Alameer A, Degenaar P, Nazarpour K.

Biologically-inspired object recognition system for recognizing natural scene categories.

In: International Conference for Students on Applied Engineering (ISCAE).

2016, Newcastle upon Tyne: IEEE

DOI link to article:

http://dx.doi.org/10.1109/ICSAE.2016.7810174

Date deposited:

12/01/2017

Natural Scene Categorization with the EN-HMAX Model

Ali Alameer, *Student Member, IEEE*, Patrick Degenaar, and Kianoush Nazarpour, *Senior Member, IEEE*

***Abstract*—Visual processing has attracted a lot of attention in the last decade. Hierarchical approaches for object recognition are gradually becoming widely-accepted. Generally, they are inspired by the ventral stream of human visual cortex, which is in charge of rapid categorization. Similar to objects, natural scenes share common features and can, therefore, be classified in the same manner. However, natural scenes generally show a high level of statistical correlation between classes. This, in fact, is a major challenge for most object recognition models. Rapid categorization of a natural scene in the absence of attention is a challenge. However, researchers have found that 150 ms is enough to categorize a complex natural scene. We tested the capability of our recent and bio-inspired En-HMAX model of visual processing for scene classification. The results show the En-HMAX model has a comparable performance to state of the art methods for natural scene categorization.**

***Index Terms*—Elastic-net regularization, hierarchical MAX, dictionary learning, object recognition, sparsity**

I. Introduction

THE last decade has witnessed great advances in the fields of machine learning and visual recognition. Computer vision techniques have achieved the state of the art performance in observing the environment and recognizing objects. However, there is still a large margin between these methods and biological systems. Consequently, biologically-inspired models have attracted the attention of many researchers in the recent years. Hierarchical models of visual cortex [1] have proved promising in many computer vision tasks. These models mimic the ventral stream of the visual cortex. The ventral stream is involved with extracting information about objects' shape and texture which are very informative for rapid categorization.

Many examples of hierarchical structures can be found in the literature such as the convolutional neural network (CNN) [2] and the Hierarchical MAX (HMAX) model [1]. The hierarchy of these models helps to imitate the shape-texture-extracting pathway of the brain. The layers of the hierarchy alternate between the weighted linear summation and the MAX operation.

At the other extreme, histogram based models have shown good performance in single tasks categories such as cars and face [3]. One limitation of these approaches is their limited capabilities in capturing the objects' texture and shape. SIFT-based features, on the other hand, have shown excellent performance in terms of detecting previously seen objects from a different angle. However, experiments have shown that they offer limited performance in generic object recognition tasks.

Fig. 1. An image from living-room class of scene categories dataset. Highlighted are the two notions of co-occurrences. On the left is ambiguity co-occurrences: image patches compatible with multiple unrelated classes. On the left are contextual co-occurrences: patches of multiple other classes related to the image class.

The second class of scene understanding involves understanding the context of the complete scene rather than the texture of objects within the scene, this is known as "obtaining gist" [4], [5]. It is important to note that humans are able to recognize the meaning of a complex scene (or gist) within $\frac{1}{2}$ of a second, regardless of the number of objects in the scene [6]. In this stage of recognition the scene (or the image) is not grouped into regions, but decoded in a holistic framework. Recent studies have shown that semantic description of the scene can be achieved without the need for image segmentation [7]. A large body of evidence in both cognitive neuroscience [8] and psychology [9] literature support this.

The literature on holistic representation shows a variety of approaches to address a context within a scene. One approach is to calculate the low-level statistics of the visual features of the whole image [7]. This is by obtaining the differential regularities of the second order statistic. One other approach is by modifying the bag of features model for representing a local low level features. A holistic context model is formed by aggregating the bags across the image [9], [10]. However, these methods tend to neglect the spatial information to favour the invariance.

In this work, we present our biological inspired model for object recognition, that is the En-HMAX model [11], [12], to address the holistic scene understanding problem.

Our model analyses both the object-centric and holistic

The work of A. Alameer is supported by the HCED (Higher Committee for Education Development in Iraq). The work of K. Nazarpour is supported by the EPSRC, UK (grants: EP/M025977/1 and EP/M025594/1).

A. Alameer is with the School of Electrical and Electronic Engineering, Newcastle University, Newcastle NE1 7RU, UK.

Patrick Degenaar and K. Nazarpour are with the School of Electrical and Electronic Engineering and the Institute of Neuroscience, Newcastle University, Newcastle NE1 7RU, UK.

