Signal Processing: *Image Communication* 38 (2015) 15–31

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

A perceptually based spatio-temporal computational framework for visual saliency estimation

Petros Koutras*, Petros Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Zografou Campus, Athens 15773, Greece

ARTICLE INFO

Available online 20 August 2015

Keywords:
Spatio-temporal visual frontend
3D Gabor filters
LAB color space
Visual saliency
Eye-tracking database

ABSTRACT

The purpose of this paper is to demonstrate a perceptually based spatio-temporal computational framework for visual saliency estimation. We have developed a new spatio-temporal visual frontend based on biologically inspired 3D Gabor filters, which is applied on both the luminance and the color streams and produces spatio-temporal energy maps. These volumes are fused for computing a single saliency map and can detect spatio-temporal phenomena that static saliency models cannot find. We also provide a new movie database with eye-tracking annotation. We have evaluated our spatio-temporal saliency model on the widely used CRCNS-ORIG database as well as our new database using different fusion schemes and feature sets. The proposed spatio-temporal computational framework incorporates many ideas based on psychological evidences and yields significant improvements on spatio-temporal saliency estimation.

1. Introduction

In biological vision systems there exist significant neurobiological and psychophysical evidences that the first stages of visual information processing include many feature detection processes. Since various stages of biological vision systems involve spatio-temporal processing and nature has a tendency to represent information in optimal ways, efficient perception-inspired spatio-temporal processing as well as easily computable features that can compactly represent salient structure in moving images should be one of the important early goals of video processing.

Visual attention is a cognitive mechanism employed by humans, animals and artificial systems for selecting the most important part of information from a visual stimulus and then perform more complex and demanding processes. This field has been for years an active research subject for psychophysics and cognitive scientists, because attention mechanisms play a dominant role in human visual system.

Attention may have two modes, a top-down expectation-driven, and a bottom-up stimulus-driven, and so there is often a confusion between attention and visual saliency. Visual attention is a wider concept, which often includes many topics, such as top-down cognitive information processing, memory, object searching, task demands or expectations. On the other hand, visual saliency is a bottom-up process and is based on the sensory cues of a stimulus that make certain image or video regions more conspicuous. In addition to its cognitive and biological nature, several computational frameworks have also been proposed for modeling visual saliency [1], because it plays a significant role in many computer vision applications, such as object and action recognition [2–5] and movie summarization [6–8].

We propose a spatio-temporal computational frontend for visual saliency, which is suitable for estimating spatio-temporal events in video streams. Its design is built upon

* Corresponding author.
E-mail addresses: pkoutras@cs.ntua.gr (P. Koutras), maragos@cs.ntua.gr (P. Maragos).

http://dx.doi.org/10.1016/j.image.2015.08.004
0923-5965/© 2015 Elsevier B.V. All rights reserved.

many ideas from biological and perceptual image processing, related to human vision modeling. During the past decades several computational approaches have been developed for visual saliency estimation in the spatial domain, which had incorporated many advanced techniques for processing the luminance and color modalities. More recently there appeared models for spatio-temporal estimation in video stimuli that are mainly based on simple motion estimation or spatio-temporal filtering rather than using only the classic static cues (intensity, color, orientation). Our approach is designed for spatio-temporal estimation and incorporates advances in both static and spatio-temporal pathways. A brief summary of biologically inspired feature detection methods as well as the spatial and spatio-temporal computational models is given in Section 2.

Our framework exhibits unification and computational economy in at least three important ways: it produces both spatio-temporal and static energy volumes by using the same multi-scale filterbank based on quadrature Gabor filters in three dimensions (space and time). In addition, the same framework can be applied for two different modalities, i.e. the image luminance and color stream modalities, producing independent spatio-temporal energy volumes. For the color stream we have incorporated many modern ideas such as LAB color space or PCA analysis. Further, our spatio-temporal framework can provide motion information in different scales and directions without having to process it as a separate cue or use a small number of frames like other video saliency approaches require. In this way, our approach achieves to detect both the fastest changes in the video stimuli (e.g. flicker) and the slowest motion changes related to action events. The produced energy maps can be integrated into a single spatio-temporal saliency map, by using different energy mixtures and fusion schemes. The complete model and the filtering details are analyzed in Section 3.

