Structural Erosion Layer: CNN-Based Binary Object Feature Extraction

Journal:	*IEEE Transactions on Neural Networks and Learning Systems*
Manuscript ID	TNNLS-2023-P-26879
Manuscript Type:	Regular Paper
Date Submitted by the Author:	13-Mar-2023
Complete List of Authors:	KUMAR, RATNESH; University of Kalyani, comuter science & engineering
Keywords:	Erosion, Dilation, Min filter, Roto-Translation, Region of Interest, ReLU, Binary Object

Structural Erosion Layer: CNN-Based Binary Object Feature Extraction

Ratnesh Kumar , Kalyani Mali,

***Abstract*—This paper addresses the design, implementation, and tools for convolutional neural networks (CNN). The CNN may give novel guidelines for preparing and extracting geometric transformation invariant binary object shape descriptors. Herein, we have developed a model based on the CNN for binary shape classification and recognition by incorporating the sequence of successive convolution layers and structural erosion layers onto the raw pixels of an input binary object. In the structural erosion layer, we have used the idea of a self-contained, concise, and comprehensive definition in place of the sub-sampling layer in the existing CNN model for roto-translation equivariant networks. The basic tools such as dilation and erosion in the language of mathematical morphology can be defined by the maximum or minimum value of the local object pixels also called MAX and MIN filters. The structural erosion layer preserves the shape complexity and simultaneously reduces the size of an object. Therefore, in our CNN model, we have used structural erosion in place of the pooling or sub-sampling layer. The sub-sampling layer in the existing CNN model generates lossy information from object pixels. Although our aim is to efficiently obtain lossless shape-preserving object descriptions, this task can be implemented by modifying the layer of the existing CNN model.**

***Index Terms*—Erosion, Dilation, Min filter, Roto-Translation, Region of Interest (ROI), ReLU, Binary Object.**

I. Introduction

THE Computer vision in image processing techniques consists of different types of problems such as segmentation, object detection, pixel localization, and image classification. Among those, binary image shape classification and recognition can be considered the most fundamental problems in computer vision and image processing techniques. Because, the shapes do not have brightness, colour and texture information and are only represented by their silhouettes. Prior to the development of convolutional neural networks (CNN), a subset of artificial neural networks (ANN), eminent researchers relied solely on traditional machine learning approaches for shape classification and recognition [1, 2]. But the scope of classification tasks and their accuracy depends on several challenges, such as information loss during sampling of the image pixels. The results of the sampling in extracted descriptors lose the information from the shape deformation, articulation and occlusion points. Another challenge in binary shape classification and object recognition is the extraction of geometric transformation invariance shape features. The diversity and richness of the early pattern recognition convolutional neural network model for pattern descriptors suffer from these challenges.

Ratnesh Kumar, Dept. of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia-741235, West Bengal, India, Email: rkratneshkumar@gmail.com

Kalyani Mali, Dept. of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia-741235, West Bengal, India, Email: kalyanimali1992@gmail.com

Convolutional Neural Network (CNN or ConvNet) is a special type of multi-layer neural network inspired by the mechanism and functional architecture of the monkey striate cortex [6]. The cortex is viewed as a system that is uniquely arranged both vertically and horizontally in the smaller receptive field. This is the inspirational model for CNN. LeCun et al. deployed the LeNet-5 [5] CNN model in 1998 to classify visual patterns utilizing their raw pixels rather than any additional feature engineering process. The CNN-based model performs remarkably well as a result of advancements in object classification and recognition research. A large deep convolution neural network deployed by Krizhevsky et al. in 2012, known as AlexNet [7], demonstrated excellent performance on the ILSVRC: ImageNet Large Scale Visual Recognition Challenge [8]. The popularity of AlexNet has served as inspiration for many CNN models such as ZFNet, VGGNet, GoogleNet, ResNET, DenseNet etc, a detailed explanation is given in the relevant article or in the review paper [9]. In all the deep CNN architectures for image recognition, the most effective tools in subsequent layers are rectified linear units (ReLU) non-linearity [10], and subsampling or pooling layers [11, 12, 13]. The rectified linear units preserve information about relative intensities as the information travels through multiple layers of feature detectors. However, the subsampling or pooling layer reduces the dimensions of an input image for reliable, robust, comparable, and efficient computational processes in the next layers. Simultaneously, pooling layers in CNNs summarize the outputs of neighbouring groups of pixels in the same kernel map. And at each subsequent layer, it lowers the size of the input image and collects the statistics of object pixels as well as background pixels. Besides the roles of the ReLU and the pooling layer in the attractive qualities of deep CNN models, the input layer and flattening layer also play an important role for Roto-Translation image datasets [3, 4]. In almost all existing deep CNN models, the preprocessing steps pre-define the size of the input image for the next layers. And it could be required to resize by scale-up or scale-down the input image. This is one of the major issues to pre-define the size of an input image by preserving its shape structure. Therefore,

