ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1125

Local Image Descriptor Inspired by Visual Cortex

Hui Wei and **Zheng Dong**[1]

Abstract. The ability of visual cortex to accomplish object recognition tasks accurately and effortlessly makes it an attractive goal of computer vision to emulate the mechanism of the cortex. The neural process of object recognition in the brain follows a hierarchical scheme. In this paper, we present a novel model inspired by the visual pathway in primate brains. This multi-layer neural network model imitates the hierarchical convergent processing mechanism of the visual pathway. We show experimentally that local image features generated by this model exhibit robust discrimination and even better generalization ability compared with some existing image descriptors. We also demonstrate the application of this model to object recognition tasks. The result provides strong support for the potential of this model.

1 INTRODUCTION

In this paper, we propose a multi-layer neural network model which imitates the neural mechanism of the ventral visual pathway [6], which is involved in object recognition. The ventral pathway travels through V1 and some other visual areas (V2 and V4) and finally reaches the inferior temporal (IT) cortex. In our model, different stages of the pathway are explicitly mapped to different layers of a feed-forward neural network. With this network, high-level local image features are extracted from images. These features exhibit competitive performance when applied to computer vision tasks.

2 MULTI-LAYER NEURAL NETWORK

Four stages in the ventral pathway are mapped to four layers in our neural network. They are composed of simple cells, complex cells, V4 neurons and IT neurons respectively (Figure 1).

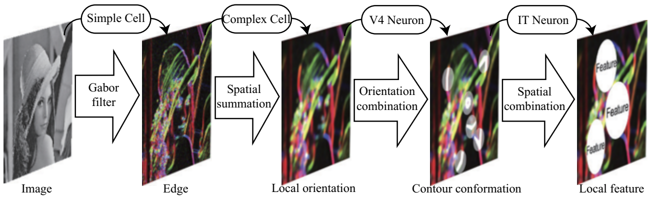


Figure 1: Multilayer neural network inspired by ventral pathway.

Simple cells [3] respond primarily to oriented edges and gratings. They can be understood as linear filters modeled as Gabor functions [2] like equation 1, where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$. In the equation, λ is the wavelength of the sinusoidal factor, θ represents the preferred orientation and σ_s approximates the radius of the receptive fields.

$$g_\theta(x, y; \lambda, \sigma_s) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma_s^2}\right)\cos(2\pi x'/\lambda), \qquad (1)$$

Complex cells differ from simple cells in that a stimulus is effective wherever it is placed in the receptive field, provided that the orientation is appropriate [3]. In our model, complex cells are modeled as weighted summation of the absolute value of simple cell output. The weight function is Gaussian function, $f(x, y; \sigma_c) = \frac{1}{2\pi\sigma_c^2}\exp\left(-\frac{x^2+y^2}{2\sigma_c^2}\right)$, where $\sigma_c = 2\sigma_s$ represents the radius of the complex receptive fields. Given $I(x, y)$ as some input image, the output of complex cells with the preferred orientation θ is calculated as the following convolution,

$$C_\theta = |I \otimes g_\theta| \otimes f. \qquad (2)$$

V4 neurons are selectively tuned for curvature and angular position of convex boundaries [8]. In the experiment demonstrated in Figure 2a and 2b, a single-layer perceptron which takes complex cells as input is trained to selectively respond to a sharp convex angle towards the top-right [8]. This experiment shows that complex cells provide sufficient information for the emergence of V4 selectivity. Therefore V4 neurons in our model combine the output of complex cells that share the same receptive field. The V4 output at a given position (x, y) is a vector defined as follows.

$$\vec{\boldsymbol{V}}(x, y) = \sum_k C_{\theta_k}(x, y) \cdot \vec{\boldsymbol{e}}_k, \qquad (3)$$

where $\{\vec{\boldsymbol{e}}_k : k = 1, 2, 3, \cdots\}$ is a group of standard basis; C_{θ_k} is the output of complex cells with preferred orientation θ_k defined by equation (2).

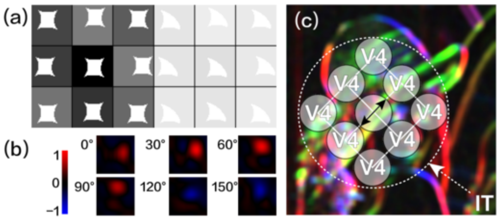


Figure 2: V4 and IT. (a) V4 perceptron was trained to distinguish a sharp convex angle towards the top-right. The gray scale of the background indicates the strength of response. (b) The weight matrix of the V4 perceptron. Each block corresponds to a specific orientation. (c) V4 neurons in 9 grid points contribute to an IT neuron. The grid is rotated to the major orientation of the central V4 unit.

The IT neuron collects the responses from a group of afferent V4 neurons and synthesizes a unified representation of shapes. In

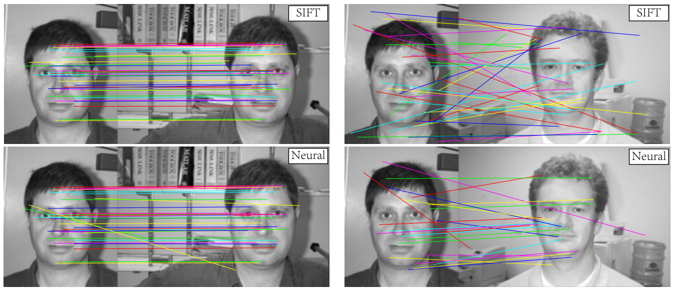
[1] Department of Computer Science, Laboratory of Cognitive Model and Algorithm, Shanghai Key Laboratory of Data Science, Fudan University, China, email: weihui@fudan.edu.cn

our model, the IT output is a vector produced by concatenating the output vectors of V4 neurons. We choose afferent V4 neurons from 3×3 rectangular grid points (Figure 2c). The gird is rotated to the strongest orientation of the central V4 input in order to achieve the invariance of feature orientations.

