Context-Aware Keypoint Extraction for Robust Image Representation

Pedro Martins[1]
pjmm@dei.uc.pt
Paulo Carvalho[1]
carvalho@dei.uc.pt
Carlo Gatta[2]
c.gatta@cvc.uab.es

[1] Centre for Informatics and Systems
University of Coimbra
Coimbra, Portugal

[2] Computer Vision Centre
Autonomous University of Barcelona
Barcelona, Spain

Abstract

Tasks such as image retrieval, scene classification, and object recognition often make use of local image features, which are intended to provide a reliable and efficient image representation. However, local feature extractors are designed to respond to a limited set of structures, e.g., blobs or corners, which might not be sufficient to capture the most relevant image content.

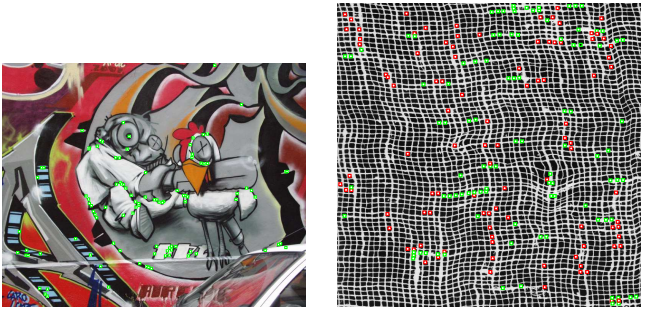
We discuss the lack of coverage of relevant image information by local features as well as the often neglected complementarity between sets of features. As a result, we propose an information-theoretic-based keypoint extraction that responds to complementary local structures and is aware of the image composition. We empirically assess the validity of the method by analysing the completeness, complementarity, and repeatability of context-aware features on different standard datasets. Under these results, we discuss the applicability of the method.

1 Introduction

Local feature extraction is a prominent and prolific research topic for the computer vision community, as it plays a crucial role in many vision tasks. Local feature-based strategies have been successfully used in numerous problems, such as wide-baseline stereo matching [1, 17, 30], content-based image retrieval [22, 25, 29], object class recognition [6, 21, 26], camera calibration [9], and symmetry detection [4]. The key idea underlying the use of local features is to represent the image content by a sparse set of salient regions or (key)points. By discarding most of the image content, we save computation and improve robustness as there are redundant local image patches rather than a limited number of global cues [28].

The desired properties of a local feature extractor are dictated by its application. For example, matching and tracking tasks mainly require a *repeatable and accurate feature extraction*. The objective is to accurately identify the same features across a sequence of images, regardless of the degree of deformation. It is not relevant if the set of features fails to cover the most informative image content. On the other hand, tasks such as object (class) recognition, image retrieval, scene classification, and image compression, require a *robust image representation* [31]. The idea is to analyse the image statistics and use local features to

© 2011. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.



(a) (b)

Figure 1: Proposed keypoint extraction: (a) Context-aware keypoints on a well-structured scene (100 most informative locations); (b) A combination of context-aware keypoints (green squares) with SFOP keypoints [9] (red squares) on a highly textured image.

capture relevant image content. Here, the requirements of repeatability and accuracy become less relevant.

This paper focuses on the problem of providing a robust feature-based image representation. Our contribution is a feature extractor aimed at covering the most informative image content. The proposed algorithm, coined as Contex-Aware Keypoint Extractor (CAKE), responds to complementary local structures and is aware of the image composition. We follow an information-theoretic approach by assuming that the so-called salient locations correspond to points within structures with a low probability of occurrence, which is in accordance with a plausible characterisation of visual saliency [3]. We are motivated by the fact that the majority of local feature extractors makes strong assumptions on the image content, which can lead to an ineffectual coverage of the content [5, 8]. Here, the idea is not to formulate any a priori assumption on the structures that might be salient. Moreover, our scheme is designed to take advantage of different local representations (descriptors).

A context-aware extraction can respond to features with a reasonable degree of complementarity as long as they are informative. For images with many types of structures and patterns, one can expect a high complementarity among the features retrieved by a context-aware extractor. Conversely, images with repetitive patterns inhibit context-aware extractors from retrieving a clear summarised description of the content. In this case, the extracted set of features can be complemented with a counterpart that retrieves the repetitive elements in the image. To illustrate the above-mentioned advantages, we depict these two cases in Figure 1. The left image shows a context-aware keypoint extraction on a well-structured scene, retrieving the 100 most informative locations. This small number of features is sufficient to provide a good coverage of the content, which includes several types of structures. The right image depicts the benefits of combining context-aware keypoints with strictly local ones (SFOP keypoints [9]) to obtain a better coverage of textured images.