numerous unsupervised strategies have been used in recent deep learning research for object recognition challenges, which has minimized the requirement for labelled samples [14 - 20]. Despite the existing deep CNN model, we have required a Translation, Rotation, and Scale (TRS) invariant CNN model that minimizes the loss of original attributes or properties from our image descriptor.

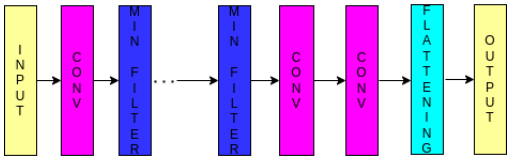


Figure 1. Our proposed model

Consequently, we have deployed a typical deep CNN model using the MIN filter, the essence of the suggested model is depicted in figure 1. The initial section of this paper will concentrate on shape analysis and recognition for binary image datasets before moving on to grey-level datasets. In the binary image, we have assumed the white pixels are the foreground or object pixels and the black pixels are the background pixels. Additionally, our area of interest for shape classification and recognition lies in the foreground pixels of the binary image. As a result, our proposed CNN model using the MIN filter concentrates only on the non-zero pixels in an image. The application of the MIN filter on the binary objects reduces the size of the region of interest (ROI), this operation is also called erosion and simultaneously preserves the shape-size complexity of an object. Although, when the 3x3 box filters are used for extracting the minimum value with stride one is equivalent to pooling so, the MIN filter can also be explained by a pooling operation. Therefore, the MAX pooling or MIN pooling is also represented by the MAX filter or the MIN filter, respectively. The architecture and motive of our proposed CNN model are different from the existing CNN model. Here, we have introduced a structural erosion layer that has been constructed using the convolution operation, followed by the MIN filter's sequence. The aim of this layer is to scale down the object contour by preserving its structure in the image for efficient shape descriptors. The final two convolutional layers expand the numerical value within an object's ROI to improve matching efficiency and discrimination. The functional framework of the flattening layer depends on the chosen supervised or unsupervised classifier's. Along with the recommended model, we have also developed a convolution kernel and a more generalized version of the formulation of the rectified linear units (ReLU) activation function. Applications of the described ReLU serve as a key component in several image processing methods, such as segmentation, classification, and others, which are illustrated in the following sections.

The rest of the paper is structured as follows. The various operations in basic mathematical morphology using MIN filters are presented in detail in Section II, the proposed deep CNN architecture for image recognition is presented in Section III, the results and discussions are presented in Section IV, and the conclusions and future work are presented in Section V.

II. EROSION AND DILATION

The elements in set theory, which represent the shape points in two-dimensional Euclidean space, provide a substructure for the concepts of mathematical morphology. Mathematical morphology is a branch of set theory and the best tool for the manifestation of binary objects or gray-tone images. The details of different types of mathematical morphology stuff that are supported by a set theory in 2-D Euclidean space are discussed in [1] and [2]. Erosion and dilation are the two basic types of operations in mathematical morphology. The other mathematical morphology operations are derived from the basic ideas of erosion and dilation. The derivations and explanations of basic operations in mathematical morphology consist of pixels that can have one of exactly two colors, usually black and white pixels. Herein, based on the intensity of colors, the eminent researchers have defined the pattern of white pixels over black pixels. The values of white pixels are represented by one and are called the foreground pixels. The values of black pixels are represented by zeros and are called background pixels. These types of images are called binary images. Hence, the erosion and dilation operations on these types of images can be defined by using local minima and local maxima. However, the local minima and local maxima over the binary images or graytone images have equivalent capabilities as the so-called min and max filters. Whichever, we have used the local minima or maxima in our model. A more detailed discussion of erosion and dilation using min and max filters are discussed in the subsequent subsection.

