Proceedings of the Fifth IASTED International Conference
VISUALIZATION, IMAGING, AND IMAGE PROCESSING
September 7-9, 2005, Benidorm, Spain

SIMPLE-SHAPE CLASSIFICATION BASED ON THE HUMAN VISUAL SYSTEM

Vassilios Vonikakis, Ioannis Andreadis and Antonios Gasteratos
Democritus University of Thrace
GR-671 00 Xanthi
Greece
{bbonik, iandread}@ee.duth.gr, agaster@pme.duth.gr

Abstract
This paper presents a new method suitable for shape classification, inspired by the early processing levels of the human visual system. It extracts a description for any simple 2-dimensional shape having a closed contour, regardless of its size, rotation and position, in affordable computational cost. The paper introduces a new approach to the modeling of the hypercolumns of the primary visual cortex, which requires significantly less computational burden and that is highly parallel. A new shape descriptor based on the relative angles of an object is also proposed. It produces close results for different shapes of the same object, it is proportion-flexible and it can identify distorted shapes correctly. Experimental results prove that the method is adequate for industrial production applications based on shape classification, as well as for shape-based image retrieval.

Key Words
Human Visual System, shape description, kernels.

1. Introduction

The Human Visual System (HVS) and generally the biological visual systems are far superior to artificial. They can discriminate among thousands of different shapes, colors, moves and textures in a variety of lighting conditions, ranging from extremely poor to ideal. More important, they provide already solutions to many partially, or totally unresolved problems that today's computer-vision science faces. For that reason it has been the centre of focus for many researchers, especially into the past decade.

Research in the field has been focused in two different directions. Firstly, there are models based on neuroscientific data that attempt to interpret the way that the HVS operates either in total or partially. These models are complicated simulations that intentionally reduce input space either by using low resolution images or by reducing the complexity of the input data [1-4]. Such models, although they give important clues about the HVS, they cannot be used in practical applications, and are more neuroscience-oriented. Typical models of this category are the Max model [5], which attempts to present an interpretation to the way that the HVS deals with the "binding problem" of the shapes. The CINNIC [2, 3] and RF-SLISSOM [6] attempt to present a solution to the integration of salient contours in the Primary Visual Cortex of the HVS.

On the other hand, other models exist that, though they are inspired by some attributes of the HVS, they concentrate adequately to performance as well as to neuroscientific aspect and, thus, they can be utilized into many computer vision applications. Such models are Fukushima's Neocogntitron [7], which deals with the shape binding problem and has already been used successfully in OCR techniques [8], and the SEEMORE [9], that focuses to object perception and was employed in image retrieval applications. Additionally, the various Retinex implementations [10, 11] can be placed into this kind of models. The Retinex is inspired by the way the HVS produces lightness records and color perception and was successfully used for image enhancement and restoration. Last, the Boundary Contour System (BCS) and Feature Contour System (FCS) that deal with shape perception, has been used in radar applications [12] and was implemented in analogue VLSI [13].

This paper proposes a new system for shape classification, belonging to the second category. Consequently, creating an accurate neuroscientific model of the HVS is out of our scope. However, the proposed model adopts many attractive characteristics of the visual pathways especially between the retina and the Primary Visual Cortex and suggests a solution for shape representation and the shape retrieval problem. Although many of the aforementioned models employ the same characteristics, they tend to achieve compatibility with the HVS rather than performance. The primary objective of the proposed model on the other hand is performance. A new simple approach for the usage of a new simple set of oriented

filters in the hypercolumns is presented. This is full parallel and it avoids complex convolutions with 2-dimensional oriented Gaussians or Gabor filters, though it exhibits the same good results. This makes the model in hand capable of handling images with resolution up to 1000×1000 pixels in real-time, when executed by a contemporary personal computer. Such a performance has not yet been reported by any of the aforementioned models, notwithstanding that in majority they process retinal images usually up to 200×200 pixels. Moreover, a shape description is employed, that relies on the angles created between the stimulated hypercolumns along the contour. This descriptor is shown to be both scale and rotation invariant, as well as immune to contour fluctuations. Furthermore, it is exhibited that it can be easily adjusted to a simple classifier, such as a conventional feed-forward neural network, for a successful classification of shapes. The paper is organized as follows: Section 2 provides a thorough description of the stages of the proposed method; section 3 presents experimental results and comparisons with existing methods; finally in section 4, conclusions, possible applications and improvements of the method are discussed.

