

Algorithmic Optimizations in the HMAX Model Targeted for Efficient Object Recognition

Ahmad W. Bitar, Mohamad M. Mansour, and Ali Chehab^(✉)

Department of Electrical and Computer Engineering, American University of Beirut,
Beirut 1107 2020, Lebanon

{ab76,mmansour,chehab}@aub.edu.lb

<http://www.aub.edu.lb>

Abstract. In this paper, we propose various approximations aimed at increasing the accuracy of the S1, C1 and S2 layers of the original Gray HMAX model of the visual cortex. At layer S1, an image is convolved with 64 separable gabor filters in the spatial domain after removing some irrelevant information such as illumination and expression variations. At layer C1, some of the minimum scales values are exploited in addition to the maximum ones in order to increase the model's accuracy. By applying the embedding space in the additive domain, the advantage of some of the minimum scales values is taken by embedding them into their corresponding maximum ones based on a weight value between 0 and 1. At layer S2, we apply clustering, which is considered one of the most interesting research areas in the field of data mining, in order to enhance the manner by which all the prototypes are selected during the feature learning stage. This is achieved by using the Partitioning Around Medoid (PAM) clustering algorithm. The impact of these approximations in terms of accuracy and computational complexity was evaluated on the Caltech101 dataset containing a total of 9,145 images split between 101 distinct object categories in addition to a background category, and compared with the baseline performance using support vector machine (SVM) and nearest neighbor (NN) classifiers. The results show that our model provides significant improvement in accuracy at the S1 layer by more than 10% where the computational complexity is also reduced. The accuracy is slightly increased for both approximations at the C1 and S2 layers.

Keywords: HMAX · Support vector machine · Nearest neighbor · Caltech101

1 Introduction

The human visual system is quite powerful. It is perhaps not too surprising that the human brain has achieved, through millions of years of evolution, a remarkable ability to recognize and differentiate among very similar objects in a selective, robust and fast manner. Modern machines can perform many apparently

complex tasks much faster, more efficiently and more precisely than humans. Some estimates indicate that the human visual system can discriminate at least tens of thousands of different object recognition. Therefore, it would be relatively easy to build a computer system that can be extremely selective by just memorizing all the pixels in several training images. Modern computers are able to translate the human ventral visual pathway (known as the “WHAT” stream) in order to achieve, in a similar manner to the human brain, an impressive trade-off between selectivity and invariance. Several scientists have attempted to model and mimic the human vision system [1].

The Hierarchical Model And X (HMAX) is an important model for object recognition in the visual cortex known for its high ability to achieve performance levels close to the human object recognition capability [2]. HMAX divides the human ventral stream into five layers: S1, C1, S2, C2 and View-Tuned (VTU).

The first layer S1 of the HMAX model relies on the Gabor filter [3], which is a linear filter used for edge detection. It differs from other filters by its capability to highlight all the features that are oriented in the direction of the filtering. The features are therefore extracted from the images by tuning the gabor filter to several different scales and orientations using fine-to-coarse approach.

Several methods have been proposed in the literature in order to improve the efficiency of the original HMAX model. An extension of the original HMAX model has been proposed in [4], emphasizing the importance of shape selectivity in area V4. A simpler radial basis function (RBF) model for object recognition was proposed in [5] to maintain a good degree of translation and scale invariance. The proposed model was considered better than the original HMAX for translation and scale invariance by changing the point of attention and decreasing the amount of visual information to be processed. In [6], they developed a new set of receptive field shapes and parameters for cells in the S1 and C1 layers. The method serves to increase position invariance in contrast to scale invariance, which is decreased. In [7], they proposed a general framework for robust object recognition of complex visual scenes based on a quantitative theory of the ventral pathway of visual cortex. A number of improvements to the base model were proposed in [8] in order to increase the sparsity. The proposed model has shown a remarkable improvement on classification performance and the resulting model is found more economical in terms of computations. In [9], they proposed several approximations at the four HMAX layers (S1, C1, S2 and C2) in order to increase the efficiency of the model in terms of accuracy and computational complexity. A semi-supervised learning algorithm for visual object categorization was proposed in [10] by exploiting unlabelled data and employing a hybrid generative-discriminative learning scheme. The method achieved good performance in multi-class object discrimination tasks. In [11], they proposed a scheme based on a kernel function for discriminative classification. The method achieved improved accuracy and reduced computational complexity compared to the baseline model.

In this paper, the goal is to perform various optimizations at the S1, C1 and S2 layers of the original HMAX model. The results demonstrate that these optimizations increase the accuracy of the HMAX model as well as reduce its computational complexity at the S1 layer. The accuracy of the final model proves the advantage of exploiting only the important features for recognition and generating the prototypes in a more efficient way.

The remainder of this paper is organized as follows. In Sect. 2, the visual system of the human brain is briefly presented. In Sect. 3, a brief overview of the original HMAX model is explained. The proposed approximations at S1, C1 and S2 layers are presented in Sects. 4, 5 and 6, respectively. Experimental results are shown in Sect. 7. Finally, Sect. 8 gives concluding remarks and some directions for future work.

2 The Visual System of the Human Brain

The light enters our eye from the pupil to the retina through the Crystalline lens. The iris is considered the colored part of the eye and it controls the amount of light that enters our eye. The pupil is the central aperture of the iris and the retina sends images to the brain through the optic nerve [18, 19].

The retina contains five types of neurons: Photoreceptors (95 % Rods and 5 % Cones), Horizontal neurons, Bipolar neurons, Amacrine neurons and Ganglion neurons. There is amazing collaboration among all 5 neuron types. Of the existing 120 millions of photoreceptors in each eye, 95 % are Rods. In fact, the Rods are located on the surface of the retina, only sensible to luminance, responsible for vision at low light and active in scotopic vision. The Cones are only sensible to chrominance, responsible for vision at normal light, active in photopic vision and located in the fovea which constitutes 1 % of the retina's surface. Interestingly, both Rods and Cones are active in mesopic vision which is considered as a combination between scotopic and photopic.

One of the most important problems in vision is that at low light, the pupil increases in size, the light reflects into the retina's surface where the Rods are present. When suddenly a high level of light comes to the eye and before the pupil decreases in size, it reflects directly into the Rods which are sensible to luminance not to chrominance.

