Signal Processing: Image Communication 44 (2016) 57–68

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Visual saliency detection: From space to frequency

Dongyue Chen*, Tong Jia, Chengdong Wu

College of Information Science and Engineering, Northeastern University, China

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form
10 March 2016
Accepted 10 March 2016
Available online 12 March 2016

Keywords:
Saliency detection
Human fixation prediction
Gabor function
Fourier transform
Multi-scale saliency

ABSTRACT

In this paper, a new saliency detection model is proposed based on a space-to-frequency transformation. Firstly, the equivalence of spatial filtering and spectral modulation is demonstrated to explain the intrinsic mechanism of typical frequency-based saliency models. Then a novel frequency-based saliency model is presented based on the Fourier Transformation of multiple spatial Gabor filters. Besides, a new saliency measurement is proposed to implement the competition between saliency maps at multiple scales and the fusion of color channels. In experiments, we use a set of typical psychological patterns and four popular human fixation datasets to test and evaluate the proposed model. In addition, a new energy-based criterion is proposed to evaluate the performance of our model and is compared with five traditional saliency metrics for validation. Experimental results show that our model outperforms most of the competing models in salient object detection and human fixation prediction.

1. Introduction

Despite rapid development in computational capability, data transfer speed and storage volume in recent years, high dimensionality, complexity and redundancy of visual signals remain a big challenge for image and video processing technologies. Recent remarkable works on computational saliency detection possibly provide an effective way to address these problems. Visual saliency is the subjective perceptual quality which makes some items in the scene pop out from their surroundings and immediately catch our attention. Under the direction of visual attention, human brain needs to dedicate relatively fewer computational resources to the detection and recognition of one or a few objects but not the whole scene at one time, which significantly enhances the ability of the human brain in visual perception and cognition. Accordingly, visual saliency detection models can be used to improve image and video processing technologies by screening out worthless or redundant visual information and extracting more valuable pieces of visual information. Actually, saliency models have been widely applied in image processing, computer vision and pattern recognition, such as object detection [1], video compression [2], watermarking [3], retargeting [4] and classification [5]. In general, human visual attention can be categorized into two classes: bottom-up and top-down. In this paper, we focus on bottom-up visual saliency detection.

The majority of earlier relevant models are inspired by some biological theories, such as feature integration theory, lateral inhibition and the Winner-Take-All principle [6–8]. Most of these models, known as the space-based model, employ a similar structure where various feature maps are obtained by spatial filterings and are combined into a saliency map. In general, a space-based saliency detection model is mainly determined by three factors: (1) spatial unit, (2) local feature, (3) saliency measurement.

Spatial units of various saliency models can be classified into three categories: pixel, patch and segment. Pixel units are suitable for the higher-resolution saliency map, however, a single pixel is too simple to differentiate between complex visual patterns. So pixels are typically used for salient object segmentation [9–12]. Local patch is the most popular spatial unit for most of saliency prediction models [7,8,13–15]. In general, patches are uniformly arranged (overlapped or not) and each of them covers a small square of the input image. Patch-based feature representations refer to complex visual patterns and promise the higher completeness and the better distinguishability although they increase the computational complexity dramatically. Segment-based units are generally obtained using some segmentation algorithms, such as image abstraction [16], superpixel [1,17], boolean map [18] and soft-segmentation [19]. All the segment pieces will comprise a more compact and reasonable representation for the input image. Nevertheless, related segmentation algorithms sometimes are time-consuming and unreliable.

Local features include color, orientation, texture, regional statistics and so on, which can be extracted using some spatial filters or statistics. Some of spatial filters can be described by pre-

* Corresponding author.
E-mail address: chendongyue@ise.neu.edu.cn (D. Chen).

http://dx.doi.org/10.1016/j.image.2016.03.003
0923-5965/© 2016 Elsevier B.V. All rights reserved.

designed functions and patterns, such as Gabor function [20], DoG function [7,8,15], edge detectors [14] and wavelets [21], while some others are obtained using learning algorithms such as ICA [13], off-line sparse coding [22,23] and deep learning [24]. Since a given spatial filter generally works only for one specific feature, it is too time-consuming to employ sufficient number of filters to cover all the feature space and scales. Regional statistical functions are designed to compute some statistics of the local visual stimuli or the adaptive combinations of low-level visual features. Typical statistical features include color histogram [25], conditional entropy [26], PCA [27], regional covariance [28], online sparse coding [29,30], etc. However, lower-dimensional statistical features like color histogram, image moments and different kinds of entropy are unable to distinguish complex visual patterns. In contrast, higher-dimensional features like PCA and online sparse coding will significantly increase the computational cost.