A. EROSION

Erosion is a morphological transformation that combines two sets using the vector subtraction of set elements. The following is the conventional definition of erosion in the language of mathematical morphology that is based on set theory and covered in [2]. With f and B as sets in Z^2, the erosion of f by B, denoted $f \ominus B$, is defined as follows:

$$f \ominus B = \{z|(B)_z \subseteq A\} \tag{1}$$

In another way, we can express erosion in following equivalent form:

$$f \ominus B = \{z|(B)_z \cap A^c = \emptyset\} \tag{2}$$

The expression in equations (1) and (2) can also be expressed in the following equivalent form that is discussed in [1].

$$f \ominus B = \min_{(i,j)\in B}\{f(x-i),(y-j)\} \tag{3}$$

f=

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

$\ominus$ B=

1	1	1
1	1	1
1	1	1

$\xrightarrow[\min_{(i,j)\in B}\{f(x-i,y-j)\}]{f\ominus B=}$

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 2. $(f \ominus B) = \min\{f(i,j), f(i,j-1), f(i,j+1), f(i-1,j), f(i-1,j+1), f(i-1,j-1), f(i+1,j), f(i+1,j-1), f(i+1,j+1)\}$

In the following discussion in equation (3), set B is assumed to be a structuring element and set f is assumed to be binary object. Herein, we have apply the erosion operation on f by B using the local minima. Figure 2, represents the results of an operation $f \ominus B$ on binary 8x8 square shape using $f(x,y) = \min\{f(i,j), f(i,j-1), f(i,j+1), f(i-1,j), f(i-1,j+1), f(i-1,j-1), f(i+1,j), f(i+1,j-1), f(i+1,j+1)\}$. The application of a MIN filter over the binary image is equivalent to erosion by a 3x3 box filter. The implementation of erosion operation can also be carried out by any kernel map. This technique may be insisted for the development of the deep CNN models.

We shall henceforth discuss the application of the MIN filter for grey-tone images. The experimental results, using the MIN filter, have been consisting of different types of analysis for grey-tone image segmentation in noise-less or noisy environments. Figure 3(a) represents the noise-less woman's face, having a size of 512x512. The outcome of the application of the MIN filter in figure 3(a) are shown in figure 3(b). Now, we have used the arithmetic operation between figure 3(c) and figure 3(d), and the results of the subtraction operator, applied between images, are shown in figure 3(e). Herein, figure 3(c) represents figure 3(a) and figure 3(d) represents figure 3(b). Figure 3(f) represents the noisy picture of the cameraman, having a size of 512x512. The outcome of the application of the MIN filter in figure 3(f) is depicted in figure 3(g). And the results of the subtraction operator (figure 3(f) - figure 3(g)), applied between images, are shown in figure 3(h). Now, we have applied the MIN filter in figure 3(g), and the result is shown in figure 3(i). And at last, the results of the subtraction operator (figure 3(g) - figure 3(i)), applied between images, are shown in figure 3(j). The use of the MIN filter on grey-tone images is supported by the aforementioned justification, and the outcomes are shown in figure 3. The MIN filter is used for the segmentation as well as to reduce the noise from

