Towards Embedded Robot Vision for Multi-scale Object Recognition
Repeatability of Interest Points Detected in Half-octave Binomial Pyramids

Peter Andreas Entschev and Hugo Vieira Neto
Graduate School of Electrical Engineering and Applied Computer Science, Federal University of Technology – Paraná
Avenida Sete de Setembro 3165, Curitiba, Brazil
peter@entschev.com, hvieir@utfpr.edu.br

Keywords: Repeatability, Interest points, Multi-scale pyramids, Embedded robot vision.

Abstract: The construction of multi-scale image pyramids is used in state-of-the-art methods that perform robust object recognition, such as SIFT and SURF. However, building such image pyramids is computationally expensive, especially when implementations in embedded systems with limited computing resources are considered. Therefore, the use of alternative less expensive approaches are necessary if near real-time operation is desired. Previous work has reported that using binomial filters to construct half-octave multi-scale pyramids consumes only $1/4$ of the processing time of the Gaussian pyramid originally used in the SIFT framework. Here we investigate how interest points detected using the binomial approach behave when compared to the Gaussian approach, focusing on repeatability. Experimental results show that in average up to 86% of interest points detected with the original SIFT pyramid building scheme are also detected when using the binomial method, despite of large gains in processing time. When rotation of image features is considered, experimental results demonstrate that slightly superior repeatability of interest points is achieved using the binomial pyramid.

1 INTRODUCTION

In the last decade, many research efforts were directed to robust object recognition in terms of invariance to scale, rotation and partial occlusion. Among some of the most well-known methods are SIFT (Lowe, 1999; Lowe, 2004) and SURF (Bay et al., 2008).

However, most methods designed for multi-scale object recognition are computationally expensive and memory consuming – therefore, implementations that run in real-time are difficult to be achieved, even in modern computer architectures. When it comes to performing multi-scale object recognition in embedded systems with limited computing resources, the difficulty in achieving near real-time performances is even more challenging.

Recently, physically small and low-power embedded systems based on ARM cores were made available at affordable costs, such as the BeagleBoard-xM (Coley, 2010) and Raspberry Pi (Halfacree and Upton, 2012), making them interesting platforms for such object recognition methods, especially when autonomous robot systems are considered. As these embedded systems consume little power at full load and are small in size, their computing power is naturally significantly smaller than the computing power available in most desktop computers.

Among the efforts being directed at enabling embedded systems to be capable of performing real-time multi-scale object recognition, there are some that are based on computationally cost-effective, but coarser approximations of already known methods available in the literature. For example, the work in (Entschev and Vieira Neto, 2013) aims at near real-time object recognition in an embedded platform, using a half-octave binomial image pyramid (Crowley et al., 2002) to represent multi-scale visual information, approximating the Gaussian image pyramid approach commonly used in object recognition methods, such as SIFT (Lowe, 2004) and other approaches (Mikolajczyk and Schmid, 2004).

Previous work shows that the use of half-octave binomial pyramids for multi-scale interest point detection is approximately four times faster than when using their Gaussian counterparts, yielding similar properties – real-time interest point detection at 25 frames per second can be achieved for images of 129×129 pixels in size (Entschev and Vieira Neto, 2013). Here, the repeatability of interest points detected using half-octave binomial pyramids is assessed in comparison to that obtained using conventional Gaussian pyramids.

2 RELATED WORK

Multi-scale representations of images using the notion of low-pass filtering and sub-sampling was first proposed in (Burt and Adelson, 1983) and has been exploited for about three decades now. One of the great advantages in the use of such a multi-scale approach is that it allows recognition of objects independently of the scale in which it appears in the scene. Much research work has been done in multi-scale object recognition since then, from which we can particularly cite SIFT, first introduced in (Lowe, 1999) and later extended and improved in (Lowe, 2004), and also SURF (Bay et al., 2008).

Both SIFT and SURF present processes to generate multi-scale image pyramids, followed by the extraction of robust local descriptors from a model object image, which can be subsequently matched to the descriptors of a scene to find the object position and affine orientation regardless of changes in scale and some degree of changes in illumination.

In SURF, a multi-scale image pyramid is obtained by first computing an integral image (Viola and Jones, 2001) and then processing it with coarse approximations of Gaussian derivative filters in different scales and orientations. As the processing effort to filter integral images is independent of the scale of the filter, overall computational cost is reduced when compared to traditional filter-and-subsample pyramid building schemes. The Hessian matrix is then applied locally at the pyramid in order to detect stable interest points, usually corresponding to blobs and corners, around which image information is encoded as a descriptor that is invariant to rotation.

