Neurocomputing 138 (2014) 61–68

Contents lists available at ScienceDirect

Neurocomputing

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

Learning to predict eye fixations for semantic contents using multi-layer sparse network

Chengyao Shen [a], Qi Zhao [b,*]

[a] NUS Graduate School for Integrative Sciences and Engineering (NGS), National University of Singapore, 117456, Singapore
[b] Department of Electrical and Computer Engineering, National University of Singapore, 117576, Singapore

ARTICLE INFO

Article history:
Received 2 August 2012
Received in revised form
29 July 2013
Accepted 16 September 2013
Available online 1 April 2014

Keywords:
Semantic saliency
Gaze prediction
Sparse coding
Deep learning

ABSTRACT

In this paper, we present a novel model for saliency prediction under a unified framework of feature integration. The model distinguishes itself by directly learning from natural images and automatically incorporating higher-level semantic information in a scalable manner for gaze prediction. Unlike most existing saliency models that rely on specific features or object detectors, our model learns multiple stages of features that mimic the hierarchical organization of the ventral stream in the visual cortex and integrate them by adapting their weights based on the ground-truth fixation data. To accomplish this, we utilize a multi-layer sparse network to learn low-, mid- and high-level features from natural images and train a linear support vector machine (SVM) for weight adaption and feature integration. Experimental results show that our model could learn high-level semantic features like faces and texts and can perform competitively among existing approaches in predicting eye fixations.

1. Introduction

Visual attention is a fundamental process of our visual system that happens in our everyday life. It enables us to allocate our limited processing resources to the most informative part of the visual scene. Visual attention has been studied in different areas such as psychology, neurosciences and computer vision. Various computational models, which are called saliency models, have been proposed according to the psychological and neurobiological findings in this area.

Most saliency models follow the "Feature Integration Theory" (FIT) [1–3] framework which suggests low-level visual feature maps such as luminance, color, orientation and motion to compute saliency map and predict human eye fixations [4,5]. These models work well to a certain extent, but are usually insufficient in predicting accurate eye fixations, especially when the scene contains strong semantic objects such as faces, texts, or other socially meaningful contents [6,7].

To overcome this so-called "semantic gap", many improved computational models [7–10] have been proposed to better predict human fixations by integrating higher-level features (e.g., a common practice is to add specific object detectors) into the original low-level feature based models. However, regarding the fact that there are thousands of object categories existing in our daily life, simply adding detectors would make the saliency models more complex and even infeasible in implementation. Hence, a unified framework that could naturally integrate features at various levels is desirable.

Recent advances on deep learning and unsupervised feature learning [11–13] provide us a useful tool for this unified feature integration. Deep learning models are usually multilayer generative networks trained to maximize the likelihood of input data. With sparse priors on the responses of each layer, hierarchies of target-relevant features or bases with increasing complexity could be learnt out in an unsupervised way from a large amount of input data through greedy layer-wise training. After feature learning, multiple levels of sparse representations can then be generated as the efficient coding of the inputs. Such properties of deep learning models are attractivethat they resemble early processing of the primate visual system [14,15].

In this paper, we build our new saliency model upon the deep learning framework in the hope to learn saliency-relevant features from natural images and predict eye fixations that is related to object and semantic contents [16]. The model is built with three layers of filtering units and pooling units stacking together followed by a linear SVM to integrate the top-level feature map into the saliency map. To mimic the images projected to the fovea during eye fixations, the model is first pre-trained on salient regions from the MIT eye tracking dataset [7] and Fixations on Faces (FIFA) [17] dataset for feature learning. Then a SVM training

[*] Corresponding author. Tel.: +65 6516 6658.
E-mail address: eleqiz@nus.edu.sg (Q. Zhao).

http://dx.doi.org/10.1016/j.neucom.2013.09.053
0925-2312/© 2014 Elsevier B.V. All rights reserved.

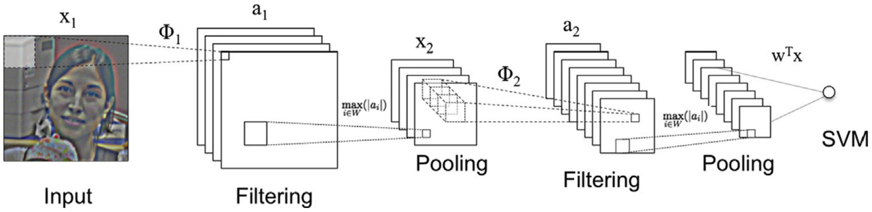


Fig. 1. Architecture of the multi-layer sparse network model, 'Filtering' layers correspond to the feature maps generated by convolutional sparse coding operations and 'Pooling' layers correspond to the feature maps generated by max-pooling operations. A linear SVM is fully connected to the output of the network to train the saliency model.

is performed on the responses of salient and non-salient regions from the datasets to learn the weight of each feature map. After training, the model is applied to test images from the same datasets and saliency maps are generated by organizing each response in a small region to a map. Experimental results show that the model is competitive among existing models to predict gaze.