The ganglion cells are the most important to study. The ganglion of type "P" are for Parvo (very small receptive field), the ganglion of type "M" are for Magno (large receptive field) and the ganglion of type "K" are for Konio or conio (also called non P-nonM). All ganglion cells' types have receptive field described as "Center ON-Peripheric OFF" or "Center OFF-Peripheric ON".

The optic nerve contains the ganglion fiber optic. Both left and right optic nerves of the left and right eyes, respectively, are crossed in a point called "Optic chiasm" which transmits the information received from the retina to the Lateral Geniculate Nucleus (LGN).

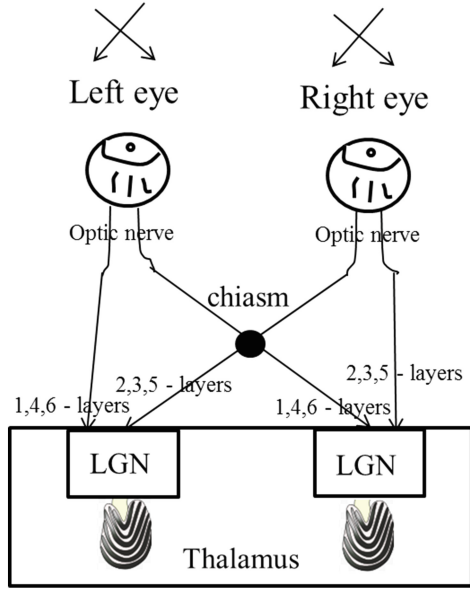


Fig. 1. Connections between the eye and LGN.

The LGNs are located in the thalamus of the brain and they are responsible for processing the information that arrives from the ganglion neurons. Each LGN is formed by six distinct layers numbered 1 to 6. The layers 1-2 contain neurons of type “M” and the layers 3-4-5-6 contain neurons of type “P”. There is LGN in each hemisphere of the brain. The right LGN receives stimulus from the left visual field and vice versa as shown in Fig. 1. Layers 1, 4, 6 of both right and left LGNs receive axons from left part of the retina of each eye (nasal hemiretine) while the layers 2, 3, 5 of both right and left LGNs receive axons from right part of retina of each eye (temporal hemiretine).

The left and right primary visual cortex V1 receive information from the left and right LGN, respectively. The primary visual cortex is the part of the cerebral cortex responsible for processing visual information and it is located in the occipital lobe of the brain. The primary visual cortex V1 is also known as “Striate cortex” or “Brodmann area 17 (BA17)”. It is located in and around the calcarine fissure (or calcarine sulcus) in the occipital lobe of the brain. Importantly, it is divided into 6 distinct layers labeled 1 through 6. Layer 4, which receives the most visual input from the LGN is further divided into 4 layers: 4A, 4B, 4C α (receives most Magnocellular inputs from the LGN) and 4C β (receives most Parvocellular inputs from the LGN). The V1 of each hemisphere transmits information to two primary pathways: Dorsal Stream and Ventral Stream. The object recognition in cortex is thought to be mediated by the ventral visual pathway running from visual cortex V1, over extrastriate visual areas V2 and V4 to Inferotemporal cortex IT. Based on physiological experiments in monkeys,

IT has been postulated to play a central role in object recognition. IT in turn is a major source of input to PFC, “the center of cognitive control” involved in linking perception to memory.

For further details, see Chap. 7 from the book “Brains: How They Seem To Work” [20].

3 HMAX Model with Feature Learning

HMAX [12] is a computational model that summarizes the organization of the first few stages of object recognition in the WHAT pathway of the visual cortex, which is located in the occipital lobe at the back of the human brain. It is considered a primordial part of the cerebral cortex responsible for processing visual information in the first 100–150 ms. Indeed, light enters our eye from the central aperture, called “Pupil”, and then passes through the “Crystalline lens” which is considered the biconvex transparent body situated behind the iris into the eye and aiming to focus light on the retina that sends images to a specific part of the brain (visual cortex) through the optic nerve. The retina contains five different types of connected neurons: Photoreceptors (95% rods and 5% cones), Horizontal, Bipolar, Amacrine and Ganglion through which the light leaves the eye. The visual cortex, located in and around the calcarine sulcus, refers to the striate cortex V1, anatomically equivalent to Brodmann area 17 (BA17), connected to several extrastriate visual cortical areas (V2, V4, V5, etc.), anatomically equivalent to Brodmann area 18 and Brodmann area 19. The right and left V1 receive information from the right and left Lateral Geniculate Nucleus (LGN), respectively. The LGNs are located in the thalamus of the brain and they receive information directly from the ganglion cells of the retina via the optic nerve and optic chiasm.

3.1 Computational Complexity

The operations of the five layers of the HMAX model are briefly summarized.

S1 Layer: All the responses of the S1 units are summarized here by simply performing 2-D convolution between 64 Gabor filters (16 scales in steps of two pixels and 4 orientations) shown in Fig. 2 and the input images in the spatial domain.

Firstly, each Gabor filter of a specific scale and orientation can be initialized as:

$$G(x, y) = \exp\left(-\frac{u^2 + \gamma^2 v^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} u\right), \quad (1)$$

where:

$$\begin{aligned} u &= x \cos \theta + y \sin \theta, \\ v &= -x \sin \theta + y \cos \theta, \\ \gamma &= 0.0036 \times \rho^2 + 0.35 \times \rho + 0.18, \\ \lambda &= \frac{\gamma}{0.8}. \end{aligned}$$

The parameter γ is the aspect ratio at a particular scale, θ is the orientation $\in [0^\circ, 45^\circ, 90^\circ, 135^\circ]$, σ represents the effective width ($=0.3$ in our case), λ is the wavelength at a particular scale, and ρ represents the scale.

Secondly, all the S1 image responses are computed by applying a two dimensional convolution between the initialized Gabor filters and the input images in the spatial domain. The S1 image responses are so-called: the Gabor features.

In fact, all the filters are arranged in 8 bands. There are two filter scales with four orientations at each band.

The S1 layer has a computational complexity of $O(N^2M^2)$ where $M \times M$ is the size of the filter and $N \times N$ is the size of the image.

