A Novel Visual Attention Framework using Unsupervised Feature Learning for Road Scene Understanding

Yanfen Mao[1, a], Qingyu Meng[1, b], Ming Chen[2, c]

[1]Automotive Engineering & Service, Sino-German College of Applied Sciences (CDHAW), Tongji University, Shanghai, 201804, China

[2]Industry 4.0 Smart Factory Laboratory, Tongji University, Shanghai, 201804, China

[a]email:maoyanfen@tongji.edu.cn, [b]email:cdhawmqy@aliyun.com, [c]email:chen.ming@tongji.edu.cn

Keywords: Visual attention; Road Scene Understanding; Deep Learning; Bayesian Framework

Abstract. Road scene understanding plays a key role in autonomous driving for intelligent vehicle. For the problem making semantic labeling with equivalent priority results in confliction between huge amounts of data and limited computation resource, this paper proposes a novel framework that efficiently fuses selective visual attention mechanism into the solution to scene perception task. Incorporating top-down and bottom-up two kinds of attention effect into an integrated Bayesian framework, total saliency map can be obtained taking use of implicit feature representation by unsupervised feature learning from natural images.

Introduction

In the case of intelligent vehicles, perception task is referred to as road scene understanding. Sensing the state of the environment surrounding the vehicle is regarded as the most difficult function for intelligent vehicles [1]. This included locating key landmarks: the road, other vehicles, pedestrians, traffic signals, road signs, and other unstructured obstacles. A complete and precise description of the state of surrounding environment is the key factor that allows the reduction of the number of false and missed alarms and provides the basis for smooth automatic driving. The perception of an outdoor environment, even if partially structured is a changing problem, not only due to the intrinsic complexity of the driving environment itself, but also due the impossibility of controlling many environmental parameters.

How to analyze and interpret traffic scene in real-time and extract useful information for autonomous driving is the difficulty of the road scene understanding. Especially in the situation that vehicles run in expressway, how to deal with a large number of rapid visual information and to extract important information is very critical. Most of the traditional methods take a comprehensive analysis of the image, and region of interest and non-interest share equal computing resources which increasing complexity of analysis processing. Selective visual attention mechanism has advantages in focusing attention on saliency and gives a clear guide description to top-down visual task, so computing resources are used to salient target analysis and to explore understanding of major significance events in the scene [2].

What is salient object catching attention in the visual field? There are many definitions from different perspectives that lead to distinct models of attention available nowadays. Those models are demonstrated successful applications in computer vision, mobile robotics, and cognitive systems [3]. One of the popular viewpoints is based on information theory in a Bayesian framework [4] [5]. It takes use of the statistics of natural images for saliency computation, and it is more comprehensive for probability estimation that is independent of the test image [6].

Framework for Scene Understanding

The Hubel-Wiesel model illustrates that visual cortex is hierarchical early in 1959 [7], and in 2006 Hinton and Salakhutdinov [8] pointed out that high-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with also a hierarchical construction

as the human visual system. The idea of deep learning or unsupervised feature learning originally developed to explain early visual processing in the brain, e.g. edge detection from a large size of natural images database [9].

