object recognition, hierarchical model,
feature extraction, visual cortex, computer vision

Rafał KOZIK[1]

A PROPOSAL OF BIOLOGICALLY INSPIRED HIERARCHICAL APPROACH TO OBJECT RECOGNITION

In this article a biologically-inspired algorithm for object recognition is presented. The approach is based on a hierarchical HMAX cortex model that was initially proposed by Riesenhuber and Poggio [12] and later extended by Serre et al [13]. The results show that despite the modification that were undertaken to simplify the HMAX model (in order to make it feasible for a real-time solutions) it is possible to achieve high effectiveness for a one-class detection problems. Moreover, it is also demonstrated how the proposed algorithm can be successfully deployed on a low-cost Android smartphone.

1. INTRODUCTION

The processing of the visual information in the human brain starts from the retina (see Fig. 1). Before the electric impulses reach the cortex at region V1 (primary visual cortex) they go through the relay centre called lateral geniculate nucleus (LGN).

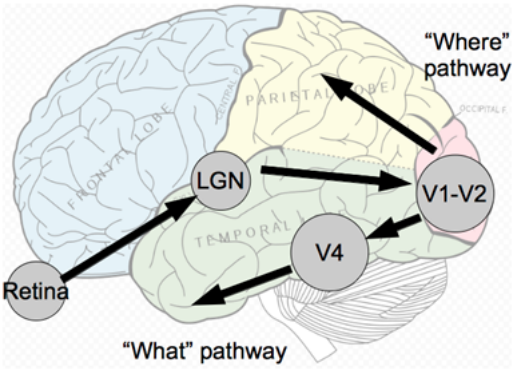

Fig. 1. Processing pathways in the human visual system proposed by Hubel and Wiesel [6].

There are multiple functions of the LGN including a temporal correlations as well as spatial correlations. The LGN is feed from M and P cells located in the retina. The P-cells play the major role in object recognition while the M-cells receive the input form large number of photoreceptors and are more sensitive in motion perception. Both M and P cells have so called On and Off centre surround inputs that map the absolute levels of illuminations to values encoding its differences for a particular neighbouring receptive fields (see Fig. 2).

[1]Institute of Telecommunications, UT&LS Bydgoszcz, Poland, e-mail: rkozik@utp.edu.pl.

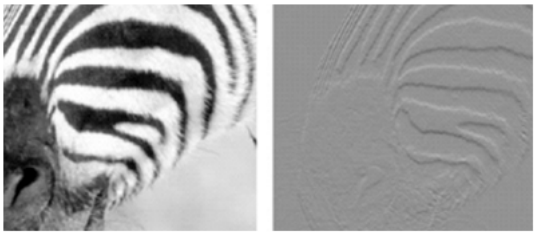

Fig. 2. Exemplar output (right image) of a differential spatial processing that takes place in a retina.

The visual information from the LGN cells projects onto V1 region (primary visual cortex), where the inputs are processed by layers of cortical neurons operating in a massively-parallel manner. V1 is the part of visual cortex that lies in the most posterior area of the occipital lobe (Fig. 1). The input form LGN reaches the bottom parts of visual cortex. Neurons in that area have concentric receptive fields. Those neurons send electrical impulses to other cortical layers. As noted by Hubel and Wiesel [6] cortical cells in these layers exhibit a transformation of the receptive field organization. Moreover, different groups of cells respond only to a particular type of stimuli such as edge or bar that has specific orientation.

The information from visual cortex is simultaneously transported to other regions of brain via ventral and dorsal pathway. So called dorsal stream takes part in object location, while the ventral stream is connected with object recognition. Whole this knowledge about the human visual system was used as an inspiration to build HMAX Visual Cortex model, which is explained in section 2.

This paper is structured as follows. First, the overview of the HMAX Visual Cortex model proposed by Riesenhuber and Poggio [12] is explained. The modifications introduced to the HMAX model are given in section 3. The conducted experiments are described and discussed in section 4. Final conclusions and remarks are given afterwards.

