Context-Aware Features and Robust Image Representations

P. Martins, P. Carvalho, C. Gatta

PII: S1047-3203(13)00186-7
DOI: http://dx.doi.org/10.1016/j.jvcir.2013.10.006
Reference: YJVCI 1279

To appear in: *J. Vis. Commun. Image R.*

Received Date: 7 March 2013
Accepted Date: 26 October 2013

Please cite this article as: P. Martins, P. Carvalho, C. Gatta, Context-Aware Features and Robust Image Representations, *J. Vis. Commun. Image R.* (2013), doi: http://dx.doi.org/10.1016/j.jvcir.2013.10.006

Context-Aware Features and Robust Image Representations

P. Martins[a,*], P. Carvalho[a], C. Gatta[b]

[a] *Center for Informatics and Systems, University of Coimbra, Coimbra, Portugal*
[b] *Computer Vision Center, Autonomous University of Barcelona, Barcelona, Spain*

Abstract

Local image features are often used to efficiently represent image content. The limited number of types of features that a local feature extractor responds to might be insufficient to provide a robust image representation. To overcome this limitation, we propose a context-aware feature extraction formulated under an information theoretic framework. The algorithm does not respond to a specific type of features; the idea is to retrieve complementary features which are relevant within the image context. We empirically validate the method by investigating the repeatability, the completeness, and the complementarity of context-aware features on standard benchmarks. In a comparison with strictly local features, we show that our context-aware features produce more robust image representations. Furthermore, we study the complementarity between strictly local features and context-aware ones to produce an even more robust representation.

Keywords: Local features, Keypoint extraction, Image content descriptors, Image representation, Visual saliency, Information theory.

1. Introduction

Local feature detection (or extraction, if we want to use a more semantically correct term [1]) is a central and extremely active research topic in the fields of computer vision and image analysis. Reliable solutions to prominent problems such as wide-baseline stereo matching, content-based image retrieval, object (class) recognition, and symmetry detection, often make use of local image features (e.g., [2, 3, 4, 5, 6, 7]).

While it is widely accepted that a good local feature extractor should retrieve distinctive, accurate, and repeatable features against a wide variety of photometric and geometric transformations, it is equally valid to claim that these requirements are not always the most important. In fact, not all tasks require the same properties from a local feature extractor. We can distinguish three broad categories of applications according to the required properties [1]. The first category includes applications in which the semantic meaning of a particular type of features is exploited. For instance, edge or even ridge detection can be used to identify blood vessels in medical images and watercourses

*Corresponding author
Email addresses: pjmm@dei.uc.pt (P. Martins), carvalho@dei.uc.pt (P. Carvalho), c.gatta@cvc.uab.es (C. Gatta)

or roads in aerial images. Another example in this category is the use of blob extraction to identify blob-like organisms in microscopies. A second category includes tasks such as matching, tracking, and registration, which mainly require distinctive, repeatable, and accurate features. Finally, a third category comprises applications such as object (class) recognition, image retrieval, scene classification, and image compression. For this category, it is crucial that features preserve the most informative image content (robust image representation), while repeatability and accuracy are requirements of less importance.

We propose a local feature extractor aimed at providing a robust image representation. Our algorithm, named Context-Aware Keypoint Extractor (CAKE), represents a new paradigm in local feature extraction: no a priori assumption is made on the type of structure to be extracted. It retrieves locations (keypoints) which are representatives of salient regions within the image context. Two major advantages can be foreseen in the use of such features: the most informative image content at a global level will be preserved by context-aware features and an even more complete coverage of the content can be achieved through the combination of context-aware features and strictly local ones without inducing a noticeable level of redundancy.

This paper extends our previously published work in [8]. The extended version contains a more detailed description of the method as well as a more comprehensive evaluation. We have added the salient region detector [9] to the comparative study and the complementarity evaluation has been performed on a large data set. Furthermore, we have included a qualitative evaluation of our context-aware features.

2. Related work

The information provided by the first and second order derivatives has been the basis of diverse algorithms. Local signal changes can be summarized by structures such as the structure tensor matrix or the Hessian matrix. Algorithms based on the former were initially suggested in [10] and [11]. The trace and the determinant of the structure tensor matrix are usually taken to define a saliency measure [12, 13, 14, 15].

The seminal studies on linear scale-space representation [16, 17, 18] as well as the derived affine scale-space representation theory [19, 20] have been a motivation to define scale and affine covariant feature detectors under differential measures, such as the Difference of Gaussian (DoG) extractor [21] or the Harris-Laplace [22], which is a scale (and rotation) covariant extractor that results from the combination of the Harris-Stephens scheme [11] with a Gaussian scale-space representation. Concisely, the method performs a multi-scale Harris-Stephens keypoint extraction followed by an automatic scale selection [23] defined by a normalized Laplacian operator. The authors also propose the Hessian-Laplace extractor, which is similar to the former, with the exception of using the determinant of the Hessian matrix to extract keypoints at multiple scales. The Harris-Affine scheme [24], an extension of the Harris-Laplace, relies on the combination of the Harris-Stephens operator with an affine shape adaptation stage. Similarly, the Hessian-

Affine algorithm [24] follows the affine shape adaptation; however, the initial estimate is taken from the determinant of the Hessian matrix. Another differential-based method is the Scale Invariant Feature Operator (SFOP) [25], which was designed to respond to corners, junctions, and circular features. The explicitly interpretable and complementary extraction results from a unified framework that extends the gradient-based extraction previously discussed in [26] and [27] to a scale-space representation.

The extraction of KAZE features [28] is a multiscale-based approach, which makes use of nonlinear scale-spaces. The idea is to make the inherent blurring of scale-space representations locally adaptive to reduce noise and preserve details. The scale-space is built using Additive Operator Splitting techniques and variable conductance diffusion.

The algorithms proposed by Gilles [29] and Kadir and Brady [9] are two well-known methods relying on information theory. Gilles defines keypoints as image locations at which the entropy of local intensity values attains a maximum. Motivated by the work of Gilles, Kadir and Brady introduced a scale covariant salient region extractor. This scheme 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 peaked. There is also an affine covariant version of this method [30].

Maximally Stable Extremal Regions (MSER)[2] are a type of affine covariant features that correspond to connected components defined under certain thresholds. These components are said to be extremal because the pixels in the connected components have either higher or lower values than the pixels on their outer boundaries. An extremal region is said to be maximally stable if the relative area change, as a result of modifying the threshold, is a local minimum. The MSER algorithm has been extended to volumetric [31] and color images [32] as well as been subject to efficiency enhancements [33, 34, 35] and a multiresolution version [36].

