AUTOMATIC THRESHOLDING OF SIFT DESCRIPTORS

Matthew R. Kirchner

Image and Signal Processing Branch, Research Office, Code 4F0000D
Naval Air Warfare Center Weapons Division, China Lake, CA 93555 USA
matthew.kirchner@navy.mil

ABSTRACT

We introduce a method to perform automatic thresholding of SIFT descriptors that improves matching performance by at least 15.9% on the Oxford image matching benchmark. The method uses a contrario methodology to determine a unique bin magnitude threshold. This is done by building a generative uniform background model for descriptors and determining when bin magnitudes have reached a sufficient level. The presented method, called meaningful clamping, contrasts from the current SIFT implementation by efficiently computing a clamping threshold that is unique for every descriptor.

Index Terms— SIFT, Clamping, A contrario method, Helmholtz principal, Gestalt theory

1. INTRODUCTION

The SIFT descriptor, introduced by Lowe in [21], is a widely used descriptor in image processing and machine learning. The descriptor and its variants have been thoroughly studied and have been shown to systematically outperform other descriptors [24]. Many extensions have been proposed, some of which include sampling on a log-polar grid [24, 28], reducing the dimension with PCA [20], and scale pooling by averaging descriptors sampled from a neighborhood around the detected scale [11]. However, little known work has been performed to study and enhance the descriptor thresholding that is presented as part of the method. This thresholding [27], also called clamping, was introduced by Lowe as an ad-hoc way to achieve robustness to non-linear illumination effects, such as sensor saturation. This would lead us to believe the clamping process would improve matching performance on image pairs that exhibit significant illumination changes; but have little effect on images with similar lighting conditions. However, Lowe’s clamping method can greatly increase matching performance (14.4% improvement on the Oxford dataset) on general image pairs even when no significant illumination changes exist.

This work proposes a novel method, which we call meaningful clamping (MC), to automatically threshold SIFT descriptors and improves on the idea of clamping by providing a rigorous process to compute the clamping threshold. This leads to significantly increased performance over the existing clamping method on a wide variety of image matching problems. The method is based on a contrario methodology for computing detection thresholds [10], and is introduced in Sec. 3. Matching results with experiments performed on the Oxford dataset [25] are shown in Sec. 5, and confirm state-of-the-art results.

We introduce a method to perform automatic thresholding of SIFT descriptors that improves matching performance by at least 15.9% on the Oxford image matching benchmark. The method uses a contrario methodology to determine a unique bin magnitude threshold. This is done by building a generative uniform background model for descriptors and determining when bin magnitudes have reached a sufficient level. The presented method, called meaningful clamping, contrasts from the current SIFT implementation by efficiently computing a clamping threshold that is unique for every descriptor.

We propose to create an extension of the SIFT descriptor, since it has been shown to systematically outperform other descriptors [24]. The SIFT descriptor is a smoothed and weighted 3D histogram of gradient orientations. For any patch $J$, we form a gradient vector field $\nabla J$. We define the grid $\Lambda$, which determines the bin centers $x_i, y_j, \theta_k$ of the histogram and has size $n(x) \times n(y) \times n(\theta)$. In typical implementations, $\Lambda$ is chosen to have $4 \times 4$ spatial bins and 8 angular bins. With $x = (x, y)$ and $\ell = (i, j, k) \in \Lambda$, a single, pre-normalized spatial bin of the SIFT descriptor can be written as the integral expression:

$$d(\ell | J) = \int g_\sigma(x) w_{\alpha}(\Lambda \nabla J(x)) w_{ij}(x) \| \nabla J(x) \| dx,$$  

where $w_{ij}(x) = w(x - x_i) w(y - y_j)$ [11] [31]. The weight function $w_{ij}$ is a bilinear interpolation with

$$w(z) = \max(0, 1 - \frac{n(z)}{2\lambda_{\text{patch}}}|z|);$$

and

$$w_{\alpha}(\theta) = \max(0, 1 - \frac{n(\theta)}{2\pi}|\theta_k - \theta \mod 2\pi|)$$

is an angular interpolation [27]. The parameter $\lambda_{\text{patch}}$ is the radius of $J$ such that the patch has dimensions $2\lambda_{\text{patch}} \times 2\lambda_{\text{patch}}$. The histogram samples are also weighted by a Gaussian density function $g_\sigma(x)$, the purpose of which is to discount the contribution of samples at the edge of the patch with the goal to reduce boundary effects. The building of SIFT descriptors using Eq. 1 for all experiments was performed with the VLFeat open source vision library [31]. For further details on how the descriptor was constructed, the reader is encouraged to review [27] [31].
2.1. Clamping

