Signal Processing: *Image Communication* 28 (2013) 642–658

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

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

RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis

Nicolas Riche*, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, Thierry Dutoit

UMONS, 31, Bld Dolez, Mons, Belgium

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
28 March 2013
Accepted 31 March 2013
Available online 6 April 2013

Keywords:
Bottom-up saliency
Comparative statistical analysis
Multi-scale rarity mechanism
Regions of interest
Saliency models evaluation
Visual attention

ABSTRACT

For the last decades, computer-based visual attention models aiming at automatically predicting human gaze on images or videos have exponentially increased. Even if several families of methods have been proposed and a lot of words like centre-surround difference, contrast, rarity, novelty, redundancy, irregularity, surprise or compressibility have been used to define those models, they are all based on the same and unique idea of information *innovation* in a given *context*.

In this paper, we propose a novel saliency prediction model, called RARE2012, which selects information worthy of attention based on multi-scale spatial rarity. RARE2012 is then evaluated using two complementary metrics, the Normalized Scanpath Saliency (NSS) and the Area Under the Receiver Operating Characteristic (AUROC) against 13 recently published saliency models. It is shown to be the best for NSS metric and second best for AUROC metric on three publicly available datasets (Toronto, Koostra and Jian Li).

Finally, based on an additional comparative statistical analysis and the effect-size Hedge' g^* measure, RARE2012 outperforms, at least slightly, the other models while considering both metrics on the three databases as a whole.

1. Introduction

There is no common definition of human attention, and it can differ depending on the domain (psychology, neuroscience or engineering) or the considered approach. But, in a general sense, human attention can be defined as the natural capacity to prioritize the incoming stimuli and selectively focus on part of them. The goal of the attentional process is to identify as quickly as possible those parts of our environment that are key to our survival. Humans but also all animals use this mechanism in their daily life and even during dreams when the rapid eye movements occur (REM stage), which are saccades and fixations on the dream scene.

The interest of attention prediction is more and more understood by the scientific community with an exponential number of papers dealing with saliency algorithms. Attention modeling has very wide applications such as machine vision, surveillance, data reduction and compression, human computer interfaces, advertising assessment or robotics. In this context, efficient attention models are of great importance for vision and signal processing algorithms improvements in the future.

In computer science, attention modeling is mainly based on the concept of "saliency maps", which provides, for each pixel, its probability to attract human attention. The idea is that the gaze of people will direct to areas which, in some way, stand out from the background. Saliency implies a competition between an objective "bottom-up" attention and a subjective "top-down" information. Bottom-up attention is a generic approach also known as stimulus-

* Corresponding author. Tel.: +32 65 37 47 25.
E-mail addresses: Nicolas.Riche@umons.ac.be (N. Riche), Matei.Mancas@umons.ac.be (M. Mancas), Matthieu.Duvinage@umons.ac.be (M. Duvinage), Bernard.Gosselin@umons.ac.be (B. Gosselin), Thierry.Dutoit@umons.ac.be (T. Dutoit).

0923-5965/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2013.03.009

driven or exogenous attention. Furthermore, it relies on the information innovation that the features extracted from the image can bring in a given spatial context. The top-down component of attention, which is also known as task-driven or endogenous attention, integrates specific knowledge that the viewer could have in specific situations (tasks, models of the kind of scene, recognized objects, etc.). The eye movements are not a direct output of the algorithms, but they can be computed from the saliency map by using winner-take-all [1] or more dynamical algorithms [2].

In this paper we present a novel attention algorithm and we focus on a fair comparison with other state of the art attention models. The algorithm proposed which we will call "RARE2012" is purely bottom-up. This is an important point for model evaluation as top-down information can drastically increase a model performance. Indeed, several models use additional post-processing which provide top-down information like centred Gaussians which leads to an artificial increase of their results. Moreover, several saliency models have a lot of parameters, which make fair comparison very difficult. Some research, like Borji and Itti [3] or Judd et al. [4], attempts to provide a benchmark between bottom-up models using several similarity measures and sometimes several datasets of images. We based our validation on Borji and Itti approach and codes [3]. A complementary statistical evaluation has also been added. The codes of the model proposed in this paper are freely available online [5].

The paper is organized as follows. Section 2 contains an overview of recent saliency models and more specifically of methods used in our comparative study. In Section 3, the architecture of our method is described in detail. The results are presented in Section 4: after a qualitative evaluation on psychophysical observations and three databases, two metrics are used to quantify the prediction of the proposed method. Section 5 details an additional two-metric based statistical analysis of the results showing the overall effectiveness of RARE2012. Finally, Section 6 provides a discussion and conclusion.

2. Related work

It is very hard to find an optimal taxonomy, which classifies all the saliency approaches. The literature is very active concerning still images saliency models. While some years ago only some labs in the world were working on the topic, nowadays a hundred different models have been published. Those models have various implementations and technical approaches despite that they all derive from the same idea of information innovation in a given context.

Some attempts of taxonomies proposed an opposition between "biologically driven" and "mathematically based" methods. Unfortunately, the biological plausibility of the methods is a difficult point to judge. Another criterion is the computational time or the algorithmic complexity, but it is very difficult to make this comparison as all the existing models do not provide cues about their complexity. Moreover, the implementations can be found in several programming languages. Finally a classification of models based on centre-surround contrast compared to information theory methods do not include different approaches as spectral residual for example. Although several taxonomies can coexist, we propose an original context-based taxonomy. In this framework, there are three classes of models with different contexts which are mostly local, global and normality.

In this section, we define the proposed saliency models categories and provide a brief overview of the recent saliency models that are used for the evaluation in this study. We focus on the models used for our evaluation and do not intend to provide an overview of all existing saliency models. For this purpose, we selected most recently published models which are also available online and classify them using the proposed taxonomy. We also focused on models which use eye-tracking as gold standard and not the models which use manual segmentation as evaluation. Some models obviously use both local and global information. In this case, the classification is made on the primary considered context.

2.1. *Local context: salient objects are contrasted compared to their surroundings*

The first approach, called local context, is about pixels surroundings: here a pixel or patch is compared with its surroundings at one or several scales like in [6]. Five models from this context are proposed for the study and described in the following subsections.

2.1.1. *AIM: Attention Based on Information Maximization (2005)*

AIM was created by Bruce and Tsotsos in 2005 [7]. The principle of this bottom-up attention model aims at maximizing information sampled from a scene. Shannon's self-information measure is computed by using patches from the image and their surrounding patches projected on a new basis obtained by performing an ICA (Independent Component Analysis) on a large sample of 7×7 RGB patches drawn from natural images. Overall, this approach quantifies how unexpected the content in a local patch is based on its surrounding.

2.1.2. *STB: Saliency ToolBox (2006)*

This toolbox [8,9] is a partial reimplementation of the Neuromorphic Vision Toolkit (iNVT) from Laurent Itti [1]. His model is composed of three steps: feature extraction, centre-surround inhibition and feature maps fusion. First, three types of static visual features are selected (colours, intensity and orientations) at several scales. The second step is the centre-surround inhibition which will provide high response in case of high contrast, and low response in case of low contrast. The third step consists in an across-scale combination, followed by normalization to form "conspicuity" maps which are single multi-scale contrast maps for each feature. Finally, a linear combination is made to achieve inter-features fusion.

2.1.3. *GBVS: Graph-Based Visual Saliency (2006)*

Harel et al. introduced the Graph-Based Visual Saliency (GBVS) model [10]. In this model, they first extracted similar feature maps to Itti's maps (see previous subsection) leading to three multi-scale feature maps (intensity,