3. Analysis and Motivation

Local feature extractors tend to rely on strong assumptions on the image content. For instance, Harris-Stephens and Laplacian-based detectors assume, respectively, the presence of corners and blobs. The MSER algorithm assumes the existence of image regions characterized by stable isophotes with respect to intensity perturbations. All of the above-mentioned structures are expected to be related to semantically meaningful parts of an image, such as the boundaries or the vertices of objects, or even the objects themselves. However, we cannot ensure that the detection of a particular feature will cover the most informative parts of the image. Figure 1 depicts two simple yet illustrative examples of how standard methods such as the Shi-Tomasi algorithm [13] can fail in the attempt of providing a robust image representation. In the first example (Fig. 1 (a)–(d)), the closed contour, which is a relevant object within the image context, is neglected by the strictly local extractor. On the other hand, the context-aware extraction retrieves a keypoint inside the closed contour as one of the most salient locations. The second example (Fig. 1 (e) and (f)) depicts the "Needle in a Haystack" im-

age and the overlaid maps (in red) representing the Shi-Tomasi saliency measure and our context-aware saliency measure. It is readily seen that our method provides a better coverage of the most relevant object.

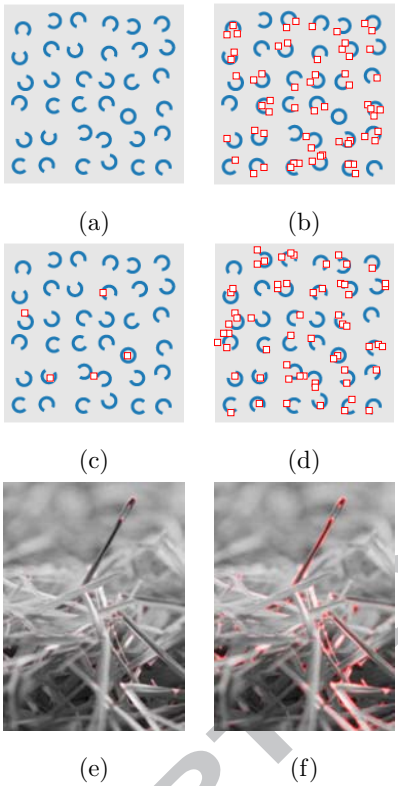

Figure 1: Context-aware keypoint extractions vs. strictly local keypoint extraction: 1. Keypoints on a psychological pattern: (a) pattern (input image); (b) 60 most salient Shi-Tomasi keypoints; (c) 5 most salient context-aware keypoints; (d) 60 most salient context-aware keypoints. 2. Saliency measures as overlaid maps on the "Needle in a Haystack" image: (e) Shi-Tomasi; (f) Context-aware. Best viewed in color.

Context-aware features can show a high degree of complementarity among themselves. This is particularly noticeable in images composed of different patterns and structures. The image in the top row of Fig. 2 depicts our context-aware keypoint extraction on a well-structured scene by retrieving the 100 most salient locations. This relatively small number of features is sufficient to provide a reasonable coverage of the image content, which includes diverse structures. However, in the case of scenes characterized by repetitive patterns, context-aware extractors will not provide the desired coverage. Nevertheless, the extracted set of features can be complemented with a counterpart that retrieves the repetitive elements in the image. The image in the bottom row of Fig. 2 depicts a combined feature extraction on a textured image in which context-aware features are complemented with SFOP features [25] to achieve a better coverage. In the latter example, one should note the high complementarity between the two types of features as well as the good coverage that the combined set provides.

4. Context-Aware Keypoints

Our context-aware feature extraction adopts an information theoretic framework in which the key idea is to use information content to quantify (and express) feature saliency. In our case, a context-aware keypoint will correspond to a particular point within a structure with a low probability of occurrence.

Shannon's measure of information [37] forms the basis for our saliency measure. If we consider a symbol s, its information is given by

$$I(s) = -\log(P(s)), \tag{1}$$

where $P(\cdot)$ denotes the probability of a symbol. For our purposes, using solely the content of a pixel

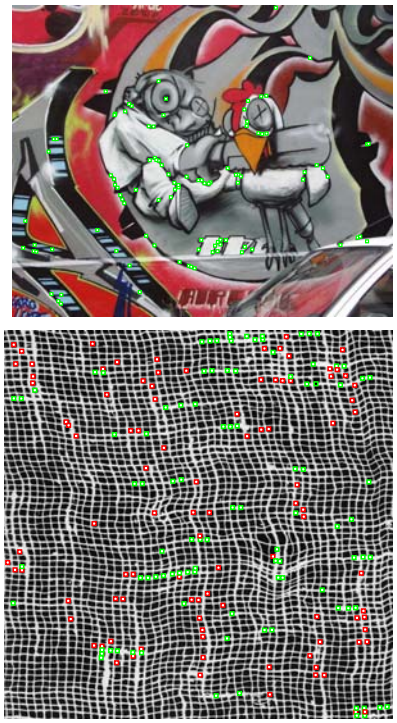

Figure 2: Proposed context-aware extraction. Top row: context-aware keypoints on a well-structured scene (100 most salient locations); bottom row: a combination of context-aware keypoints (green squares) with SFOP keypoints [25] (red squares) on a highly textured image. Best viewed in color.

$\mathbf{x}$ as a symbol is not applicable, whereas the content of a region around $\mathbf{x}$ will be more appropriate. Therefore, we will consider any local description $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^n$ that represents the neighborhood of $\mathbf{x}$ as a viable codeword. This codeword will be our symbol s, which allows us to rewrite Eq. (1):

$$I(\mathbf{x}) = -\log(P(\mathbf{w}(\mathbf{x}))). \tag{2}$$

However, in Shannon's perspective, a symbol should be a case of a discrete set of possibilities, whereas we have $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^n$. As a result, to estimate the probability of a certain symbol, a frequentists approach might be used. In this case, one should be able to quantize codewords into symbols. It is clear that the frequentists approach becomes inappropriate and the quantization becomes a dangerous process when applied to a codeword, since the quantization errors can induce strong artifacts in the $I(\mathbf{x})$ map, generating spurious local maxima.

