a good estimate of the elliptical shape, we will keep the t_l fixed during the optimization. In the adaptation, we will maximize the two criteria, H(region entropy) and W(inter-scale saliency), by interleaving *scale saliency selection* with *location refinement*.

We start each seed with a *scale saliency selection*. If the initial seed is scale salient (has local H maximum), it will undergo *local adaptation*; otherwise, it will be discarded. For seeds passing the initial scale saliency test, *local adaptation* will end when either maximum H and W are found or the iteration limit is encountered.

Scale Saliency Selection When choosing the optimal scale of a seed region $f_l = \{\mathbf{x}_l, s_l, t_l\}$, we look for a local maximum of $H(s_l, \mathbf{x}_l)$ by changing the scale s_l while keeping the location $\mathbf{x}_l$ fixed. If there exists a local maximum at scale s'_l, we update this seed's scale to

s_l' and proceed with location refinement. If no maximum is obtained, this seed is regarded as non-salient and discarded.

Since we have already obtained a rough scale in the seeding step, we can search more efficiently thanks to two simplifications. First, the search range of s_l' can be set to be small. This is in contrast to the original scale-saliency method [6], where a large search space is needed in order to capture all possible *salient regions*. Second, we can stop searching once we encounter the first local maximum H. This is because we are already working in a predefined narrow range of scale and the first characteristic salient scale already gives us a tight bound of the interest region.

Position Refinement Once the seed's optimal scale is obtained, we maximize the seed's $W(s,\mathbf{x})$ by looking for the nearest neighbor that has a higher $W(s,\mathbf{x})$. Within a certain range, if there is a region at $\mathbf{x}_l'$ that has a larger $W(s,\mathbf{x})$, we move the seed to this position (by updating $\mathbf{x}_l$ with $\mathbf{x}_l'$). After position adjustment, we go back to the previous step to see if any scale adaption is needed. If, on the other hand, no neighbor has a better $W(s,\mathbf{x})$, we stop the iteration and take the current $\mathbf{x}_l$ as the optimal position.

3.4 Robust Histogram Estimation and Extension To Color Image

Region intensity histogramming is used for estimating the local *pdf* over the elliptical sampling window. For an 8-bit grayscale image, for example, a 256-bin histogram is used to count the number of occurrences of pixels with gray levels from 0 to 255. We find, however, that the region's local intensity histogram is very sensitive to noise. This sensitivity is more evident when the region is small, since only a small number of pixels are used in filling the histogram and small gray-level deviations of some of them will change the overall histogram significantly.

We tackle this problem by applying Gaussian smoothing and sub-sampling to the initial intensity histogram. The smoothing window size is related to the sub-sampling factor. Here, for grayscale images we use a sub-sampling factor of 4 by representing the smoothed histogram with a 64-bin histogram. This procedure makes salient region intensity *pdf* estimation more robust to noise.

More importantly, this robust estimation makes SGSR's extension to color images practical. The original formulation of scale saliency is applicable to color images. In practice, however, one would have to work on a histogram of dimension 16777216 ($256 \times 256 \times 256$) with a normal RGB image. This demands prohibitive resources and the representation will be very sensitive to noise. With a sub-sampling factor of 16 for RGB color images, we will end up working with 4096-dimensional ($16 \times 16 \times 16$) histograms.

4 Performance Evaluation on Planar Scenes

The objective of performance tests on planar scenes is to evaluate the extent to which SGSRs commute with viewpoint. We use the testing methodology and *graffiti* image set proposed in [11]. In testing performance under viewpoint changes, we ran the SGSR detector on a set of images of the same planar scene acquired from different viewpoints. The homographies between the images are given as ground truth.