2 Related Work

We will provide a brief, yet illustrative, description of solutions that have contributed towards establishing local feature extraction as a mature research topic. For a more complete review, we will refer to the work of Tuytelaars and Mikolajczyk [31] and references within.

A large family of local feature extractors is based on local differential geometry. The Harris-Stephens keypoint extractor is a well-known example of an algorithm based on local differential geometry: it takes the spectrum of the structure tensor matrix to define a saliency measure. The Scale Invariant Feature Operator (SFOP) [9] also relies on the structure tensor matrix. It responds to corners, junctions and circular features. This explicitly interpretable and complementary extraction is a result of a unified framework that extends the gradient-based extraction previously discussed in [7] and [23] to a scale-space representation [15]. The Hessian matrix is often used to extract *blob-like structures*, either by using its determinant– it attains a maximum at *blob-like keypoints* – or by searching for local extrema of the Laplacian operator, i.e., the trace of the Hessian matrix. The Harris-Laplace [18] is a scale covariant region extractor that results from the combination of the Harris-Stephens scheme with a Laplacian-based automatic scale selection. The Harris-Affine scheme [19], an extension of the Harris-Laplace, relies on the combination of the Harris-Laplace operator with an affine shape adaptation stage [16]. Similarly, the Hessian-Affine extractor [19] follows the same affine shape adaptation, however, the initial estimate is taken from the determinant of the Hessian matrix. Some extractors rely solely on the intensity image, such as the Maximally Stable Regions (MSERs) extractor [17], which retrieves stable regions (with respect to intensity perturbations) that are either brighter or darker than the pixels on their outer boundaries. In [10], Gilles proposes an information-theoretic algorithm: keypoints correspond to image locations at which the entropy of local intensity values attains a maximum. Kadir and Brady [11] introduce a scale covariant salient region extractor, which estimates the entropy of the intensity values distribution inside a region over a certain range of scales. Salient regions in the scale space are taken from scales at which the entropy is at its peak. In [12], Kadir et al. propose an affine covariant version of the extractor introduced in [11]. Dickscheid et. al [5] suggest a local entropy-based extraction, which uses the entropy density of patches to build a scale-space representation. This density is expressed using a model for the power spectrum that depends on the image gradient and noise.

3 Context-Aware Keypoint Extraction

Our context-aware keypoint extraction is formulated in an information theoretic framework. We define saliency in terms of information content: a keypoint corresponds to a particular image location within a structure with a low probability of occurrence (high information content). We follow Shannon's definition of information [27]: if we consider a symbol s, its *information* will be given by $I(s) = -\log(p(s))$, where $p(\cdot)$ denotes the probability of a symbol. In the case of images, defining symbols is complex and the content of a pixel $\mathbf{x}$ is not very useful, whereas the content of a region around the point would be more appropriate. We can consider $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^D$, any viable local representation, e.g, the Hessian matrix, as a "codeword" that represents $\mathbf{x}$. We can see the image codewords as samples of a multivariate probability density function (PDF). In the literature, there is a number of methods to estimate an unknown multi-variate PDF with a sufficient number of samples. Among all, we have decided to use the Parzen density estimator [24], also known as *Kernel Density Estimator* (KDE). The KDE is appropriate for our objective since it is non-parametric, which will allow us to estimate any PDF, as long as there is a sufficient number of samples. Using

the KDE, the probability of a codeword $\mathbf{w}(\mathbf{y})$ is

$$\widehat{p}(\mathbf{w}(\mathbf{y})) = \frac{1}{Nh}\sum_{\mathbf{x}\in\Phi} K\left(\frac{d(\mathbf{w}(\mathbf{y}),\mathbf{w}(\mathbf{x}))}{h}\right), \tag{1}$$