We abandon the frequentist approach in favor of a Parzen Density Estimation [38], also known as Kernel Density Estimation (KDE). The Parzen estimation is suitable for our method as it is non-parametric, which will allow us to estimate any probability density function (PDF), as long as there is a reasonable number of samples. Using the KDE, we estimate the probability of a codeword $\mathbf{w}(\mathbf{y})$ as follows:

$$\hat{P}(\mathbf{w}(\mathbf{y})) = \frac{1}{Nh} \sum_{\mathbf{x} \in \Phi} K\left(\frac{d(\mathbf{w}(\mathbf{y}), \mathbf{w}(\mathbf{x}))}{h}\right), \tag{3}$$

where K denotes a kernel, d is a distance measure, h is a smoothing parameter called bandwidth and $N = |\Phi|$ is the cardinality of the image domain Φ. The key idea behind the KDE method is to smooth out the contribution of each sample $\mathbf{x}$ by spreading it to a certain area in $\mathbb{R}^n$ and with a certain shape as defined by the kernel K. There is a number of choices for the kernel. Nonetheless, the most commonly used and the most suitable is a multi-dimensional Gaussian function with zero mean and standard deviation σ_k. Using a Gaussian kernel, (3) can be rewritten as

$$\tilde{P}(\mathbf{w}(\mathbf{y})) = \frac{1}{N\Gamma} \sum_{\mathbf{x} \in \Phi} \mathrm{e}\left(-\frac{d^2(\mathbf{w}(\mathbf{y}), \mathbf{w}(\mathbf{x}))}{2\sigma_k^2}\right), \tag{4}$$

where h has been replaced by the standard deviation σ_k and Γ is a proper constant such that the estimated probabilities are taken from an actual PDF.

Summarizing, our saliency measure will be given by

$$m(\mathbf{y}) = -\log\left(\frac{1}{N\Gamma}\sum_{\mathbf{x}\in\Phi}\mathrm{e}\left(-\frac{d^2(\mathbf{w}(\mathbf{y}),\mathbf{w}(\mathbf{x}))}{2\sigma_k^2}\right)\right), \tag{5}$$

and context-aware keypoints will correspond to local maxima of m that are above a given threshold T.

For a complete description of the proposed method, we have to define a distance measure d and set a proper value to σ_k. Due to relevance of these two parameters in the process of estimating the PDF, they will be discussed in two separate subsections (4.1 and 4.2). Nonetheless, the KDE has an inherent and significant drawback: the computational cost. To estimate the probability of a pixel, we have to compute (4), which means computing N distances between codewords, giving a computational cost of $\mathcal{O}(N^2)$ for the whole image. The computational complexity of the KDE is prohibitive for images, where N is often of the order of millions. Different methods have been proposed to reduce the computation of a KDE-based PDF. Many methods rely on the hypothesis that the sample distribution forms separated clusters, so that it is feasible to approximate the probability in a certain location of the multivariate space using a reduced set of samples. Other methods have been devised for the purpose of a Parzen classifier, so that the cardinality of the training sample is reduced, without changing significantly the performance of the reduced Parzen classifier. In our case, none of the two aforementioned strategies can be straightforwardly used since (i) we cannot assume that the multivariate distribution forms different clusters, and (ii) we do not have ground truth labels to use the same strategy as the one defined for Parzen classifiers. We propose an efficient method that reduces the number of samples by approximating the full $\mathcal{O}(N^2)$ PDF in (4) with a $\mathcal{O}(N\log N)$ algorithm. A detailed explanation of the speed-up strategy can be found in Appendix B.

4.1. The distance d

To completely define a KDE-based approach, we have to define (i) the distance d, (ii) the kernel K, and (iii) the bandwidth h. These three parameters are interrelated since they will form the final "shape" of the kernel. As for the distance function d, we consider the Mahalanobis distance:

$$d(\mathbf{w}(\mathbf{x}),\mathbf{w}(\mathbf{y})) = \sqrt{(\mathbf{w}(\mathbf{x})-\mathbf{w}(\mathbf{y}))^T\Sigma_W^{-1}(\mathbf{w}(\mathbf{x})-\mathbf{w}(\mathbf{y}))}, \tag{6}$$

where $W = \bigcup_{\mathbf{x}\in\Phi}\mathbf{w}(\mathbf{x})$ and Σ_W is the covariance matrix of W. Using this distance, any affine covariant codeword will provide an affine invariant behavior to the extractor. In other words, any affine transformation will preserve the order of P. This result is summarized in the following theorem:

Theorem 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,*

$$P^{(2)}(\mathbf{x}_l) \leq P^{(2)}(\mathbf{x}_m) \iff P^{(1)}(\mathbf{x}_l) \leq P^{(1)}(\mathbf{x}_m), \forall \mathbf{x}_l, \mathbf{x}_m \in \Phi.$$

Proof: (See Appendix A).

4.2. The smoothing parameter σ_k

A Parzen estimation can be seen as an interpolation method, which provides an estimate of the continuous implicit PDF. It has been shown that,

for $N \to \infty$, the KDE converges to the actual PDF [38]. However, when N is finite, the bandwidth h plays an important role in the approximation. In the case of a Gaussian kernel, σ_k is the parameter that accounts for the smoothing strength.

The free parameter σ_k can potentially vanish the ability of the proposed method to adapt to the image context. When σ_k is too large, an over-smoothing of the estimated PDF occurs, canceling the inherent PDF structure due to the image content. If σ_k is too small, the interpolated values between different samples could be low, such that there is no interpolation anymore. We propose a method, in the case of univariate distribution, to determine an optimal sigma $\sigma_k^\star$, aiming at *sufficient blurring* while having the *highest sharpen* PDF between samples. We use univariate distributions since we approximate the KDE computations of a D-dimensional multivariate PDF by estimating D separate univariate PDFs (see Appendix B). From N samples w, we define the optimal σ_k for the given distribution as

$$\sigma_k^\star = \arg\max_{\sigma>0} \int_{w_i}^{w_{i+1}} \frac{1}{\sqrt{2\pi}\sigma} \frac{\mathrm{d}\left(e^{\frac{-(w-w_i)^2}{2\sigma^2}} + e^{\frac{-(w-w_{i+1})^2}{2\sigma^2}}\right)}{\mathrm{d}\,w} \,\mathrm{d}\,w, \tag{7}$$

where w_i and w_{i+1} is the farthest pair of consecutive samples in the distribution. It can be shown that, by solving (7), we have $\sigma_k^\star = |w_i - w_{i+1}|$. It can be also demonstrated that for $\sigma < |w_i - w_{i+1}|/2$, the estimated PDF between the two samples is concave, which provides insufficient smoothing. Using $\sigma_k^\star$ as defined above, we assure that we have *sufficient blurring* between the two farthest samples, while, at the same time, providing the *highest sharpen* PDF.

5. CAKE Instances

Different CAKE instances are constructed by considering different codewords. As observed by Gilles [29] and Kadir and Brady [9], the notion of saliency is related to rarity. What is salient is rare. However, the reciprocal is not necessarily valid. A highly discriminating codeword will contribute in turning every location into a rare structure; nothing will be seen as salient. On the other hand, with a less discriminating codeword, rarity will be harder to find. We will present a differential-based instance, which is provided by a sufficiently discriminating codeword. The strong link between image derivatives and the geometry of local structures is the main motivation to present an instance based on local differential information.

