Robotics and Autonomous Systems 75 (2016) 221–232

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Towards a unified multiresolution vision model for autonomous ground robots

John Sustersic [a,*], Brad Wyble [b], Siddharth Advani [c], Vijaykrishnan Narayanan [c]

[a] The Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16804, USA
[b] Department of Psychology, The Pennsylvania State University, University Park, PA 16804, USA
[c] Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16804, USA

HIGHLIGHTS

- Generalize the CORF operator for color images for contrast invariant edge detection.
- Unify center–surround differencing with Serre's color image descriptor.
- Cropped Gaussian Pyramid as a piece-wise linear approximation for foveated vision.
- Shown competitive performance in visual salience at reduced computational costs.
- Enabling more complex image processing in real-time, ideal for FPGA implementation.

ARTICLE INFO

Article history:
Received 14 February 2014
Received in revised form
20 July 2015
Accepted 29 September 2015
Available online 22 October 2015

Keywords:
Computer vision
Gaussian pyramid
Visual salience
Parallel algorithms

ABSTRACT

While remotely operated unmanned vehicles are increasingly a part of everyday life, truly autonomous robots capable of independent operation in dynamic environments have yet to be realized — particularly in the case of ground robots required to interact with humans and their environment. We present a unified multiresolution vision model for this application designed to provide the wide field of view required to maintain situational awareness and sufficient visual acuity to recognize elements of the environment while permitting feasible implementations in real-time vision applications. The model features a kind of color-constant processing through single-opponent color channels and contrast invariant oriented edge detection using a novel implementation of the Combination of Receptive Fields model. The model provides color and edge-based salience assessment, as well as a compressed color image representation suitable for subsequent object identification. We show that bottom-up visual saliency computed using this model is competitive with the current state-of-the-art while allowing computation in a compressed domain and mimicking the human visual system with nearly half (45%) of computational effort focused within the fovea. This method reduces storage requirement of the image pyramid to less than 5% of the full image, and computation in this domain reduces model complexity in terms of both computational costs and memory requirements accordingly. We also quantitatively evaluate the model for its application domain by using it with a camera/lens system with a 185° field of view capturing 3.5M pixel color images by using a tuned salience model to predict human fixations.

1. Introduction

Models of visual saliency, as a surrogate measure of attention to indicate visually interesting image features, have been an active field of research in both computational neuroscience as well as in computer vision; sometimes symbiotically feeding on the other's findings, and sometimes moving ahead independently [1–5]. These saliency models have used low-level features to build information maps, which are then fused together to form what is popularly called a saliency map. A compilation of how these models performed, when compared against human fixations, shows that they do remarkably well [6]. The Human Visual System (HVS) exhibits multi-resolution characteristics, where the fovea is at the highest resolution while the resolution tapers off towards the periphery. In [7], the authors looked at adapting a model [4] to develop a bi-

* Corresponding author. Tel.: +1 814 863 3015.
E-mail addresses: jps263@psu.edu (J. Sustersic), bpw10@psu.edu (B. Wyble), ska130@cse.psu.edu (S. Advani), vijay@cse.psu.edu (V. Narayanan).

http://dx.doi.org/10.1016/j.robot.2015.09.031
0921-8890/

ologically inspired multi-resolution framework for salience, which showed both quantitative and qualitative advantages. However the current approach offers a major advantage relative to [7], which is that the processing of the visual input through a series of maps can be precisely controlled. This top-down control permits parametric manipulation of the computation that can be modulated based on the task at hand. The result is a computational algorithm that is both tunable, and computationally efficient, thus making it ideal for real-time, on-demand processing in an autonomous system.

The proposed algorithm is based on simple cells in the primary visual cortex that are believed to extract local contour information from a visual scene. This information serves as the building block of early vision, and is important for such tasks as determining salience, contour processing, object recognition, and scene gist determination. The CORF (Combination of Receptive Fields) model [8] considers a computational model of a simple cell as an alternative to the 2D Gabor function (GF) model [9]. The Scale Invariant Feature Transform (SIFT) is a well-known method used for object recognition using a Difference of Gaussian (DoG) approach to produce translation invariant feature vectors [10]. These models process visual input in a similar fashion as the earliest stages of other models that describe hierarchical processing of objects, such as HMAX [11], which simulates the visual ventral pathway and HOP [12], which recognizes complex objects by combining simpler parts into more complex representations. In [13], the authors show that recognition rates improve when incorporating attention in cluttered and crowded conditions. In [14], the authors discuss various state-of-the-art models with regards to a rapid visual categorization task and postulate that attention is important in un-trained environments.

The model described here incorporates a novel formulation of the CORF algorithm that is neuromorphic, tunable, and can support both attention and object recognition. We describe how this model may be used to determine high-salience regions of visual images as a form of attention, which is important in object recognition and scene understanding in humans.