The main contributions of our work are as follows:

1. We learn meaningful high-level visual features using the principled framework of deep networks by modeling the way humans sample the visual scene and we show that this way of sampling plays an important role in the learning of these features.
2. We propose a unified feature integration framework for saliency detection that could integrate low-, mid- and high-level features learned from natural images.

The rest of the paper is organized as follows. In Section 2, we review related works on saliency detection and deep network. We then present the model of multi-layer sparse network and the way of training and testing the model in Section 3. In Section 4, experiments are conducted on MIT eye tracking and FIFA datasets and both quantitative and qualitative results are given. Section 5 concludes the paper.

2. Related works

In recent years, due to the limitation of classical saliency model based on low-level features [6,7], there have been growing interests in modeling eye fixations by integrating mid-/high-level features [17,9,7,18]. Cerf et al. [17] refine Itti and Koch's model [4] by adding a face detector. Zhao and Koch [9] further improve Itti and Koch's model [4] by using a least square technique to learn the weights of face and low-level feature maps from different eye tracking datasets. In Judd et al.'s work [7], low-level features including statistics of local orientations, luminance and colors, mid-level features such as a horizon line detector, and high-level features such as a face detector and a person detector are integrated by a linear SVM to predict where humans look. Based on Judd et al.'s work, Lu et al. [18] further improve the saliency computation by including Gestalt cues such as convexity, symmetry and surroundedness into their model. All these works indicate that mid-/high-level features play an important role in predicting human fixations, but there still lacks a unified framework that could integrate various low-, mid- and high-level features that have been mentioned or not mentioned above.

Also closely related are deep learning models that could learn higher-level features from natural images. In one seminal work [11], Lee et al. show that, by training on well-aligned images from the Caltech 101 dataset [19], hierarchies of representations which correspond to object parts and objects could be learned with a convolutional Restricted Boltzmann Machine (RBM). In [12], Zeiler et al. propose a hierarchical sparse network in which each layer reconstructs the input and shows that edges, junctions, and even object parts can be learned out from the images that contain objects. In one recent work [13], Le et al. build a three-layer deep auto-encoder and prove that neurons representing faces, human bodies, and cats can be learned out in a fully unsupervised way on images sampled from 10 million YouTube videos. These models all validate that, by training on natural images, meaningful high-level features can be learned out using a deep network. However, none of them has considered the influence of visual attention on the feature learning in deeper levels. Furthermore, compared with existing works, our model is able to learn out meaningful high-level neurons in relatively few samples with the aid of eye fixations.

3. The multi-layer sparse network framework

In this section, we describe the hierarchical model that is used to learn features from natural images and predict visual saliency. The general structure of the model is shown in Fig. 1, which is composed of multiple layers of filtering and pooling sublayers stacking together (here we only show two layers for the brevity of illustration) and a linear SVM at the end to generalize the responses of the network to visual saliency.

This hierarchical model can be seen as a natural extension of previous hierarchical models such as Neocognitron [20], HMAX [21,22] and Convolutional Neural Network [23] that aim to model the hierarchical structure of ventral stream[1] in the visual cortex. This structure is also a common structure employed by many recent deep learning models [11–13].

3.1. Sparse coding and unsupervised feature learning

Sparse coding is an unsupervised scheme that learns to represent input data using a small set of bases (or features). It is the core computational algorithm in our model and constitutes the basic unit for the filtering layer.

The idea of sparse coding originates from Barlow's principle of redundancy reduction [25], which states that a useful goal of sensory coding is to transform the input in such a manner that reduces the redundancy of the input stream. In its original form of modeling image patches [26], it can be described as a generative image model as

$$E = \|\mathbf{x} - \Phi \mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1 \tag{1}$$

[1] The division of "Ventral Stream" and "Dorsal Stream" is a widely accepted concept to the function of primate visual cortex. The ventral stream (also called "What Pathway") is related to object recognition and form representation and is found to have a hierarchical structure with larger receptive field size, and more complexity along the stream from V1 to AIT [24]. The"Dorsal Stream" (also called "Where Pathway") deals with the guidance of actions and localization of object.

where $\mathbf{x}$ is the input data, Φ denotes the bases or features learnt from the data, $\mathbf{a}$ is the sparse codes for the data, and λ is the penalty constant for sparsity. Here $\|\mathbf{a}\|_p = (\sum_m |a_m|^p)^{1/p}$ is called L_p norm.

In (1), if we see $\mathbf{x}$ as an image, the first item $\|\mathbf{x} - \Phi\mathbf{a}\|_2^2$ can be seen as the difference between the original image and the reconstructed image and the second item $\lambda\|\mathbf{a}\|_1$ can be seen as the sparse penalty which regularizes the sparseness of the output codes. The features Φ and the sparse codes $\mathbf{a}$ can be found by iteratively minimizing the energy function:

$$\Phi = \arg\min_{\Phi} \left\langle \min_{\mathbf{a}} E \right\rangle \tag{2}$$

In our model, we update $\mathbf{a}$ with coordinate descent [27] by fixing Φ and updating Φ with the Lagrange dual method [28] by fixing $\mathbf{a}$. In our experiment, the stop learning condition for the unsupervised feature learning is the gradient of E is less than $\epsilon = 10^{-9}$.