In an effort to design a descriptor to be robust to non-linear contrast changes, Lowe proposed to threshold the bin magnitudes of the descriptor. Lowe defines this threshold as

\[ d_c(\ell) = \min(d(\ell), c ||d||), \]

with the parameter \( c = 0.2 \) set experimentally, and this is the default setting in [31]. This is followed by an additional normalization to ensure unit length of the descriptor is preserved after thresholding. It is important to note that the thresholding in Eq. 2 maintains invariance to affine contrast changes. The thresholding process, or clamping, is thought to reduce the effects of camera saturation or other illumination effects. However, we will show empirically in Sec. 3.3 that clamping also increases the general matching performance of the descriptor, observed to be 14.4% compared to the performance without clamping on the Oxford dataset. This occurs even when there exists consistent lighting conditions between image pairs. The threshold parameter of \( c = 0.2 \) is set rather arbitrarily; and is fixed for every descriptor. By applying an automatic threshold that is allowed to vary for every descriptor, we can significantly improve the performance of the SIFT descriptor for image matching problems.

3. MEANINGFUL CLAMPING

The bins of the SIFT descriptor represent the underlying content of a local image patch. We wish to detect when geometric structure is present in the patch; and this is indicated by the observation of large descriptor bin values. This amounts to detecting significant bins by computing a perception threshold for each descriptor and using that as the clamping limit. The idea is that once bins reach the perception threshold, little information is gained by exceeding this value. A contrario methodology is proposed to compute descriptor perception thresholds, and is based on applying a mathematical foundation to the concept of the Helmholtz principal, which states “we immediately perceive whatever could not happen by chance” [10]. It has been shown to be highly successful for many problems in image processing such as detecting line segments [18], change detection [14], contrasted boundaries [8], vanishing points [1], and modes of histograms [6,15].

Instead of trying to define a priori the structure of the underlying image content, an impossible task for general natural images, we instead define what it means to have a lack of structure. Using the Helmholtz principal, lack of structure is simply modeled as uniform randomness, which we call the uniform background model, or the null hypothesis \( H_0 \). We assume the descriptor has been generated from \( H_0 \), and claim a detection, i.e. significant geometric content is present, when there is a large deviation from \( H_0 \). If the observed event is extremely unlikely to have been generated from this background model, we claim the event as meaningful because it could not have occurred by random chance.

Let \( \Lambda \) be the histogram grid associated with the descriptor \( d \), which represents a set of \( L = n(x)n(y)n(\theta) \) connected bins such that every bin \( \ell = (i, j, k) \in \Lambda \) contains a number of sample counts \( d(\ell) \), and a neighborhood \( C_\ell \subset \Lambda \) of bins for which \( \ell \) is connected. Introducing a neighborhood set for each bin allows us to have circular connected angular histograms, while spatial dimensions are rectangular. We also let \( M = \sum_\ell d(\ell) \) be the total number of samples of the descriptor and \( p(\ell) \) be the probability that a random sample is drawn in bin \( \ell \), which leads to the definition of the null hypothesis for the descriptor \( d \).

\[ NFA(\varepsilon) = N^B(M, d(\ell), p(\ell)), \]

where \( N \) is the number of tests, and is typically defined as the number of all possible connected subsets of the histogram. \( N \) can be seen as a Bonferroni correction [17, 19] for the expected value in Eq. 3. Which leads to the following definition of a meaningful bin.

Definition 3.2. A bin \( \ell \in \Lambda \) of the SIFT descriptor \( d \) is an \( \varepsilon \)-meaningful bin if

\[ NFA(\varepsilon) = N^B(M, d(\ell), p(\ell)) < \varepsilon. \]

This leads to the question of what to use for \( \varepsilon \)? We can follow the work of Desolneux, et al. [7], and always set \( \varepsilon = 1 \), since including the number of tests, \( N \), allows the threshold to scale automatically with histogram size. The setting of \( \varepsilon = 1 \) can be interpreted as setting the threshold so as to limit the expected number of false detections under a uniform background model to less than one. This has two important consequences. First, for some applications, it is important for the algorithm to correctly give zero detections when no object exists. Second, this strategy gives detection thresholds that are similar to that of human perception [9]; and the dependence on \( \varepsilon \) is logarithmic and hence very weak [18]. We will hereafter refer to an \( \varepsilon \)-meaningful bin as just a meaningful bin.