E-mail for correspondence: a.m.a.alameer@newcastle.ac.uk.

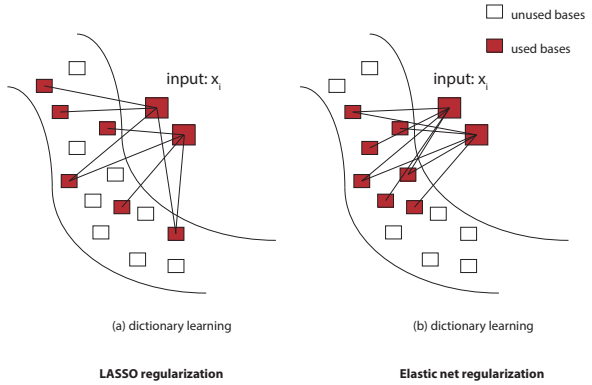

Fig. 2. Compression between LASSO regularization and Elastic net regularization. The selected bases are highlighted in red.

information to make the final decision. This is by extracting a mixture of low-level and high-level features throughout the hierarchy of the model. This enables the model to extract the object-centric information within the image. At the same time, using the elastic net in different layers of the hierarchy enable the model to better understand the highly correlated ambiguous data and extract meaningful features that regard to the global scene representation.

II. Sparsity and Dictionary learning in En-HMAX

Sparse coding is a linear model that can obtain linear statistical regularities from the input data. It has been shown that sparse coding can be used to generate an optimal dictionary in terms of computational resources and parameter tuning. There are many approaches that include dictionary learning for visual recognition; such as, the statistical approach for online dictionary learning has been proposed in [13].

The En-HMAX model [12] utilizes independent component analysis (ICA) to pre-process the input data. This is by projecting the input data to the whitening space, in order to extract only the informative decorrelated data. Elastic net, on the other hand, is used to sparsify the data and to create dictionaries that extract high level features though the hierarchy of the model. Both of the mentioned coding methods were utilized in both lower and higher layers of En-HMAX. The filters in the first layer were learned from natural images using ICA instead of the using Gabor filters. Dictionary learning was proposed using a combianation of the l_1 norm regularizer or LASSO (Least Absolute Shrinkage and Selection Operator) [14] and the l_2 norm or ridge regression regularizer. LASSO is a regression method that include penalizing the absolute size of the regression coefficients, to extract both low-and high-level features from the input images. While ridge regression is method that include all input data into the solution (sparse-free).

Natural scene images comprise nonlinear statistical regularities [15]. This outcome has motivated several findings to obtain non-linear statistical regularities from natural scene images [16]–[18]. It has also been shown that statistical regularities from images can be extracted using linear sparse coding, as a result of using the non-linear max pooling process in the C layers.

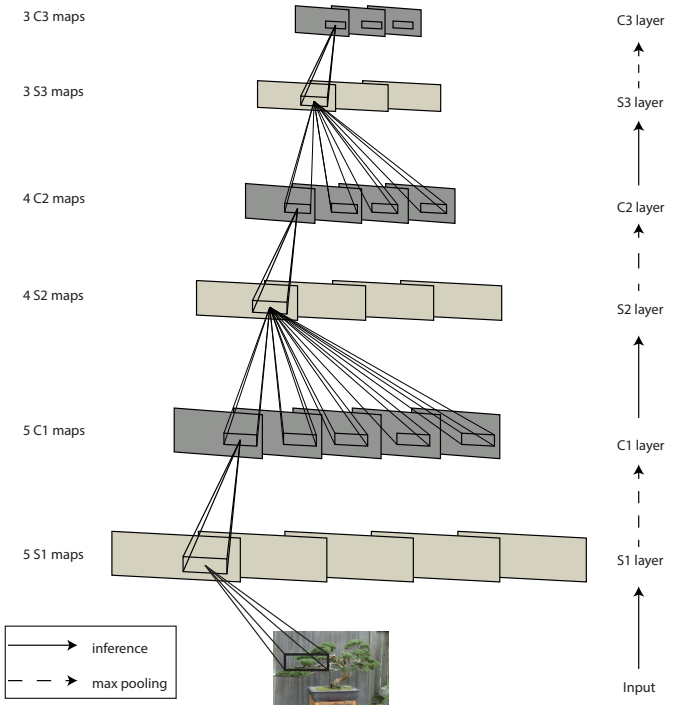

Fig. 3. Illustration of the En-HMAX model. The model has six layers in its hierarchy, three of them are simple and three of them are complex. Elastic net regularizer is used in the simple layers, while, norm-pooling is used in complex layers.