C1 Layer: The C1 units are considered to have larger receptive field sizes and a certain degree of position and scale invariance. For each band, each C1 unit response (image response) is computed by taking the maximum pooling between the gabor features of the two scales at the same orientation. The main role of the maximum pooling function is to subsample the number of the S1 image responses and increase tolerance to stimulus translation and scaling. Then, the pooling over local neighborhood using a grid of size $n \times n$ is performed. From band 1 to 8, the value of n starts from 8 to 22 in steps of two pixels, respectively. Furthermore, a subsampling operation can also be performed by overlapping between the receptive fields of the C1 units by a certain amount $\Delta_s (= 4_{\text{band1}}, 5_{\text{band2}}, \dots, 11_{\text{band8}})$, given by the value of the parameter C1Overlap. The value C1Overlap = 2 is mostly used, meaning that half the S1 units feeding into a C1 unit were also used as input for the adjacent C1 unit in each direction. Higher values of C1Overlap indicate a greater degree of overlap. This layer has a computational complexity of $O(N^2M)$.

S2 Layer: The original version of HMAX was the *standard model* in which the connectivity from C1 to S2 was considered *hard-coded* to generate several combinations of C1 inputs. The model was not able to capture discriminating features to distinguish facial images from natural images. To improve that, an extended version was proposed [1], and is called *HMAX with feature learning*. In this model, each S2 unit acts as a Radial Basis Function (RBF) unit, which serves to compute a function of the distance between the input and each of the stored prototypes learned during the feature learning stage. That is, for an image patch X from the previous C1 layer at a particular scale, the S2 response (image response) is given by:

$$S2_{\text{out}} = \exp^{(-\beta \|X - P_i\|^2)}, \quad (2)$$

where β represents the sharpness of the tuning, P_i is the i th prototype and $\|\cdot\|$ represents the Euclidean distance. This layer has a computational complexity of $O(PN^2M^2)$, where P is the number of prototypes.

C2 Layer: It is considered the layer at which the final invariance stage is provided by taking the maximum response of the corresponding S2 units over all

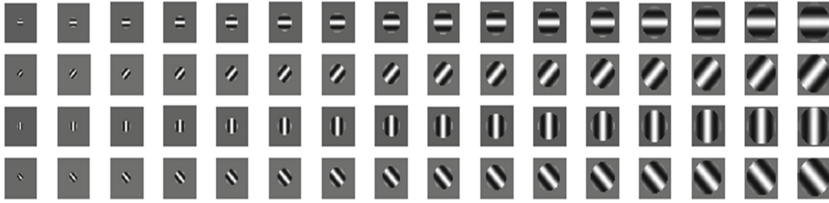


Fig. 2. 64 Gabor filters (16 scales in steps of two pixels $[7 \times 7$ to $37 \times 37] \times 4$ orientations $[0^\circ, 45^\circ, 90^\circ, 135^\circ]$).

scales and orientations. The C2 units provide input to the VTUs. This layer has a computational complexity of $O(N^2MP)$.

VTU Layer: At runtime, each image in the database is propagated through the four layers described above. The C1 and C2 features are extracted and further passed to a simple linear classifier. Typically, support vector machine (SVM) and nearest neighbor (NN) classifiers are employed.

The learning stage: The learning process aims to randomly select P prototypes used for the S2 units. They are selected from a random image at the C1 layer by extracting a patch of size 4×4 , 8×8 , 12×12 , or 16×16 at random scale and position (Bands 1 to 8). For an 8×8 patch size for example, it contains $8 \times 8 \times 8 = 512$ C1 unit values instead of 64. This is expected since for each position, there are units representing each of the four orientations $[0^\circ, 45^\circ, 90^\circ, 135^\circ]$.

4 S1 Layer Approximations

At the S1 layer, several approximations are investigated in order to increase the efficiency of the original HMAX model in terms of accuracy and computational complexity. Each approximation has been evaluated independently using SVM and NN classifiers.

4.1 Combined Image-Based HMAX Using 2-D Gabor Filters

In this approximation, all unimportant information such as illumination and expression variations are eliminated from the image and hence its salient features become richer [13]. To achieve this, four main steps are applied to the original image A of size $h \times a$:

Step 1 – Adaptive Histogram Equalization: In order to handle the large intensity values to some extent, adaptive histogram equalization is applied to the original image A :

$$\text{Adapted_Image} = \text{AdaptHistEq}(A) \tag{3}$$

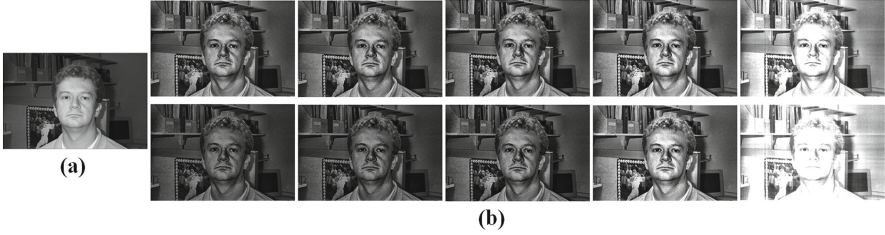


Fig. 3. (a) The original image and (b) Combined images using $\alpha = 0.25, 0.5, 0.75, 1$ and 1.25 , respectively. c is equal to 0.25 and 0.75 on the top and bottom, respectively.

Step 2 – SVD Decomposition: Singular value decomposition (*SVD*) is applied to the image after equalization. The concept behind *SVD* is to break down the image into the product of three different matrices as:

$$SVD(\text{Adapted_Image}) = \mathbf{L} \times \mathbf{D} \times \mathbf{R}^T \tag{4}$$

where \mathbf{L} is the orthogonal matrix of size $h \times h$, \mathbf{R}^T is the transpose of an orthogonal matrix \mathbf{R} of size $a \times a$ and \mathbf{D} is the diagonal matrix of size $h \times a$. This decomposition helps the computations to be more immune to numerical errors, as well as to expose the substructure of the original image more clearly and orders their elements from most amount of variation to the least.

Step 3 – Reconstruction Image: According to the values of \mathbf{L} , \mathbf{D} and \mathbf{R} , the reconstructed image is computed as follows:

$$\text{Reconstructed_Image} = \mathbf{L} * \mathbf{D}^\alpha * \mathbf{R}^T, \tag{5}$$

where α is a magnification factor that varies between 1 and 2. The idea to have the value of α vary between one and two in order to magnify the singular values of \mathbf{D} is to make them invariant to illumination changes. When α equals to 1, the reconstructed image is equivalent to the equalized image. When α is chosen between $]1\ 2]$, then the singular values greater than unity will be magnified. Thus, the combination between the reconstructed image and the equalized image will be a fruitful step to making the model more robust against illumination and expression variations.