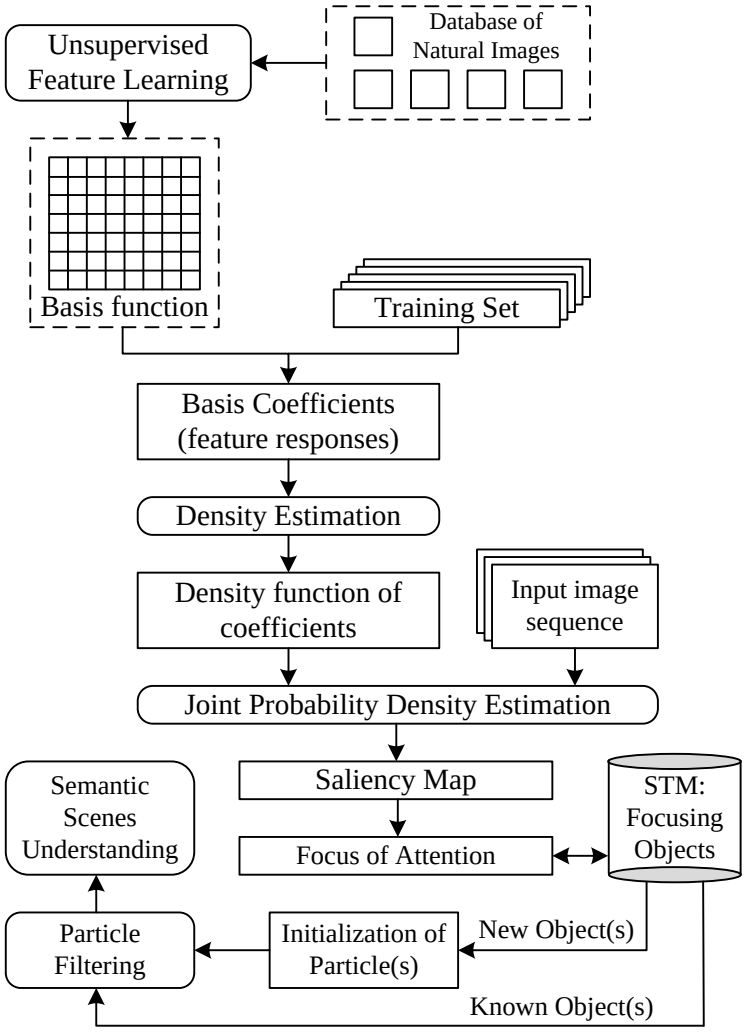

Fig.1. Framework for semantic scenes understanding. Rounded rectangles illustrate processing.

From the viewpoint of biology, an image patch is regarded as a receptive field of simple cells in the primary visual cortex, and each basis is the weight vector to extract a related feature in the receptive field. The basis functions are obtained by learning large numbers of patches, randomly sampled from thousands of natural images.

For any input image, each pixel is projected onto the bank of basis to obtain the independent coefficients taken just as the features of the image under consideration (Fig.1). And the probability density of the coefficient for the given value can be obtained from the probability density function.

After the saliency map is formed from joint density function estimation, focus of attention can be laid to salient objects. Focusing objects are stored in short-term memory (STM) while new and known objects traced using particle filtering.

Total Saliency Map Computation

Bayesian modeling is used for combining sensory evidence with prior constraints, and both top-down and bottom-up attention are computed in an integrated manner. In this model, prior knowledge and sensory information are probabilistically combined according to Bayes' rule [3].

Let $\mathbf{x}$ denote a pixel in the image, and L indicates whether or not a pixel belongs to a salient class. Let $\mathbf{f}$ be the visual features of a point. The saliency $S_{\mathbf{x}}$ of a point $\mathbf{x}$ is defined as

$$\begin{aligned} S_{\mathbf{x}} &\propto p\left(L_{\mathbf{x}}=1 \middle| \mathbf{f}=\mathbf{f}_{\mathbf{x}}\right) \\ &\overset{Bayes\ '\ Rule}{=\!=\!=} \frac{p(\mathbf{f}=\mathbf{f}_{\mathbf{x}} \middle| L_{\mathbf{x}}=1)\, p(L_{\mathbf{x}}=1)}{p(\mathbf{f}=\mathbf{f}_{\mathbf{x}})} \end{aligned} \tag{1}$$

Since the logarithm function is monotonically ascending, take the logarithm function for Equation (1):

$$\begin{aligned} \log S_{\mathbf{x}} &= \log p(\mathbf{f}=\mathbf{f}_{\mathbf{x}} \middle| L_{\mathbf{x}}=1) + \log p(L_{\mathbf{x}}=1) - \log p(\mathbf{f}=\mathbf{f}_{\mathbf{x}}) \\ &= \log p(\mathbf{f}=\mathbf{f}_{\mathbf{x}} \middle| L_{\mathbf{x}}=1) - \log p(\mathbf{f}=\mathbf{f}_{\mathbf{x}}) + const_1 \end{aligned} \tag{2}$$

$p(L_x=1)$ can be set as a constant in the case of one single target class. $p(\mathbf{f}=\mathbf{f}_{\mathbf{x}} \middle| L_{\mathbf{x}}=1)$ is a log-likelihood term that favors the feature vector consistent with the knowledge of the target's presence at point $\mathbf{x}$, and $p(\mathbf{f}=\mathbf{f}_{\mathbf{x}})$ is the self-information represents the salience of the bottom-up part when the feature vector $\mathbf{f}$ takes $\mathbf{f}_{\mathbf{x}}$.