We propose the use of the Hessian matrix as a codeword to describe the local shape characteristics. We will consider components computed at multiple scales, which will allow us to provide an instance with a quasi-covariant response to scale changes. The codeword for the multiscale Hessian-based instance is

$$\begin{aligned}\mathbf{w}(\mathbf{x}) = \Big[\; & t_1^2 L_{xx}(\mathbf{x};t_1) \quad t_1^2 L_{xy}(\mathbf{x};t_1) \quad t_1^2 L_{yy}(\mathbf{x};t_1) \\ & t_2^2 L_{xx}(\mathbf{x};t_2) \quad t_2^2 L_{xy}(\mathbf{x},t_2) \quad t_2^2 L_{yy}(\mathbf{x},t_2) \\ & \dots \\ & t_M^2 L_{xx}(\mathbf{x};t_M) \quad t_M^2 L_{xy}(\mathbf{x};t_M) \quad t_M^2 L_{yy}(\mathbf{x};t_M) \; \Big]^T,\end{aligned} \tag{8}$$

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 $i = 1, \dots, M$, represents the scale.

6. Experimental Results and Discussion

Our experimental validation relies on a comparative study that includes both context-aware features and strictly local ones. We recall that our context-aware features were designed to provide a robust image representation, with or without the contribution of strictly local features.

We compare the performance of our Hessian-based instance, here coined as [HES]-CAKE, with some of the most prominent scale covariant algorithms: Hessian-Laplace (HESLAP) [22], Harris-Laplace (HARLAP) [22], SFOP [25], and the scale covariant version of the Salient Region detector (Salient) [9]. The MSER algorithm [2], which has an affine covariant response, is also included in the evaluation. All the implementations correspond to the ones provided and maintained by the authors.

We follow the evaluation protocol proposed by Dickscheid et al. [39] to measure the completeness and the complementarity of features. Completeness can be quantified as the amount of image information preserved by a set of features. Complementarity appears as a particular case of completeness: it reflects the amount of image information coded by sets of potentially complementary features. Measuring such properties is crucial as the main purpose of context-aware features is to provide a robust image representation, either in an isolated manner or in combination with strictly local features.

The metric for completeness is based on local statistics, which totally excludes the bias in favor of our context-aware features, since our algorithm is based on the analysis of the codeword distribution over the whole image. In fact, this evaluation gives a hint on the quality of the trade-off between the context-awareness and the locality of context-aware features. However, it does not provide a hint on how features cover informative content within the image context. If we take the "Needle in a Haystack" image depicted in Fig. 1 as an example, we can claim that strictly local features can show high completeness scores without properly covering the most interesting object in the scene. Note that such image representation, despite its considerable robustness, might be ineffectual if the goal is to recognize the salient object. Therefore, for a better understanding of the performance of our method, we complement the completeness analysis with a qualitative evaluation of the context-awareness.

Repeatability is also considered in our validation. We measure it through the standard evaluation protocol proposed by Mikolajczyk et al. [40]. Although the presence of repeatable features may not always be a fundamental requirement for tasks demanding a robust image representation, their existence is advantageous: a robust representation with repeatable features provides a more predictable coverage when image deformations are present. In addition, repeatable features allow a method to be used in a wider range of tasks.

Both evaluation protocols deal with regions as local features instead of single locations. To obtain regions from context-aware keypoints, a normalized Laplacian operator, $\nabla^2 L_n = t^2(L_{xx} + L_{yy})$, is used. The characteristic scale for each keypoint corresponds to the one at which the operator attains an extremum. This scale defines the radius of a circular region centered about the keypoint. Note that the CAKE instance does not solely respond to blob-

like keypoints: it captures other structures where scale selection can be less reliable (e.g., edges). Nevertheless, the combination of [HES]-CAKE with the normalized Laplacian operator provides robustness to scale changes, despite the resulting method not being entirely scale covariant. Figure 3 depicts the extraction of [HES]-CAKE regions using the Laplacian operator for scale selection.

Figure 3: [HES]-CAKE regions under viewpoint changes.

6.1. Completeness and complementarity evaluation

To measure the completeness of features, Dickscheid et al. [39] compute an entropy density $p_H(\mathbf{x})$ based on local image statistics and a feature coding density $p_c(\mathbf{x})$ derived from a given set of features. The measure of (in)completeness corresponds to the Hellinger distance between the two densities:

$$d_H(p_H, p_c) = \sqrt{\frac{1}{2}\sum_{\mathbf{x}\in\Phi}(\sqrt{p_H(\mathbf{x})} - \sqrt{p_c(\mathbf{x})})^2}, \quad (9)$$

where Φ is the image domain. When p_H and p_c are very close, the distance d_H will be small, which means the set of features with a coding density p_c effectively covers the image content (the set of features has a high completeness). Such metric penalizes the use of large scales (a straightforward solution to achieve a full coverage) as well as the presence of features in pure homogeneous regions. On the other hand, it will reward the "fine capturing" of local structures or superimposed features appearing at different scales (the entropy density takes into consideration several scales). The dataset contains six of the seven categories used in the original evaluation (Fig. 4). It comprises four categories of natural scenes [41, 42], the Brodatz texture collection [43] as well as a set of aerial images. The seventh category, which is comprised of different cartoon images, was not made publicly available.

The cardinality of the sets influences the completeness scores, as sparser sets tend to be less complete. While it is interesting to analyze the completeness of sets with comparable sparseness, one cannot expect similar cardinalities when dealing with different features types. We take such facts into consideration and, as a result, we perform two different tests. The first one corresponds to the main completeness test, which does not restrict the number of features. The second one allows us to make a direct comparison between our method and the salient regions algorithm by using the same number of features. Let $\mathcal{F}_{[HES]-CAKE}(g)$ and $\mathcal{F}_{Salient}(g)$ be the respective sets of [HES]-CAKE regions and Salient regions extracted from an image g. From each set, we extract the n highest ranked features (both methods provide a

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

Figure 4: Example images from the categories in the completeness and complementarity dataset.

well-define hierarchy among features), where $n = \min\{|\mathcal{F}_{[HES]-CAKE}(g)|, |\mathcal{F}_{Salient}(g)|\}$.

The parameter settings for our algorithm are outlined in Table 1. For the remaining algorithms, default parameter settings are used.

Table 1: Parameter settings for [HES]-CAKE.