We can now select a clamping threshold for \( d \) as the minimum descriptor bin value needed to be detected as a meaningful bin. For a given descriptor \( d \), with corresponding properties \( M \) and \( p(\ell) = 1/L \), we define this threshold as

\[ t_d = \min \{ k : N^B(M, k, p(\ell)) < 1 \}. \]

We then proceed to create the new clamped descriptor as

\[ d_c(\ell) = \min(t_d, d(\ell)), \]

for every bin \( \ell \in \Lambda \).

4. IMPLEMENTATION DETAILS

The a contrario threshold in Eq. 5 has dependence on \( N \), which is defined as the number of all possible connected subsets of \( \Lambda \). However, for any histogram greater than dimension one, we cannot explicitly compute this, and instead use the number of aligned rectangular regions

\[ N_{Rect} = \frac{1}{8} n(x)n(y)n(\theta) (n(x) + 1) (n(y) + 1) (n(\theta) + 1). \]
Proposition 4.1. Let \( d_1 \) be a SIFT descriptor clamped with the approximate threshold \( \tilde{t}_a \) given in Eq. 9 and \( t_a \) is the exact threshold given in Eq. 3. Then, as \( M \) grows large, \( d_1(\ell) \leq t_a \) for all \( \ell \in \Lambda \) such that either (a) \( p \leq 1/4 \) and \( p \leq \frac{1}{4M} \), or (b) \( p \leq \frac{1}{4M} \leq 1 - p \). As \( M \) grows large, the \( O\left(\frac{\ln M}{M}\right) \) term becomes small and Eq. 8 converges to the central limit approximation. Using this we can solve for the detection threshold as

\[
\tilde{t}_a = M p + \alpha(\mathcal{N}_{\text{Rect}}) \sqrt{Mp(1-p)},
\]

with \( \alpha(\mathcal{N}_{\text{Rect}}) = -\ln(1/\mathcal{N}_{\text{Rect}}) \). From this we can compute a new clamped descriptor, \( d_1(\ell) \), with Eq. 6 using the bin threshold \( t_b \). Using the approximation \( t_a \) still maintains the property from Eq. 6 that \( d_1(\ell) \leq t_a \).

5. RESULTS

We present image matching results applying the newly developed meaningful clamping method, and compare it to the clamping procedure proposed by Lowe. For reference, we also compare both clamping methods to descriptors with which no clamping was performed.

### 5.1. Dataset

To evaluate matching performance, we use the Oxford dataset \([25]\), which is comprised of 40 image pairs of various scene types undergoing different camera poses and transformations. These include viewpoint angle, zoom, rotation, blurring, compression, and illumination. The set contains eight categories, each of which consists of image pairs undergoing increasing magnitudes of transformations. Included with each image pair is a homography matrix, which represents the ground truth mapping of points between the images. The transformations applied to the images are real and not synthesized as in \([13]\). The viewpoint and zoom-rotation categories are generated by focal length adjustments and physical movement of the camera. Blur is generated by varying the focus of the camera; and illumination by varying the aperture. The compression set was created by applying JPEG compression and adjusting the image quality parameter.

### 5.2. Metrics

To evaluate the performance of local descriptors with respect to image matching, we follow the methods of \([24]\). Given a pair of images we extract SIFT features from both images. A match between two descriptors is determined when the Euclidean distance is less than some threshold \( t \). Any descriptor match is considered a correct match if the two detected features correspond as defined in \([25]\). We compute the AP for every image pair in the Oxford dataset, each category of the Oxford dataset. SIFT detector parameter FirstOctave is set to 0.

| Category  | No Clamping | Low Clamping | MC |
|-----------|-------------|--------------|----|
| Graffiti  | 0.123       | 0.161        | 0.205 |
| Wall      | 0.327       | 0.371        | 0.405 |
| Boats     | 0.301       | 0.341        | 0.375 |
| Bark      | 0.111       | 0.119        | 0.120 |
| Trees     | 0.207       | 0.288        | 0.366 |
| Bikes     | 0.414       | 0.371        | 0.496 |
| Leuven    | 0.387       | 0.538        | 0.635 |
| UBC       | 0.558       | 0.588        | 0.615 |
| All images| 0.303       | 0.347        | 0.402 |

Table 1. Mean average precision for each category of the Oxford dataset. SIFT detector parameter FirstOctave is set to 0.
A new method to threshold SIFT descriptors was presented. This method significantly improves mAP for image matching on the standard Oxford dataset. Future work is to study the impact meaningful clamping has on other problems, such as large scale image retrieval. Also of interest is the study of why meaningful clamping (and also clamping in general) has such a large impact on image matching.