Here, we test SGSR against the state-of-the-art detectors reported in [11]: Hessian-Affine detector, Harris-Affine detector, MSER detector, Intensity Extrema-based Region

detector, and Edge-based Region detector. We compare them on four performance indicators: the number of correspondences, the repeatability, the number of correct matches, and the matching score (as defined in [11]):

- *The number of correspondences* is the absolute number of region pairs (between the reference image and the matching image) which are repeatably detected. Two regions are deemed to be repeatably detected if the overlap error ε_O is sufficiently small (in this experiment, we choose $\varepsilon_O \leqslant 40\%$). The overlap error is defined as the error in the feature areas when the two corresponding regions are converted to a common coordinate frame according to the homography:

$$\varepsilon_O = 1 - \frac{R_{\mu_a} \bigcap R_{H^T \mu_b H}}{R_{\mu_a} \bigcup R_{H^T \mu_b H}}, \tag{5}$$

  where H is the homography relating the two images, and $(R_{\mu_a} \bigcap R_{H^T \mu_b H})$ and $(R_{\mu_a} \bigcup R_{H^T \mu_b H})$ represent the area of intersection and union of the regions respectively.

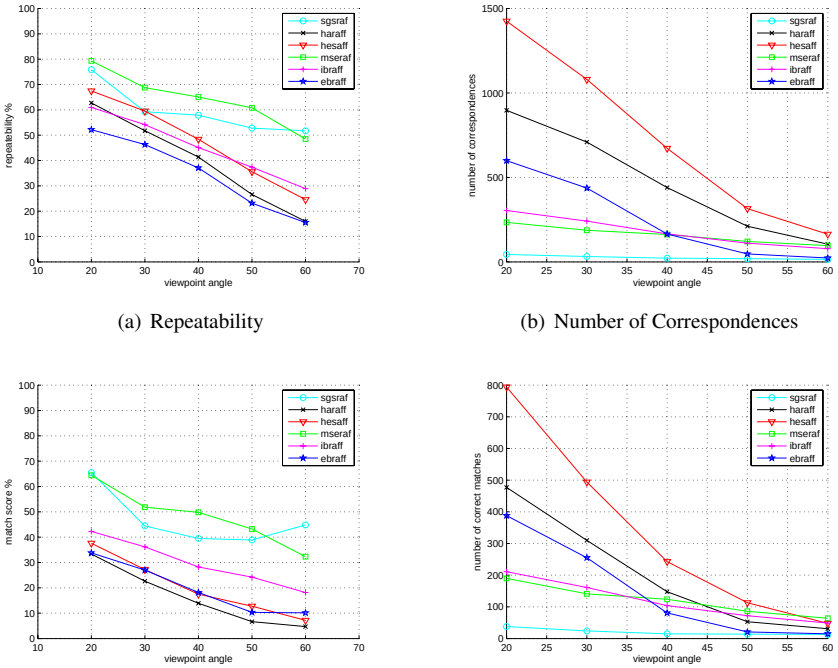
- *Repeatability* is the ratio between *the number of correspondences* and the smaller of the number of detected regions in the pair of images.

- *The number of correct matches* is the total number of correct matches among the correspondences. A region correspondence is deemed correct if the overlap error is minimal and less than a predefined threshold ($\varepsilon_O \leqslant 40\%$). This is the ground truth for correct matches in the *matching score* comparison.

- *The matching score* is meant as an indication of the distinctiveness of features detected by a particular detector. The idea is to see how well the regions can be matched, when all are represented by SIFT descriptors [8]. A match is the nearest neighbour in the descriptor space according to their Euclidean distance. The *matching score* is defined as the ratio between the number of correct matches (obtained using SIFT descriptors) and the smaller number of detected regions in the pair of images. The results are indicative rather than quantitative, since they depend on many factors, one of which is the type of descriptor that is used in representing the feature.

Comparison Results The repeatability comparison results are reported in Figure 1(a), showing repeatability as a function of viewpoint change. SGSRs achieve competitive performance for most viewing angles, but rely on a relatively small number of features (Figure 1(b)). When represented by SIFT descriptors, SGSRs' matching scores are close to that of the best performer, MSERs, for smaller viewpoint angle changes, and 10% better than MSERs for a viewpoint change of 60° (Figure 1(c)). Again, this is achieved using a much smaller number of features (Figure 1(d)).