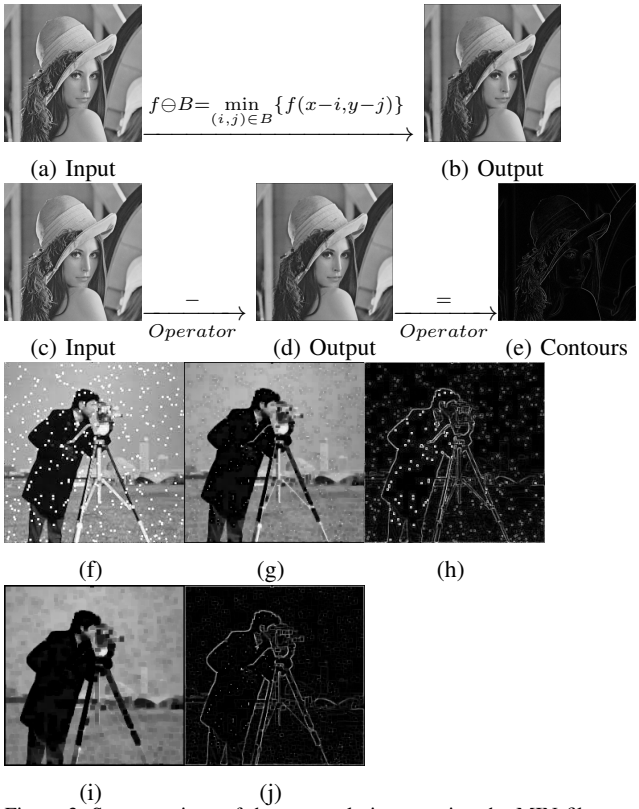


Figure 3. Segmentations of the greyscale image using the MIN filter.

corrupted images. The segmentation of noisy images using the MIN filter technique supports our proposed deep CNN model. Now, the MIN filter can be used to remove out-layer object pixels or noisy contour pixels from a binary image. Because the single pixels close to an object contour or the multiple pixels in out-layers do not constitute an object. In general, the binary object shape descriptors are affected by the out-layer pixels or the noisy contour pixels. These justifications set the stage for the introduction of the structural erosion layer, which has the ability to preserve the shape structure and simultaneously remove the noise from the region of interest.

B. DILATION

Dilation is the morphological dual to erosion, which combines the two sets using vector addition of set elements. The following is the conventional definition of dilation in the language of mathematical morphology that is based on set theory and covered in [1, 2]. If f and B as sets in Z^2, the dilation of f by B, denoted $f \oplus B$, is defined as follows:

$$f \oplus B = \{z|(\hat{B})_z \cap A \neq \emptyset\} \tag{4}$$

In another way, we can express erosion in following equivalent form:

$$f \oplus B = \{z|[(\hat{B})_z \cap A] \subseteq A\} \tag{5}$$

The expression in equations (4) and (5) can also be expressed in the following equivalent form that is discussed in [1].

$$f \oplus B = \max_{(i,j)\in B}\{f(x+i),(y+j)\} \tag{6}$$

f=

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

$\oplus$ B=

1	1	1
1	1	1
1	1	1

$\xrightarrow[\max_{(i,j)\in B}\{f(x+i,y+j)\}]{f\oplus B=}$

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

Figure 4. $(f \ominus B) = \max\{f(i,j), f(i,j-1), f(i,j+1), f(i-1,j), f(i-1,j+1), f(i-1,j-1), f(i+1,j), f(i+1,j-1), f(i+1,j+1)\}$

In the following discussion in equation (6), set B is assumed to be a structuring element and set f is assumed to be binary object. Herein, we have apply the dilation operation on f by B using the local maxima. Figure 4, represents the results of an operation $f \oplus B$ on binary 8x8 square shape using $f(x,y) = \max\{f(i,j), f(i,j-1), f(i,j+1), f(i-1,j), f(i-1,j+1), f(i-1,j-1), f(i+1,j), f(i+1,j-1), f(i+1,j+1)\}$. The application of a MAX filter over the binary image is equivalent to dilation by a 3x3 box filter. The implementation of dilation operation can also be carried out by any kernel map. This technique may be insisted for the development of the deep CNN models.