where d is a distance function, K is a kernel, h is a smoothing parameter called *bandwidth*, Φ is the image domain, and N represents the number of pixels. The idea of the KDE method is to blur the contribution of each sample $\mathbf{x}$ by spreading it to a certain area in $\mathbb{R}^D$ with a certain shape, which is defined by K. If K has compact support, or decreases as a function of the distance d, then codewords in dense areas of the multi-variate distribution will have higher probability than isolated samples. There are several choices for the kernel. The most commonly used and the most appropriate for our method is a multidimensional Gaussian function with zero mean and standard deviation σ_k. Using a Gaussian kernel, (1) becomes

$$\tilde{p}(\mathbf{w}(\mathbf{y})) = \frac{1}{N\Gamma}\sum_{\mathbf{x}\in\Phi} \mathrm{e}\left(-\frac{d^2(\mathbf{w}(\mathbf{y}),\mathbf{w}(\mathbf{x}))}{2\sigma_k^2}\right), \tag{2}$$

where h has been substituted by the standard deviation σ_k and Γ is a proper constant such that the estimated probabilities are taken from an actual PDF. Having defined the probability of a codeword, we can define the saliency measure as follows:

$$m(\mathbf{y}, I(\Phi)) = -\log\left(\frac{1}{N\Gamma}\sum_{\mathbf{x}\in\Phi} e\left(-\frac{d^2(\mathbf{w}(\mathbf{y}),\mathbf{w}(\mathbf{x}))}{2\sigma_k^2}\right)\right), \tag{3}$$

where I denotes the image. In this case, context-aware keypoints will correspond to local maxima of $m(\cdot, I(\Phi))$ that are beyond a certain threshold T.

To conclude, we need to define a distance function d and set a proper value to σ_k. AQUI!!!

The distance d We define W, the set of all the multi-variate samples as $W = \bigcup_{\mathbf{x}\in\Phi}\mathbf{w}(\mathbf{x})$. Being Σ_W, the covariance matrix of W, we define the following *Mahalanobis distance* $d_M(\mathbf{w}(\mathbf{x}),\mathbf{w}(\mathbf{y})) = \sqrt{(\mathbf{w}(\mathbf{x})-\mathbf{w}(\mathbf{y}))^T\Sigma_W^{-1}(\mathbf{w}(\mathbf{x})-\mathbf{w}(\mathbf{y}))}$. By considering the aforementioned distance, the following property can be drawn:

Property 1. *Let* $\mathbf{w}^{(1)}$ *and* $\mathbf{w}^{(2)}$ *be codewords such that* $\mathbf{w}^{(2)}(\mathbf{x}) = T(\mathbf{w}^{(1)}(\mathbf{x}))$*, where* T *is an affine transformation. Let* $p^{(1)}$ *and* $p^{(2)}$ *be the probability maps of* $\mathbf{w}^{(1)}$ *and* $\mathbf{w}^{(2)}$*, i.e.,* $p^{(i)}(\cdot) = p(\mathbf{w}^{(i)}(\cdot))$*,* $i = 1,2$*. In this case,*

$$p^{(2)}(\mathbf{x_l}) \leq p^{(2)}(\mathbf{x_m}) \iff p^{(1)}(\mathbf{x_l}) \leq p^{(1)}(\mathbf{x_m}), \forall \mathbf{x_l}, \mathbf{x_m} \in \Phi.$$

The smoothing parameter σ_k The parameter σ_k can potentially vanish the ability of the method to adapt to the image context. If σ_k is too large, the KDE over-smoothes the estimated PDF, which cancels the inherent PDF structure due to the image content. On the other hand, if σ_k is too small, the interpolated values between different samples could be very low, such that there is no interpolation anymore. We propose a strategy, in the case of an univariate distribution, to determine $\sigma_k^\star$, an optimal σ_k, aiming at *sufficient blurring* while having the *highest sharpen* PDF between samples. For our purposes, we can use univariate distributions, since we approximate the KDE computations of a D-dimensional multi-variate

PDF by estimating D separate univariate PDFs (see Appendix). Having a series of N samples w, we define the optimal σ_k for the given distribution as

$$\sigma_k^\star = \arg\max_{\sigma>0} \int_{w_i}^{w_{i+1}} \frac{1}{\sqrt{2\pi}\sigma} \left| \frac{\mathrm{d}\left(e^{\frac{-(w-w_i)^2}{2\sigma^2}} + e^{\frac{-(w-w_{i+1})^2}{2\sigma^2}} \right)}{\mathrm{d}\,w} \right| \mathrm{d}\,w, \tag{4}$$

where w_i and w_{i+1} is the farthest pair of consecutive samples in the distribution. It can be demonstrated that, by solving (4), we have $\sigma_k^\star = |w_i - w_{i+1}|$. It can also be demonstrated that for $\sigma < |w_i - w_{i+1}|/2$, the estimated PDF between the two samples is concave, which provides insufficient smoothing.

