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]
cgatta@cvc.uab.es

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

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

We introduce a context-aware keypoint extractor, coined as CAKE, aimed at capturing the most informative image content. We find this algorithm particularly useful in tasks such as image retrieval, scene classification, and object (class) recognition, in which local features are mainly used to provide a robust and efficient image representation. We are motivated by the fact that the majority of local feature extractors are designed to respond to a reduced number of structures. Furthermore, we observe that the existent complementarity among feature sets is often neglected. Our context-aware algorithm is designed to respond to complementary features as long as they are informative. In the particular case of images with different types of structures, one can expect a high complementarity among the features retrieved by a context-aware extractor. By contrast, images with repetitive patterns will inhibit our method from retrieving a clear summarised description of the image content. Nonetheless, the extracted set of features can be complemented with a counterpart that retrieves the repetitive elements in the image. These two cases are depicted in Figure 1. The upper image shows a context-aware keypoint extraction on a well-structured scene, which retrieves the 100 most informative keypoints. This small number of features is sufficient to provide a good coverage of the content, which includes different types of structures. The lower image illustrates the advantages of combining context-aware keypoints with strictly local ones (SFOP keypoints [2]) to obtain a better coverage of images with repetitive patterns.

An information theoretic framework is used to formulate our context-aware keypoint extraction. A keypoint will correspond to a certain image location within a structure with a low probability of occurrence (high information content). For each image location $\mathbf{x}$, we consider $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^D$, any viable local representation (e.g, the Hessian matrix or the structure tensor matrix) as a "codeword" that represents the neighbourhood of $\mathbf{x}$. To define the saliency measure, we regard the image codewords as samples of a multivariate probability density function. We compute the probability of a codeword $\mathbf{w}(\mathbf{y})$ using a Kernel Density Estimator [4] in which the kernel is a multidimensional Gaussian function with zero mean and standard deviation σ_k:

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

where d is a distance function, K is a kernel, Φ is the image domain, N represents the number of pixels, and Γ is a proper constant such that the estimated probabilities are taken from an actual PDF. From Eq. (1), the saliency measure at $\mathbf{y}$ is defined as

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

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. We use the *Mahalanobis distance* as the distance function d. Since this distance is invariant under affine transformations, we can draw the following result:

Property 1. *Let $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ be image 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,*

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

We propose a strategy that includes approximating the KDE computations of a D-dimensional multi-variate PDF by estimating D separate univariate PDFs, which simplifies the computation of distances. Furthermore, we reduce the number of samples: samples that are close to each other are replaced by a new one that summarises the previous ones.

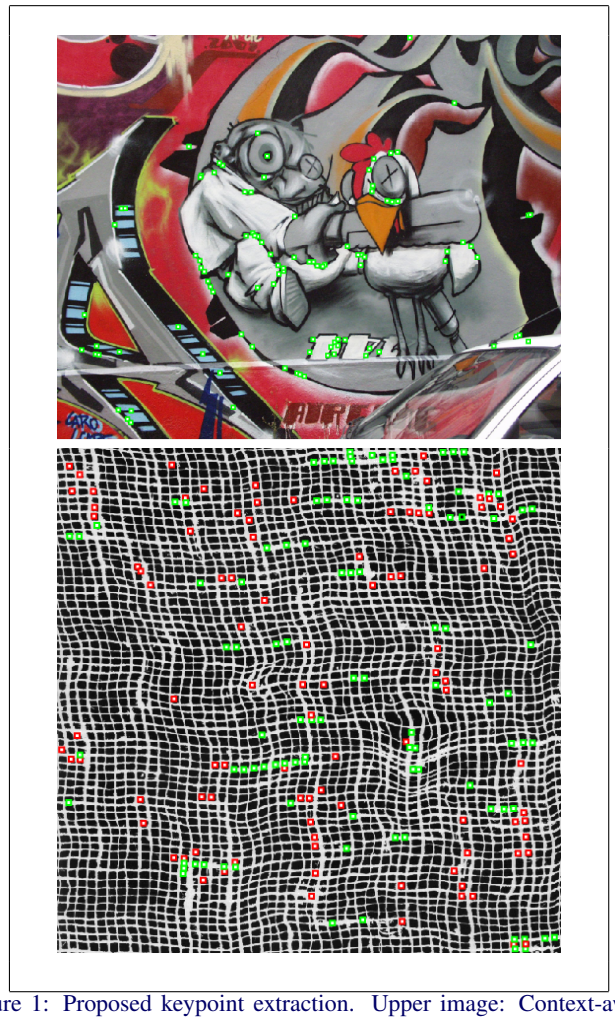

Figure 1: Proposed keypoint extraction. Upper image: Context-aware keypoints on a well-structured scene (100 most informative locations). Lower image: a combination of context-aware keypoints (green squares) with SFOP keypoints [2] (red squares) on a textured image.

The context-aware keypoint extractor is evaluated in terms of repeatability, completeness, and complementary. A multi-scale Hessian matrix is used as the codeword. Repeatability is evaluated following the standard protocol proposed by Mikolajczyk et al. [3]. Completeness and complementarity are evaluated on the benchmark proposed by Dickscheid et al. [1].

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

[2] 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.

[3] 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.

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