Number of scales	3
t_{i+1}/t_i (ratio between successive scale levels)	1.19
t_0 (initial scale)	1.4
Non-maximal suppression window	3×3
Threshold	None
σ_k	optimal
N_R (number of samples)	200

Figure 5 is a summary of the main completeness evaluation. Results are shown for each image category, in terms of the distance $d_H(p_H, p_c)$. The plot includes the line $y = \sqrt{\frac{1}{2}}$, which corresponds to an angle of 90 degrees between $\sqrt{p_H}$ and $\sqrt{p_c}$. For a better interpretation, the average number of features per category is also shown. Regardless of the image collection, our context-aware instance retrieves more features than the other algorithms, which contributes to achieve the best completeness scores. The exception is the Brodatz category, which essentially contains highly textured images. For this category, salient regions achieve a better completeness score despite the lower number of regions.

The additional test computes the completeness scores of context-aware regions and salient regions for the first 20 images in each category using the same number of features. The results are summarized in Fig. 6. Here, salient regions achieve better results; however, the difference between scores is not significant. Aerial and Kitchen are the categories where context-aware features exhibit the lowest scores. This is explained by the strong presence of homogeneous regions, which might be part of salient objects within the image context, such as roads, rooftops (Aerial category), home appliances, and furniture (Kitchen category).

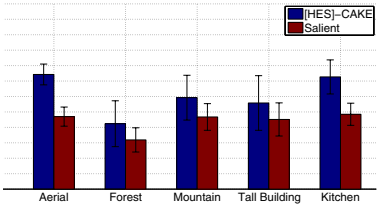

Figure 6: Average dissimilarity measure $d_H(p_H, p_c)$ for the different sets of features extracted over the categories of the dataset (20 images per category).

Complementarity was also evaluated on the first 20 images of each category by considering combinations of two different feature types. The results are summarized in Table 2. As expected, any combina-

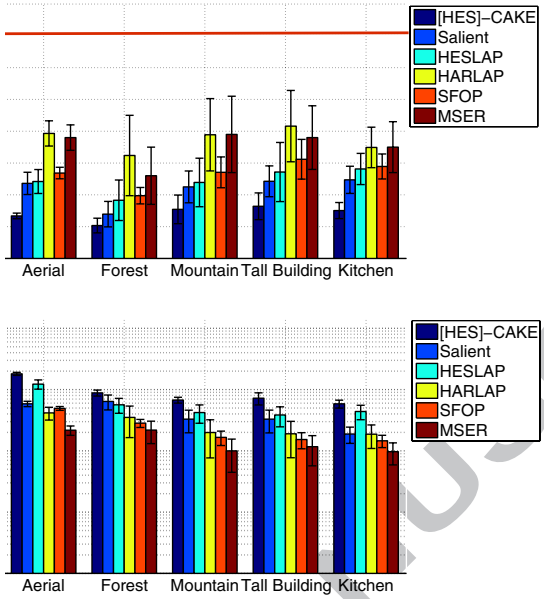

Figure 5: Completess results. Top row: Average dissimilarity measure $d_H(p_H, p_c)$ for the different sets of features extracted over the categories of the dataset. Bottom row: Average number of extracted features per image category.

tion that includes [HES]-CAKE regions achieves the best completeness scores. We give particular emphasis to the complementarity between HESLAP and [HES]-CAKE: both methods are Hessian-based and yet they produce complementary regions. The combination of [HES]-CAKE and Salient regions is also advantageous: the latter provides a good coverage of "busy" parts composed of repetitive patterns.

6.2. Context-awareness evaluation

For a qualitative evaluation of the context-awareness of [HES]-CAKE regions, we use three images typically used in the validation of algorithms for visual saliency detection [44]. Each one of the test images shows a salient object over a background containing partially salient elements. Fig-

Table 2: Average dissimilarity measure $d_H(p_H, p_c)$ for different sets of complementary features (20 images per category). Δ represents the difference $d_H(p_H, p_c) - \min\{d_1, d_2\}$, where d_1 and d_2 denote the average dissimilarity measures of the two different sets.

[HES]-CAKE	HESLAP	HARLAP	SFOP	MSER	SALIENT	$d_H(p_H, p_c)$	Δ
•	•					0.1028	-0.0352
•					•	0.1214	-0.0166
•		•				0.1242	-0.0138
•			•			0.1246	-0.0134
•				•		0.1253	-0.0127
	•				•	0.1375	-0.0579
			•		•	0.1550	-0.0404
	•		•			0.1725	-0.0493
		•			•	0.1765	-0.0189
				•	•	0.1789	-0.0165
	•			•		0.1983	-0.0235
	•	•				0.2187	-0.0031
		•	•			0.2274	-0.0366
		•		•		0.2895	-0.0311
			•	•		0.2052	-0.0588

ure 7 depicts the test images, the corresponding information maps given by the CAKE instance, as well as the coverage provided by the context-aware regions when 100 and 250 points are used. In all cases, our algorithm succeeds in covering distinctive elements of the salient objects. With 250 [HES]-CAKE regions, the coverage becomes a relatively robust image representation in all cases.

6.3. Repeatability Evaluation

The repeatability score between regions extracted in two image pairs is computed using 2D homographies as a ground truth. Two features (regions) are deemed as corresponding and, therefore, repeated, with an overlap error of $\epsilon_R \times 100\%$ if

$$1 - \frac{\left|\mathcal{R}_{\mu_1} \cap \mathcal{R}_{(H^T \mu_2 H)}\right|}{\left|\mathcal{R}_{\mu_1} \cup \mathcal{R}_{(H^T \mu_2 H)}\right|} < \epsilon_R, \qquad (10)$$

where R_μ denotes the set of image points in the elliptical region verifying $\mathbf{x}^T \mu \mathbf{x} \leq 1$ and H is the homography that relates the two input images. For a given pair of images and a given overlap error, the repeatability score corresponds to ratio between the number of correspondences between regions and the smaller of the number of regions in the pair of images. Only regions that are located in parts of the scene that are common to the two images are considered. The benchmark is supported by the Oxford image, which comprises 8 sequences of images, each one with 6 images, under different photometric and geometric transformations (Table 3). The repeatability of regions is computed within an overlap error of 40%, using the first image as a reference.

Table 4 outlines the parameter settings for [HES]-CAKE. Our method retrieves more features than its counterparts. Thus, for a fair evaluation of repeatability, we defined a threshold that avoids a discrepancy between the number number of features retrieved by [HES]-CAKE and the remaining algorithms.

Table 3: Image sequences in the Oxford dataset.

Sequence	**Scene type**	**Transformation**
Graffiti	well-structured	viewpoint change
Wall	textured	viewpoint change
Boat	well-structured + textured	zoom + rotation
Bark	textured	zoom + rotation
Bikes	well-structured	blur
Trees	textured	blur
Leuven	well-structured	lighting conditions
Boat	well-structured + textured	JPEG compression

Table 4: Parameter settings for [HES]-CAKE.