The SIFT framework also builds a scale space, but using the traditional Gaussian filter-and-subsample approach at multiple scales. After filtering and subsampling, differences between adjacent levels of the Gaussian pyramid are computed in order to obtain a difference-of-Gaussian pyramid, which is a finer representation of derivatives when compared to the coarser method used in SURF. The sub-sampling process is used to avoid the use of unnecessarily large Gaussian filters, whose sizes grow with scale. The set of pyramid levels in which the scale of filtering is doubled is conventionally named an octave (Crowley et al., 2002).

The difference-of-Gaussian constitutes an approximation of the Laplacian of Gaussian (Burt and Adelson, 1983) and the main purpose of using it resides in its reduced computational cost when compared to the direct computation of the Laplacian of Gaussian (Crowley and Stern, 1984). Both difference-of-Gaussian and Laplacian of Gaussian approaches result in high-pass filtering of the original image, enhancing edges as their result.

A more computationally efficient way to build multi-scale image pyramids than the one used in the original SIFT approach uses binomial kernels instead of Gaussian kernels (Crowley et al., 2002). Although it is not possible to generate all possible scales using binomial filters, their resolution is enough to build image pyramids with scales sufficiently spaced by $\sigma = \sqrt{2}$ (Crowley and Riff, 2003).

More recently, the method presented in (Crowley et al., 2002) was used in (Entschev and Vieira Neto, 2013) to build SIFT image pyramids more efficiently. In this approach, the time spent to compute the whole multi-scale pyramid was reduced to up to one fourth of the time spent to build the same multi-scale pyramid using the original approach of SIFT.

The work presented in (Entschev and Vieira Neto, 2013) proposes the use of binomial filters to build image pyramids for SIFT-like object recognition, aiming at near real-time performance in a BeagleBoard-xM embedded development kit. That work explains how the scale space for the SIFT framework can be built more efficiently, focusing in execution time performance, but did not consider the repeatability of interest points. The present work intends to fill the gap left in (Entschev and Vieira Neto, 2013), assessing the repeatability of interest points detected using binomial image pyramids.

2.1 Binomial Pyramid

As described in (Entschev and Vieira Neto, 2013), there are two kernels of special interest for the construction of a binomial pyramid, $\frac{1}{4} \times [1\ 2\ 1]$ and its auto-convolution $\frac{1}{16} \times [1\ 4\ 6\ 4\ 1]$, which are approximations to Gaussian kernels of $\sigma^2 = \frac{1}{2}$ and $\sigma^2 = 1$, respectively (Crowley et al., 2002). The reason for such importance is that with just these two kernels, it is possible to build full scale spaces separated by $\sigma = \sqrt{2}$.

The relationship for cascaded convolutions of binomial filters is given in:

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}. \quad (1)$$

From Equation 1, it is possible to deduce that with two consecutive convolutions with the $\frac{1}{16} \times [1\ 4\ 6\ 4\ 1]$ kernel, an image with scale $\sigma = \sqrt{2}$ is obtained. The same scale is also obtainable by applying four consecutive convolutions with the $\frac{1}{4} \times [1\ 2\ 1]$ kernel. This important relationship determines that multiple convolutions of binomial kernels result in a scale space

separated by $\sigma = \sqrt{2}$, the so-called half-octave binomial pyramid.

Every time that the scale doubles, 1:2 nearest neighbour sub-sampling is performed in each dimension for the next octave, in order to avoid increasing the size of the filters and therefore saving computational resources (Crowley et al., 2002). In order to facilitate the computation of differences between pyramid levels, in (Entschev and Vieira Neto, 2013) neighbouring levels in adjacent octaves are either up-sampled or down-sampled to match image dimensions in each particular pyramid octave, in a similar fashion to what is done in (Lowe, 2004).

In Figure 1, the construction model of a binomial pyramid with two octaves and four levels per octave is shown. The first level of each octave (except for the first octave) is obtained by nearest neighbour down-sampling of the third level of the previous octave. The fourth level is obtained by up-sampling via bilinear interpolation of the second level of the next octave.

3 EXPERIMENTAL SETUP

The experiments to assess interest point repeatability reported in this work involve comparisons between the conventional SIFT pyramid building scheme and the half-octave pyramid described in (Entschev and Vieira Neto, 2013), regarding both stability in the detection of interest points (keypoints) and their robustness to rotations.