Interestingly, when the singular values are scaled in the exponent, a non-linearity is introduced. Therefore for a specific database (Caltech101 for example), scaling down the magnification factor α may be helpful.

Step 4 – Combined Image: The combined image is produced by simply combining the reconstructed image and the equalized image as shown in Fig. 3, using a combination parameter c which varies between 0 and 1.

$$I_{\text{Comb}} = \frac{\text{Adapted_Image} + (c * \text{Reconstructed_Image})}{1 + c} \tag{6}$$

By applying this approximation, the computations in this layer become faster as shown in Fig. 6 since only the significant information are used for recognition. In

addition, the approximation can significantly improve the model’s accuracy. It can be explained by the fact that when the model uses a challenge database such as Caltech101 or Caltech256 in which there are a lot of unimportant information such as illumination and expression variations, it will be interesting to exploit only the most important features in the images in order to make the recognition easier and more robust where the accuracy is increased by 10 % using SVM while by more than 13 % when using NN classifier. There are no related works yet that approximate the S1 layer.

4.2 Combined Image-Based HMAX Using Separable Gabor Filters

In this approximation, all the combined images of the previous approximation are convolved with 64 Gabor filters in a separable manner ($G(x, y) = f(x)g(y)$), instead of just performing the 2-D convolution. In this case, the Gabor features are computed using two 1-D convolutions corresponding to convolution by $f(x)$ in the x-direction and $g(y)$ in the y-direction. Based on the definition of separable 2-D filters, the Gabor filters are parallel to the image axes ($\theta = k\pi/2$). In order to be applied to an image along diagonal directions, they have been extended to further work with $\theta = k\pi/4$. The main issue of these techniques is that they will not work with any other desired direction. To handle this problem, Eq. (1) can be rewritten using the isotropic version ($\gamma = 1$, circular) in the complex domain [9]. In this case, $u^2 + v^2 = (x \cos \theta + y \sin \theta)^2 + (-x \sin \theta + y \cos \theta)^2 = x^2 + y^2$.

$$\begin{aligned}
 G(x, y) &= e^{-\frac{x^2+y^2}{2\sigma^2}} \times \cos\left(\frac{2\pi}{\lambda}(x \cos(\theta)+y \sin(\theta))\right) \\
 &= \text{Re}(f(x)g(y))
 \end{aligned}$$

where

$$\begin{aligned}
 f(x) &= e^{-\frac{x^2}{2\sigma^2}} \times e^{ix \cos(\theta)}, \\
 g(y) &= e^{-\frac{y^2}{2\sigma^2}} \times e^{iy \sin(\theta)}.
 \end{aligned}$$

Finally, the convolution using this approximation can therefore be expressed as:

$$I_{\text{Comb}} * G(x, y) = I_{\text{Comb}}(x, y) * f(x) * g(y) \tag{7}$$

By exploiting the separability of Gabor filters and convolving them with the original image, the computational complexity is reduced from $O(N^2M^2)$ to $O(tN^2M)$ where $t=8$ due to complex valued arithmetic. But since in this approximation, the separable Gabor filters are convolved with the combined image I_{Comb} , the complexity is being more reduced since only the significant information are used for recognition. The accuracy is not increased by more than 10.5 % for SVM (between 10.4 % and 10.5 %) while is increased by more than 14 % for the NN classifier.

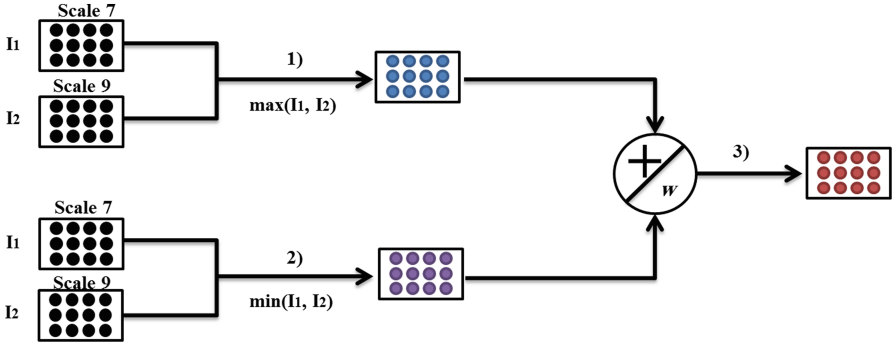


Fig. 4. Scheme example of the C1 approximation.

4.3 Combined Image-Based HMAX Using Haar Wavelet Transform

The foundation of the discrete wavelet transform (DWT) goes back to 1976 when Crochiere et al. for the first time introduced sub-band coding [15]. In 1983, Burt defined a technique very similar to sub-band coding and named it pyramidal coding which is also known as multi resolution analysis [16]. Later in 1989, Vetterli and Le Gall made some improvements to the sub-band coding scheme and removed the existing redundancy in the pyramidal coding scheme [17]. DWT definition is based on sub-band coding and multi-resolution analysis.

In this approach, we add one more step to the Subsect. 4.1 in order to have a total of 5 steps. Hence, to handle efficiently the condition variations, the wavelet transform DWT of LEVEL1 decomposition can be used to segment the image into four sub-bands: Low frequency component (LL), and High frequency components (LH, HL and HH). Thus, to help the recognition process to fully focus on important features, the LL sub-band has been considered ineffective with illumination changes and expression variations.

4.4 Baseline Model Using Haar Wavelet Transform

In this approach, we only added the Multi-resolution approach step to the baseline model.

5 C1 Layer Approximations

Concerning the C1 layer, a pooling between the S1 responses over scales within each band is performed by simply taking the maximum response between them. By testing what can be the result of the minimum pooling that has not been exploited at this layer, it was noticed that all the minimum scales values are very close to their corresponding maximum ones. Some of them are equal, otherwise the most of minimum scales values are not smaller more than 6 or 7%. As such, it will be important to further consider some of the minimum scales values

when taking the maximum pooling. In other words, some of the minimum scales values can be exploited in addition to the maximum ones in order to increase the model's accuracy. But the remaining question to be solved is "How to take advantage of minimum and maximum scales values at the same time". So that under a specific conditions, some of the minimum scales values can be embedded into their corresponding maximum ones. The easiest way to achieve that is to apply the embedding in the additive domain. A general scheme of this approximation is shown in Fig. 4. In this figure, two S1 image responses I_1 and I_2 of the same orientation at the first band (band1) are considered and which are belong to the filter scale 7 and 9, respectively. The circles shown within the images correspond to their pixels. In step 1, the maximum pooling (max function) is performed between I_1 and I_2 . The pixels of the resulting image correspond to the maximum scales values (shown with blue circles). In step 2, the minimum pooling (min function) is performed between I_1 and I_2 . The pixels of the resulting image correspond to the minimum scales values (shown with violet circles) that are then embedded into their corresponding maximum ones in the additive domain under specific conditions as shown in step 3. In other words, each minimum scale value is added into the maximum one that has the same (x, y) coordinates. w is the weight of the embedding.