Scales	12
t_{i+1}/t_i	1.19
t_0 (initial scale)	1.19
Non-maximal suppression window	3×3
Threshold	12 (or 3000 points)
σ_k	optimal

Figure 8 reports the results in terms of average repeatability scores (top plot) and number of correspondences (bottom row) for each sequence. Among scale covariant features, HESLAP regions exhibit a slightly better overall repeatability score, namely in well-structured scenes (e.g., Bikes) where blob-like features are more present and well-defined.

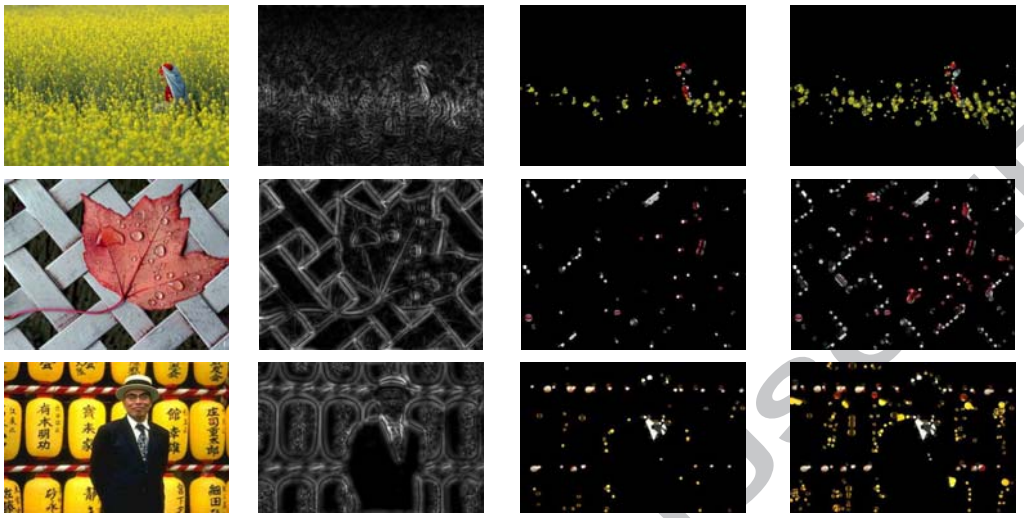

Figure 7: [HES]-CAKE information maps and extraction results in terms of coverage.

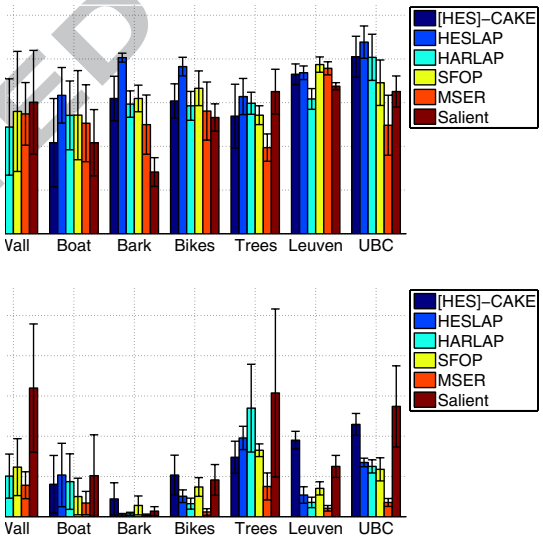

Figure 8: Repeatability score and number of correspondences with an overlap error of 40% on the Oxford dataset. Top row: Average repeatability. Error bars indicate the standard deviation. Bottom row: Average number of correspondences. Error bars indicate the standard deviation.

HARLAP has a similar performance, yielding the most repeatable results in textured scenes. The repeatability scores of SFOP and [HES]-CAKE are very similar, yet the latter responds to a higher number of features. Aside from viewpoint changes, the repeatability of MSER tends to be lower than its counterparts. In a direct comparison of information theoretic-based methods, we observe that [HES]-CAKE features are more repeatable than salient regions. The only two exceptions to this observation are the results for Trees and Wall sequences. Such results are explained by the fact that both sequences depict highly textured scenes, providing denser sets of salient regions. As for scale changes (Boat and Bark sequences), [HES]-CAKE regions show a sufficiently robust behavior. In the case of the Bark sequence, only HESLAP features are more repeatable than the proposed regions.

7. Conclusions

We have presented a context-aware keypoint extractor, which represents a new paradigm in local feature extraction. The idea is to retrieve salient locations within the image context, which means no assumption is made on the type of structure to be extracted. Such scheme was designed to provide a robust image representation, with or without the contribution of other local features.

The algorithm follows an information theoretic approach to extract salient locations. The possible shortcomings of such approach were analyzed, namely the difficulties in defining sufficiently discriminating descriptors and estimating the information of the inherent distributions in an efficient way.

The experimental evaluation has shown that relying on image statistics to extract keypoints is a winning strategy. A robust image representation can be easily achieved with context-aware features. Furthermore, the complementarity between context-aware features and strictly local ones can be exploited to produce an even more robust representation.

As for the applicability of the method, we believe that most of the tasks requiring a robust image representation will benefit from the use of context-aware features. In this category, we include tasks such as scene classification, image retrieval, object (class) recognition, and image compression.

Appendix A. Proof of Theorem 1

Proof. Let us suppose that $P^{(2)}(\mathbf{x}_l) \leq P^{(2)}(\mathbf{x}_m)$ (the reasoning will be analogous if we consider the other inequality). From the definition of probability, we have

$$\sum_{j=1}^{N} \mathrm{e}\,(-\frac{(\mathbf{w}^{(2)}(\mathbf{x}_l)-\mathbf{w}^{(2)}(\mathbf{x}_j))^T \Sigma^{-1}_{W^{(2)}} (\mathbf{w}^{(2)}(\mathbf{x}_j)-\mathbf{w}^{(2)}(\mathbf{x}_l))}{2\sigma_k^2}) \leq$$
$$\leq \sum_{j=1}^{N} \mathrm{e}\,(-\frac{(\mathbf{w}^{(2)}(\mathbf{x}_m)-\mathbf{w}^{(2)}(\mathbf{x}_j))^T \Sigma^{-1}_{W^{(2)}} (\mathbf{w}^{(2)}(\mathbf{x}_j)-\mathbf{w}^{(2)}(\mathbf{x}_m))}{2\sigma_k^2}).$$