3.2. Spatial pooling

Spatial pooling is an operation that integrates the responses of nearby feature detectors into one. It is often used in image recognition models to obtain a more compact representation that preserves the important information in the input signal while discarding noises and irrelevant details.

In our model, we implement the max-pooling in the pooling layer. We use the max-pooling here mainly because of its good performance for sparse codes and simplicity in implementation [29].

Given a disjoint local neighborhood W of size $l \times l$ in the sparse response maps, the max-pooling responses $\mathbf{z}$ can be obtained by

$$\mathbf{z} = \max_{i \in W}(|a_i|) \tag{3}$$

here a_i indicates the local neighborhood of sparse responses in $\mathbf{a}$.

After this operation, the sparse responses of the layer would shrink in a scale of l (as indicated in Fig. 1) and become more tolerant to minor translation and scaling.

3.3. Multi-layer architecture

To model the hierarchical structure of the ventral stream, we stack multiple layers of filtering units and pooling units together to construct a multi-layer sparse network.

3.3.1. Preprocessing

The input data of the network is sampled from natural images. Before reaching the first layer, the raw data $\mathbf{x}$ is whitened with local contrast normalization to have zero mean and unit variance:

$$\mathbf{x}_1 = \frac{\mathbf{x} - \overline{\mathbf{x}}}{\mathrm{var}(\mathbf{x})} \tag{4}$$

This operation approximates the visual processing in retina and LGN [30,31] and is important for fast convergence in unsupervised feature learning.

3.3.2. Dimensionality reduction

The outputs of each layer are usually high-dimensional and redundant. To speed up the unsupervised feature learning algorithms in the next stage, we apply Principle Component Analysis (PCA) to perform dimensionality reduction

$$\mathbf{x}_{i+1} = \mathbf{P}_k^T \mathbf{z}_i \tag{5}$$

where $\mathbf{P}$ is the PCA projection matrix where k principle components are preserved to retain the 95% of variance in the data, $\mathbf{z}_i$ is the pooled responses of the current layer, and $\mathbf{x}_{i+1}$ the inputs to the next layer.

3.3.3. Feature learning

In the feature learning stage, the network is pre-trained with sparse coding and greedy layer-wise training. The features are all learned out from the image regions where most human fixations are on.

To accomplish this, we first collect salient regions from each image in the datasets. Particularly we convolve a Gaussian mask with an accumulated fixation map from all the subjects on that image and crop square bounding boxes of size 100 × 100 centered at positions of local maxima. In this way, regions with dense fixations are extracted and we are able get a large number of salient image regions for hierarchical feature learning.

With the input data, a greedy layer-wise training is then implemented and the entire network is trained layer by layer with sparse coding to learn features from salient regions. In each layer, a large number of patches in the size of the features are extracted randomly from the inputs of the current layer and features are learned by alternatively updating Φ and $\mathbf{a}$ according to the rule derived from sparse coding. The sparse codes of the current layer are pooled with max-pooling and then used as the inputs to the next layer. In this way, hierarchy of features with increasing complexity is learned out from a low-level to a high-level.

3.3.4. SVM training

After feature learning, we then integrate the hierarchy of features learned from the greedy layer-wise training to predict visual saliency. Here we take an approach similar to Judd et al. [7], using a linear SVM to learn optimal weights for feature integration. To train the linear SVM, we collect salient and non-salient regions from images and use their responses from the highest layer as positive and negative samples. Particularly, salient regions are collected as described in the last section and non-salient regions are randomly sampled from non-fixated area of the training set. A linear SVM is then trained as a two-class classification problem based on these positive and negative responses and weights are learned to denote the contributions of each high-level features to saliency.

3.3.5. Inference and saliency computation

In the inference stage, full images are used as the input of the network and a hierarchy of sparse codes are computed in a convolutional way by the features learned in the previous stage. The saliency map is then constructed by the output value of the linear SVM on each local region:

$$s = g \circ \max(\mathbf{w}^T \mathbf{x}, 0) \tag{6}$$

here $\mathbf{w}$ denotes the learned weights of the linear SVM, $\mathbf{x}$ represents the vectorized highest level feature responses for the local region, and g is a Gaussian mask used to blur the saliency map. To compensate the boundary loss after stages of convolution, a zero-value boundary is added according to the effective receptive field size of the highest-level neuron in the input space.

Since there is a strong bias for human fixations to be near the center of the image [7,9], we also model this center bias in our final saliency map explicitly by multiplying a Gaussian mask centered in the middle of the image on the final saliency map. The standard deviation of this Gaussian mask is decided by the average fixation map from the entire dataset.

4. Experiments

This section reports experimental results to validate our model. We first discuss the learned higher-level features with visualization results, and then train a saliency model using the learned features and compare it quantitatively with existing models.