An elastic-net regularizer for dictionary learning was presented in [19]. The benefit of using the elastic net as a regularizer is taking the advantage of the sparsity of the l_1 norm as well as the grouping effect. Figure 2 illustrates how the elastic net regularizer select atoms from a dictionary with highly correlated data.

State of the art HMAX performance has been achieved using the l_1 norm regularizer [20]. However, if the atoms in the dictionary are correlated, the l_1 norm regularizer is confronted by a number of problems regarding its mechanism of selecting atoms. We therefore proposed the En-HMAX model in [12].

III. The architecture of the En-HMAX model

The architecture of En-HMAX (shown in Fig. 3) consists of the same simple (S_1, S_2) and complex (C_1, C_2) layers used in the original HMAX. However, En-HMAX model extended to a three-stage model, comprising an additional S_3 and C_3 layers to achieve more abstract representation of the input images. In line with the traditional HMAX, the proposed model uses the pooling operator in the C layers to achieve invariance. Using pooling operator helps the En-HMAX model to downsample the data through the hierarchy of the model, creating position invariant features. Norm pooling is proposed in the complex layers in the En-HMAX model to enhance the specificity of the features. On the other hand, elastic net regularizer is used in the higher layers of the model in order to achieve both sparsity and grouping in the generated dictionaries.

Inspired by [19], we augmented the dictionary learning approach in S_2 and S_3 by using both ℓ_1 and ℓ_2 norms of the sparse coefficients matrix as penalizing terms. In other words, let $\mathbf{x_i} \in R^m$ be an image patch in S_2 or S_3, $\mathbf{d_i} \in R^m$ be a set of bases, and s_j be elements of the sparse vector $\mathbf{s}$. Then, $\mathbf{x_i} = \sum_{j=1}^{p} \mathbf{d_i}\, s_j$, where p denotes the size of the dictionary represent sparse coding. In the matrix notation, this converts to $\mathbf{X} = \mathbf{DS}$, where $\mathbf{X}$ contains n-dimensional local descriptors extracted from the input images, $\mathbf{D}$ is a p-dimensional dictionary matrix with a set of $\mathbf{d_i}$ in its columns. Each column of $\mathbf{S}$ is a vector $\mathbf{s_i} \in R^p$ holding the sparse coefficients of the p bases. Therefore elastic-net regularization dictionary learning will be

$$\begin{aligned} \underset{\mathbf{D},\mathbf{S}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 + \lambda_2 \|\mathbf{S}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{d_i}\|_2 \leq 1, \forall i = 1, \cdots, p. \end{aligned} \tag{1}$$

where $\|.\|_F$ denotes the Frobenius norm and λ_1 and λ_2 are the regularization parameters that regulate the trade-off between sparsity and the sensitivity of basis selection. When $\lambda_1 = 1$ and $\lambda_2 = 0$, (1) reduces to the ℓ_1 coding method described in [20], [21], hereafter called LASSO-HMAX and when $\lambda_1 = 0$ and $\lambda_2 = 1$, (1) reduces to another extreme case, which we call Ridge-HMAX. The notions of LASSO and Ridge regressions are borrowed from [21].

A. Image Database

Fifteen scene categories in the dataset presented in [5] were selected. These classes were: bedroom, CAL suburb, industrial, kitchen, living room, MIT coast, MIT forest, MIT highway, MIT inside city, MIT mountain, MIT open country, MIT street, MIT tall building, PAR office, and store. Figure 4 shows examples of each class of the images. It also shows the number of images in each class.

Fig. 4. Example images from the scene category database. In this type of data sets, the important features are located on the peripheral rather than the center of the images.

B. Classification

The extracted features were classified using a linear support vector machine (SVM) [22]. The use of multi-class linear SVM has been supported in [23], that carried out a complete comparison between different classifiers in a visual recognition task. Similar investigations [24], [25] have suggested that, in comparison to the use of nonlinear kernels, standard linear SVMs have a reduced risk of over-fitting data. We have therefore used LIBLINEAR [23] library for classification. In order to solve the multi-class problem we used the one-vs-the-rest method, as implemented in LIBLINEAR. Additionally, SVM remained the top choice for our En-HMAX model because of its computational simplicity and speed.