Let A be the matrix that represents the transformation T (we assume no translation). Since $\Sigma_{W^{(2)}} = A\Sigma_{W^{(1)}}A^T$, the numerators from the exponents in the first and second members of the inequality can be rewritten as

$$(A(\mathbf{w}^{(1)}(\mathbf{x}_l)-\mathbf{w}^{(1)}(\mathbf{x}_j)))^T (A\Sigma_{W^{(1)}}A^T)^{-1} (A(\mathbf{w}^{(1)}(\mathbf{x}_j)-\mathbf{w}^{(1)}(\mathbf{x}_l)))$$

and

$$(A(\mathbf{w}^{(1)}(\mathbf{x}_m)-\mathbf{w}^{(1)}(\mathbf{x}_j)))^T (A\Sigma_{W^{(1)}}A^T)^{-1} (A(\mathbf{w}^{(1)}(\mathbf{x}_j)-\mathbf{w}^{(1)}(\mathbf{x}_m)),$$

respectively. By simplifying the previous expressions, we have

$$((\mathbf{w}^{(1)}(\mathbf{x}_l)-\mathbf{w}^{(1)}(\mathbf{x}_j)))^T \Sigma^{-1}_{W^{(1)}} (\mathbf{w}^{(1)}(\mathbf{x}_j)-\mathbf{w}^{(1)}(\mathbf{x}_l)))$$

and

$$((\mathbf{w}^{(1)}(\mathbf{x}_m) - \mathbf{w}^{(1)}(\mathbf{x}_j)))^T \Sigma^{-1}_{W^{(1)}} (\mathbf{w}^{(1)}(\mathbf{x}_j) - \mathbf{w}^{(1)}(\mathbf{x}_m))).$$

Thus,

$$P^{(2)}(\mathbf{x}) = \frac{1}{|\det A|} P^{(1)}(\mathbf{x}), \forall \mathbf{x} \in \Phi.$$

From the hypothesis, we have $P^{(1)}(\mathbf{x}_l) \leq P^{(1)}(\mathbf{x}_m)$.

□

Appendix B. Reduced KDE

As shown by Theorem 1, applying an affine transformation to the codewords does not change the result of the extractor. We take advantage of this, and perform a principal component analysis (PCA) to obtain a new codeword distribution W_P, where elements are denoted by $\mathbf{w}_P(\mathbf{x})$. In this case, the inverse of the covariance matrix $\Sigma^{-1}_{W_P}$ is a diagonal matrix, where the elements on the diagonal contain the inverse of the variance of every variable of W_P. Consequently, we can rewrite the Gaussian KDE in (4), using the Mahalanobis distance $d(\cdot,\cdot)$, as another Gaussian KDE with Euclidean distance as

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

where $a_i = \sqrt{\Sigma^{-1}_{W_P}(i,i)}$, i.e., the square root of the i^{th} diagonal element of the inverse of covariance matrix. Equation (Appendix B.1) can be rewritten as

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

By assuming that each dimension i provides a PDF that is independent of other dimensions, Equation (Appendix B.2) can be approximated as follows:

$$\begin{aligned}\tilde{p}(\mathbf{w}_P(\mathbf{y})) &\simeq \frac{1}{N\Gamma} \prod_{i=1}^{D} \sum_{\mathbf{x}\in\Phi} \mathrm{e}^{\left(-\frac{a_i (w_{P,i}(\mathbf{y}) - w_{P,i}(\mathbf{x}))^2}{2\sigma_k^2}\right)} \\ &\simeq \frac{1}{N\Gamma} \prod_{i=1}^{D} \tilde{p}_i(w_{P,i}(\mathbf{y})).\end{aligned} \quad \text{(Appendix B.3)}$$

Note that this approximation is only valid if PCA is able to separate the multivariate distribution into independent univariate distributions. This is not always verified. However, the proposed approximation works sufficiently well for convex multivariate distributions, which is the case in all the experiments we have conducted in the paper. Therefore, we have to compute D one dimensional KDEs $\tilde{p}_i(w_{P,i}(\mathbf{y}))$, using the Euclidean distance, which reduces a multivariate KDE to D univariate problems. This step simplifies the computation of distances between codewords, but still does not reduce the number of basic product-sum computations. Nevertheless, we can approximate the D one dimensional KDEs to speed-up the process. The fact that we have univariate distributions will be profitably used. 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:

$$p_R(w(\mathbf{y})) = \sum_{\mathbf{x}\in\Phi_R} v(\mathbf{x})\, \mathrm{e}^{\left(-\frac{(w(\mathbf{y}) - w(\mathbf{x}))^2}{2\sigma_k^2}\right)}, \quad \text{(Appendix B.4)}$$

where $\Phi_R \subset \Phi$. This formulation can be seen as a 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 idea is to fuse samples that are close each other into a new sample that "summarizes" them. Given a desired num-

ber of samples N_R, the algorithm progressively fuses pairs of samples that have a minimum distance:

1: $\Phi_R \leftarrow \Phi$
2: $v(\mathbf{x}) \leftarrow \mathbf{1},\ \forall \mathbf{x} \in \mathbf{\Phi}$
3: **while** $|\Phi_R| > N_R$ **do**
4: $\quad \{\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1\} \leftarrow \arg \min_{\mathbf{x_0}, \mathbf{x_1} \in \mathbf{\Phi_R},\, \mathbf{x_0} \neq \mathbf{x_1}} |w(\mathbf{x_0}) - \mathbf{w}(\mathbf{x_1})|$
5: $\quad v(\mathbf{x_{01}}) \leftarrow \mathbf{v}(\tilde{\mathbf{x}}_\mathbf{0}) + \mathbf{v}(\tilde{\mathbf{x}}_\mathbf{1})$
6: $\quad w(\mathbf{x_{01}}) \leftarrow \frac{\mathbf{v}(\tilde{\mathbf{x}}_\mathbf{0})\mathbf{w}(\tilde{\mathbf{x}}_\mathbf{0}) + \mathbf{v}(\tilde{\mathbf{x}}_\mathbf{1})\mathbf{w}(\tilde{\mathbf{x}}_\mathbf{1})}{\mathbf{v}(\tilde{\mathbf{x}}_\mathbf{0}) + \mathbf{v}(\tilde{\mathbf{x}}_\mathbf{1})}$
7: $\quad \Phi_R \leftarrow (\Phi_R \setminus \{\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1\}) \cup \{\mathbf{x_{01}}\}$
8: **end while**

The algorithm uses as input the N samples of the univariate distribution (line 1), giving constant weight 1 to all the samples (line 2). While the number of points is greater than the desired number N_R (line 3 to 8), the algorithm selects the pair of samples that show 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 the value w is a weighted convex linear combination of the previous samples. The two selected samples are then removed by the set Φ_R and replaced by the new one (line 7).