4.1. Datasets

We evaluate our model on the MIT eye tracking dataset [7] and the FIFA dataset [17] which contain fixations on strong semantic contents such as faces and texts. The MIT eye fixation dataset [7] includes 1003 landscape and portrait images mostly in 36° × 27° and the images in the dataset contain a variety of objects like cars, people, faces, animals, etc. These images are randomly collected from Flickr creative commons and LabelMe dataset and the fixation data were collected from 15 subjects with 3-s-long “free-viewing”.

The FIFA dataset [17] contains 181 colored natural images (28° × 21°) with fixation data. The fixation data were collected from 8 subjects with 2 s long“free-viewing” and most of the images in FIFA dataset contain faces in various sizes with different postures.

For feature learning, we collect 2178 salient regions from MIT eye fixation dataset and 424 salient regions from FIFA dataset and use them to pre-train the three-layer sparse coding network with greedy layer-wise training and collect 2797 and 658 non-salient regions respectively for the purpose of SVM training.

4.2. Parameters

The parameters of the network are listed in Tables 1 and 2. The size parameters of the network are fixed for all the datasets and the sparsity parameters λ are set according to the quality of the features learned from the images.[2]

4.3. Results and performance

We evaluate our model using the ROC curve. The ROC curve is obtained by varying the threshold of saliency map and calculating the true positive rate with respect to fixations across all subjects. The thresholds are set at $n=5$, 10, 15, 20, 25 and 30 percent of the area of the saliency map which is usually distinctive across different saliency models and the first fixation for each image is eliminated as it is always the center of the image.

In our experiments, the ROC curve of inter-subject variability is provided as the baseline for comparison. This curve is computed by iterating all the subjects and averaging the ROC curve on whether the fixations of this subject can be predicted by the saliency map generated by the other $n-1$ subjects.

MIT eye tracking dataset: For the MIT eye fixation dataset, we divide it into 501 training images and 502 testing images and train a linear SVM based on the third layer responses.

We then compare our algorithm with classical saliency algorithms based on low-level features [4,5] and the benchmark algorithms on MIT eye fixation dataset [7] which combines classical low-level features, mid-level features (a horizon detector) and high-level features (face and people detectors). From Fig. 2, we can see that our model outperforms the models based on low-level features, and work comparably well to the benchmark algorithm.

FIFA dataset: For the FIFA dataset, we divide it into 90 training images and 91 testing images and train a linear SVM based on the third layer responses.

For comparison, we also compute the saliency maps using classical saliency algorithms based on low-level features [4,5], as well as the one with an additional face channel and learned weight for each channel [9,10].[3] The ROC curves for all the algorithms are shown in Fig. 3. From Fig. 3, we can see that, our model outperforms all the previous models and is much closer to the human performance.

Table 1
Parameters of the network.

Property name	Layer 1	Layer 2	Layer 3
Feature size	6 × 6 × 3	6 × 6 × 25	6 × 6 × 100
Pooling size	3 × 3	3 × 3	1 × 1
Number of features	25	100	225

Table 2
Sparsity penalty λ for the two datasets.

Sparsity penalty	Layer 1	Layer 2	Layer 3
λ_{MIT} [7]	0.15	0.10	0.05
λ_{FIFA} [17]	0.15	0.12	0.07

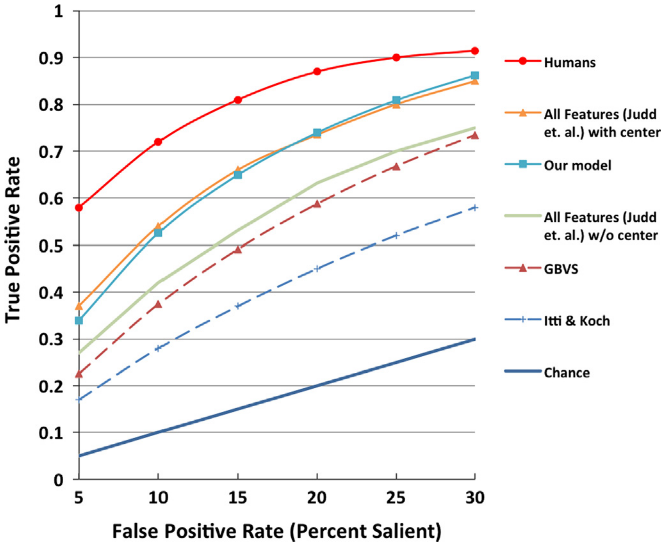


Fig. 2. ROC curve of different saliency models on the MIT eye fixation dataset [7].

To better illustrate and analyze the results, we also visualize the saliency maps generated by our model on the two datasets in Figs. 4 and 5 respectively as qualitative results and illustrate the influence of the standard deviation σ of the Gaussian center bias mask on the Area Under ROC Curve in Fig. 6.

4.4. Feature visualization

We then visualize the features learned in our multi-layer sparse network to verify the pattern they represent. For the first layer features, since they are connected to the whitened input space, we visualize their weights in direct to inspect their properties. For the higher layer features, since we cannot tell what their weights actually represent, we choose to validate them by visualizing their most responsive stimuli in the effective receptive field. The effective receptive field is computed by remapping one unit in the deeper layer to the input pixel space.