The author conjectures that clamping effects the distribution of large point sets of descriptors. If the descriptor is not clamped, then a small number of descriptor bins would dominate all other bins. This would constrain the points to lie mostly along the axes. Performing nearest neighbor-type searches could become ambiguous, since many points would exist with a similar spatial distance. By clamping, we are thresholding bin magnitudes; and this causes the points to ‘spread out’, and more uniformly occupy the $\mathbb{R}^2$ space in which the descriptors lie. This conjecture is supported by the observation in the presented experiments that the improvement drastically increased when attempting to match larger sets of points extracted from the image pairs when the SIFT detector parameter FirstOctave was changed to -1, generating many more features for each image.

### Acknowledgments

The author would like to thank UCLA professor Stefano Soatto for many great discussions on the topic of clamping SIFT descriptors, and his student Jingming Dong for direction on descriptor evaluation. The author would also like to thank Lawrence Peterson, Shawnasie Kirchner, and the anonymous reviewers for spending valuable time to provide a careful and thorough review of this paper.
7. REFERENCES

[1] A. Almansa, A. Desolneux, and S. Vamcheck. Vanishing point detection without any a priori information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):502–507, 2003.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[3] C. Berger, T. Géraud, R. Levillain, N. Widynski, A. Baillard, and E. Bertin. Effective component tree computation with application to pattern recognition in astronomical imaging. In *IEEE International Conference on Image Processing (ICIP)*, 2007.

[4] F. Cao, J.-L. Lisani, J.-M. Morel, P. Musé, and F. Sur. *A Theory of Shape Identification*, volume 1948. Springer, 2008.

[5] J. Delon, A. Desolneux, J.-L. Lisani, and A. B. Petro. Histogram analysis and its applications to fast camera stabilization. In *International Workshop on Systems, Signals and Image Processing*, pages 431–434, 2004.

[6] J. Delon, A. Desolneux, J.-L. Lisani, and A. B. Petro. A non-parametric approach for histogram segmentation. *IEEE Transactions on Image Processing*, 16(1):253–261, 2007.

[7] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[8] A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.

[9] A. Desolneux, L. Moisan, and J.-M. Morel. Computational gestalts and perception thresholds. *Journal of Physiology-Paris*, 97(2):311–324, 2003.

[10] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, volume 34. Springer, 2007.

[11] J. Dong and S. Soatto. Domain size pooling in local descriptors: DSP-SIFT. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[12] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[13] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor matching with convolutional neural networks: A comparison to SIFT. *arXiv preprint arXiv:1405.5769*, 2014.

[14] A. Flenner and G. Hewer. A Helmholtz principle approach to parameter free change detection and coherent motion using exchangeable random variables. *SIAM Journal on Imaging Sciences*, 4(1):243–276, 2011.

[15] A. Flenner, G. Hewer, and C. Kenney. Two dimensional histogram analysis using the Helmholtz principle. *Inverse Problems and Imaging*, 2(4):485–525, 2008.

[16] W. Forstner, T. Dickscheid, and F. Schindler. Detecting interpretable and accurate scale-invariant keypoints. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2256–2263. IEEE, 2009.

[17] R. G. Von Gioi and J. Jakubowicz. On computational Gestalt detection thresholds. *Journal of Physiology-Paris*, 103(1):4–17, 2009.

[18] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732, 2010.

[19] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, New York, 1987.

[20] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II:506–513. IEEE, 2004.

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

[22] 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, 2004.

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

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

[25] 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.

[26] L. Najman and M. Couprie. Building the component tree in quasi-linear time. *IEEE Transactions on Image Processing*, 15(11):3531–3539, 2006.

[27] I. R. Otero and M. Delbracio. Anatomy of the SIFT method. *Image Processing On Line*, 4:370–396, 2014.

[28] J. Rabin. *Approches robustes pour la comparaison d’images et la reconnaissance d’objets*. PhD thesis, Télécom ParisTech, 2009.

[29] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009.

[30] E. V. Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, pages 404–412, 1977.

[31] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, pages 1469–1472. ACM, 2010.