Embedding in the Additive Domain: This kind of embedding is very straightforward to implement since the minimum scales values (after applying the minimum pooling over scales within each band) can be directly embedded into their corresponding maximum values by simply using the addition operator.

Generally, the embedding process at a particular pixel coordinate (x, y) in the additive domain can be expressed as:

$$I_{\text{Embed}}(x, y) = \max_{\text{scale}}(x, y) + w * \min_{\text{scale}}(x, y), \quad (8)$$

where $I_{\text{Embed}}(x, y)$ represents the final result after the embedding process, \max_{scale} is the maximum scale value, \min_{scale} is the minimum scale value, and $w \in [0, 1]$ represents the weight of the embedding.

Two different conditions are considered to embed the minimum scales values into their corresponding maximum ones:

Condition 1: At each band, after computing the maximum pooling over scales of the same orientation, the minimum pooling is also performed and then all the minimum scales values are embedded into their corresponding maximum ones. In this case, w is set to 1.

Condition 2: Each minimum scale value is embedded if and only if its corresponding maximum value belongs to the interval $[0\% \ 5\%[$. The values within the interval specifies how much a maximum scale value is greater than its corresponding minimum one. In fact, the interval $[0\% \ 5\%[$ is divided into two groups: $[0\% \ 2\%[$ and $[2\% \ 5\%[$, and two distinct sub-conditions are thus considered:

- *Sub-condition 1:* The embedding is performed by setting w to 1 for [0% 2%[and 0.5 for [2% 5%[.
- *Sub-condition 2:* The embedding is performed by setting w to 0.5 for [0% 2%[and 0.1 for [2% 5%[.

The accuracy is not increased by more than 1% in all conditions when SVM is used, while the opposite for NN classifier. However, the computational complexity at this layer is slightly increased due to the embedding process.

6 S2 Layer Approximations

At the S2 layer, the focus is to enhance the manner by which all the prototypes are selected during the feature learning stage. In the original model, P prototypes are randomly selected from the training images at the C1 layer. If more than P prototypes are used, the model's accuracy will increase at the expense of additional computational complexity. That is why our motivation is to learn the same number of prototypes P but in an efficient way in order to decrease the model's false classification rate while keeping the same computational complexity.

In order to achieve this, clustering is exploited, which is considered one of the most important research areas in the field of data mining. It aims to divide the data into groups, (clusters) in such a way that data of the same group are similar and those in other groups are dissimilar. Clustering is considered useful to obtain interesting patterns and structures. That is why, one of the existing clustering algorithms, more specifically the Partitioning Around Medoid (PAM) clustering algorithm [14] has been exploited in this approximation to generate the prototypes.

Furthermore, one of the important issues to consider, is the redundancy of some prototypes especially those selected from the homogeneous areas of the image (prototypes' pixels are being equal to zero). That is why, our contribution also aims to generate a non-redundant P prototypes and force the model not to generate any unimportant prototype. Accordingly, each of the selected prototypes will be important and aims to increase the model's accuracy.

PAM is characterized by its robustness to the presence of noise and outliers. Its complexity is defined by $O(i(b - q)^2)$ where i is the number of iterations, q is the number of clusters, and b represents the total number of objects in the data set.

To generate 2000 prototypes in a more efficient way and use them in our model instead of the traditional ones, the PAM algorithm is performed and it consists of 6 different steps:

Step 1 –5 medoids of 4×4 pixels at four orientations of each training category (total of 30 images) from the total 102 categories are randomly initialized.

Step 2 –For each category, the Frobenius distance between each of the C1 response of each image with all the selected medoids is then computed in order to associate each data image to the closest medoid.

Step 3 –For a random cluster, a non-medoid image patch is randomly selected in order to be swapped with the original medoid of the cluster in which the non-medoid is selected.

Step 4 –steps 2 and 3 are repeated until the total cost of swapping becomes greater than zero. The total cost of swapping can be defined as follows:

$$\text{Cost}_{\text{swapping}} = \text{Current Total Cost} - \text{Past Total Cost}$$

Step 5 –All the previous steps are also performed for all the other remaining three sizes of the medoids (8×8 , 12×12 and 16×16) in order to have a total of 20 medoids in each category.

Step 6 –Finally, a total of 2040 medoids are being selected to be used as prototypes. 10 prototypes are dropped from each size in order to end up with only 2000 prototypes.

This algorithm is complex since there are six steps to perform in order to generate the prototypes. But in fact, the run of the HMAX model relies on two parts. The first part is responsible to generate and reserve all the necessary prototypes by only running the first two layers S1 and C1. The second part consists of running the whole model and use the prototypes that have been generated and reserved for the S2 layer. Interestingly, the complexity of the model depends only on the second part, which means that the large complexity of our algorithm does not affect the computational complexity of the model, more precisely, of the S2 layer. That is why, the computational complexity at the S2 layer of our model remains $O(PN^2M^2)$, where P is the number of prototypes. By applying this approximation, the accuracy of the model increases by 0.68% approximately using the SVM classifier.

