Image and Vision Computing 67 (2017) 43–51

Contents lists available at ScienceDirect

Image and Vision Computing

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

Texture features for object salience[☆]

Kasim Terzić[a, c, *], Sai Krishna[b, c], J.M.H. du Buf[c]

[a] School of Computer Science, University of St Andrews, Scotland, United Kingdom
[b] Department of Science and Technology, Centre for Applied Autonomous Sensor Systems, Örebro University, Sweden
[c] Department of Electronic Engineering and Computer Science, University of the Algarve, Portugal

ARTICLE INFO

Article history:
Received 15 November 2016
Received in revised form 2 July 2017
Accepted 18 September 2017
Available online 22 September 2017

Keywords:
Texture
Colour
Salience
Attention
Benchmark

ABSTRACT

Although texture is important for many vision-related tasks, it is not used in most salience models. As a consequence, there are images where all existing salience algorithms fail. We introduce a novel set of texture features built on top of a fast model of complex cells in striate cortex, i.e., visual area V1. The texture at each position is characterised by the two-dimensional local power spectrum obtained from Gabor filters which are tuned to many scales and orientations. We then apply a parametric model and describe the local spectrum by the combination of two one-dimensional Gaussian approximations: the scale and orientation distributions. The scale distribution indicates whether the texture has a dominant frequency and what frequency it is. Likewise, the orientation distribution attests the degree of anisotropy. We evaluate the features in combination with the state-of-the-art VOCUS2 salience algorithm. We found that using our novel texture features in addition to colour improves AUC by 3.8% on the PASCAL-S dataset when compared to the colour-only baseline, and by 62% on a novel texture-based dataset.

1. Introduction

The seminal work by Itti, Koch and Niebur [26,27] included an orientation component from responses of oriented Gabor filters. However, since then, texture has largely been ignored in computational salience models. Most recent work on salience has focused on the pop-out effect primarily caused by colour and intensity, and widely-used benchmarks in this field mostly feature prominent, brightly coloured objects. Colour and intensity are undoubtedly very important cues, but texture can also evoke a pop-out effect; see Fig. 1. Any observer immediately experiences the striking effect in the left image, but most state-of-the-art salience models will fail to identify the salient region. The remarkable success of these models on challenging datasets has unfortunately led to a neglect of texture as an attentional cue.

In this paper, we revisit the Itti and Koch model and examine which types of features are well-suited to detecting salient regions on the basis of texture. We then propose a simple set of features on top of complex cells. We combine these features with the recent state-of-the-art VOCUS2 algorithm, which evolved from the Itti and Koch framework, in order to demonstrate the effectiveness of our approach. We evaluate our approach on a set of standard datasets, and on a novel dataset which specifically addresses texture.

We see this work as a first step towards a texture-based salience mechanism based on a fast model of cortical cells in V1 [54]. To the best of our knowledge, this is the first model of its kind, which can provide a baseline for further work in this area. We do not expect that a purely texture-based approach will ever outperform colour-based approaches. Rather, we are convinced that an additional salience channel can improve existing algorithms in situations where object and background colours are similar.

2. Related work

Visual salience has become one of the central topics in computer vision over the past few decades, and considerably longer in the field of human and biological vision. In order to deal with the inherent complexity of the visual world, biological systems have evolved a way to prioritise information by identifying objects, or parts, which stand out from the rest, and which are likely to characterise the essence of the surrounding scene. The concept of Bayesian surprise has been explored to model this process [25]. Psychophysical experiments have shown that texture is perceived in a pre-attentive

[☆] This paper has been recommended for acceptance by Michael Felsberg.
[*] Corresponding author at: School of Computer Science, University of St Andrews, Scotland, United Kingdom.

E-mail addresses: kt54@st-andrews.ac.uk (K. Terzić), sai.krishna@oru.se (S. Krishna), dubuf@ualg.pt (J.M.H. du Buf).
URLs: http://kt54.host.cs.st-andrews.ac.uk (K. Terzić), http://mpi.aass.oru.se/people/sai-krishna/ (S. Krishna), http://w3.ualg.pt/~dubuf/ (J.M.H. du Buf).

https://doi.org/10.1016/j.imavis.2017.09.007
0262-8856/© 2017 Elsevier B.V. All rights reserved.

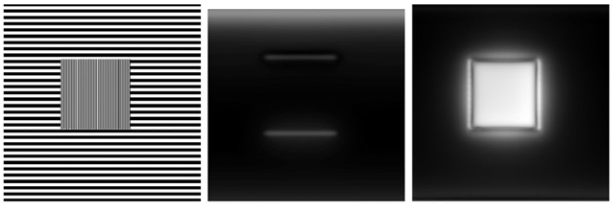

Fig. 1. An example of texture salience. The textured region in the left image leads to a strong pop-out effect, despite it having the same average colour and intensity as the surrounding region. Blob detection based on colour therefore fails in this case (middle image). However, blob detection based on texture features, as described in this paper, detects the salient blob (right image).

fashion [48]. Pre-attentive means bottom-up and data-driven, which is also referred to as covert attention, in contrast to overt, consciously directed attention.