4 Hessian-based CAKE instance

In this section, we introduce an instance of the context-aware keyponit extractor. Different instances are given by considering different local representations.

The notion of saliency is related to rarity [10, 11]. What is salient is rare. However, the reciprocal is not necessarily valid. The discriminating power of the codewords has to be analysed. A highly discriminating codeword will turn every location into a rare structure; nothing will be regarded as salient. On the other hand, with a less discriminating codeword, rarity will be harder to find.

The Hessian matrix (Eq. (5)) appears as a suitable intrinsic codeword for an instance that provides . Furthermore, the inclusion of multi-scale components provides a framework to design an instance characterized by a quasi-scale-covariant extraction. We define the codeword for the multi-scale Hessian-based instance as

$$\begin{aligned} \mathbf{w}(\mathbf{x}) = \big[\; & t_1^2 L_{xx}(\mathbf{x};t_1) \quad t_1^2 L_{xy}(\mathbf{x};t_1) \quad t_1^2 L_{yy}(\mathbf{x};t_1) \quad t_2^2 L_{xx}(\mathbf{x};t_2) \quad t_2^2 L_{xy}(\mathbf{x},t_2) \quad t_2^2 L_{yy}(\mathbf{x},t_2) \\ & \cdots \qquad\qquad t_L^2 L_{xx}(\mathbf{x};t_M) \quad t_L^2 L_{xy}(\mathbf{x};t_M) \quad t_L^2 L_{yy}(\mathbf{x};t_M) \; \big]^T, \end{aligned} \tag{5}$$

where where L_{xx}, L_{xy} and L_{yy} are the second order partial derivatives of L, a Gaussian smoothed version of the image, and t_i, with $i = 1,\ldots,M$, represents the scale.

5 Experimental Validation and Discussion

We have evaluated and compared the performance of the Hessian-based CAKE instance (which we will refer to as [HES]-CAKE) using three criteria: completeness, complementarity, and repeatability. We can quantify completeness as the amount of image information that is preserved by a set of features. Complementarity is a particular case of completeness analysis: it reflects the amount of image information coded by sets of potentially complementary features [5]. The repeatability score is a measure that ascertains how precisely an extractor responds to the same locations under several image transformations, which reflects the level of covariance and robustness. The experiments were performed on the Oxford dataset [20] and on the dataset used by Dickscheid et al. for completeness evaluation [5]. The latter comprises four categories of natural scenes [13, 14], the Brodatz texture collection [2] as well as a set of aerial images. Example images from this dataset are depicted in Figure 2.

With the aim of comparison, we have also evaluated the performance of some of the leading algorithms on scale or affine covariant feature extraction: the Hessian-Laplace (HESLAP), the Harris-Laplace (HARLAP), the Scale Invariant Feature Operator (SFOP), and the Maximally Stable Extremal Regions (MSER) extractor. The implementations are the ones given and maintained by the authors and default parameters were used.

Table 1 outlines the parameter settings for the CAKE instance. We note that our extractor retrieves more features than its counterparts. For a fair evaluation of repeatability, we have defined a threshold to avoid a considerable discrepancy in the number of features.

Table 1: [HES]-CAKE parameter settings.

[HES]-CAKE	
Number of scales	12
t_{i+1}/t_i (ratio between successive scale levels)	1.19
t_0 (initial scale)	1.4
Non-maximal suppression window	3×3
T (threshold)	12 or 3000 keypoints (for the Oxford dataset)/ None (otherwise)
σ_k	optimal
N_R (number of samples, see Appendix)	200

The evaluation protocols require regions rather than keypoints. We use the normalized Laplacian operator, $\nabla^2 L_n = t^2(L_{xx} + L_{yy})$, to determine the characteristic scale for each detected keypoint, which, in this case, corresponds to the one at which the normalized Laplacian attains an extremum. This scale defines the radius of a circular region centered about the keypoint.