One distinction of the SGSR detector is that it achieves competitive results using the most compact set of features. This can be advantageous when applications (such as object or landmark recognition) require a compact representation, as we find that most detectors' performances decline when they are asked to detect a smaller set of repeatable features. It is shown in figure 21(c) of [11] that most detectors' repeatability falls with decreasing number of features used.



(a) Repeatability

(b) Number of Correspondences

(c) Matching Score

(d) Number of correct matches

Figure 1: **Performance Comparison** The detectors are compared on the *graffiti* image set; we show the 4 performance measurements of the detectors SGSR (denoted sgsraf), Hessian-Affine detector (hesaff), Harris-Affine detector (haraff), MSER, Intensity extrema-based Region detector (ibraff), and Edge-based Region detector (ebraff).

5 Performance Evaluation on 3D Scenes

The aim here is to measure our method's performance at detecting features in images of 3D scenes for the purpose of wide baseline matching. We will use one pair of images of the same 3D scene acquired from different angles and distances (we call it the *J-scene*, see Figure 2). It represents a typical outdoor scene containing mainly man-made structures, common to the pair are two buildings with walls (roughly) perpendicular to each other. For comparison, we apply three different feature detectors, Hessian-Affine, MSER, and SGSR, on the *J-scene* image set. To gauge the quality of features localized by each of the detectors, SIFT descriptors are used as a common basis for matching.

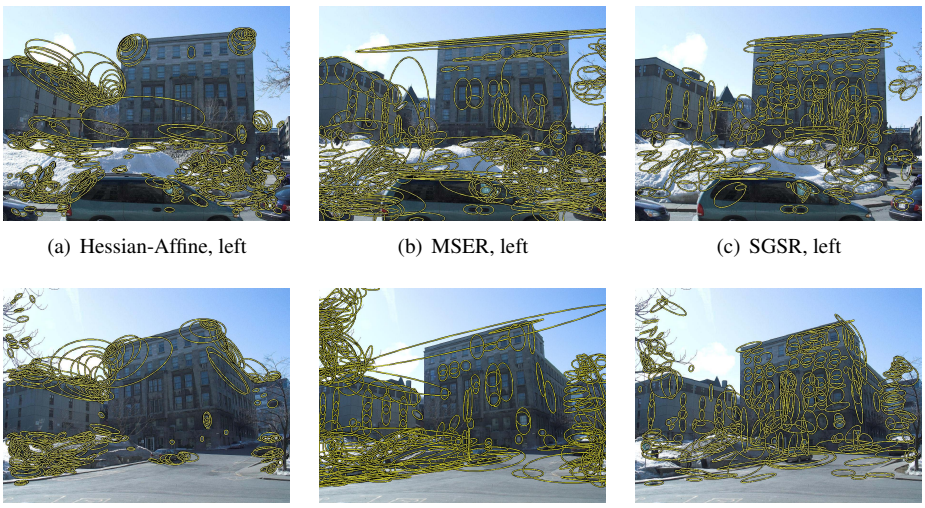
Feature Detection Results Figure 3 shows the features detected by the detectors. The Hessian-Affine features occur mainly in two places: corners and edges of buildings, where surface discontinuities occur; and snow-banks, which are densely textured and full of noise. In comparison, fewer MSERs occur on building edges and corners and more of them are detected on the building walls. MSERs are also densely detected on the snow-

(a) left (b) right

Figure 2: Stereo images of the *J-scene*

banks and tree branches. SGSR detector mainly captures blob structures on the building walls and much fewer of them occur in noisy parts of the scene such as snow-banks and tree branches.

The figure shows that the Hessian-Affine detector failed to detect structures such as windows and bricks on the wall. These blobs are close to each other and create a regular repetitive pattern. If we look at the only window detected (on the upper part of the front building in Figure 3(a)), it is isolated from its neighbors with distinct intensity. The MSER detector was able to extract some high-contrast blobs, but it also responded positively to many noisy regions. The SGSR detector captured most of the blob patterns on the walls and also discarded many noisy regions.