In their influential work, Laurent Itti, Christof Koch and Ernst Niebur [26,27] introduced a filtering approach to covert attention. Their model, which inspired countless others, extracts salience by a combination of centre-surround DoG filters. They applied these filters to feature maps which consist of colour channels and the responses of oriented Gabor filters, thus mimicking early biological vision. Their model was designed for explaining sequential saccadic eye movements, from the most conspicuous image point to other points with decreasing order of conspicuity and inhibition of return. The recent algorithm VOCUS2 by Frintrop et al. has extended the same principle to detecting larger salient regions instead of points [14], demonstrating the continued usefulness of the concept. The original Itti and Koch model has been extended numerous times, for example by weighting the different feature maps after identifying useful features [24] and by exploring the role of salience in overt attention [46]. In addition, eye fixation maps have been combined with traditional segmentation methods in order to model the segmentation of salient regions [35]. The idea of contrasting the centre of a region against its surround has also been applied using different similarity measures. Bruce and Tsotsos used information content of the two regions [6] for their AIM model, while Klein and Frintrop used the KL divergence of feature statistics [30] and later multivariate probability distributions [31].

Much research in recent years has moved towards detecting entire salient objects in scenes. For testing the methods, there exist several high-profile benchmarks of natural images where the task is to segregate a prominent object. Most of the current approaches try to segment an entire object, and regions can be modelled according to their colour and luminance [1], contrast [8,9] or dissimilarity [13]. Another approach is to learn a correct foreground object segmentation from a set of training images [38]. This object-based salience can be very important for providing top-down feedback for scene understanding in artificial intelligence [43,51] and cognitive robotics [32,53]. Yet other methods try to represent the scene in terms of visual perception [17], graph-based visual salience [21], and object-based salience features [20]. Additionally, salience has also been modelled as a discriminant process [16] and as a regression problem [28]. Multi-scale processing has been shown to improve salience on small-scale, high-contrast patterns [59].

Despite the vast variety of developed methods, almost all are based on colour and intensity. These feature channels are very convenient: an object with largely constant colour which differs from the colour of its background will generate a strong response from an appropriately-sized centre-surround filter. However, the prevalence of colour-based features is also partly due to the way that modern benchmarks have been designed: most images feature brightly-coloured objects that are particularly suited to being identified by colour. Unfortunately, this benchmarking aspect has contributed to the neglect of other important feature channels. The result is that a completely trivial example as shown in Fig. 1 defeats nearly every available salience algorithm. This example creates a pop-out effect solely on the basis of texture, not colour nor intensity, and only very few salience methods explicitly employ spatial frequency or texture. The original Itti and Koch model included responses of oriented Gabor filters as one of the feature channels, so at least local orientation could play a role. However, this feature was found not to contribute strongly to the final results, and in recent variations of the Itti and Koch model this channel is ignored altogether [14]. Achanta et al. [2] used bandpass filtering to obtain uniform regions with sharp boundaries, but their features were still based on colour. Texture models have typically been used for texture segmentation, and are often built on top of Gabor filter responses, followed by further processing such as spatial averages of local neural responses [39]. Alternatively, a bank of matched filters for specific textures can be used [33], but performance becomes limited by the representativeness of the chosen filters. Wavelets have also been used to successfully classify different textures [4]. Typically, texture segmentation is based on some kind of feature gradient (or feature contrast), and the maxima represent texture boundaries. Although texture models (and especially Gabor-based texture models) have been extensively benchmarked [18] and successfully used for texture segmentation [44] and classification [4], comparatively few authors have explored their use for salience and attention models.

The earliest work on texture-based salience was probably by Sayeda-Mahmood [48]. The algorithm produces four binary maps from the image, and constructs a number of features, including the total number of holes in a region, the area occupied by holes in a white region, and the shape and distribution of the holes. A heuristic algorithm then combines these into a salience score. The features are complex to compute because they involve region growing, counting and computing convex hulls, and they were only tested on artificial images in a segmentation context. Building on the Itti and Koch model, Li's method [37] employed responses of V1 cells directly to detect pop-out effects in simple textures consisting of oriented textons. This work has been extended to multi-spectral features and a large number of textons [56], although it was only tested on a novel multi-spectral dataset. In [7], texture features are used to detect edges and combined with an object model to fill the rest of the salient object. Kalinke et al. [29] used co-occurrence matrices in order to extract texture-based features for creating hypotheses in an intelligent vehicle scenario. Powerful texture models for video [10,58] are often difficult to use within the centre-surround filtering context, but they can be used within a discriminative framework [15]. More recently, the eye fixation model of Momtaz and Daliri uses human fixations to train a salience model using features like orientation and spatial frequency [41]. However, most of the above approaches are either difficult to apply within a centre-surround filtering context, or they do not aim to be general enough for salient region detection in natural images.

There are several approaches which build salience maps from the frequency spectrum of the image. The method of Hou and Zhang is based on the global Fourier transform [23]. They subtract the average log-spectrum of many images from the log-spectrum of a specific image. This produces a residual spectrum. When this spectrum is transformed back to the spatial domain, it indicates salient regions which potentially correspond to objects. Guo et al. [19] built on this concept, but argued that the phase, not the amplitude, of the spectrum is key to finding salient regions. They extended this concept to the Quaternion Fourier Transform which can represent intensity, colour and motion of each pixel. A more recent take on quaternion-based salience was proposed by Schauerte and Stiefelhagen [49], whose method achieved state-of-the-art results on predicting human eye fixations. These methods are not biologically plausible, nor are they based explicitly on texture, but our