The non-optimised Matlab implementation of [HES]-CAKE takes, on average, 272.3 seconds to process an 800× 600 image.

Brodatz (30 images) | Aerial (28 images) | Forest (328 images) | Mountain (374 images) | Tall building (356 images) | Kitchen (210 images)

Figure 2: Example images from the categories in the dataset.

5.1 Completeness and complementarity

To measure completeness, Dickscheid et al. [5] compute an entropy density, $p_H(\mathbf{x})$, based on local image statistics, which is not context aware, and a feature coding density, $p_c(\mathbf{x})$, derived from a given set of features. The (in)completeness measure corresponds to the *Hellinger distance* between the two distributions: $d_H(p_H, p_c) = \sqrt{\frac{1}{2}\sum_{\mathbf{x}\in\Phi}(\sqrt{p_H(\mathbf{x})} - \sqrt{p_c(\mathbf{x})})^2}$. When p_H and p_c are very close, d_H will be small, which means that the set with a coding density p_c efficiently covers the image content, i.e., the set has a high completeness. We note that this measure penalises the use of large scales (which is a straightforward solution to achieve full coverage of the content) as well as the presence of features in pure homogeneous regions. On the other hand, it rewards the "fine capturing" of local structures or superimposed features appearing at different scales.

Figure 3 outlines the completeness results for each image category (Figure 2), in terms of the distances between distributions p_H and p_c. As in [5], we have plotted the line $y = \sqrt{\frac{1}{2}}$, which corresponds to an angle of 90 degrees between p_H and p_c. Any distance beyond this

threshold will correspond to a pair of “sufficiently different" densities. Regardless of the image collection, the [HES]-CAKE achieves the best completeness scores. The complete-

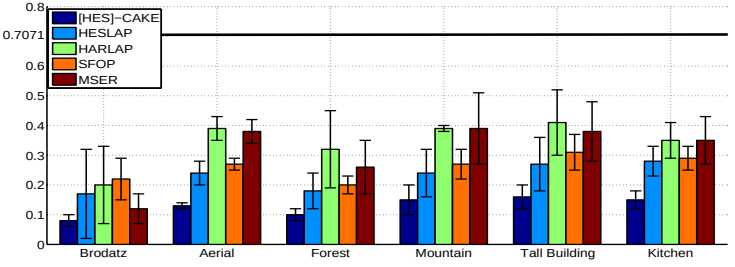


Figure 3: Average dissimilarity measure $d_H(p_H, p_c)$ for the different sets of features extracted over the categories of the dataset.

ness and complementarity scores obtained from the third image in each Oxford sequence are reported in Figure 4.

Figure 4 also shows the results of our complementarity evaluation. An important conclusion to be drawn is that combining features such as SFOP and [HES]-CAKE provides more than sufficient coverage of relevant content in the Oxford dataset. The combination of MSER with SFOP or [HES]-CAKE features is equally advantageous. SFOP mainly extracts corners and junctions, whereas the MSER algorithm usually extracts more homogeneous regions. This complementarity is particularly discernible when analysing the results for sequences such as “Trees” or “Wall”.

In [5], Dickscheid et al. report the low complementarity scores among Laplacian-based sets of features, e.g., HESLAP and HARLAP, which can also be inferred from the additional experiment with the Oxford dataset. On the other hand, the combination of the new set of features with the above mentioned ones does not imply a final set with redundancy. The combination of HESLAP and [HES]-CAKE yields the most complete set for “Bark", whereas the set of features retrieved by the former is the one with the lowest completeness score. We give a particular emphasis to this result since it shows that we are not extracting the same keypoints as the HESLAP. In fact, we are building a set of features that can be complementary to the one retrieved by the other method.

5.2 Repeatability

Although CAKE was designed for robust image representation, we have also evaluated the repeatability using the benchmark proposed in [20]. Figure reports the average repeatability score of the different algorithms for each sequence in the Oxford dataset, with an overlap error of 40%. In most of the sequences, SFOP and CAKE show a comparable performance, which is slightly worse than the one of HESLAP.

6 Conclusions and Perspectives

We have presented a keypoint extraction scheme that makes use of local information and is aware of the image composition. The notion of saliency is explicitly used in the algorithm: we compute a descriptor of the neighbouring area for every pixel and keypoints correspond to locations at which the information content is a local maximum.