Firstly, the binomial pyramids were constructed exactly as explained in (Entschev and Vieira Neto, 2013) and the reference Gaussian pyramids were constrained to have the same number of octaves as the binomial pyramids. In order to achieve this, the SIFT implementation in the OpenCV library (Bradski and Kaehler, 2008) was adapted to generate image pyramids with five scales per octave. The number of octaves for both pyramid types is equal to $\lfloor \log_2 d \rfloor - 2$, where $\lfloor . \rfloor$ indicates the floor function and d is the smallest dimension of the original input image.

For a detected test keypoint K_t to be considered a repetition of a reference keypoint K_r, the euclidean distance between K_t and K_r must not be larger than the scale s of K_r. The scale s of the keypoint being tested must also satisfy the condition $\sqrt{2}s - s \leq s \leq \sqrt{2}s + s$. When robustness to rotation is concerned, all keypoints that lie outside a circle with radius equal to the smaller dimension of the image are disregarded, because some of them will not be present when the image is synthetically rotated. The orientation of keypoints was not considered in the experiments reported here.

To test if the keypoints are repeated, we can not consider the exact coordinates of keypoints K_t and K_r. Due to the nature of the scale space, the coordinates of keypoints might shift slightly, and that is the reason to consider a small surrounding region as valid for location of repeated keypoints. When keypoints are found in different octaves, their location must be estimated according to the original size of the input image, which might also vary for different approaches of scale space construction. Finally, as proposed in (Brown and Lowe, 2002) and used in SIFT (Lowe, 2004), the interpolated location of keypoints is determined by fitting a quadratic 3D function to the scale space.

All tests were performed using a dataset containing 12 images taken from the Affine Covariant Regions Dataset provided by the Visual Geometry Group of the University of Oxford (Visual Geometry Group, 2004) and from the original dataset of images of SIFT, provided along with the SIFT demo program from David Lowe(Lowe, 2005). The images were not used in their full size, but were cropped to dimensions of $2^N + 1$, with both horizontal and vertical dimensions being the same.

In a first experiment, the repeatability of keypoints detected using the binomial pyramid is first compared to the ones detected using the conventional Gaussian pyramid. Then, in a second experiment, the robustness of keypoint detection using both approaches is assessed regarding rotation – for this purpose, the original input images are synthetically rotated around their central pixel using bicubic interpolation, and the repeatability of keypoints computed.

4 RESULTS

In this section some peculiarities that arise from the use of binomial filters to construct image pyramids are discussed, as well as the similarity between Gaussian and binomial pyramids with regard to keypoint repeatability.

The process of up-sampling the subsequent octave to obtain scales that cannot be achieved by a binomial filter of small size has a coarser effect than that of using larger Gaussian filters directly. This coarseness of the method seems to yield more extrema when performing non-maximum suppression in scale space, which are prone to result in unstable keypoints. To prevent the existence of too many unstable keypoints with corresponding low curvatures, we performed tests with non-maximum suppression using smaller curvature threshold values than the original curvature threshold value of 10 proposed in (Lowe,