2. HMAX VISUAL CORTEX MODEL

The HMAX Visual Cortex model proposed by Riesenhuber and Poggio [12] exploits a hierarchical structure for the image processing and coding. It is arranged in several layers that process information in a bottom-up manner. The lowest layer is fed with a grayscale image. The higher layers of the model are either called ”S” or ”C”. These names correspond to simple the (S) and complex (C) cells discovered by Hubel and Wiesel [6]. Both type of cells are located in the striate cortex (called V1), which is the part of visual cortex that lies in the most posterior area of the occipital lobe.

The simple cells located in ”S” layers apply local filters that responses form a vector of texture features. As noted by Hubel and Wiesel the individual cell in the cortex respond to the presence of edges. They also discovered that these cells are sensitive to edge orientation (some of cells fire only when a given orientation of an edge is observed).

The complex cells located in ”C” layers calculate in a limited range of a previous layer the strongest responses of a given type (orientation). That way more complex combination of simple features are obtained combined from three simple features).

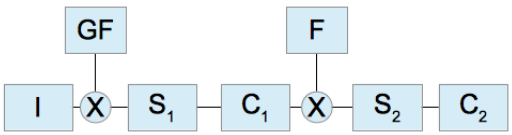

Fig. 3. The structure of a hierarchical model proposed by Mutch and Lowe [11]. (used symbols: I - image, S - simple cells, C - complex cells, GF - Gabor Filters, F - prototype features vectors, X - convolution operation).

The hierarchical HMAX model proposed by Mutch and Lowe [11] is a modification of the model

presented by Serre et al in [13]. It introduces two layers of simple cells (S_1 and S_2) and two layers of complex cells (see Fig. 3). It uses set of filters designed to emulate V1 simple cells. The C layers are computed using a hard max filter. It means that the "C" cells responses are the maximum values of the associated "S" cells. As it is shown in the Fig. 3 the images are processed by the subsequent simple and complex cells layers and reduced to feature vectors, which are further used in the classification process. The set of features (F) is shared across all images and object categories. Features are computed hierarchically in subsequent layers built from the previous one by alternating the template matching and the max pooling operations.

The S_1 layer in the Mutch and Lowe [11] model adapts the 2D Gabor filters computed for four orientations (horizontal, vertical, and two diagonal) at each possible position and scale. The Gabor filters are 11x11 in size, and are described by:

$$G(x,y) = e^{-(X^2+\gamma Y^2)/(2\sigma^2)} cos(\frac{2\pi}{\lambda}) \tag{1}$$

where $X = x\cos\phi - y\sin\phi$ and $Y = x\sin\phi + y\cos\phi$; x, y $\in< -5;5 >$, and $\phi \in< 0;\pi >$. The aspect ration (γ), effective width (σ), and wavelength (λ) are set to 0.3, 4.5 and 5.6 respectively. The response of a patch of pixels X to a particular filter G is computed using the formula (2).

$$R(X,G) = abs(\frac{\sum X_i G_i}{\sqrt{\sum X_i^2}}) \tag{2}$$

The complex cells located in C_1 layer pool associated units in the S_1 layer. For each orientation, the S_1 responses are convolved with a max filter, that is 10x10 of size in x, y dimension (position) and has 2 units of deep in scale.

As it is shown in the Fig. 3, the intermediate S_2 layer is formed by convolving the C_1 layer response with a set of intermediate-level features (depicted as F in Fig. 3). The set of intermediate-level features is established during the learning phase. For a given set of learning images C_1 responses are computed. The most meaningful features are selected using SVM weighting. Mutch and Lowe [11] suggested to sub-sample the C_1 responses before feature selection. Therefore, authors select at random positions and scales of the C_1 layer patches of size 4x4, 8x8, 12x12, and 16x16. Selected and weighted features compose so called prototypes that are used in the Mutch and Lowe model as filters which responses create the S_2 layer. The C_2 layer composes a global feature vector which particular element corresponds to the maximum response to a given prototype patch. In order to identify the visual object on the basis of feature vector a classifier is learnt (e.g. SVM).

3. PROPOSED APPROACH

The proposed approach follows the idea of HMAX model proposed by Riesenhuber and Poggio [12]. The modifications aim at reducing computational complexity of the original algorithm without decreasing the effectiveness of object recognition.

The hierarchical fed-forward processing approach is basically the same. In contrast to Mutch and Lowe [11] model an additional layer that mimics the "retina codding" mechanism is added. The results showed that this step increases the robustness of the proposed method.