3 EMPIRICAL EVALUATION

We demonstrate the performance of the features generated by our model in feature matching and object recognition tasks.

We provide a qualitative evaluation of our features by matching features between different images. The SIFT features are used for comparison. The VLFeat library [10] were utilized for the SIFT implementation and the feature matching algorithm.



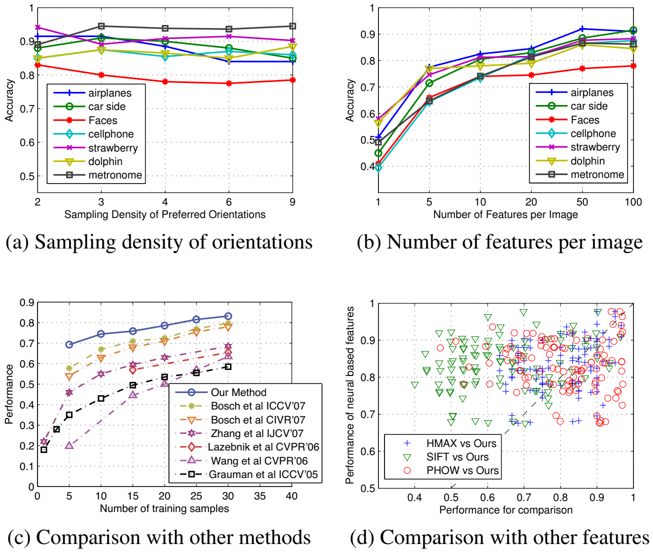
(a) Matching between the same face (b) Matching between different faces

Figure 3: Matching face images.

As shown in Figure 3a, SIFT features match perfectly between face images of the same person. The neural features generated by our model achieve competitive distinctiveness. However, Figure 3b shows that matching of SIFT features between face images of different people is quite disordered. SIFT features excel in matching highly distinctive features under image transformations but lack the generalization ability to capture variations in objects appearance of the same category. By contrast, the neural features show robust performance in matching features from different faces. The result suggests that our model exhibits better generalization ability.

We also evaluate the performance of the proposed model in object recognition tasks with the Caltech 101 dataset [1]. We use the bag of words approach for object recognition. Key-points are detected with the salient region detector proposed by Kadir and Brady [4]. Features at the key-points are extracted with our model and passed to an SVM classifier. We investigate the selection of sampling density of preferred orientations, i.e., how many orientations a V4 output vector contains. The result averaged over 8 independent runs is shown in Figure 4a. As is shown, the selection of sampling density does not exhibit obvious impact on the performance. In the succeeding experiments, a sampling density of 3 is used. To investigate the contribution of the number of features on performance, we vary the number of features chosen from each training image. The result shown in Figure 4b is also averaged over 8 independent runs. With 50 features per image the result approaches the best performance. In the succeeding experiments, 50 features are extracted from each image.

The performance of our model is compared with published result in Figure 4c. The performance is averaged over all categories and the comparison was taken over different number of training images. Figure 4d shows the performance of our model over the 101 categories compared with SIFT features [7], PHOW features [5], and biologically inspired HMAX model [9]. The results are obtained with 30 training images randomly selected from each category. Our model exhibits competitive performance compared with other models.



(a) Sampling density of orientations (b) Number of features per image

(c) Comparison with other methods (d) Comparison with other features

Figure 4: Performance with different parameters.

ACKNOWLEDGEMENTS

This work was supported by the 973 Program (Project No. 2010CB327900), the NSFC project (Project No. 61375122, 81373556), and the National "Twelfth Five-Year Plan" for Science and Technology (Project No. 2012BAI37B06).

REFERENCES

[1] Li Fei-Fei, Rob Fergus, and Pietro Perona, 'Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories', *Computer Vision and Image Understanding*, **106**(1), 59–70, (2007).

[2] Dennis Gabor, 'Theory of communication. part 1: The analysis of information', *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, **93**(26), 429–441, (1946).

[3] David H Hubel and Torsten N Wiesel, 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *The Journal of physiology*, **160**(1), 106, (1962).

[4] Timor Kadir and Michael Brady, 'Saliency, scale and image description', *International Journal of Computer Vision*, **45**(2), 83–105, (2001).

[5] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, 'Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories', in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pp. 2169–2178. IEEE, (2006).

[6] Sidney R Lehky and Anne B Sereno, 'Comparison of shape encoding in primate dorsal and ventral visual pathways', *Journal of neurophysiology*, **97**(1), 307–319, (2007).

[7] David G Lowe, 'Object recognition from local scale-invariant features', in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pp. 1150–1157. Ieee, (1999).

[8] Anitha Pasupathy and Charles E Connor, 'Shape representation in area v4: position-specific tuning for boundary conformation', *Journal of neurophysiology*, **86**(5), 2505–2519, (2001).

[9] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, 'Robust object recognition with cortex-like mechanisms', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29**(3), 411–426, (2007).

[10] Andrea Vedaldi and Brian Fulkerson, 'Vlfeat: An open and portable library of computer vision algorithms', in *Proceedings of the international conference on Multimedia*, pp. 1469–1472. ACM, (2010).