[2] Whether the second layer features would represent junctions, parallel line or other mid-level features and whether the third layer would represent object-parts or object-like features.

[3] We use 0.027 for color, 0.024 for intensity, 0.222 for orientation and 0.727 for face channel according to Table 1 in [10].

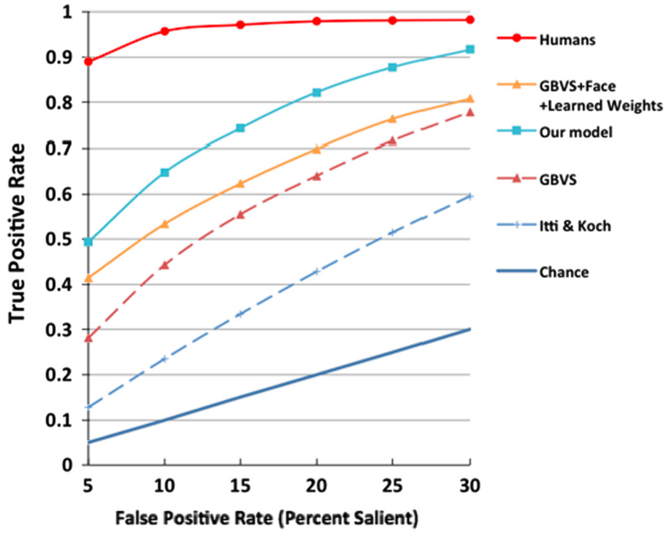


Fig. 3. ROC curve of different saliency models on the FIFA dataset [17].

To ensure that optimal stimuli in the input space were better found, we traverse the whole response space of second layer and third layer for the images in MIT eye tracking and FIFA datasets.

MIT eye tracking dataset: The visualization of features learned from MIT eye fixation dataset is shown in Figs. 7 and 8. Through visualization, we found that, by training on salient regions from MIT eye fixation dataset, neurons in the second-layer encode mid-level features like T-junctions, corners, textures, and parallelism (as shown in Fig. 7) and neurons in the third-layer are able to learn high-level concepts like faces, texts, man-made structures, and circle shapes (as shown in Fig. 8).

The visualization of features learned from FIFA dataset is shown in Fig. 9. Through visualization, we found that, by training on salient regions from FIFA dataset, neurons in the second-layer would encode not only junctions, contours, textures, parallelism but also face parts (as shown in Fig. 9). We also found that neurons in the third-layer tend to learn faces with different sizes and postures (as illustrated in Fig. 9), which matches the property of the FIFA dataset well.

To further verify the role of salient regions in the results of feature learning, we train the network by sampling random

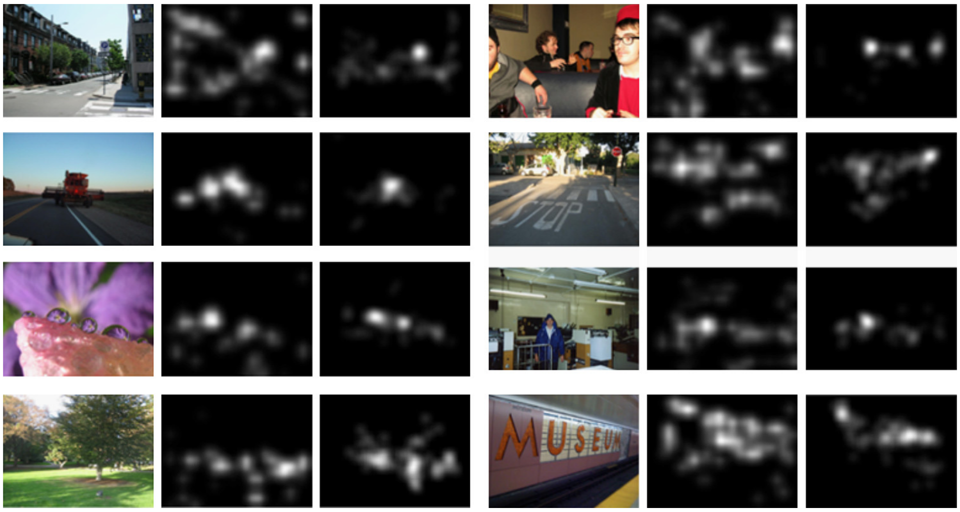


Fig. 4. Qualitative results from MIT eye tracking dataset, left: original image, middle: saliency prediction results of our algorithm, right: ground-truth map by convolving all the fixations with a Gaussian mask with σ of about one visual degree.