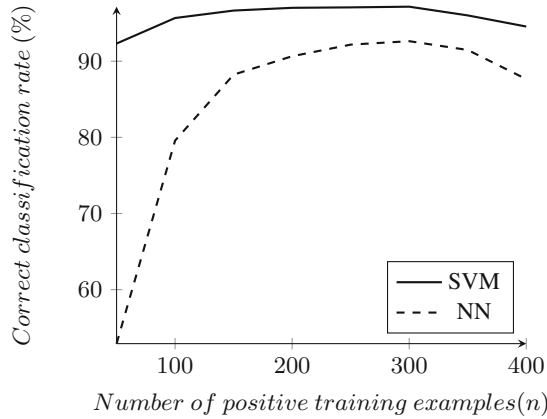
7 Experimental Results

The proposed optimizations at the S1, C1 and S2 layers were implemented using MATLAB in order to evaluate their accuracy and computational complexity using experimental simulations. The S1, C1 and S2 approximations were evaluated using the Caltech101 database, which contains a total of 9,145 images split between 101 distinct object categories in addition to a background category. All the results of our approximations were the average of 3 independent runs. For each run, the following steps were performed:

1. A set of 30 images are randomly chosen from each category for training, while all the remaining images are used for testing. All the images are normalized to 140 pixels in height and the width is rescaled accordingly so that the image aspect ratio is preserved.

Table 1. Simulation results for SVM and NN on face category.

Positive training	SVM	NN
50	92.325 %	52.903 %
100	95.678 %	79.578 %
150	96.656 %	88.240 %
200	97.018 %	90.658 %
250	97.075 %	92.181 %
300	97.165 %	92.634 %
350	96.019 %	91.480 %
400	94.558 %	87.649 %


Fig. 5. SVM and NN accuracies on face category.

2. C1 sub-sampling ranges do not overlap in scales.
3. The prototypes are learned at random scales and positions. They are extracted from all the eight bands.
4. C2 vectors are built using the training set.
5. Training applied using both SVM and Nearest-Neighbor classifiers.
6. C2 vectors for the test set are built, and then the test images are classified.

7.1 Performance of SVM and NN

The performance of both SVM and NN are performed on the face category extracted from the caltech101 database and which contains 435 face images.

The results show that the accuracy decreases when the number of training becomes greater than 300. This is expected because the data becomes unbalanced.

The images were rescaled to 160×160 pixels, the C1 sub-sampling ranges overlap in scale (C1Overlap = 2) and the prototypes are chosen only from Bands

1 and 2. The classifiers were trained with $n = 50, 100, 150, 200, 250, 300, 350$ and 400 positive examples and 50 negative examples from the background class, while they are tested with all the remaining positive examples and 50 examples from the negative set as shown in Table 1 and Fig. 5. 1000 prototypes (250 patches) \times (4 sizes) are used in the S2 layer.

7.2 Evaluations at the S1 Layer - Part 1

At this layer, the computational complexity and correct classification rates (accuracies) for each of the proposed approximations (**Approx**) are compared to the baseline model.

- **Approx1:** Combined Image-based HMAX using 2-D Gabor filters.
- **Approx2:** Combined Image-based HMAX using separable Gabor filters.

Interestingly, to avoid any confusion, the performances of the two approaches “Combined Image Based HMAX using Haar Wavelet Transform” and “Baseline model using Haar Wavelet Transform” are tested in an independent subsection (see Subsect. 7.3).

In order to compute the speed of the approximations at this layer, the total time complexity of the S1 layer is measured on a specific face image from the face category. All the evaluations were done on a core i7 2.4 GHZ machine. The simulations were repeated five times. Figure 6 illustrates an average of the results. It shows that both Approx1 and Approx2 are faster than the baseline (blue curve) for all the tested image sizes. It has been noticed that for an image of size between 100×100 and 160×160 , Approx1 is always faster than Approx2. For example, Approx1 is faster than Approx2 by 3.23 % for an image of size 100×100 .

For other image sizes greater than or equal to 160×160 , Approx2 always shows lower timing than Approx1. For example, for an image of size 160×160 , Approx1 is faster than the baseline by 2.95 % while by 3.42 % for Approx2.

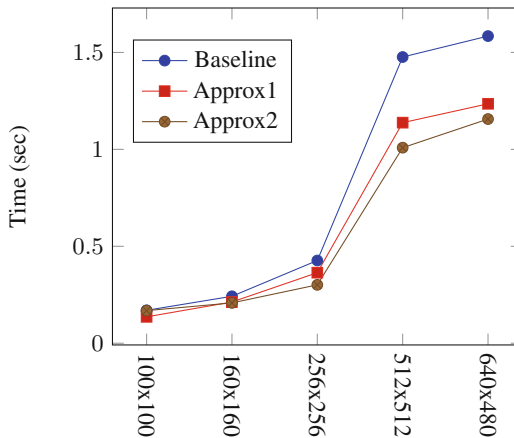


Fig. 6. Timing comparison (in sec).

Table 2. Classification accuracies of Approx1 approximation.

Approx1	Classifier	$c = 0.25$	$c = 0.75$
$\alpha = 0.25$	SVM	34.36 %	33.59 %
	NN	20.28 %	20.28
$\alpha = 0.5$	SVM	45.20 %	45.16 %
	NN	31.23 %	31.23
$\alpha = 0.75$	SVM	49.02 %	48.27 %
	NN	35.01 %	34.45 %
$\alpha = 1$	SVM	47.74 %	47.74 %
	NN	31.36 %	31.36 %
$\alpha = 1.25$	SVM	39.35 %	40.74
	NN	23.99 %	24.08 %

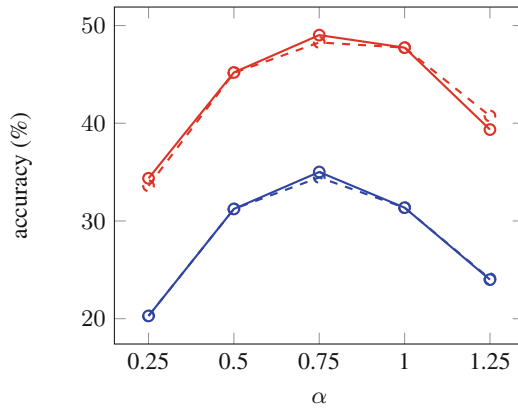


Fig. 7. Approx1 accuracies under different values of α and c (Color figure online).

For an image of size 256×256 , Approx1 is faster than the baseline by 6.29 % while by 12.56 % for Approx2.

In order to assess the correct classification rates, both SVM and NN classifiers were used. The average accuracies of Approx1 under different values of α and c are shown in Table 2. From all the following experiments, 2000 prototypes (500 patches) \times (4 sizes) are used and all the images were rescaled to 140 in height. Recall that C1 sub-sampling ranges do not overlap in scales and the prototypes are extracted from all the eight bands. The performance of the original model reaches 39 % and 21.2 % when using 30 training examples per class averaged over 3 repetitions under SVM and NN, respectively. Table 2 proves our significant contribution at the S1 layer especially for $\alpha = 0.75$ and $c = 0.25$ where the accuracy is increased by 10.02 % and 13.811 % using SVM and NN, respectively.