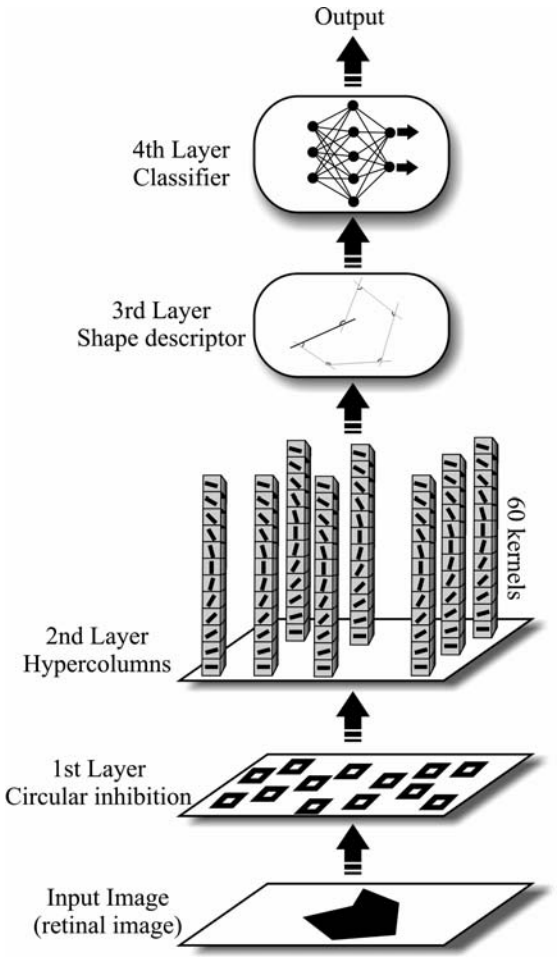

Figure 1: Overview of the 4 layers of the proposed method.

2. Description of the method

The proposed system consists of four processing levels. The first one, being the circular shaped inhibition layer, approximates the effects of ganglion and Lateral Geniculate Nucleus' (LGN) cells to the input retinal image. The second one is the hypercolumns of the Primary Visual Cortex, which processes the output of the previous layer, in accordance to the HVS. Next, the shape descriptor is applied to the output of the hypercolumns, producing a shape signature for a given contour. Last, the shape signature is fed to a neural network, which is the last level and classifies the shape to one of the output classes created during a training period. Figure 1 illustrates an overview of the method, with a clear distinction of the 4 processing layers. It is worth noting that layers 1 and 2 have a retinotopical organization, whereas layers 3 and 4 do not.

2.1 First Layer: Circular Inhibition

Circular inhibition in the HVS takes place in the ganglion and LGN cells. These cells have circular receptive fields with centre-surround opponency, meaning that they are either inhibited by their surround and excited by their centre, or excited by their surround and inhibited by their centers [14-16]. This kind of receptive fields is sensitive only to differences in intensity, which results to the extraction of edges in the retinal image.

Ganglion and LGN receptive fields are usually modeled as 2-dimensional Gaussians (see Figure 2a) over a large number of pixels, in order to create the circular receptive field and maintain the different grades of the Gaussian curve. In our method, we included the centre-surround opponency, avoiding however the use a 2-dimensional Gaussian. We particularly used a rough approximation of the Gaussian, as presented in Figure 2b, which can be implemented by a 3×3 pixel neighborhood.

In this approach, the central pixel represents the positive centre and the 8 surrounding pixels the negative surround. For simplicity we did not include cells with inhibitory centre and excitatory surround. The above approximation reduces significantly the computational burden of this layer. The output of this layer is the convolution of the input image with the following mask:

$$\frac{1}{8}\times\begin{bmatrix}-1 & -1 & -1\\ -1 & 8 & -1\\ -1 & -1 & -1\end{bmatrix}$$

Taking into account that the input is a grey-scale image, the output of every cell of the layer will be also a grey-scale image, with intensity values depending on its local contrast. Additionally, every cell of the layer corresponds to a single pixel of the input image, thus, maintaining the initial resolution of the image.