(a) Hessian-Affine, left (b) MSER, left (c) SGSR, left

(d) Hessian-Affine, right (e) MSER, right (f) SGSR, right

Figure 3: **Features Detected** The images show features detected by three different affine detectors: Hessian-Affine, MSER, and SGSR, on the *J-scene* images, .

Detector	# Features Detected (left-right)	# Total Matches	# Outlier Matches
Hessian-Affine	569-382	2	2
MSER	311-271	4	2
SGSR	266-258	15	2

Table 1: Feature matching comparison

Feature Matching Results For each detector, we perform a feature matching experiment with the following procedure. First, the features are normalized to a fixed-sized circular region and their SIFT descriptors are extracted. Second, we obtain the initial set of matches by nearest neighbour matching in the descriptor space. Finally, outliers are rejected by global consistency checking using RANSAC [4].

Table 1 gives the number of detected features, the number of matched features and the number of outlier matches found by the three detectors. We can see that SGSRs perform best for wide baseline matching of the *J-scene*. In contrast, MSERs and Hessian-Affine regions are poorly matched. Hessian-Affine regions are either not distinct enough (building corners will have similar SIFT descriptors) or not repeated in the scene (lower part of the images, such as noisy snow-banks and cars). Thus, no correct matching is found. Although the MSER detector repeatably captured some high contrast regions such as windows, their SIFT descriptor is not distinct enough due to large region sizes and different light reflectance of the corresponding window glasses (see windows on the side building in Figure 2). Finally, we show the correctly matched SGSR regions in Figure 4. Notice how their shapes are adapted to the scale and viewing angle changes.

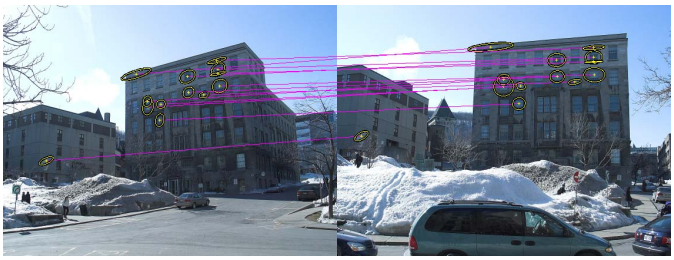


Figure 4: Matching result obtained with SGSRs

6 Conclusion

In this paper, we have presented a novel affine covariant feature detector based on entropy-based saliency theory. Our method is different from the original salient region detector of Kadir et al. [6] in both its initial seeding procedure and subsequent local region adaptation. We also introduced robust histogram smoothing and sub-sampling to cope with image noise and to extend SGSR's tractability to color images. This method's competitive performance is demonstrated in both planar and 3D scenes.

References

[1] A.M. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of Computer Vision and Pattern Recognition*, pages 774–781, 2000.

[2] H. Bay, V. Ferraris, and L. Van Gool. Wide-baseline stereo matching with line segments. In *Proceedings of Computer Vision and Pattern Recognition*, pages 329–336, 2005.

[3] D. Chetverikov, Z. Megyesi, and Z. Janko. Finding region correspondences for wide baseline stereo. In *Proceedings of International Conference on Pattern Recognition*, pages 276–279, 2004.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[5] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, November 2001.

[6] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proceedings of European Conference on Computer Vision*, pages 228–241, 2004.

[7] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[8] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, September 2004.

[10] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, October 2004.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[12] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85, August 2004.

[13] T. Tuytelaars and L.J. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proceedings of British Machine Vision Conference*, pages 412–425, 2000.

[14] J.J. Xiao and M. Shah. Two-frame wide baseline matching. In *International Journal of Computer Vision*, pages 603–609, 2003.

[15] J. Xie and H.T. Tsui. Wide baseline stereo matching by corner-edge-regions. In *Proceedings of International Conference on Image Analysis and Recognition*, pages 713–720, 2004.