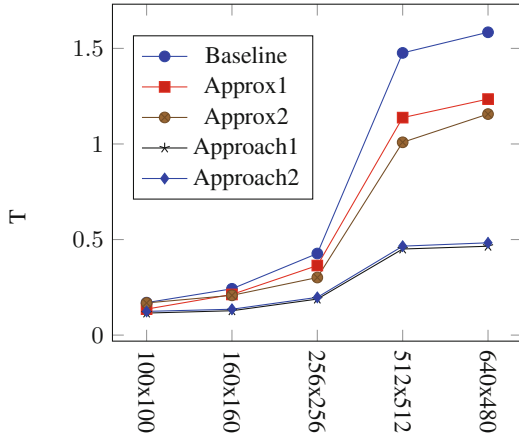


Fig. 8. Timing comparison (in sec).

Figure 7 illustrates the results shown in Table 2. It shows 4 different curves. The red and blue solid curves represent the accuracy values for $c = 0.25$ under SVM and NN, respectively.

While the red and blue dashed curves are for $c = 0.75$ under SVM and NN, respectively. Finally, the separability of Gabor filters is exploited and applied to the combined image with $\alpha = 0.75$ and $c = 0.25$. Approx2 shows an accuracy equal to 49.471% and 35.372% for SVM and NN, respectively.

7.3 Evaluations at the S1 Layer - Part 2

In this subsection, we aim to test the performance of the two approaches “Combined Image-Based HMAX using Haar Wavelet Transform” and “Baseline Model using Haar Wavelet Transform” in terms of speed and correct classification rates. To facilitate the notations, we name the first approach by “Approach1” while the second by “Approach2”.

- **Approach1:** Combined Image-Based HMAX using Haar Wavelet Transform.
- **Approach2:** Baseline Model using Haar Wavelet Transform.

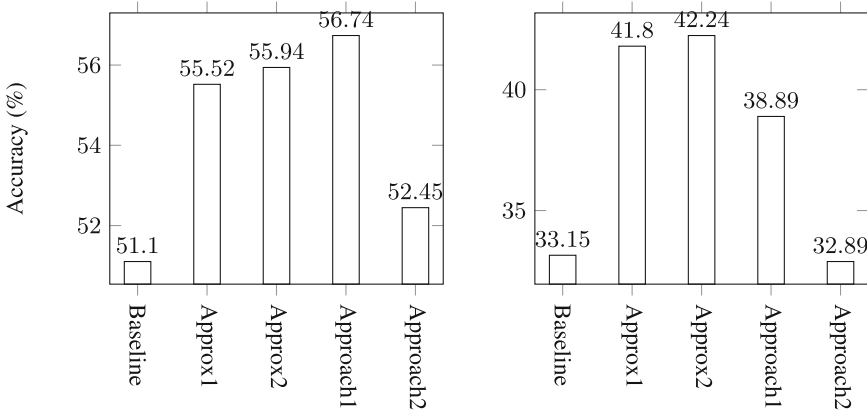
In order to compute the speed of the two approaches, we perform exactly the same computations as we did in Fig. 6. In other words, we just added the two approaches to Fig. 6 without any modification to the inputs of the model in order to get Fig. 8 that contains a total of 5 curves.

The simulations were also repeated five times and Fig. 8 illustrates the average of the results. It shows that both Approach1 and Approach2 are significantly faster than the baseline, Approx1 and Approx2 for all the tested images. From Fig. 8, we also notice that Approach1 is always approximately faster than Approach2. For more details about the numerical results of Fig. 8, refer to Table 3.

By normalizing the images to 140 pixels in height where the width is rescaled accordingly so that the image aspect ratio is preserved, some images become too

Table 3. Timing comparison (in sec) for all the proposed approximations at the S1 layer versus the baseline model.

Size	Baseline	Approx1	Approx2	Approach1	Approach2
100×100	0.170	0.136	0.168	0.115	0.124
160×160	0.242	0.213	0.208	0.128	0.135
256×256	0.427	0.364	0.301	0.189	0.198
512×512	1.476	1.138	1.009	0.450	0.465
640×480	1.584	1.235	1.156	0.465	0.483



(a) Classification accuracies using SVM classifier. (b) Classification accuracies using NN classifier.

Fig. 9. Classification accuracies.

small after applying Level1 Haar wavelet decomposition. Hence, choosing the prototypes from all bands becomes impossible. That is why, our contribution is to rescale only in this part all the images to 160×160 . As in Subsect. 7.2, the correct classification rates are measured in the same way. The performance of the original model reaches 51.1% and 33.1% when using 30 training examples per class averaged over 3 repetitions under SVM and NN, respectively. Figure 9 (a) shows that by using the SVM classifier, Approach1 is the best and that Approach2 is only better than the baseline by 1.35%. Approach2 is worse than (Approx1, Approx2 and Approach1) by (3.07%, 3.49% and 4.29%), respectively.

Fig. 9(b) shows that by using the NN classifier instead of SVM, Approx2 becomes the best and Approach2 the worst. Approx2 reaches 42.24% while Approach1 and Approach2 reach 38.89% and 32.89%, respectively.

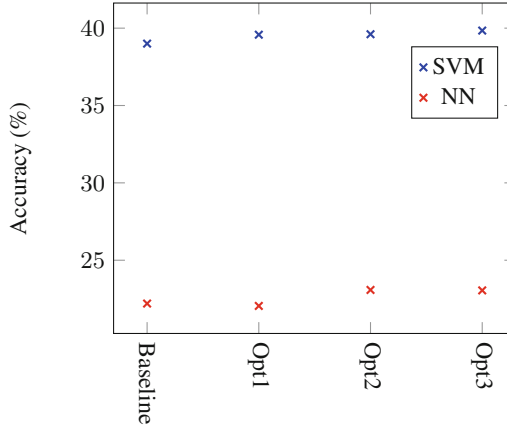


Fig. 10. Average accuracies of C1 approximations (Color figure online).

Table 4. Accuracy of S2 approximation.