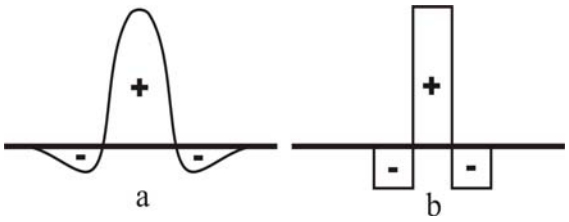

Figure 2: (a) The typical modelling of Ganglion and LGN cells and (b) the approximation used in the proposed method.

2.2 Second Layer: Hypercolumns

The next processing level of the HVS takes place in the Primary Visual Cortex. LGN cells are combined to form simple oriented cells, complex oriented and end-stopped or hypercomplex cells [14-16]. All these cells are organized in retinotopic columnar formations known as hypercolumns. Every hypercolumn processes a very small and specific part of the visual field, convolving it with all of its cells.

The simple oriented cells are modeled in the literature as 2-dimensional Gabor filters having specific orientations, which are then convolved with the input image [1-4, 6, 9, 12, 17]. Usually 12 different orientations are used, meaning that there is an oriented cell approximately every 15° [2, 3]. Convolving an image with 12 different 2-dimensional oriented Gabor filters, which comprise of a sinusoidal multiplied by a Gaussian, is computationally expensive and, consequently, many of the models adopting this approach avoid to process high resolution images.

Instead of using 2-dimensional oriented Gabor filters, we propose the use of a set of simpler oriented filters that are less computational intensive, and have approximately the same results. These filters, which will approximate the operation of the hypercolumns, are binary kernels with a straight oriented segment within their receptive field. Contrary to Gabor filters, they comprise only 2 values: a positive one, i.e. the oriented segment corresponding to the excitatory region and a negative one, i.e. the background corresponding to the inhibitory region. There are 12 different kernel groups, each one with a certain orientation: 0º, 15º, 30º, 45º, 60º, 75º, 90º, 105º, 120º, 135º, 150º and 165º. For every group, all the possible positions of the segment, within its receptive field, are included as different instances of the same orientation. This is to say that every group of kernels with the same orientation operates as a complex cortical cell, since it detects lines with a particular orientation, in every position of its receptive field. Key-role to the functionality of the set of kernels plays the ratio of the dimension of kernels, over the width of the excitatory segment. It directly affects the number of total kernels in the set, needed to include all possible positions of the excitatory segment. The larger the number of kernels in the set, the greater is the computational burden. After extensive experimental search, we concluded that the optimum size of the receptive field, that would result to minimum number of total kernels and at the same time would maintain adequate accuracy, should be 10×10 pixels, while the width of the oriented excitatory segment should be 2 pixels. This values result to a set of 60 kernels, divided into 12 different orientation groups. The complete set of kernels is presented in Table 1.

Group 1	0º	
Group 2	15º	
Group 3	30º	
Group 4	45º	
Group 5	60º	
Group 6	75º	
Group 7	90º	
Group 8	105º	
Group 9	120º	
Group 10	135º	
Group 11	150º	
Group 12	165º	

Table 1: The 60 kernel set utilized in the proposed method. Excitatory regions are white, whereas inhibitory regions are black.

The main advantage of these kernels is that they are used more as tiles than classical convolution kernels. Most models use only one kernel in each orientation for a particular spatial resolution. This kernel is shifted to all possible positions on the image and convolved with each one. Our "tilling" approach convolves all the kernels of the set with non-overlapping regions. This means that the image of the circular inhibition layer is divided into 10×10-pixel regions, which are non-overlapping and all kernels of the set are convolved with every region. This approach makes the detection of the orientation of any image region feasible, whilst it reduces the number of required convolutions significantly.

For an image of 1000×1000 pixels, the conventional convolution of 12 10×10 kernels would result to a total number of $12 \times (1000)^2 = 12{,}000{,}000$ convolutions (assuming zero-padding), whereas the tilling of 60 kernels in non-overlapping regions of a 1000×1000-picture, would result to $60 \times (1000/10)^2 = 600{,}000$. This proves that the tilling approach requires 20 times less convolution operations than the classical approach. Furthermore, there are no dependences in the analysis of every image region, which makes the tilling approach highly parallel.