C. Opening and Closing

The essential use of erosion and dilation motivates the significance of describing the mathematical morphological operations, opening and closing. The opening of image f by structuring element B is denoted by $f \circ B$ and is defined as follows:

$$f \circ B = \max_{(i,j)\in B}\{\min_{(i,j)\in B}\{f(x+i),(y+j)\}\} \tag{7}$$

The closing of image f by structuring element B is denoted by $f \bullet B$ and is defined as follows:

$$f \bullet B = \min_{(i,j)\in B}\{\max_{(i,j)\in B}\{f(x+i),(y+j)\}\} \tag{8}$$

Equations (7) and (8) describes the opening and closing in terms of the MIN filter and MAX filter. This technique may be insisted for the development of the deep CNN models.

III. Proposed deep CNN model

The ability of the existing deep convolutional neural network architecture for image recognition tasks, mainly, consists of two parts: 1) convolutional Networks, and 2) a non-trainable or trainable classifier (Neural Networks). The first part of the architecture described the convolutional networks. Functionally, the convolutional network has been categorized in two ways: 1) trainable parameter convolution network, and 2) non-trainable parameter convolution network. In general the convolutional networks also called image descriptor layer, that ensure the three types of architectural ideas for translation, scale, distortion, and rotation invariance in shape classification model: 1) local receptive fields, 2) spatial or temporal subsampling, and 3) Shared weights. The idea of connecting units to local receptive fields comes from Hubel and Wiesel's discovery of locally sensitive, orientation-selective neurons in the cat's visual system [21]. The cortex is viewed as a system that is organized vertically and horizontally in completely different ways, according to another research paper by Hubel and Wiesel's [6]. Inspired by the way of cortex organization, we have constructed different types of kernel maps (receptive fields) for image descriptors that have identical weights. The concept of non-trainable and identical-weights feature map descriptors for image classifiers is shown in figure 5. The coefficient of non-trainable parameters that have been used in our proposed work is different from the currently used complete convolutional layers, which are made up of many feature maps, with trainable bias and trainable coefficients in each feature map. On the basis of our intuition, the convolution network can also be called the image descriptor layer. In this layer, we have implemented the activation, and pooling by the generalized ReLUs activation function that has been discussed in section III(B).

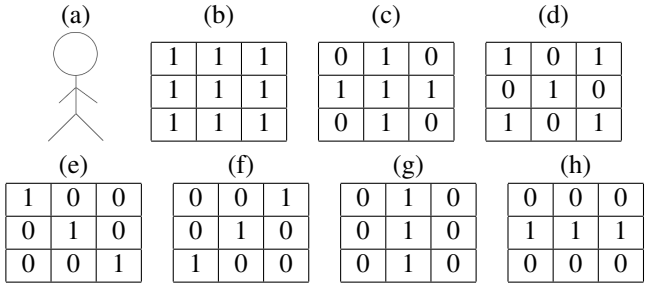


Figure 5. The various types of kernel functions (receptive field).

The other's part of the existing CNN architecture described the trainable or non-trainable classifier, which categorizes the resulting feature vectors into classes. In the trainable classifier, standard, fully connected multilayer networks can be used as classifiers.

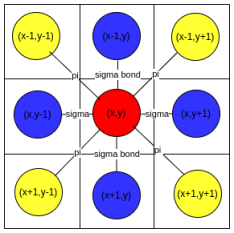
Figure 5(a) represents the single individual, the structure of the individual constitutes the two forward slashes, two backward slashes, a vertical line, and a circle. The computer can recognize the backward slash by kernel map shown in figure 5(e), forward slash by kernel map shown in figure

5(f), vertical line by kernel map shown in figure 5(g), and the circle by the kernel map shown in figure 5(c). The computer can also recognize the structure of the cross sign, and the horizontal line by the kernel map shown in figures 5(d), and 5(h) respectively. Whichever, the kernel map identifies the structure of a single individual. But, to deploy our model to rotation invariant, a single equivalent kernel function will be required for the structure like the lines, shown in figure 5 (e-h).