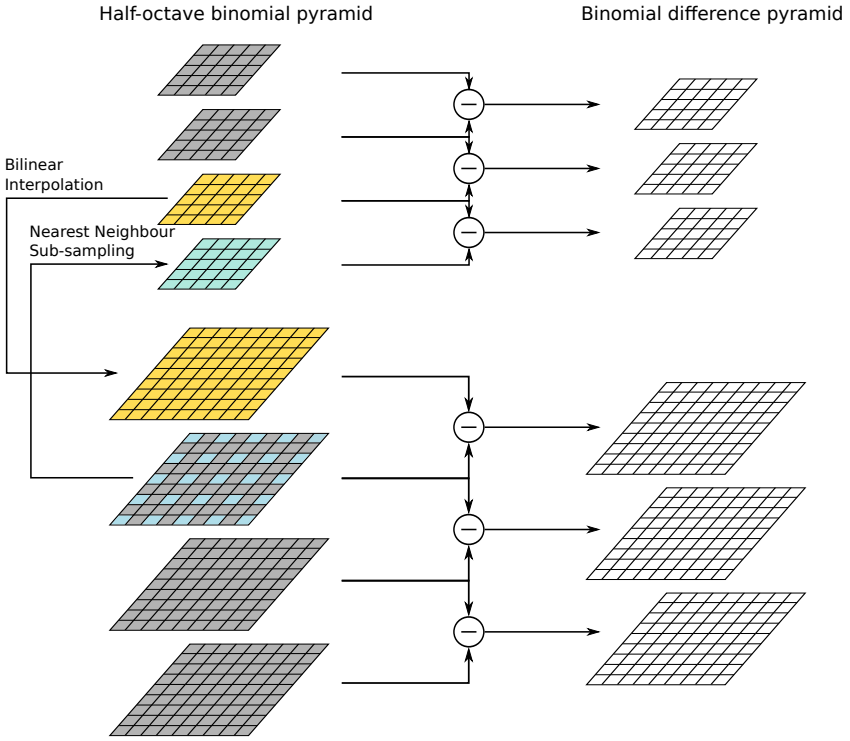

Figure 1: Construction of a half-octave binomial pyramid with two octaves and corresponding difference pyramid. Neighbouring levels in adjacent octaves are either up-sampled or down-sampled to match image dimensions and facilitate computation of differences between pyramid levels, resulting in four levels per octave (Entschev and Vieira Neto, 2013).

2004).

4.1 Keypoint Repeatability

Figure 2 shows the relative amount of keypoints detected within images processed with a binomial scale space that are also detected in the Gaussian scale space. According to the explanation given in section 3, K_t are the test keypoints found using the binomial approach, and K_r are the reference keypoints, found using the Gaussian approach. The repeatability of keypoints is shown in the graph by the ascending continuous line, where average samples are marked ($\times$) and vertical bars represent the standard deviation obtained with the test dataset.

A pyramid constructed with binomial filters and a curvature threshold of 5 repeats in average 78% of the keypoints found in a pyramid cosntructed with Gaussian filters, as shown in Figure 2. The repeatability of keypoints grows up to about 86% when using a curvature threshold of 10. Repeatability stabilises if the curvature threshold is raised any further.

This experiment shows that even using a much less computationally expensive approach as the one in (Entschev and Vieira Neto, 2013) to build the scale-space, most of the relevant keypoints are preserved,

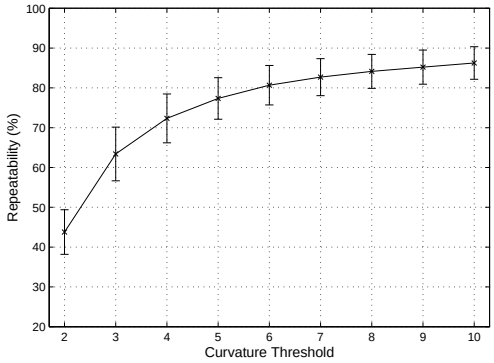

Figure 2: Average repeatability of keypoints found in a binomial pyramid in relation to the ones found in a Gaussian pyramid as a function of curvature threshold in the binomial pyramid – vertical bars indicate standard deviations. The number of average repeated keypoints grows as the curvature threshold is raised.

while reducing processing time to up to $1/4$.

An example of the resulting keypoints for a test image using both Gaussian and binomial pyramids, with curvature threshold values of 10 and 5, respectively, is shown in Figure 3.

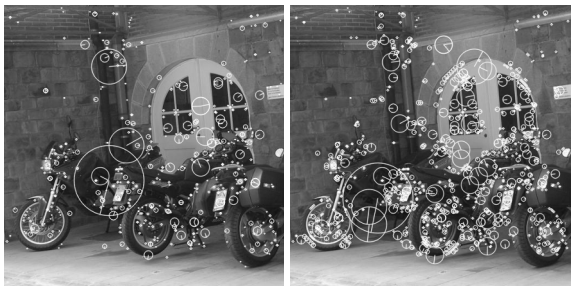

Figure 3: Example of detected keypoints with a Gaussian pyramid and curvature threshold value of 10 (left) and with a binomial pyramid and curvature threshold value of 5 (right). All detected keypoints in both approaches are shown, according to the constraints defined in section 3. More keypoints are detected when a binomial pyramid is used, even with lower curvature threshold values are used

4.2 Keypoint Ratio

When a binomial pyramid is used, more keypoints are found due to the coarseness that results from upsampling the levels from subsequent octaves. The ratio between the number of detected keypoints in a binomial pyramid and the ones detected in a Gaussian pyramid can be seen in Figure 4 – the average ratio as a function of curvature threshold, now considering the total amount of keypoints found, results in an ascending continuous line with increasing standard deviation, represented by the vertical bars.