Consequently, our approach is less computationally expensive, allowing the processing of higher resolution images.

Every hypercolumn contains all 60 kernels of the set and convolves them independently with only one image region. The convolution of a kernel with a particular image region gives the stimulation of this kernel in this exact region of the image. The kernel with the highest stimulation in every hypercolumn is the closest to the orientation of the region. A hypercolumn produces a valid output only if the kernel with the highest stimulation differs significantly from the second highest. This rule is used in order to ensure that the correct kernel is always selected. If the first kernel with the highest stimulation has a similar value to the one with the second highest, then the hypercolumn produces no output. The results of the tilling approach are depicted in Figure 3. Obviously, there are some inconsistencies when an edge is located exactly at the boundaries of two regions, but as it can be seen, the overall result is adequate.

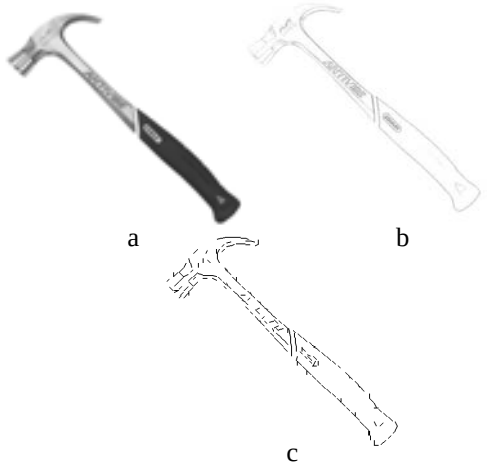

Figure 3: (a) The input image, (b) circular inhibition layer and (c) the output of the hypercolumns.

2.3 Third Layer: Shape descriptor

The outputs of the hypercolumns are input to the shape descriptor. The shape description is a chain of the angles formed by the excitatory segments of the kernels, when moving along the contour in a clockwise direction. To ensure correct angle calculations, only the kernels with a very high stimulation are used in the calculations. Its novelty is that no proportions are taken into account, but only angles. As it is demonstrated in the experimental results section, this is an important advantage for shape-based image retrieval, because it allows different and even abstract versions of the same object to be retrieved.

The angles that are recorded are the ones formed between the axes of the current kernel and the next kernel. When the formed angle is on the left of the current kernel, according to the direction of move, it is negative. Figure 4 shows the formation of a chain in a simple contour. The angles are calculated in terms of steps, which in our case are 15°. Consequently, an angle having value '-2' means "30° on the left".

In order for the description to be independent of the starting kernel, normalization is used. Let the description of object A be:

$$A: S_1, S_2, ..., S_M$$
$$\text{where } S_m \in \{-11,-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7,8,9,10,11\}$$
$$\text{and } m = 1...M$$

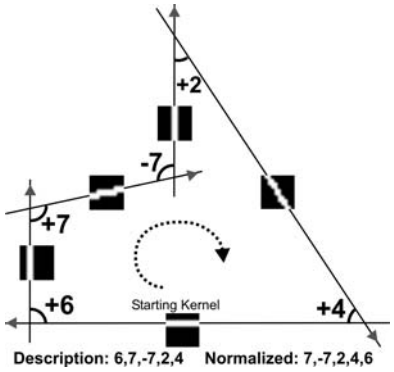

Figure 4: Description of a simple shape.

Then, chain A is scanned in pairs of nearby angles, in order to detect the greatest transition as shown by the following equation:

$$\max \left[\left| S_n - S_{n+1} \right| \right] \quad \text{where } n = 1...(M-1) \quad (1)$$

If the greatest transition occurs more than once, the same procedure is applied for the second highest transition. Once the starting point is redefined, the chain is shifted to it giving the final normalized description. It is important to mention that the normalized description is already rotation invariant, since only the relative angles and not the real ones are recorded. For the same reason, it is also scale invariant, since only angles and no proportions are recorded.

2.4 Fourth Layer: Classifier

The normalized shape description is in a form adequate for classification. A typical feed-forward neural network was used and it was trained in the back-propagation fashion. After extensive experimentation we concluded that one hidden layer is sufficient for shape retrieval applications, giving to the network the adequate generalization ability. The activation function of the neurons was selected to be the logarithmic sigmoid function, as it gives the better results. The number of neurons in the input layer must be at least the same with the longest shape description, because each angle of the shape description is entered to one neuron of the input layer. The output layer must have the same number of neurons with the number of possible classes. These neurons can have output values ranging in the interval [0,1], resulting to a membership function for every possible class.