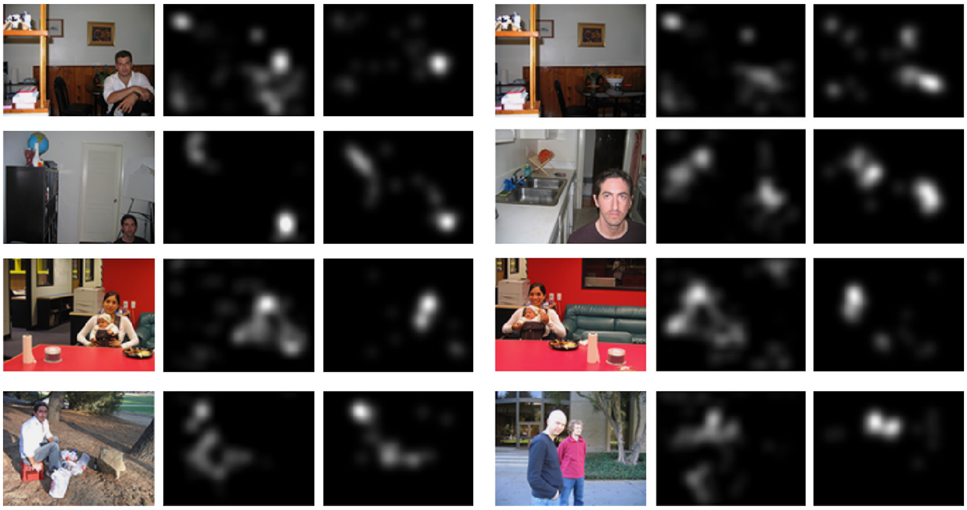


Fig. 5. Qualitative results from FIFA dataset, left: original image, middle: saliency prediction results of our algorithm, right: ground-truth map by convolving all the fixations with a Gaussian mask with σ of about one degree.

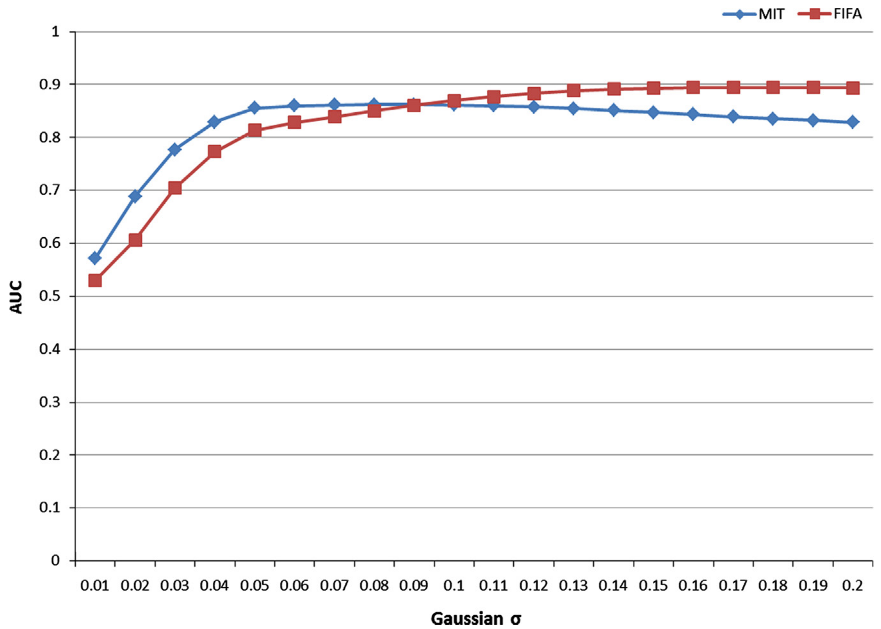


Fig. 6. The change of AUC score under different standard deviation σ of Gaussian mask (in image width).

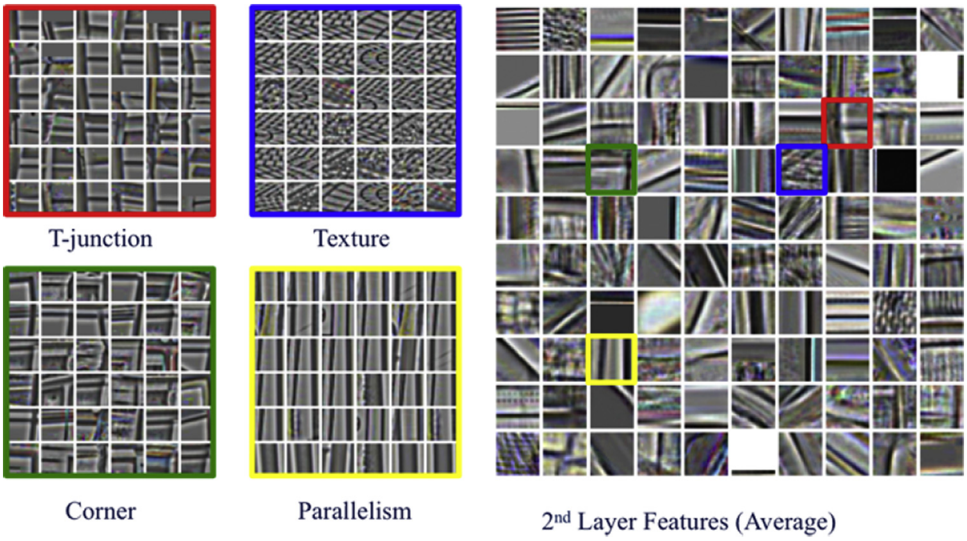


Fig. 7. Left: Illustration of second layer neurons that encode mid-level features like T-junctions, corners, textures, and parallelism. Right: Average of top 36 stimuli for all the second-layer neurons (MIT eye fixation dataset).