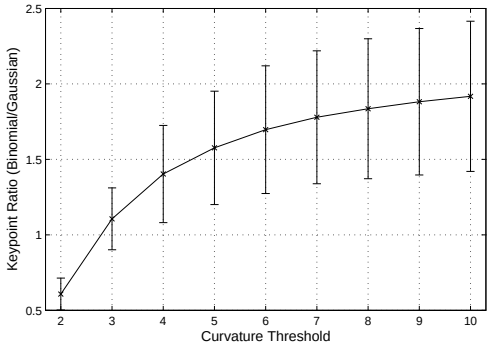

Figure 4: Ratio between the number of keypoints found in a binomial pyramid and the number found in a Gaussian pyramid — vertical bars indicate the standard deviation. For the same curvature threshold value of 10, the binomial approach yields almost twice the number of keypoints than the Gaussian approach.

As shown in Figure 4, when non-maximum suppression is performed over a scale space built with binomial filters, with a curvature threshold of 5, about 50% more keypoints are found in average, compared to non-maximum suppression performed over a scale space built with Gaussian filters. The average number of keypoints found almost doubles when the curvature threshold is set to 10. As the curvature threshold is increased, the standard deviation of the ratio of keypoints also increases.

4.3 Repeatability and Rotation

In this subsection, the repeatability achieved when input images are subject to rotation will be analysed. The results that follow were all gathered with synthetic rotations of the same 12 image samples used to obtain the results previously presented.

Figure 5 shows a polar graph with the repeatability obtained for a 360 degree rotation window, in steps of five degrees. For the Gaussian scale space, a curvature threshold of 10 was used, in accordance to the original value proposed in (Lowe, 2004), and for the binomial scale space, a curvature threshold of 5 was used, based on the results obtained in subsections 4.1 and 4.2.

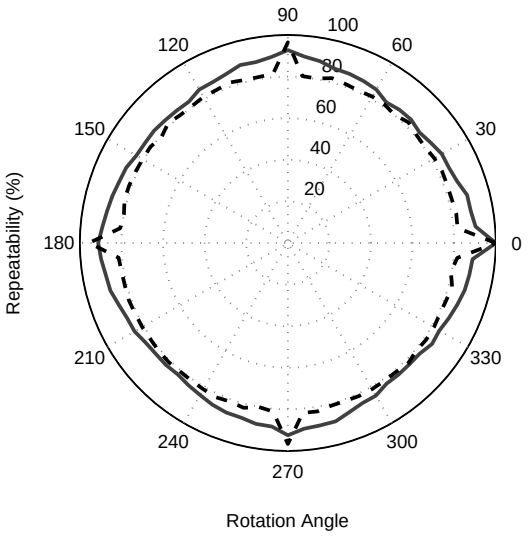

Figure 5: Average rotation repeatability of keypoints. The black dashed line represents the repeatability of keypoints found in the original SIFT scheme and the solid gray line represents the repeatability of keypoints found using the binomial approach. The binomial approach results in a slightly superior overall repeatability of keypoints, except for rotation angles that are multiples of 90 degrees.

As can be seen in Figure 5, the original scale space proposed in (Lowe, 2004) has maximum repeatability when the rotation angle is a multiple of 90 degrees. This is not surprising, as there is no need for interpolation to generate the synthetic pixel values of the rotated image, but only transposition of intensity values in the image. For different angles, the repeatabil-

ity of keypoints for the original SIFT scheme almost remains constant at a rate of 80%.

Using the binomial scale space pyramid proposed in (Entschev and Vieira Neto, 2013), the repeatability is also maximum for 90 degrees multiples, similarly to the scale space built with Gaussian filters, achieving approximately 90%. The farther rotation gets from a 90 degree multiple and closer it gets to an odd 45 degree multiple, the repeatability decreases, the smallest repeatability being around 80%.