2. Theory

2.1. Single opponent color difference of Gaussian model

We begin by describing a formulation for the computation of Difference of Gaussian (DoG), defining center ($G_{c,\sigma}$) and surround ($G_{s,\sigma}$) Gaussians conventionally and constraining the ratios of inner and outer Gaussians to be 2 as found in electrophysical studies in mammalian LGN cells [8]:

$$G_{c,\sigma}(x,y) \stackrel{\text{def}}{=} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

$$G_{s,\sigma}(x,y) \stackrel{\text{def}}{=} \frac{1}{2\pi\,(2\sigma)^2} e^{-\frac{x^2+y^2}{2(2\sigma)^2}} \tag{2}$$

where σ is the standard deviation of the inner Gaussian. Let $\mathrm{P} \in \mathbb{R}_+$, $\mathrm{P} \leq \mathrm{P}_{\max}$ and let the input image $\mathcal{J} : \mathbb{N}^2 \to \mathrm{P}^3$. As in [15] we consider opponent color channels Red–Green (R–G), Red–Cyan (R–C), Yellow–Blue (Y–B), and White–Black (W–K), so we must project from RGB color space into the higher-dimension RGBCYW color space to compute these opponent color channels. Let $\Omega(r,g,b) = \langle r, g, b, 0.5g + 0.5b, 0.5g, +0.5r, 0.299r + 0.587g + 0.11b\rangle$ where $r, g, b \in \mathrm{P}$. Thus $\Omega : \mathrm{P}^3 \to \mathrm{P}^6$ and describes how yellow, cyan and white (intensity) channels are derived from an RGB pixel (note that this differs from [15]). Thus $\Omega(\mathcal{J})$ denotes the image $\mathcal{J}$ in RGBCYW color space. We then convolve $\Omega(\mathcal{J})$ with both center and surround Gaussians to yield center and surround responses $G_{c,\sigma} * \Omega(\mathcal{J})$ and $G_{s,\sigma} * \Omega(\mathcal{J})$, respectively. Thus the input image $\mathcal{J}$ may be thought of as a discrete function mapping from pixel coordinate space ($\mathbb{N}^2$) to RGB, and $G * \Omega(\mathcal{J})$ is a function mapping pixel coordinate space to RGBCYW color space. Mathematically, $G * \Omega(\mathcal{J}) : \mathbb{N}^2 \to \mathrm{P}^6$.

Let $[\cdot]$ be the index operator that selects a color channel by name (e.g. $\langle 255, 0, 0, 0, 0, 0\rangle[R] = 255$). The center-on or positive DoG_σ^+ may then be defined as:

$$DoG_\sigma^+ \stackrel{\text{def}}{=} \left\langle \begin{array}{l} |G_{c,\sigma} * \Omega(\mathcal{J})[R] - G_{s,\sigma} * \Omega(\mathcal{J})[G]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{J})[R] - G_{s,\sigma} * \Omega(\mathcal{J})[C]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{J})[Y] - G_{s,\sigma} * \Omega(\mathcal{J})[B]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{J})[W] - G_{s,\sigma} * \Omega(\mathcal{J})[W]|^+ \end{array} \right\rangle \tag{3}$$

where $|\cdot|^+$ designates half-wave rectification ($|x|^+ = x\,\forall x \geq 0,\ 0\ otherwise$). This defines center-on DoG_σ^+ as a four-dimension color space consisting of red channel center minus the green channel surround, red center minus cyan surround, etc. Thus the DoG_σ^+ function may be thought of as a 3D convolution that transforms an image from RGB color space into a single-opponent color space ⟨R–G, R–C, Y–B, W–K⟩.

Unlike previous work [8], $DoG_\sigma^+ \neq -DoG_\sigma^+$ and is defined:

$$DoG_\sigma^- \stackrel{\text{def}}{=} \left\langle \begin{array}{l} |G_{c,\sigma} * \Omega(\mathcal{J})[G] - G_{s,\sigma} * \Omega(\mathcal{J})[R]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{J})[C] - G_{s,\sigma} * \Omega(\mathcal{J})[R]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{J})[B] - G_{s,\sigma} * \Omega(\mathcal{J})[Y]|^+ \\ |G_{s,\sigma} * \Omega(\mathcal{J})[W] - G_{c,\sigma} * \Omega(\mathcal{J})[W]|^+ \end{array} \right\rangle. \tag{4}$$

Note that the definition of the white channel is conceptually the same as that used in the CORF model [8]; therefore one may think of this single-opponent color model as a generalization of CORF and we call it CCORF (Color CORF). To illustrate what is generated from this model, we present DoG responses for the test image as illustrated in Fig. 1.

2.2. Cropped Gaussian pyramid

Fig. 2 illustrates the concept of transforming an image into a 'Cropped Gaussian Pyramid' (CGP) in which pyramid levels are progressively cropped such that full resolution is preserved only in a narrow field of view fovea, and there is a progressive loss of acuity with increasing eccentricity due to increasingly sparse representations in larger levels of the pyramid. This yields a multi-resolution representation of the image in which the number of pixels in the CGP is much less than the input image. Conversely, traditional Gaussian Pyramids [16] (GPs) have more pixels than the image. For example, an imaging system producing 1920 × 1920 (3.51 Mpix) images can be represented in a 6-level CGP in only 166.78 Kpix, with 45% of those in the fovea,[1] while a 6-level GP model would require 4.91 Mpix.

Traditionally, GPs are constructed by successive (sequential) Gaussian blur and downsampling operations. In the downsampling step for imagery data, fully 3/4 of the computational work done in blurring is discarded in current practice, but in a CGP an even larger percent of computational work will be discarded in the progressive cropping. We therefore separate computations of the levels and combine this with previously described (DoG) computation to limit total computational work to only required computations while enhancing parallelizability—significantly reducing the time complexity of CGP pyramid construction at the tradeoff of progressively reduced visual acuity in peripheral regions of the image. Thus the CGP may be thought of as a piece-wise linear approximation of biological foveated vision.

Each level of the pyramid is computed independently, directly from the source image. The 3D convolution of the source image

[1] In humans, ∼50% of cortical resources processing retinal information are allocated to the fovea.