Our computational approach is evaluated in two different ways. At first we employed simple spatio-temporal stimuli, where our method manages to detect time-varying events that static saliency method cannot find. The second application is the prediction of human eye fixations while the subjects watch video stimuli, using a single spatio-temporal saliency map. We use two databases with eye-tracking data annotation: the CRCNS-ORIG [9] and our newly created *Eye-Tracking Movie Database (ETMD)*. The latter was collected for the purposes of the presented study and comprises short video clips from Hollywood movies along with eye tracker data for 10 subjects. In Section 4 we describe the evaluation procedure and our experimental results in both these databases. In general, our method for spatio-temporal saliency estimation is quite promising as it achieves higher performance than many other state-of-the-art saliency models.

2. Background/related work

Assuming that visual information processing by several classes of optical neurons can be modeled by linear operators, there was a hot debate in the perceptual and neurophysiological research community during the 1960s and 1970s as to whether the early stages of visual information processing in primates can be modeled as spatial local feature detectors or as filterbanks in the frequency domain. From the side of spatial processing, Hubel and Wiesel [10,11] found in cat's and monkey's visual cortex simple cells whose behavior they described as approximately linear *feature detectors* with line-, edge- or bar-shaped receptive fields that exhibited scale and orientation selectivity. Since then these results have been confirmed and refined by many other researchers [12–14]. From the side of frequency domain, several researchers have argued, based on psychophysical experiments, that the early visual system can be approximately modeled using Fourier analysis ideas [15–17], mainly in one-dimension (1D) until Daugman [18] proposed a two-dimensional (2D) spatial filtering and Fourier analysis. In another experimental direction, Pollen and Ronner [19] found that adjacent simple cells in the visual cortex are tuned to the same spatial frequency and orientation, but their responses are in *quadrature*.

Daugman [20] also observed that, from a mathematical viewpoint, the antagonism between the spatial and frequency domain interpretations of visual information processing is illusionary, since neurons in the retina or visual cortex can both resemble filterbanks of bandpass filters, or, equivalently, convolutions with neuron responses that have excitatory or inhibitory regions in their center-surround receptive fields. He further extended the existing 1D Gabor theory [21,22] and proposed the 2*D oriented Gabor filters* as optimal models for simple cell impulse responses, where ‘optimality’ here means having minimal *space–frequency uncertainty*. Since then, Gabor filters in quadrature pairs have been extensively used in many early computer vision tasks, e.g. in 2D spatial texture analysis [23] and in spatio-temporal models for motion [24] and optical flow estimation [25,26].

For the modeling of receptive fields (RFs) of cells in the visual system in parallel to the use of Gabor filters, a few other approaches were also proposed such as *Difference of Gaussians* (*DOG*) filters by Wilson and Bergen [27] and the *Derivatives of Gaussians (GD)* (or in discrete form *Difference of Offsets of Gaussians* (*DOOG*)) by Young [28,29]. The former has limited applicability and was used mainly for isotropic center-surround RFs and edge detection [30]. The latter found a wider acceptance and was used for modeling the RFs of simple cells in primate visual systems, by applying *Gaussian Derivatives* up to tenth order instead of Gabor filters. As Koenderink and van Doorn [31] proved, for high orders the Gaussian derivatives become approximate Gabor filters. Later, the GD model was extended to spatio-temporal vision [32].

In addition, all the above filter models come at *multiple scales* (corresponding to the various frequency channels) and may be either isotropic (e.g. in the retina) or *oriented* (e.g. in the visual cortex). Moreover, there have also been other perception-inspired models for feature detection that are nonlinear and based on ideas of phase congruency and quadrature energy, as in Morrone et al. [33,34]. Filterbanks with 2D spatial filters in quadrature pairs of the Gabor, GD, or similar type followed by nonlinear operations like energy computation or