Learning from a large-size natural images database, bank of basis is obtained utilizing unsupervised feature learning with training set [10]. The feature responses, coefficients of basis functions are taken as the implicit features representation of the image. It is confirmed that these probability densities for band-pass features in natural images can be well approximated by Generalized Gaussian Distributions (GGDs).

Considering a color image as the input of the computational model, the saliency map is computed using Equation (3) in each pixel of the image.

$$\log S_{\mathbf{x}} = \prod_{k=1}^{n} p(L_{\mathbf{x}}=1 \middle| f = f_{\mathbf{x},k}) = \sum_{k=1}^{n} \left| \frac{f_{\mathbf{x},k}}{\alpha_k} \right|^{\beta_k} + const_2, \quad (k=1,2,\cdots,n) \tag{3}$$

Here the joint probability density can be calculated in this simplified way due to the probability densities of each coefficient are mutually independent. α_k and β_k are the parameters for scale and shape of gamma function in the GGD for feature k, respectively.

Conclusion

A problem for road scene understanding is the confliction of large size of visual data and limited computing resource. To this problem, the proposed novel framework introduces deep learning into visual attention computational model in Bayesian framework for road scene understanding application which is one of active research areas over decades. Being one promising direction in machine learning for the age of big data, deep learning extract targets' features directly from nature images that differs from the traditional tedious extraction by experienced engineers. Bayesian framework combines bottom-up and top-down attention computation in an integrated mathematics manner. The performance of the new model is under verification using Matlab simulation.

Acknowledgement

In this paper, the research was sponsored by the National Natural Science Foundation of China (Project No. 61203250) and CDHAW educational project of the Chinese Ministry of Education (MoE) and the German Federal Ministry of Education and Research (BMBF).

References

[1] Broggi A, Zelinsky A, Parent M, Thorpe C. Springer Handbook of Robotics, Part F, chap51. Intelligent Vehicles, 2008, 1179-1184.

[2] Hoiem D. Seeing the World behind the Image: Spatial Layout for 3D Scene Understanding, Technical Report CMU-RI-TR-07-28, Robotics Institute, Carnegie Mellon University, USA, 2007.

[3] Ali Borji, Laurent Itti. State-of-the-Art in Visual Attention Modeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013 35(1) 185-207.

[4] Zhang Ling-yun, Matthew H. Tong, Tim K. Marks, Shang Hong-hao & Garrison W. Cottrell. SUN: A Bayesian Framework for Saliency using Natural Statistics [J]. Journal of Vision, 2008, 8(7):32, 1-20.

[5] Bruce, N.D.B., Tsotsos, J.K.. Saliency, Attention, and Visual Search: An Information Theoretic Approach [J]. Journal of Vision, 2009 9(3), 1-24.

[6] Zhang Li-ming, Lin Wei-si. Selective Visual Attention: Computational Models and Applications, Wiley-IEEE Press, 2013.

[7] D. H. Hubel and T. N. Wiesel. Receptive Fields of Single Neurones in the Cat's Striate Cortex [J]. The Journal of Physiology, 1959 148(3) 574–591.

[8] Hinton, G. E. and Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks [J]. Science, 2006,vol.313. no.5786, 504-507.

[9] Yoshua Bengio, Learning Deep Architectures for AI, Foundation and Trends in Machine Learning, 2009 2(1) 1-127.

[10]Quoc V. Le, Re Karpenko, Jiquan Ngiam, Andrew Y. Ng. ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning. NIPS 2011.