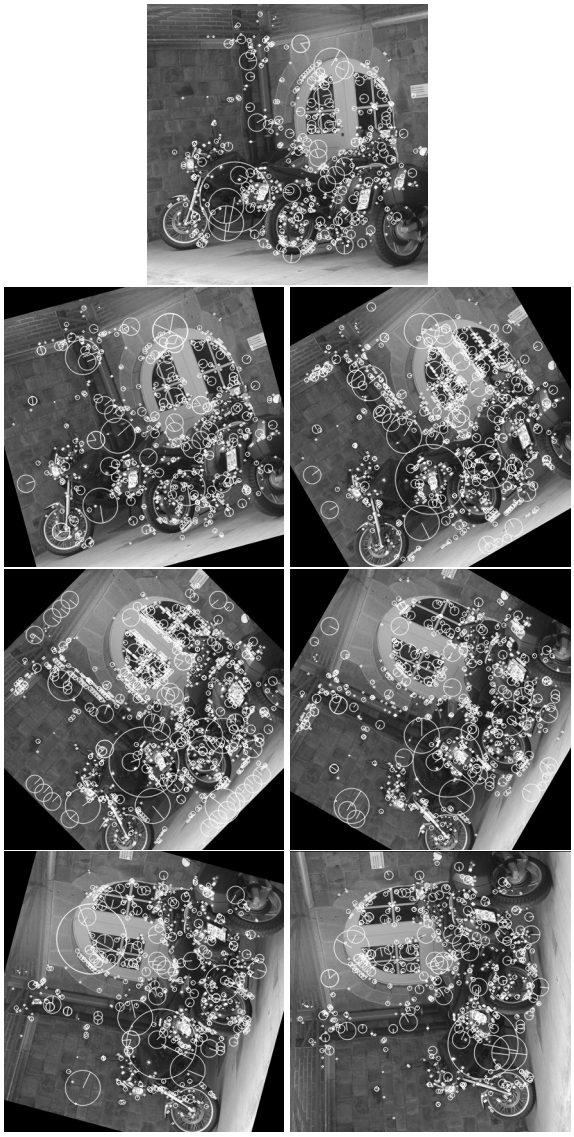

Figure 6: Example of detected keypoints using a binomial pyramid as a function of rotation. The original input image appears at the top – then, from left to right and top to bottom, rotations of 15 to 90 degrees in steps of 15 degrees are shown. The centre of the white circles represent keypoint locations and their radius represent their corresponding scale.

In Figure 6, the detected keypoints from 0 to 90 degrees, in steps of 15 degrees, are presented for one of the image samples used in the experiments. Because the average repeatability is similar, only samples for the first 90 degrees are shown.

5 CONCLUSION

The experiments presented here show that the approach proposed in (Entschev and Vieira Neto, 2013) yields good keypoint repeatability rate when compared to the original SIFT method proposed in (Lowe, 2004). When compared to the conventional Gaussian approach, the average keypoint repeatability rate for a binomial scale space is 78% for a curvature threshold value of 5 and reaches about 86% with a curvature threshold value of 10.

With a curvature threshold value of 5, the keypoint repeatability of the binomial scale space regarding rotations is maximum at multiples of 90 degrees and reaches more than 90%, while the worst cases involve rotations of odd multiples of 45 degrees, no less than 82%.

The most computationally expensive step of the method is considered to be upscaling of levels from subsequent octaves by bilinear interpolation during the pyramid construction. Further studies on this specific topic, including experiments with alternative methods to build binomial pyramids in time-efficient manners are still in sight.

Additional comparisons concerning changes in scale of the input image are also necessary, as well as performance assessment involving actual matching of SIFT descriptors generated for keypoints found when using the approach proposed in (Entschev and Vieira Neto, 2013) to construct the scale space – these are subjects of future research.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359.

Bradski, G. and Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV library*. O'Reilly Media.

Brown, M. and Lowe, D. G. (2002). Invariant features from interest point groups. In *Proceedings of the 13th British Machine Vision Conference*.

Burt, P. and Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540.

Coley, G. (2010). Beagleboard-xm system reference manual, revision A2. Beagleboard.org.

Crowley, J. L. and Riff, O. (2003). Fast computation of scale normalised gaussian receptive fields. In *Scale Space Methods in Computer Vision*, pages 584–598. Springer.

Crowley, J. L., Riff, O., and Piater, J. (2002). Fast computation of characteristic scale using a half octave pyramid. In *Proceedings of the International Workshop on Cognitive Vision*.

Crowley, J. L. and Stern, R. M. (1984). Fast computation of the difference of low-pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):212–222.

Entschev, P. A. and Vieira Neto, H. (2013). Efficient construction of SIFT multi-scale image pyramids for embedded robot vision. In *Proceedings of TAROS 2013: Towards Autonomous Robotic Systems*.

Halfacree, G. and Upton, E. (2012). *Raspberry Pi User Guide*. Wiley.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE international Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE.

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

Lowe, D. G. (2005). Demo software: SIFT keypoint detector. University of British Columbia. http://www.cs.ubc.ca/ lowe/keypoints/.

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

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518.

Visual Geometry Group (2004). Affine covariant regions dataset. University of Oxford. http://www.robots.ox.ac.uk/∼vgg/data/data-aff.html.