A. Kernel Map Using Lattice

In this section, we have discussed the rotation invariant kernel function that recognizes the boundary structure of an object. This kernel map is also known as convolutional filters that are capable of finding features of roto-translation images. The authors in [14,15,16], mapped the image pixels with the atoms of crystalline solid-state substances, also known as lattices. The central atom with neighbour atoms in a lattice constitutes the sigma and pi bond. In the 2-D lattice, the boundary atom overlaps with the central atom in two possible ways: face-on or diagonally. Intuitively, the face overlap is stronger than the diagonal overlap. In figure 6 (a), we have been showing a 3x3 two-dimensional lattice consisting of 4-faced overlapping represented by blue and 4-diagonal overlapping represented by yellow, with the centre atom represented by red. Herein, we have mapped the sigma bond by the face pixels and the pi bond by diagonal pixels, and we have related the bond energy of respective pixels by a relation that is inversely proportional to the euclidean distance from the centre pixels. As a result, the larger the distance between the centre pixel and the boundary pixel, the shape structure becomes less stable.

(a) (b)



$2+\frac{1}{\sqrt{2}}$	$3+\frac{2}{\sqrt{2}}$	$2+\frac{1}{\sqrt{2}}$
$3+\frac{2}{\sqrt{2}}$	$4+\frac{4}{\sqrt{2}}$	$3+\frac{2}{\sqrt{2}}$
$2+\frac{1}{\sqrt{2}}$	$3+\frac{2}{\sqrt{2}}$	$2+\frac{1}{\sqrt{2}}$

Figure 6. The structure of atoms and their kernel map in a 2-D lattice.

Figure 6(a), shows the face and diagonal pixels and their arrangement in terms of the 2-D lattices. For each pixel in a 3x3 image, we computed the kernel map in terms of the energy needed to break the sigma and pi connection, as shown in figure 6(b). This has directly proportional to the sum of the inverse of the euclidean distance of the neighbouring pixels. The kernel map corresponding to blue pixels constructs the circle shape as shown in figure 5(c) and the kernel map corresponding to yellow pixels constructs the cross shape as shown in figure 5(d). This description of the kernel using the proportionality of the pixel's bond dissociation energy also maintains the non-linearity in the formulation of the generalized function for the rectified linear units (ReLUs), that have been discussed in the subsequent section.

B. Generalization of ReLU Nonlinearity

Rectified linear units (ReLUs), allow each unit to express more information for an unsupervised learning algorithm that builds a non-linear generative model for pairs of face images from the same individual [10, 22]. The non-linearity $f(x) = max(0, x)$ refers to neurons, with this non-linearity as Rectified Linear Units (ReLUs). In deep CNN models, the application of ReLU on non-trainable parameters avoids the negative results for each neuron. The generalization of ReLU function is described as follows:

$$f(x) = max\{0, min(x_1, x_2, ..., x_n)\} \tag{9}$$

Equation (9) summarizes the n-neighbouring neurons with respect to the kernel map shown in figure 5. The result of the MIN value of n-adjacent neighbouring neurons for non-negative and non-trainable parameters implements the ReLU activation function and simultaneously reduces the size of object pixels by preserving the structure of the contour. In this paper, we have divided the deep CNN architectures into two parts: the convolution network and classifier networks. The generalized ReLU function has been used in convolution networks for extracting translation, rotation and scale (TRS) invariant binary shape feature descriptors.

C. The Architecture

This subsection describes the proposed architecture for the convolutional network used in the experiments in more detail. The architecture of our convolution network is summarized in figure 7. Here, we have divided the deep CNN architecture into three parts: 1) the Structural erosion layer, 2) two Convolution Layers, and 3) the classifier's. Convolution is the first operation of the structural erosion layer, followed by several MIN filters. This is also a similar composition to the existing two layers architecture: the first layer does a convolution of the input image with a filter and the second layer down-samples the result of the first layer, using the MIN filters operation in places of the pooling. Convolution in the structural erosion layer is crucial for deploying the structure of the input image by computing the statistics of the neighbouring pixels. The input of the proposed deep CNN architecture is an m x n pixel image. The architecture does not depends on dimension of the image. In the case of binary shape classification, it reduces the size of the "region of interest" (object pixels) by preserving the image dimension and shape complexity.