At first sight, the reduction algorithm seems may appear computationally expensive ($\sim \mathcal{O}(N^3)$), since a minimum distance over N_R^2 pairs of points has to be found. However, $w \in R$, so that $w(\mathbf{x})$ can be ordered at the beginning of the algorithm (with cost $\mathcal{O}(\lceil N \log N \rceil)$), and the pairs of minimum distance can be computed in N subtractions. Consequently, for each sample $\mathbf{x}$, we have the respective sample at minimal distance $\mathbf{x}_m$ and their distance $d_m(\mathbf{x}) = |w(\mathbf{x}) - w(\mathbf{x}_m)|$. This data can be represented using a self-balancing tree [45] allowing us to perform deletion and insertions (line 7), in $\log N$ time. Since the samples are ordered both in terms of $w(\mathbf{x})$ and $d_m(\mathbf{x})$, updating the distances after deletions and insertions can be done in $\mathcal{O}(1)$. Summarizing, we need to perform $2(N - N_r)$ deletions and $N - N_r$ insertions, so that the total cost of the reduction algorithm is proportional to $\lceil N \log N \rceil + 3(N - N_r) \log N$, thus being $\mathcal{O}(N \log N)$. The total cost to compute $p_R(w(\mathbf{y}))$ linearly depends on the desired N_R and the number of dimensions D.

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× further speed-up.

References

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

[2] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust Wide Baseline Stereo from Maximally Stable Extremal Regions, in: British Machine Vision Conference 2002 (BMVC'02), 2002, pp. 384–393.

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

[4] M. Mirmehdi, R. Periasamy, CBIR with Perceptual Region Features, in: Proc. of the British Machine Vision Conference 2001 (BMVC'01), 2001.

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

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

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

[8] P. Martins, P. Carvalho, C. Gatta, Contex Aware keypoint Extraction for Robust Image Representation, in: Proceedings of the 2012 British Machine Vision Conference (BMVC'12), 2012.

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

[10] W. Förstner, E. Gülch, A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features, in: ISPRS Conference on Fast Processing and Photogrammetric Data, 1987, pp. 281–305.

[11] C. Harris, M. Stephens, A combined corner and edge detector, in: Proc. of the 4th ALVEY Vision Conference, 1988, pp. 147–151.

[12] A. Noble, Descriptions of image surfaces, Ph.D. thesis, Department of Engineering Science, University of Oxford (1989).

[13] J. Shi, C. Tomasi, Good features to track, in: Proc. of the 1994 IEEE Computer Socitey Conference on Computer Vision and Pattern Recognition (CVPR'94)), 1994, pp. 593–600.

[14] K. Rohr, On 3d differential operators for detecting point landmarks, Image Vision Comput. 15 (3) (1997) 219–233.

[15] C. S. Kenney, B. S. Manjunath, M. Zuliani, G. Hewer, A. V. Nevel, A condition number for point matching with application to registration and post-registration error estimation, IEEE Trans. on Pattern Analysis and Machine Intelligence 25 (11) (2003) 1437–1454.

[16] A. P. Witkin, Scale-space filtering, in: Proceedings of the Eighth International Joint Conference on Artificial Intelligence - Volume 2, 1983, pp. 1019–1022.

[17] J. Koenderink, The structure of images, Biological Cybernetics 50 (5) (1984) 363–370.

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

[19] T. Lindeberg, 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.

[20] A. Baumberg, Reliable Feature Matching Across Widely Separated Views, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00), Vol. 1, 2000, pp. 1774–1781.

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

[22] K. Mikolajczyk, C. Schmid, An affine invariant interest point detector, in: European Conference on Computer Vision (ECCV'02), Vol. I, 2002, pp. 128–142.

[23] T. Lindeberg, Feature Detection with Automatic Scale Selection, International Journal of Computer Vision 30 (1998) 79–116.

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

[25] W. Förstner, T. Dickscheid, F. Schindler, Detecting interpretable and accurate scale-invariant keypoints, in: IEEE International Conference on Computer Vision (ICCV'09), Kyoto, Japan, 2009, pp. 2256–2263.

[26] W. Förstner, A Framework for Low Level Feature Extraction, in: European Conference on Computer Vision (ECCV'94), Vol. 3, 1994, pp. 383–394.

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

[28] P. F. Alcantarilla, A. Bartoli, A. J. Davison, KAZE features, in: Proceedings of the 12th European Conference on Computer Vision (ECCV'12), 2012.

[29] S. Gilles, Robust description and matching of images, Ph.D. thesis, University of Oxford (1998).

[30] T. Kadir, A. Zisserman, M. Brady, An Affine Invariant Salient Region Detector, in: European Conference on Computer Vision (ECCV'04), 2004, pp. 228–241.

[31] M. Donoser, H. Bischof, 3d segmentation by maximally

stable volumes (msvs), in: Proc. of the 18th International Conference on Pattern Recognition (ICPR'06), Vol. 1, 2006, pp. 63–66. doi:10.1109/ICPR.2006.33.

[32] P.-E. Forssén, Maximally stable colour regions for recognition and matching, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, IEEE, Minneapolis, USA, 2007.

[33] E. Murphy-Chutorian, M. Trivedi, N-tree Disjoint-Set Forests for Maximally Stable Extremal Regions, in: Proceedings of the British Machine Vision Conference, 2006.

[34] M. Donoser, H. Bischof, Efficient maximally stable extremal region (mser) tracking, in: Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 553–560.

[35] D. Nistér, H. Stewénius, Linear time maximally stable extremal regions, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 183–196.

[36] P.-E. Forssen, D. Lowe, Shape Descriptors for Maximally Stable Extremal Regions, in: Proc. of IEEE 11th Int. Conf. on Computer Vision, 2007, pp. 1–8.

[37] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423.

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

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

[40] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. V. Gool, A Comparison of Affine Region Detectors, International Journal of Computer Vision 65 (1/2) (2005) 43–72.

[41] F. Li, P. Perona, A Bayesian Hierarchial Model for Learning Natural Scene Categories, in: IEEE Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, 2005, pp. 524–531.

[42] S. Lazebnik, C. Schmid, 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), 2006, pp. 2169–2178.

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

[44] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans. on Pattern Analysis and Machine Intelligence.

[45] E. B. Koffman, P. A. T. Wolfgang, Objects, Abstraction, Data Structures and Design: Using C++, John Wiley & Sons, Inc., New York, NY, USA, 2007, Ch. 12.

Highlights:

- No a priori assumption is made on the type of structure to be extracted.
- Suitable for robust image representation.
- Different instances of the method can be created.
- In some cases, context-aware features can be complemented with strictly local features without inducing redundancy.
- Repeatability scores are comparable to state-of-the-art methods.