3. Experimental results

A shape retrieval example demonstrating the capabilities of the proposed method follows. A neural network of one hidden layer with 20 neurons was used. The network was trained in order to identify four classes of different patterns. The training set is indicated in Table 2.

After the training of the neural network, various test shapes were assessed, in order to prove the classification capabilities of the system. Table 3 presents the output results of the neural network.

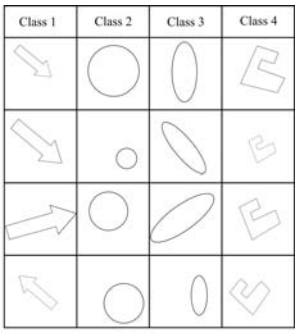

Table 2: The training sets for the four classes.

Test shapes	Classes			
	1	2	3	4
a	0.954	0	0.014	0.102
b	0.954	0	0.014	0.101
c	0.963	0	0.015	0.104
d	0.957	0	0.015	0.102
e	0.954	0	0.014	0.101
f	0	0.98	0.019	0.001
g	0.003	0.053	0.889	0
h	0.02	0.001	0.091	0.939
i	0.037	0.001	0.067	0.924

Table 3: Results of classification on 9 test shapes.

It is clear from the results that the method achieves rotation and scale invariance, since all test shapes differ in size, rotation or position in the image, compared to the training set. Of particular interest are the shapes a, b and c. They are different instances of the same object, but with a dissimilar shape. Many shape descriptors such as chain codes [18, 19], or shape numbers [20] would identify these shapes as different objects, because of their different length proportions. Our method identifies that these objects are the same, due to the fact that it is based only in relative angles. Particularly shape c, which is a freehand drawing with significant distortions in its contour, is identified accurately. The robustness of the proposed method is also demonstrated with shapes d and e, which are two distorted versions of the same training object. Shape d is compressed to the 0.65 of the original width while e is compressed to the 0.65 of the original length. As it can be seen, both shapes are classified correctly. The same conclusions are also visible in shapes h and i. Shape h is stretched in the x axis, while compressed in the y axis. The classification seems not to be affected by warping. Shape i, has many non uniform distortions, resulting to a total change of the original relative angles, and yet classified correctly. Shape g, is an ellipse with higher eccentricity than those which were used in the training set and it is also successfully classified.

The proposed method was implemented in C code and executed by an Intel Celeron Processor, running at 1 GHz with 512MB RAM, under Windows XP. The typical execution time for an image of 1000×1000 pixels, is 0.5 seconds and for images of 700×700 pixels the execution time is 0.2 seconds.

4. Discussion and Conclusions

The experimental results demonstrate that the proposed method is invariant in size and rotation changes. Furthermore it is proven that it can tolerate significant distortions in the contour of the objects (d, e, g, h), while it maintains correct classification ratios. More importantly, it is not bound by length proportions and thus can identify different instances of the same object, in various representations (a, b, c, i).

Low computational demands of the method, allow the manipulation of high resolution images even with an average personal computer. Clearly the most computationally intensive layer of the method is the second one (i.e. the hypercolumns), where convolution with multiple kernels occurs. Since this level is highly parallel, a parallel execution of this level would minimize much more the overall execution time. The low computational demands in addition to its tolerance to contour distortions make the proposed method adequate for shape retrieval purposes.

The most significant drawback of the method is that it can handle only closed contours. Cross like shapes, for example, constructed by a very thin line do not have a closed contour and thus cannot be identified correctly by the present method. A possible solution to this problem could be a pre-processing of the original shape by a dilation operation, which would thicken the lines. That would result to the appearance of a perimeter in the circular inhibition level, and thus of a closed contour. Also the use of 10×10 pixel kernels, prevents features of smaller size to be extracted. Although this is an important drawback in some cases, it stands also as an advantage in others. It particularly prevents contour fluctuations introduced by the pixel grid, in cases of rotations or noise. These fluctuations are a problem to shape descriptors that focus on the pixel level, such as chain codes or shape numbers [18-20].