The second modification includes a different method for calculating the S_1 layer response. The responses of 4 Gabor filters (two diagonals, horizontal and vertical) responses are approximated using horizontal and vertical Prewitt filters.

In contrast to the author previous work [7] the S_2 and C_2 layers are replaced with machine-learned classifier. Such approach significantly simplifies the learning process and decreases the amount of computations. However, this impacts the invariance to shape changes of a recognised object. The fact explaining that phenomenon is that the role of the S_2 layer is similar to a bag-of-words model applied to image classification. It encodes (without information about position) presence of a given visual feature

within the sliding window. Therefore, the object will be still recognized when the spatial relationships among the visual feature will change.

3.1. RETINA CODDING

Human retina shows remarkable and interesting properties of image enhancement. From a general point of view, the the retina serves as a first step of visual informations processing. In the literature there are several models explaining the basic mechanism the retina uses to encode visual information before it reaches visual cortex [2],[5]. There is also an implementation of such model in C/C++ code available in OpenCV library [1]. Basically, this model works as a filter that whitens the image spectrum and corrects luminance thanks to local adaptation. It has also the ability to filter out spatio-temporal noise and enhance the image details.

More simplistic approach to retina-based image enhancement was proposed in [3]. Authors adapted a local method that is based on a contrast equalisation. Within the sliding window authors normalises the luminance in the way, that the mean value is set to zero while the Euclidean norm is set to 1. This allows the authors enhance image details and reduce the noise.

In this work Difference of Gaussians (DoG) filter is used to mimic retina behaviour. It allows for feature enhancement and it involves the subtraction of two images blurred with different Gaussians filters (different standard deviation). It can be expressed with equation 3, where "$*$" represents convolution operation and σ_1 and σ_2 mentioned above standard deviations.

$$DoG_{\sigma_1\sigma_2}(x,y) = I * \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x^2+y^2)/(2\sigma_1^2)} - I * \frac{1}{\sigma_2\sqrt{2\pi}} e^{-(x^2+y^2)/(2\sigma_2^2)} \tag{3}$$

3.2. SIMPLE CELLS AND COMPLEX CELLS LAYERS

In order to achieve scale invariance the processing in S_1 layer is applied for three scales (an original image, and two images scaled by a factor of 0.7 and 1.5). For each scale in the S_1 layer there are NxMx4 simple cells arranged in a grid of size NxM blocks. In each block there are 4 cells. Each cell is assigned a receptive field (pixels inside the block). Each cell activates depend on a stimuli. In this case there are four possible stimulus, namely vertical, horizontal, left diagonal, and right diagonal edges. As a result the S_1 simple cells layer output has dimensionality of a size 4 (x,y,scale and 4 cells). In order to compute responses of all four cells inside a given block (receptive field), an algorithm 1 is applied.

The algorithm computes the responses of all cells using only one iteration over the whole input image I. For each pixel at position (x, y) a vertical and horizontal gradients are computed (G_x and G_y). Given the pixel position (x, y) and gradient vector $[G_x, G_y]$ the algorithm indicates the block position (n, m) and type of cell ($active$) that response has to be incremented by $|G|$. In order to classify given gradient vector $[G_x, G_y]$ as horizontal, diagonal or vertical the $get\_cell\_type(\cdot,\cdot)$ uses the wheel shown in Fig. 4. If a point $(|G_x|, |G_y|)$ is located between line $y = 0.3 \cdot x$ and $y = 3.3 \cdot x$ it is classified as a diagonal. If G_y is positive then the vector is classified as a right diagonal (otherwise it is a left diagonal). In case the point $(|G_x|, |G_y|)$ is located above line $y = 3.3 \cdot x$ the gradient vector is classified as vertical and as horizontal when it lies below line $y = 0.3 \cdot x$.

Data: Grayscaled image I of WxH size
Result: S_1 layer of size NxMx4
Assign G_{min} the low-threshold for gradient magnitude.
for *each pixel* (x, y) *in image* I **do**
  Compute horizontal G_x and vertical G_y gradients using Prewitt operator;
  Compute gradient magnitude $|G_{x,y}|$ in point (x, y);
  n $\leftarrow x \cdot \frac{N}{W}$; m $\leftarrow y \cdot \frac{M}{W}$; **if** $|G| < G_{min}$ **then**
    go to next pixel;
  else
    $active \leftarrow get\_cell\_type(G_x, G_y)$;
    $S_1[n, m, active] \leftarrow S_1[n, m, active]$ + $|G_{x,y}|$;
  end
end

Algorithm 1: Algorithm for calculating S_1 response.