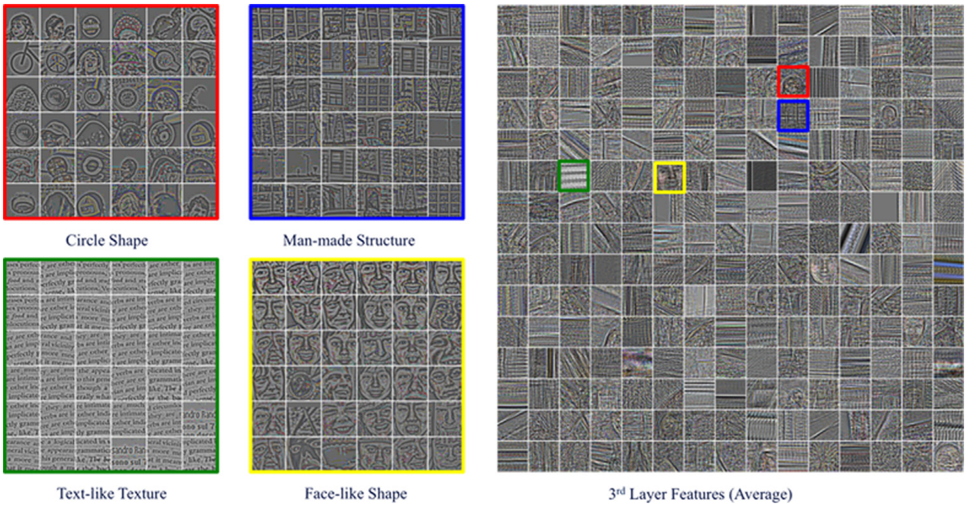


Fig. 8. Left: Illustration of second layer neurons that encode high-level concepts of circle shape, text, man-made structure and face. Right: Average of top 36 stimuli for all the third-layer neurons (MIT eye fixation dataset).

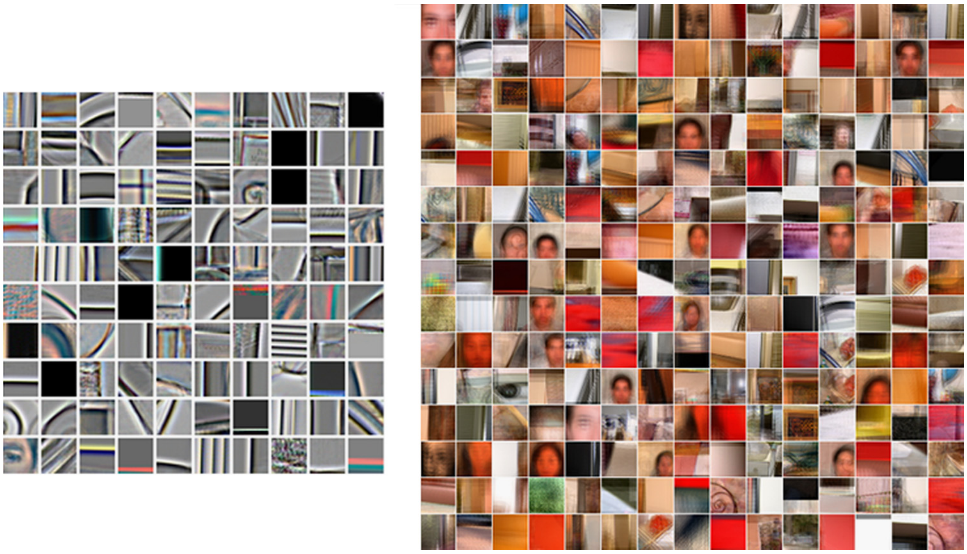


Fig. 9. Average of top 36 stimuli for all the second-layer neurons (left) and third layer neurons (right) (FIFA dataset).

patches and visualize the second-level and third-level features learned. We found that without salient region sampling, the second-level neurons tend to learn features like long edges and the third-level neurons fail to learn out meaningful features after optimization.

5. Conclusion

This paper presents a new saliency model based on the deep learning framework and demonstrates its capability in semantic saliency computation. As far as we know, this model is the first saliency model that attempts to utilize hierarchies of features learned directly from natural images and naturally integrate these features to tackle the problem of object/social saliency. Without extensive high-level features or detectors designed for specific object detection, this model can still perform competitively on two datasets with a lot of semantic content. The good performance of the model and the visualization of higher level features also indicate that, through unsupervised learning, it is possible to learn semantic-related features with a hierarchical architecture and link them with saliency by a simple linear classifier.