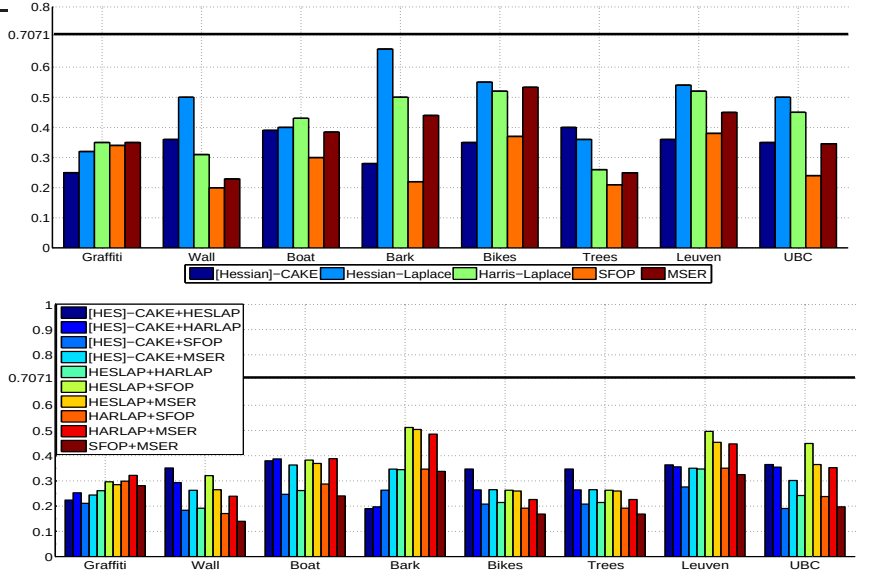


Figure 4: Dissimilarity measure $d_H(p_H, p_c)$ for different sets of features extracted from the third image in each Oxford sequence.

We have analysed the possible shortcomings that characterize our approach, especially the problem in estimating the probability of the inherent distributions in a way that it could reduce the computational complexity of the method and simultaneously make it comparable to state-of-the-art solutions in terms of repeatability rate and completeness.

While context-aware features are not be the most suitable option for tasks such as wide-baseline stereo matching and tracking, they become an appropriate choice for robust image representation, as the algorithm implicitly tries to cover the parts of the image that are informative, and this is reflected into the excellent completeness scores according to the metric defined in [5]. In fact, the results of our method are comparable to the best ones outlined in the evaluation. The complementarity between context-aware features and traditional ones is high, especially in textured scenes. We believe, therefore, that the applicability of our method can be found in tasks such as scene classification, image retrieval, object (class) recognition, and image compression.

Appendix Reduced KDE

Property 1 tells us that applying an affine transform to a codeword does not change the result of the extractor. We take advantage of this, and perform a principal component analysis (PCA). Let W_P be the new codeword distribution, in which $\mathbf{w_P}(\mathbf{x})$ are its elements. In this case, the inverse of the covariance matrix Σ_{W_P} is a diagonal matrix, where the diagonal elements contain the inverse of the variance of every variable of the multivariate distribution W_P. Therefore, we can write the Gaussian KDE in (2), using the Mahalanobis distance $d_M(\cdot,\cdot)$, as another Gaussian KDE with Euclidean distance as

$$\tilde{p}(\mathbf{w}_P(\mathbf{y})) = \frac{1}{N\Gamma} \sum_{\mathbf{x}\in\Phi} e\left(-\frac{\sum_{i=1}^{D} a_i (w_{P,i}(\mathbf{y}) - w_{P,i}(\mathbf{x}))^2}{2\sigma_k^2} \right), \tag{6}$$

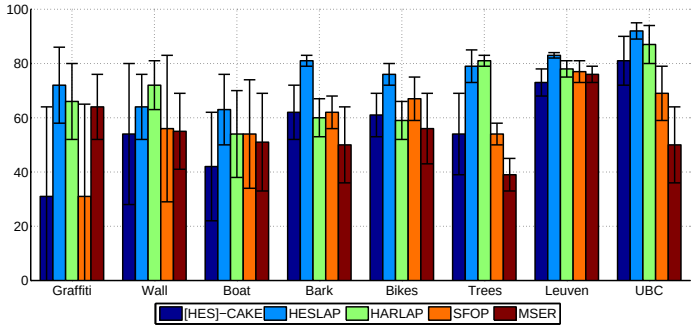


Figure 5: Average repeatability with an overlap error of 40% on the Oxford dataset.