The C_1 complex layer response is computed using max-pooling filter, that is applied to S_1 layer. The filter is a three dimensional one and is of a size 3x3x3 (x, y, scale).

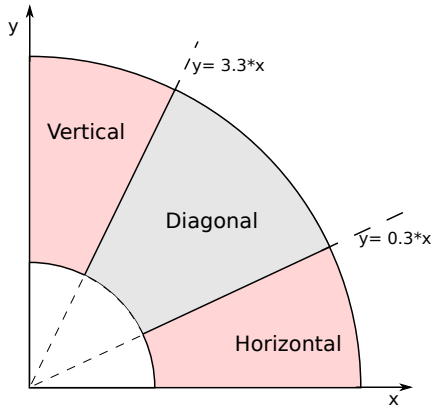

Fig. 4. A part of wheel that is used by $get\_cell\_type(\cdot, \cdot)$ to recognise a given gradient vector $[G_x, G_y]$ as horizontal, diagonal or vertical.

4. RESULTS

There are three experiments described in this section. First two experiments concern effectiveness evaluation, while the last one presents and discusses results obtained with the proposed algorithm deployed on an Android device. For the evaluation purposes a MIT CBCL [4] pedestrians data base is used. This dataset contains 924 images of pedestrians. Additionally, this dataset was extended with images obtained from surveillance system (see Fig. 5).

Fig. 5. Pedestrians samples from MIT CBCL [4] data base (on left) and example of testing sample obtained from surveillance system (big image on right).

The first experiment aimed at evaluating the influence of retina codding process on overall object recognition process. For that purpose MIT CBC pedestrians database is used. Results are shown in Fig. 6. In this experiment a Random Trees classifier is used. It can be noticed that retina coding allows for object recognition effectiveness improvement. For the experiment purposes the σ_1 parameter of DoG filter was constant and set to 1, while σ_2 was changed in range from 1.5 to 5.0.

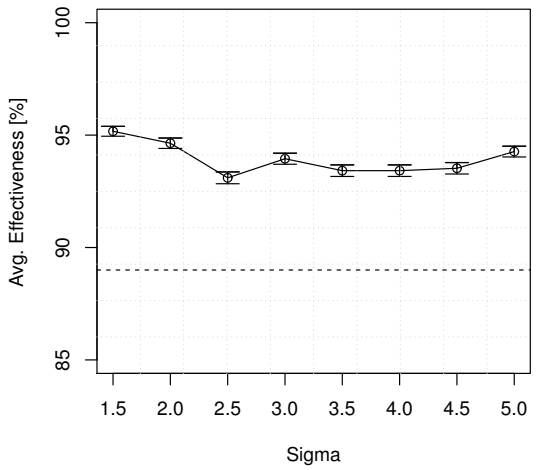

Fig. 6. Influence of retina codding on object recognition effectiveness. The dashed line indicates method without retina codding, while the solid line indicates method with retina coding enabled. Results are reported for $\sigma_1 = 1.0$ and varying σ_2.

The second experiment evaluates the influence of number of simple cells in S_1 layer on object recognition process. Results are shown in Fig. 7. During the experiment the size of S_1 was changed in range from 2x2 (4 cells) to 30x30 (900 cells). Additionally, for each experiment the number of trees in the classifier was also varying in range from 1 to 20. The best performance was observed for 20 trees and S_1 having 100 cells (99%).

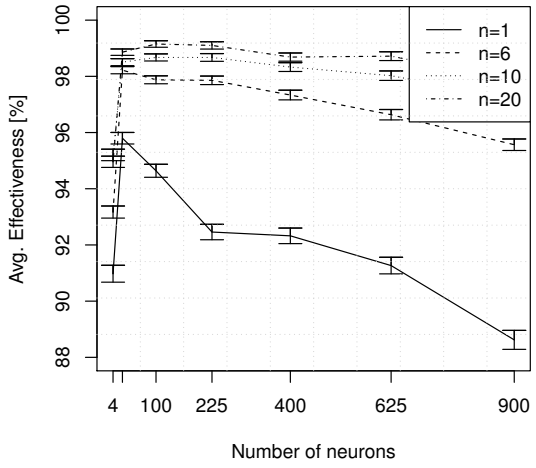

Fig. 7. Influence of number of simple cells in S_1 layer on object recognition process (n indicates number of trees in the classifier).