Most of existing saliency measurements can be categorized in three classes: (1) contrast-based, (2) similarity-based, (3) probability-based. The contrast-based saliency can be divided into local contrast and global contrast. Local contrasts are usually extracted using DOG functions [7,8] or the regional integration of differences between features [31]. Global contrasts are generally evaluated based on the global energy [7], background subtraction [11], color histogram [12] and image whitening [32]. The similarity-based saliency is evaluated by the similarity between the target and its surroundings, which can be defined as the weighted feature distance [27,33], statistical likelihood [34], graph-based dissimilarity [14,35] and PDE-based relevance [17]. For the probability-based saliency, the rarity or uniqueness of the spatial units are generally defined as the monotonic functions of the probability density, which can be obtained through normalization [18], information maximization [13], Bayesian estimation [19], kernel density estimation [36] and some trained classifiers [37,38]. Although these saliency measurements vary in detail, most of them involve some complex calculations, such as convolution, matrix decomposition or iterative processing, which generally bring a large computational cost especially for the models that contain multiple features and scales.

In spite of the biological plausibility and the flexible structure, most of the space-based models are still plagued with the trade-off between the feature completeness and the computational cost. In recent years, a new category of saliency models based on frequency modulation has been developed, which bring a novel perspective and solution to visual saliency detection [39,40,11,41,42]. In the frequency-based models, saliency evaluation is implemented based on the amplitude spectrum which is a complete and reconstructive frequency-based representation of the input image and can be obtained rapidly through some transformations like Fast Fourier Transform, Discrete Cosine Transform and Quaternion Fourier Transform, etc.

As the earlier frequency models, Spectrum Residual (SR for short) [39] and Phase Quaternion Fourier Transform (PQFT for short) [40] show the great performance in saliency detection although their principles are not clarified. Then some related models are proposed to improve the performance or to explore the intrinsic mechanisms of the frequency-based models. Hou et al. use the theories of sparse signal mixing to explain the principles of the Image Signature model (IS for short) [42]. Chen et al. propose a new scheme to balance the frequency-based saliency between all the scales [41]. Nevertheless, frequency-based models have been often criticized for the dimness of their intrinsic mechanism and their deviations from the biological facts. That is because the quantitative relationship and the biological equivalence between the frequency-based models and the space-based models were not addressed clearly in existing literatures. In this paper, we follow the frequency-based approach and present a new model by giving a quantitative analysis for the equivalence between the space models and the frequency models.

Contributions of this paper can be summarized as follows: (1) we demonstrate the principles and the biological plausibility of frequency-based saliency models by analyzing the equivalence of the spatial filtering and the spectral modulation, (2) a new frequency-based saliency model is proposed based on the space-to-frequency transformation, and experimental results show that the proposed model is more precise and stable than competing models in human fixation prediction, (3) a new saliency measurement called spatial weighted energy is presented to realize the competition between multiple scales and the weighted summation of multiple color channels, (4) a new saliency metric is presented to evaluate the similarity between the saliency map and the fixation map. Residual parts of this paper are organized as follows. Section 2 demonstrates the relationship between frequency-based models and space-based models. In Section 3, the structure and details of our model are presented. Section 4 describes experimental results. Conclusions are given in Section 5 at last.

2. Space-to-frequency transform of saliency models

2.1. Unified framework of model transformation

At first, a unified framework is proposed to describe some popular frequency-based saliency models and explain their intrinsic principles. Suppose that the input image is $I(x, y)$. The transformation function from space to frequency is denoted by $\mathcal{T}(\cdot)$. Then most frequency models can be summarized as

$$\mathrm{Sal}(x, y) = z\left(\mathcal{T}^{-1}\left\{ [A(u, v)M(u, v)]P(u, v)\right\}\right) \tag{1}$$

where $\mathrm{Sal}(x, y)$ is the saliency map, $z(\cdot)$ describes post-processing steps, $A(u, v)$ and $P(u, v)$ are respectively the amplitude spectrum and the phase spectrums. $M(u, v)$ is a frequency filter comprised a group of sparse frequency components $\phi_i(u, v)$, $i = 1, 2, \dots, k$, as showed in

$$M(u, v) = \sum_{i=1}^{k} \alpha_i \phi_i(u, v) \tag{2}$$

Substituting Eq. (2) into Eq. (1), the saliency map can be written as:

$$\begin{aligned}\mathrm{Sal}(x, y) &= z\left(\sum\nolimits_{i=1}^{k} \alpha_i \mathcal{T}^{-1}\left\{ A(u, v)P(u, v)\phi_i(u, v)\right\}\right)\\ &= z\left(\sum\nolimits_{i=1}^{k} \alpha_i \left[I(x, y)*\psi_i(x, y)\right]\right)\end{aligned} \tag{3}$$

where

$$\psi_i(x, y) = \mathcal{T}^{-1}\{\phi_i(u, v)\}, \quad i = 1, 2, \dots, k \tag{4}$$

Eq. (3) indicates that spectral modulations of most frequency models can be equivalent to the combination of multiple spatial filters. In other words, the frequency-based model can be transformed into a typical space-based model in which $\psi_i(x, y)$, $i = 1, 2, \dots, k$, are spatial filters and α_i, $i = 1, 2, \dots, k$, are corresponding saliency weights. The transformed model is compliant to the feature integration theory [6] and resembles some popular space-based models such as NVT [7] and STB [8].

2.2. Model analysis

Let us take the IS model [42], for example, to illustrate the frequency-to-space transformation of the model. According to Eq.