Now, we will be discussing the view of architecture and the working principle of each layer in the proposed deep CNN. The architecture takes input as an image having dimensions m x n, and after the convolution operation on the input image by a kernel successively applies the MIN filters, and we have been naming these layers in combined called structure erosion layers. The Generalization of ReLUs

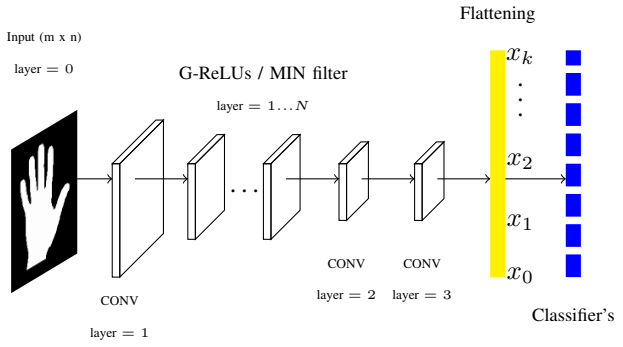


Figure 7. The proposed model: Structural Erosion, followed by the Convolutional layer

(G-ReLUs) or MIN filter layer $1...N$ shown in architecture is a variable that depends on our region of interest. In the case of, a bigger region of the object requires a more number of MIN filter operations (N) compared to the smaller region of an identical object in an image for the scale-invariant image recognition model. The results of the structural erosion layer are used for the input of convolutional layers 2 and 3. Herein, the convolution layers 2 and 3 stretch the range of the numerical value of the shrunk region of an object pixel. This stretched integer over the region of interest in a binary image tolerates the geometric differences of objects from the same class. At the same time, it gives better discriminative power to differentiate between different shape categories. Although, the framework of the flattening layer plays a crucial role in the rotational invariant feature as input for the classifier's. In this proposed model, the modelling of the flattening layer depends on the working principle of the classifier's. Considering, the situation for measuring the distance between histograms of two shapes. In this case, we have used the non-zero integer value in the shrunk regions of an object as user-defined bins in increasing order. Another situation for the working principle of the flattening layer is to develop the co-occurrence matrix classifier corresponding to the outputs of the preceding layer, that's quantified the non-homogeneity surrounding the shape boundary. The non-zero integer values corresponding to the output of convolution layer 3 have efficient power for constructing the grey-level co-occurrence matrix (GLCM) [28]. There are several classifier's that have been discussed in the subsequent subsection and their corresponding flattening layer models.

D. The Classifier's

Another crucial stage in the recognition of the object pixel's structures is the data classifier, after the feature's extraction by the convolutional neural networks from the raw binary input shape. In this section, we will be discussing the varieties of unsupervised and supervised classifiers used in our model after the flattening layer for measuring the superiority of one over others. The matching through the principal component analysis (PCA) [23], Hu's seven moments [29], dynamic programming with shape context [24][25], bipartite graph matching [24], measuring the distance between histograms (Bag-of-Word (BoW), local and global histogram) [26], Earth Mover's Distance classifier [27], statistical analysis of pattern spectrum using mathematical morphology (smoothness, roughness, energy and entropy of a boundary) [1], and others similarity (Cosine, Dice, Jaccard, Overlap Similarity) and dissimilarity (Euclidean distance, Minkowski Distance, Mahalanobis Distance, Manhattan distance) measures framework are the examples of unsupervised classifier's. The baseline nearest neighbor classifier (K-NN), linear and pairwise linear classifier, single layer neural network (RBF Network), multilayer neural network, and SVM classifer framework are the examples of supervised classifer's [5].

IV. Results and Discussions

The aforementioned version of the system has been fully constructed and tested using both conventional databases and hand-crafted binary images. The following is the structure of the suggested methodology for binary geometrical shapes.

(a)

0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0

(b)

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

(c)

0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0

(d)

0	0	0	0	0	0	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

Figure 8. The various types of hand-crafted binary geometrical shapes.

Figure 8(a) represents a 3x4 rectangular binary object, whereas its scaled-up and rotated equivalent is built in figure 8(b). Similarly, we have constructed square shapes in figures 8(c) and (d). Now, we will be comparing the differences in their shapes. To analyse the shape difference by the proposed model, the input image is first convolved by convolving kernels. The results of convolutions on input binary objects by a 3x3 box kernel are shown in figure 9 (a), (b), (c) and (d). Herein, the convolutional operator preserves the structures of the objects boundary for the next layer in structural erosion layers.