where $a_i = \sqrt{\Sigma^{-1}_{w_P}(i,i)}$. Equation (6) can be rewritten as

$$\tilde{p}(\mathbf{w}_P(\mathbf{y})) = \frac{1}{N\Gamma}\sum_{\mathbf{x}\in\Phi}\prod_{i=1}^{D} e\left(-\frac{a_i(w_{P,i}(\mathbf{y}) - w_{P,i}(\mathbf{x}))^2}{2\sigma_k^2}\right). \tag{7}$$

Assuming that each dimension i provides a PDF that is independent of other dimensions, we can approximate (7):

$$\tilde{p}(\mathbf{w}_P(\mathbf{y})) \simeq \frac{1}{N\Gamma}\prod_{i=1}^{D}\sum_{\mathbf{x}\in\Phi}\exp\left(-\frac{a_i(w_{P,i}(\mathbf{y}) - w_{P,i}(\mathbf{x}))^2}{2\sigma_k^2}\right) \simeq \frac{1}{N\Gamma}\prod_{i=1}^{D}\tilde{p}_i(w_{P,i}(\mathbf{y})). \tag{8}$$

The approximation is only valid if PCA is able to separate the multivariate distribution into independent univariate distributions. While this is not always verified, the proposed approximation works sufficiently well for convex multivariate distributions, which is the case in all the experiments that we have conducted in this paper.

We have reduced a multivariate KDE to D univariate problems, which simplifies the computation of distances. Furthermore, we can approximate the D one dimensional KDEs to speed-up the process. For the sake of compactness and clarity, in the next part of the section we will refer to $\tilde{p}_i(w_{P,i}(\mathbf{y}))$ as $p(w(\mathbf{y}))$. We will also omit the constant $1/N\Gamma$ and the constants a_i.

We can extend the concept of KDE, by giving a weight $v(\mathbf{x}) > \mathbf{0}$ to each sample, so that the univariate KDE can be rewritten as a reduced KDE:

$$p_R(w(\mathbf{y})) = \sum_{\mathbf{x}\in\Phi_R} v(\mathbf{x}) e\left(-\frac{(w(\mathbf{y}) - w(\mathbf{x}))^2}{2\sigma_k^2}\right), \tag{9}$$

where $\Phi_R \subset \Phi$. This formulation can be seen as an hybrid between a Gaussian KDE and a Gaussian Mixture Model. The former has a large number of samples, all of them with unitary weight and fixed σ_k, while the latter has a few number of Gaussian functions, each one with a specific weight and standard deviation. The goal of our speed-up method is to obtain a set Φ_R with $|\Phi_R| = N_r \ll N$ samples that approximate the $\mathcal{O}(N^2)$ KDE. The main strategy is to fuse samples that are close each other into a new sample that "summarizes" them. Given a desired number of samples N_R, the algorithm progressively fuses pair of samples that has the minimum distance, as follows:

1: $\Phi_R \leftarrow \Phi$
2: $v(\mathbf{x}) \leftarrow 1,\ \forall \mathbf{x} \in \Phi$
3: **while** $|\Phi_R| > N_R$ **do**
4: $\quad \{\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1\} \leftarrow \arg\min_{\mathbf{x}_0,\mathbf{x}_1\in\Phi_R, \mathbf{x}_0\neq\mathbf{x}_1} |w(\mathbf{x}_0) - w(\mathbf{x}_1)|$

5: $v(\mathbf{x}_{01}) \leftarrow v(\tilde{\mathbf{x}}_0) + v(\tilde{\mathbf{x}}_1)$
6: $w(\mathbf{x}_{01}) \leftarrow \frac{v(\tilde{\mathbf{x}}_0)w(\tilde{\mathbf{x}}_0)+v(\tilde{\mathbf{x}}_1)w(\tilde{\mathbf{x}}_1)}{v(\tilde{\mathbf{x}}_0)+v(\tilde{\mathbf{x}}_1)}$
7: $\Phi_R \leftarrow (\Phi_R \setminus \{\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1\}) \cup \{\mathbf{x}_{01}\}$
8: **end while**

The algorithm uses as input the N samples of the univariate distribution (line 1), giving constant weight 1 to the samples (line 2). While the number of points is greater than N_R (line 3 to 8), the algorithm selects the pair of samples that shows the minimal distance in the set Φ_R (line 4), and a new sample is created (lines 5 and 6), whose weight v is the sum of the pair's weights and w is a weighted convex linear combination of the previous samples. The two selected samples are then removed from Φ_R and replaced by the new one (line 7).

