J. Vis. Commun. Image R. 25 (2014) 339–348

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

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

ARTICLE INFO

Article history:
Received 7 March 2013
Accepted 26 October 2013
Available online 5 November 2013

Keywords:
Local features
Keypoint extraction
Image content descriptors
Image representation
Visual saliency
Information theory
Kernel estimators
Complementarity

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.

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., [2–7]).

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 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

* Corresponding author.
E-mail addresses: pjmm@dei.uc.pt (P. Martins), carvalho@dei.uc.pt (P. Carvalho), cgatta@cvc.uab.es (C. Gatta).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.10.006

can be summarized by structures such as the structure tensor matrix or the Hessian matrix. Algorithms based on the former were initially suggested in [10,11]. The trace and the determinant of the structure tensor matrix are usually taken to define a saliency measure [12–15].

The seminal studies on linear scale-space representation [16–18] 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 [26,27] to a scale-space representation.

The extraction of KAZE features [28] is a multiscale-based approach, which makes use of non-linear 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 [33–35] 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

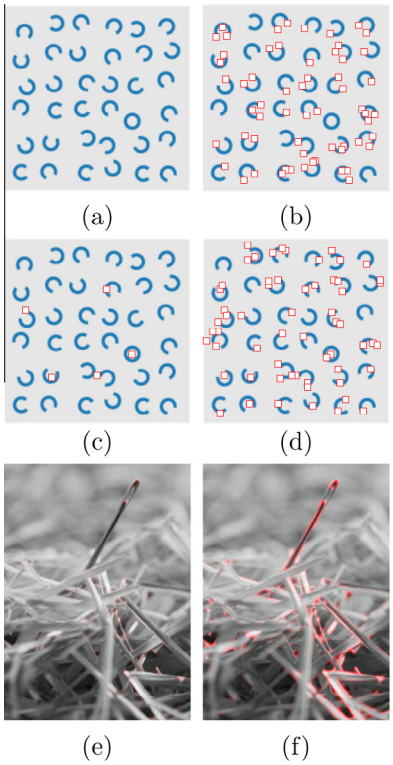

Fig. 1. Context-aware keypoint extractions vs. strictly local keypoint extraction: 1. Keypoints on a psychological pattern: (a) pattern (input image); (b) 60 most salient Shi–Tomasi keypoints; (c) 5 most salient context-aware keypoints; (d) 60 most salient context-aware keypoints. 2. Saliency measures as overlaid maps on the "Needle in a Haystack" image: (e) Shi–Tomasi; (f) Context-aware. Best viewed in color. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

themselves. However, we cannot ensure that the detection of a particular feature will cover the most informative parts of the image. Fig. 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))[1] depicts the "Needle in a Haystack" image 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.

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

[1] For interpretation of color in Figs. 1–3 and 5–8, the reader is referred to the web version of this article.