We used the same settings used in other methods (shown in Table I) to have a fair performance comparison. In particular, we have used 100 images per class for training and the rest of images for testing. In addition, to ensure that the classification scores were not biased by the random choice of training samples, we repeated the classification for 20 independent runs. We report the average classification score together with the standard deviations.

TABLE I
CLASSIFICATION RESULTS FOR THE SCENE CATEGORY DATABASE

Feature types / recognition model	Classification performance
Our method [12]	**76.4 ± 0.005**
BSC [26]	72.5 ± 0.3
Rasiwasia [27]	72.2 ± 0.2
Liu [28]	63.32
Bosch [29]	72.7

IV. Results

Table I shows the complete results of the classification performance using 100 images per class for training and the rest for testing (the same setup as [26]–[29]). Average classification results across 20 independent runs and the standard deviations are reported. Our classification rate is 76.4%, which is much higher than the best results of 72.5 %, achieved in [26].

V. Discussion

This paper reported the behavior of the recently-proposed En-HMAX model in a scene understanding problem. The model, which basically operates using a combination of sparse coding and norm pooling, showed promising results. The En-HMAX model was originally designed to recognize objects (inspired by the ventral visual stream), however, it showed an acceptable level of scene understanding. In particular, our results on the scene dataset highlight the increased selectivity of the En-HMAX model as well as the invariance to local geometrical correlation. Furthermore, using the elastic net in different layers of En-HMAX enhanced the discriminative power toward highly correlated. This is perhaps because the En-HMAX model can extract meaningful features that correspond to the global scene representation.

It is important to note that spatial pyramid pooling used in the final layer has been essential to capturing the important features, and the discriminative dominant edges and lines. This enhanced the model ability to capture discriminative features that relate to the same class of images. For example, the office image class contains white documents stuck to the wall, the dark border of the kitchen cabinet door that belongs to the kitchen image class, and the dark window frames that correspond to the inside city image class.

References

[1] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[3] H. Schneiderman and T. Kanade, "A statistical method for 3D object detection applied to faces and cars," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1. IEEE, 2000, pp. 746–751.

[4] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[6] H. Intraub, "Visual scene perception," *Encyclopedia of cognitive science*, 2002.

[7] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int J. Comp. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[8] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[9] A. Oliva and P. G. Schyns, "Diagnostic colors mediate scene recognition," *Cognitive psychology*, vol. 41, no. 2, pp. 176–210, 2000.

[10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 524–531.

[11] A. Alameer, G. Ghazaei, P. Degenaar, and K. Nazarpour, "An elastic net-regularized HMAX model of visual processing," in *Proc. 2nd IET Int. Conf. Intell. Signal Process*, 2015.

[12] A. Alameer, G. Ghazaei, P. Degenaar, J. A. Chambers, and K. Nazarpour, "Object recognition with an elastic net-regularized hierarchical max model of the visual cortex," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1062–1066, Aug 2016.

[13] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[15] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Springer Science & Business Media, 2009, vol. 39.

[16] A. Hyvärinen and P. Hoyer, "Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[17] Y. Karklin and M. S. Lewicki, "A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals," *Neural computation*, vol. 17, no. 2, pp. 397–423, 2005.

[18] A. Hyvärinen, M. Gutmann, and P. O. Hoyer, "Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2," *BMC Neurosci.*, vol. 6, no. 1, p. 12, 2005.

[19] B. Shen, B.-D. Liu, and Q. Wang, "Elastic net regularized dictionary learning for image classification," *Multimedia Tools App.*, pp. 1–14, 2014.

[20] X. Hu, J. Zhang, J. Li, and B. Zhang, "Sparsity-regularized HMAX for visual recognition," *PLoS One*, vol. 9, no. 1, p. e81813, 2014.

[21] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Stat. Soc. Series B*, vol. 58, pp. 267–288, 1994.

[22] V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[24] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 530–546.

[25] X. Sun and W. Xu, "Fast implementation of delongs algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, 2014.

[26] N. Rasiwasia and N. Vasconcelos, "Holistic context modeling using semantic co-occurrences," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1889–1895.

[27] N. Ras and N. Vas, "Scene classification with low-dimensional semantic spaces and weak supervision," pp. 1–6, June 2008.

[28] J. Liu and M. Shah, "Scene modeling using co-clustering," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–7.

[29] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," in *European conference on computer vision*. Springer, 2006, pp. 517–530.