The target application of the proposed method is shape classification for industrial production systems. In those cases, objects on conveyor belts are selected and categorized by robots mainly according to their shape. The proposed method is appropriate for this application since its low computational cost permits real-time execution by a personal computer. Furthermore, the drawbacks that have been mentioned do not impose considerable problems for the particular application, since objects always have a close contour, are located in a non-cluttered background and the resolution of the processed image is adequate for the use of 10×10 pixel kernels.

Improvements could include the use of more characteristics of the HVS. Layer 3 of the shape descriptor could be replaced by a more human-based approach, such as silent contour integration by lateral connections in the primary visual cortex, extraction of more complicated features such as corners and junctions and their invariant binding by higher visual areas, using a Hebbian learning rule.

References

[1] M. C. M. Elliffe, E. T. Rolls, S. M. Stinger, Invariant recognition of feature combinations in the visual system, *Biological Cybernetics*, vol. 86, 2002, pp. 59-71.

[2] T. N. Mundhenk, L. Itti, CINNIC: A new computational algorithm for modelling of early visual contour integration in humans, *Neurocomputing*, vol. 52-54, June 2003, pp. 599-604.

[3] T. N. Mundhenk, L. Itti, A model of contour integration in early visual cortex, In: *Second International Workshop on Biologically Motivated Computer Vision*, 2002, pp. 80-89.

[4] S. M. Stringer, E. T. Rolls, Position invariant recognition in the visual system with cluttered environments, *Neural Networks*, vol. 13, 2000, pp. 305-315.

[5] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, *Nature Neuroscience*, vol. 2, 1999, pp. 1019-1025.

[6] Y. Choe, R. Miikkulainen, Contour Integration and Segmentation with Self-Organized Lateral Connections, In: *Technical Report AI-00-286*, 2000.

[7] K. Fukushima, Neocognitron: a hierarchical neural network capable of visual pattern recognition, *Neural Networks*, vol. 2, 1988, pp. 119-130.

[8] K. Fukushima, K. Nagahara, H Shouno, M. Okada, Training neocognitron to recognize handwritten digits in the real world, In: *WCNN'96(World Congress on Neural Networks, San Diego, CA)*, 1996, pp. 21.

[9] B. W. Mel, SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition, *Neural Computation*, vol. 9, 1997, pp. 777-804.

[10] D. J. Jobson, Z. Rahman, G. A. Woodell, A multiscale retinex for bridging the gap between color images and the human observation of scenes, *IEEE Transactions on Image Processing*, vol. 6, 1997, pp. 965-976.

[11] A. Moore, J. Allman, R. M. Goodman, A real-time neural system for color constancy, *IEEE Transactions on Neural Networks*, vol. 2, 1991, pp. 237-247.

[12] S. Grossberg, E. Mingolla, J. Williamson, Synthetic aperture radar processing by a multiple scale neural system for boundary and surface representation, *Neural Networks*, vol. 8, 1995, pp. 1005-1028.

[13] G. Cauwenberghs, J. Waskiewicz, Analog VLSI Cellular Implementation of the Boundary Contour System, In: *Conference on Advances in Neural Information Processing Systems II*, 1998, pp. 657-663.

[14] D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cats visual cortex, *Journal of Physiology*, vol. 160, 1962, pp. 106-154.

[15] D. H. Hubel, T. N. Wiesel, Receptive fields and functional architecture in nonstriate areas (188 and 19) of the cat, *Journal of Neurophysiology*, vol. 28, 1965, pp. 229-289.

[16] D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture of monkey striate cortex, *Journal of Physiology*, vol. 195, 1968, pp. 215-243.

[17] R. P. Wurtz, T. Lourens, Corner detection in color images through a multiscale combination of end-stopped cortical cells", *Image and Vision Computing*, vol. 18, 2000, pp. 531-541.

[18] E. Bribiesca, A new chain code, *Pattern Recognition*, vol. 32, 1999, pp. 235-251.

[19] S.Loncaric, A survey of shape analysis techniques, *Pattern Recognition*, vol. 31, 1998, pp. 983-100.

[20] E. Bribiesca, A. Guzman, How to describe pure form and how to measure differences in shapes using shape numbers, *Pattern Recognition*, vol. 12, 1980, pp. 101-112.