The early prototype of proposed algorithm was deployed on an Android device. The code was written in pure Java and tested on Samsung Galaxy Ace device. This device is equipped with 800 MHz CPU, 278 MB of RAM and Android 2.3 operating system. Current version implements brute force Nearest

Neighbour classifier, operates only for one scale (PC scans three scale - an original image, and two images scaled by a factor of 0.7 and 1.5), and achieves about 10 FPS when less than 10 training samples are provided. Some examples of object detection are shown in Fig. 8. During the testing it was noticed that the algorithm is able to correctly recognise an object on a cluttered background even if only few learning samples are provided.

Fig. 8. Example of object detection (mug and doors) with proposed algorithm deployed on an Android device.

5. CONCLUSIONS

In this article a simplified model of biologically inspired cortical mechanisms for object recognition was presented. The proposed approach was based on the HMAX hierarchical cortex model that was proposed by Riesenhuber and Poggio [12] and later extended by Serre et al [13]. The experiments show that the introduced algorithm allows for efficient feature extraction and a visual information coding. Moreover, it was shown that it is also possible to deploy proposed approach on a low-cost mobile device.

BIBLIOGRAPHY

[1] A bio-medic human retina model. OpenCV project homepage. http://docs.opencv.org/trunk/modules/contrib/doc/retina/#retina-a-bio-mimetic-human-retina-model.

[2] BENOIT A., CAPLIER A., DURETTE B., HERAULT, J., Using Human Visual System Modeling For Bio-Inspired Low Level Image Processing, Elsevier, Computer Vision and Image Understanding 114, 2010, pp. 758-773, DOI.

[3] BRUMBY S. P., GALBRAITH A. E. , HAM M., KENYON G., GEORGE J. S., Visual Cortex on a Chip: Large-scale, real-time functional models of mammalian visual cortex on a GPGPU, GPU Technology Conference (GTC) 2010, 2010, pp. 20-23.

[4] CBCL PEDESTRIAN DATABASE. http://cbcl.mit.edu/software-datasets/PedestrianData.html.

[5] HERAULT J., Vision: Images, Signals and Neural Networks: Models of Neural Processing in Visual Perception (Progress in Neural Processing), ISBN: 9814273686. WAPI (Tower ID): 113266891.

[6] HUBEL D. H., WIESEL T. N., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J Physiol, 1996, 160:106-154.

[7] KOZIK R., A Simplified Visual Cortex Model for Efficient Image Codding and Object Recognition. Image Processing and Communications Challenges 5. Advances in Intelligent Systems and Computing S. Choras, Ryszard 10.1007/978-3-319-01622-1, 2013, pp. 271-278.

[8] KOZIK R., Rapid Threat Detection for Stereovision Mobility Aid System , In: T. Czachorski et al. (Eds.): Man-Machine Interactions 2, AISC 103, 2011, pp. 115-123.

[9] KOZIK R., Stereovision system for visually impaired. Burduk, Robert (ed.) et al., Computer recognition systems 4. Berlin: Springer (ISBN 978-3-642-20319-0/pbk; 978-3-642-20320-6/ebook). Advances in Intelligent and Soft Computing 95, 2011, pp. 459-468.

[10] MAX pooling. http://ufldl.stanford.edu/wiki/index.php/Pooling.

[11] MUTCH J., LOWE D. G., Object class recognition and localization using sparse features with limited receptive fields. International Journal of Computer Vision (IJCV), October 2008, 80(1), pp. 45-57.

[12] RIESENHUBER M., POGGIO T., Hierarchical models of object recognition in cortex, 1999.

[13] SERRE T., KREIMAN G., KOUH M., CADIEU C., KNOBLICH U., POGGIO T., A quantitative theory of immediate visual recognition. In: Progress in Brain Research, Computational Neuroscience: Theoretical Insights into Brain Function, 2007, pp. 33-56.