For the future work, we plan to improve the model by modeling the complex image transforms in the pooling layer and train the model with much more data to provide a more natural way to explain saliency in different levels.

Acknowledgments

The work is supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (No.R-263-000-648-133).

References

[1] A. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[2] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (4) (1985) 219–227.
[3] Q. Zhao, C. Koch, Learning saliency-based visual attention: a review, Signal Process. 93 (6) (2013) 1401–1407.
[4] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vis. Res. 40 (10–12) (2000) 1489–1506.
[5] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, Adv. Neural Inf. Process. Syst. 19 (2007) 545.
[6] W. Einhäuser, M. Spain, P. Perona, Objects predict fixations better than early saliency, J. Vis. 8(14) (2008).
[7] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: 2009 IEEE 12th International Conference on Computer Vision, Koyto, Japan, 2009, pp. 2106–2113.
[8] M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, Adv. Neural Inf. Process. Syst. 20 (2008).
[9] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, J. Vis. 11(3) (2011).
[10] Q. Zhao, C. Koch, Learning visual saliency, in: 2011 45th Annual Conference on Information Sciences and Systems (CISS), IEEE, Baltimore, USA, 2011, pp. 1–6.
[11] H. Lee, R. Grosse, R. Ranganath, A. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, Montreal, Canada, 2009, pp. 609–616.
[12] M. Zeiler, G. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, Barcelona, Spain, 2011, pp. 2018–2025.
[13] Q. Le, R. Monga, M. Devin, G. Corrado, K. Chen, M. Ranzato, J. Dean, A. Ng, Building High-Level Features using Large Scale Unsupervised Learning, Arxiv Preprint arxiv:1112.6209.
[14] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, T. Poggio, A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex, Technical Report, DTIC Document, 2005.
[15] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.
[16] C. Shen, M. Song, Q. Zhao, Learning high-level concepts by training a deep network on eye fixations, in: Deep Learning and Unsupervised Feature Learning Workshop, In Conduction with NIPS, Lake Tahoe, USA, December 2012.
[17] M. Cerf, E. Frady, C. Koch, Faces and text attract gaze independent of the task: experimental data and computer model, J. Vis. 9(12) (2009).
[18] Y. Lu, W. Zhang, C. Jin, X. Xue, Learning attention map from images, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Rhode Island, USA, 2012, pp. 1067–1074.
[19] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Comput. Vis. Image Underst. 106 (1) (2007) 59–70.
[20] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (4) (1980) 193–202.
[21] M. Riesenhuber, T. Poggio, et al., Hierarchical models of object recognition in cortex, Nat. Neurosci. 2 (1999) 1019–1025.
[22] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 411–426.
[23] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[24] R. VanRullen, Visual saliency and spike timing in the ventral visual pathway, J. Physiol.—Paris 97 (2–3) (2003) 365–377.
[25] H. Barlow, Possible principles underlying the transformation of sensory messages, Sens. Commun. (1961) 217–234.

[26] B. Olshausen, D. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1? Vis. Res. 37 (23) (1997) 3311–3325.

[27] J. Yang, M. Yang, Learning hierarchical image representation with sparsity, saliency and locality, in: Jesse Hoey, Stephen McKenna, Emanuele Trucco (Eds.), Proceedings of the British Machine Vision Conference, Dundee, Scotland, 2011, pp. 19.1–19.11.

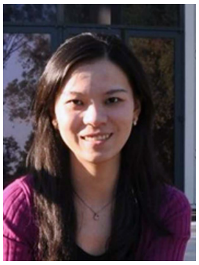
[28] H. Lee, A. Battle, R. Raina, A. Ng, Efficient sparse coding algorithms, Adv. Neural Inf. Process. Syst. 19 (2007) 801.

[29] Y. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in: International Conference on Machine Learning, Haifa, Israel, 2010, pp. 111–118.

[30] Y. Dan, J. Atick, R. Reid, Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory, J. Neurosci. 16 (10) (1996) 3351–3362.

[31] A. Hyvärinen, J. Hurri, P. Hoyer, Natural Image Statistics: A Probabilistic Approach to Early Computational Vision, vol. 39, Springer, 2009.

Chengyao Shen received the BS degree in microelectronics from Shanghai Jiaotong University, China, in 2010. He is currently a Ph.D. candidate in the Vision and Machine Learning Lab, National University of Singapore. His research interests included computer vision, machine learning and natural image statistics. He has been a recipient of NGS Scholarship since 2010.



Qi Zhao received the BS degree in computer science from Zhejiang University, China, in 2004 and Ph.D. Degree from the Computer Engineering Department at the University of California, Santa Cruz, in 2009. From 2009 to 2011, she works as a Postdoctoral Researcher at Computation & Neural Systems and Division of Biology, California Institute of Technology. Since June 2011, she has been with the Department of Electrical and Computer Engineering at National University of Singapore, where she is now an assistant professor. Her research interests include computer vision, pattern recognition, machine learning, computational neuroscience, multimedia systems, and bio-signal processing and biological image analysis. She is a member of the IEEE.