Approximation	SVM
Baseline	39 %
PAM	39.68 % (+0.68 %)

7.4 Evaluations at the C1 Layer

Figure 10 shows the average accuracies of the SVM (blue x points) and NN (red x points) on several C1 optimization options (**Opt**). The accuracy of the model is increased a little bit when three cases of the additive method are applied. For example, using SVM, the accuracy is increased by 0.577 %, 0.607 %, 0.843 % on Opt1, Opt2 and Opt3, respectively. On the other hand, the increase is 0.846 %, 1.88 %, 1.85 % using NN.

- **Opt1:** Embedding all pixels ($\alpha = 1$).
- **Opt2:** [0 %, 2 %[, ($\alpha = 0.5$); [2 %, 5 %[, ($\alpha = 0.1$)
- **Opt3:** [0 %, 2 %[, ($\alpha = 1$); [2 %, 5 %[, ($\alpha = 0.5$)

7.5 Evaluations at the S2 Layer

Table 4 shows the average accuracy of the SVM classifier based on the S2 approximation.

This approximation has a big advantage on the model since the selected prototypes are non-redundant and generated in more intelligent way. Therefore, each prototype serves to slightly increase the accuracy. The accuracy of the model is increased approximately by 0.68 %.

7.6 Combined Classification Accuracies

Figure 11 shows the average accuracies of SVM on the combination of the approximations “Approx2” + “Opt3” + “PAM”. Our model shows an accuracy equal to 51 % when using only 2000 prototypes while it shows 53.8 % when using higher number of prototypes (4080) as used in [1, 8, 10, 11].

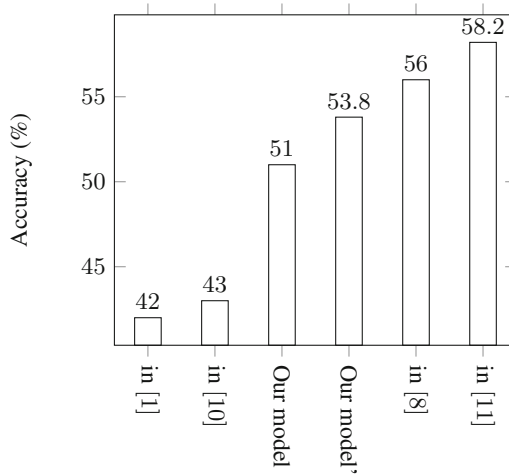


Fig. 11. The accuracies of the final models.

8 Discussion and Future Work

In this work, the complexity of all the five different layers of the original model of object recognition in the visual cortex, HMAX, is presented. Different approximations were added to the first three layers S1, C1 and S2.

The results have shown that removing all unimportant information such as illumination, expression variations and occlusions, to be a fruitful approach to improving performance. The idea behind separability of Gabor filters has been also exploited in order to be applied on the combined images generated after keeping only the important features for recognition. The change of the main concept at the C1 layer is further applied by exploiting the advantage of some of the minimum scales values and using them to be embedded into the extracted maximum scales values. The accuracy was slightly increased when the embedding process has been applied using the additive method. Our model serves also to always use an intelligent version of selected prototypes at the S2 layer in order to remove all the possibilities of having an unimportant prototype aiming to decrease the model’s accuracy.

As for future enhancements, a natural extension would be to adapt our work to the HMAX model in color mode. In addition, several new approximations will be applied and tested on more challenging databases.

References

1. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), pp. 994–1000 (2005b)
2. Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T.: A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. CBCL Paper #259/AI Memo #2005-036, Massachusetts Institute of Technology, Cambridge, MA (2005a)
3. Amayeh, G., Tavakkoli, A., Bebis, G.: Accurate and efficient computation of gabor features in real-time applications. In: Bebis, G., et al. (eds.) ISVC 2009, Part I. LNCS, vol. 5875, pp. 243–252. Springer, Heidelberg (2009)
4. Cadieu, C., Kouh, M., Riesenhuber, M., Poggio, T.: Shape representation in v4: Investigating position-specific tuning for boundary conformation with the standard model of object recognition. *J. Vis.* **5**(8), 671 (2005)
5. Bermudez-Contreras, E., Buxton, H., Spier, E.: Attention can improve a simple model for object recognition. *Image Vis. Comput.* **26**, 776–787 (2008)
6. Serre, T., Riesenhuber, M.: Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Massachusetts Institute of Technology, Cambridge, MA. CBCL, Paper 239/AI Memo 2004–017 (2004)
7. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortexlike mechanisms. In: IEEE Conference on Pattern Analysis and Machine Intelligence, vol. 29, pp. 411–426 (2007b)
8. Mutch, J., Lowe, D.G.: Multiclass object recognition with sparse, localized features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 11–18 (2006)
9. Chikkerur, S., Poggio, T.: Approximations in the hmax model. MIT-CSAIL-TR-2011-021, CBCL-298, p. 12 (2011)
10. Holub, A., Welling, M.: Exploiting unlabelled data for hybrid object classification. In: Advances in Neural Information Processing Systems (NIPS 2005) Workshop in Inter-Class Transfer (2005)
11. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1458–1465 (2005)
12. Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., Poggio, T.: A quantitative theory of immediate visual recognition. *Prog. Brain Res. Comput. Neurosci. Theor. Insights Brain Funct.* **165**, 33–56 (2007a)
13. Sharif, M., Anis, S., Raza, M., Mohsin, S.: Enhanced SVD based face recognition. *J. Appl. Comput. Sci. Math.* **12**, 49 (2012)
14. Kumar, P., Wasan, S.K.: Comparative study of k-means, pam and rough k-means algorithms using cancer datasets. In: Proceedings of CSIT: 2009 International Symposium on Computing, Communication, and Control (ISCCC) Singapore, 2011, pp. 136–140 (2011)
15. Crochiere, R., Webber, S., Flanagan, J.: Digital coding of speech in sub-bands. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 233–236 (1976)
16. Burt, P., Adelson, E.: The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **31**(4), 532–540 (1983)

17. Vetterli, M., Le Gall, D.: Perfect reconstruction FIR filter banks: Some properties and factorizations. *IEEE Trans. Acoust. Speech Sig. Process.* **37**(7), 1057–1071 (1989)
18. Hubel, D.H., Freeman, W.H.: *The Human Eye: Structure and Function*. Sinauer Associates, Sunderland (1999)
19. Oyster, C.W.: *Eye, Brain and Vision*. vol. 12(1), pp. 40–41 (1989)
20. Purves, D.: *Brains: How They Seem To Work*. FT Press, Upper Saddle River (2010)