To further speed-up the approximation, we can use a reduced number of dimensions $\tilde{D} < D$ such that the first $\tilde{D}^{th}$ dimensions of the multivariate distribution W_P cover 95% of the total distribution variance. This is a classical strategy for dimensionality reduction that has provided, in our tests, an average of $3\times$ further speed-up.

References

[1] A. Baumberg. Reliable Feature Matching Across Widely Separated Views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, volume 1, pages 1774–1781, 2000.

[2] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover, New York, NY, USA, 1966.

[3] N. Bruce. Features that Draw Visual Attention: An Information Theoretic Perspective. *Neurocomputing*, 65:125–133, 2005.

[4] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro. Principal curvature-based region detector for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, 2007.

[5] T. Dickscheid, F. Schindler, and W. Förstner. Coding images with local features. *International Journal of Computer Vision*, 94(2):154–174, 2011.

[6] G. Dorkó and C. Schmid. Selection of Scale-Invariant Parts for Object Class Recognition. In *IEEE International Conference on Computer Vision (ICCV'03)*, pages 634–640, 2003.

[7] W. Förstner. A Framework for Low Level Feature Extraction. In *European Conference on Computer Vision (ECCV'94)*, volume 3, pages 383–394, 1994.

[8] W. Förstner, T. Dickscheid, and F. Schindler. On the Completeness of Coding with Image Features. In *British Machine Vision Conference (BMVC'09)*, London, UK, 2009.

[9] W. Förstner, T. Dickscheid, and F. Schindler. Detecting interpretable and accurate scale-invariant keypoints. In *IEEE International Conference on Computer Vision (ICCV'09)*, Kyoto, Japan, 2009.

[10] S. Gilles. *Robust Description and Matching of Images*. PhD thesis, University of Oxford, 1998.

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

[12] T. Kadir, A. Zisserman, and M. Brady. An Affine Invariant Salient Region Detector. In *European Conference on Computer Vision (ECCV'04)*, pages 228–241, 2004.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for recognizing Natural Scene Categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2169–2178, 2006.

[14] F. Li and P. Perona. A Bayesian Hierarchial Model for Learning Natural Scene Categories. In *IEEE Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531, 2005.

[15] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[16] T. Lindeberg and J. Garding. Shape-adapted Smoothing in Estimation of 3-d depth Cues from Affine Distortions of Local 2-d Structures. *Image and Vision Computing*, 15, 1997.

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *British Machine Vision Conference 2002 (BMVC'02)*, pages 384–393, 2002.

[18] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision (ECCV'02)*, volume I, pages 128–142, 2002.

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

[20] 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.

[21] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple Object Class Detection with a Generative Model. In *IEEE Computer Vision and Pattern Recognition (CVPR'06)*, 2006.

[22] M. Mirmehdi and R. Periasamy. CBIR with Perceptual Region Features. In *Proc. of the British Machine Vision Conference 2001 (BMVC'02)*, 2001.

[23] L. Parida, D. Geiger, and R. Hummel. Junctions: Detection, Classification and Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7): 687–698, 1998.

[24] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[25] C. Schmid and R. Mohr. Local Grayvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[26] P. Schnitzspan, S. Roth, and B. Schiele. Automatic Discovery of Meaninful Objects Parts with Latent CRFs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 121–128, 2010.

[27] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, July 1948.

[28] Bill Triggs. Detecting keypoints with stable position, orientation and scale under illumination changes. In *European Conference on Computer Vision (ECCV'04)*, pages IV 100–113, 2004.

[29] T. Tuytelaars and L. Van Gool. Content-based Image Retrieval based on Local Affinely Invariant Regions. In *International Conference on Visual Information Systems*, pages 493–500, 1999.

[30] T. Tuytelaars and L. Van Gool. Matching Widely Separated Views based on Affine Invariant Regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[31] T. Tuytelaars and K. Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.