(a)

0	0	0	0	0
0	4	6	4	0
0	6	9	6	0
0	6	9	6	0
0	4	6	4	0
0	0	0	0	0

(b)

0	0	0	0	0	0	0	0
0	4	6	6	6	6	4	0
0	6	9	9	9	9	6	0
0	6	9	9	9	9	6	0
0	6	9	9	9	9	6	0
0	4	6	6	6	6	4	0
0	0	0	0	0	0	0	0

(c)

0	0	0	0	0
0	4	6	4	0
0	6	9	6	0
0	4	6	4	0
0	0	0	0	0

(d)

0	0	0	0	0	0	0
0	4	6	6	6	4	0
0	6	9	9	9	6	0
0	6	9	9	9	6	0
0	6	9	9	9	6	0
0	4	6	6	6	4	0
0	0	0	0	0	0	0

Figure 9. The results of convolutions by a box kernel on hand-crafted binary geometrical shapes.

The purpose of the next layer is to scale-down the object to extract the scale-invariant features. The subsequent layers in the structural erosion layer have been implemented by the generalized ReLU function discussed in subsection III(B). The results of the ReLU activation function defined in equation 9 are shown in figure 10. The ReLU/MIN filter scale-down the region of interest (ROI) by preserving the shape boundary in an object.

(a)

0	0	0	0	0
0	4	6	4	0
0	6	9	6	0
0	6	9	6	0
0	4	6	4	0
0	0	0	0	0

(b)

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	4	6	6	4	0	0
0	0	6	9	9	6	0	0
0	0	4	6	6	4	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

(c)

0	0	0	0	0
0	4	6	4	0
0	6	9	6	0
0	4	6	4	0
0	0	0	0	0

(d)

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	4	6	4	0	0
0	0	6	9	6	0	0
0	0	4	6	4	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Figure 10. The results of ReLU, (a) No, ReLU (b) Yes, ReLU (c) No, ReLU (d) Yes, ReLU.

The range of the number of pixels in figure 10 (a) and (c) is the reference shape to scale-down the ROI of the objects. Herein, the number of pixels (non-zero integer) in ROI belonging between the pre-defined range is the stopping condition for the operation's ReLU/MIN activation function. The pre-defined range of the ROI depends on the object datasets, and it may vary from one dataset to another's. Therefore, the objects in figure 10 (a) and (c) do not need to pass through the MIN filter layer. Because the number of pixels in ROI falls between 9 and 12. But, in the case of figure 10 (b) and (d) have needed to pass through the respective layer for customizing the region of interest for the next layer. For the scale-invariant shape features, ReLU/MIN activation function must be applied until the number of pixels in ROI falls between the pre-defined range. The proposed algorithm changes the number of object pixels, preserving the dimensions of the image. The results of the next subsequent layers are shown in figure 11 (a), (b), (c) and (d), which is the output after operations of the convolutional layers 2 and 3.

(a)

0	0	0	0	0
0	151	211	151	0
0	211	337	211	0
0	211	337	211	0
0	151	211	151	0
0	0	0	0	0

(b)

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	151	211	211	151	0	0
0	0	211	337	337	211	0	0
0	0	151	211	211	151	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

(c)

0	0	0	0	0
0	151	211	151	0
0	211	337	211	0
0	151	337	151	0
0	0	0	0	0

(d)

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	151	211	151	0	0
0	0	211	337	211	0	0
0	0	151	211	151	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Figure 11. The results of convolution layers 2 and 3.

Now, the extracted features in figure 11 (a), (b), (c), and (d) from hand-crafted geometrical rectangular and square shapes are ready for flattening and classifier layers. The framework of the flattening layer depends on the classifier's. Herein, we have used global histogram as a classifier, in this case the flattening layer passes the data and their frequency for classification. And finally, the classifier computes the

difference between the two shape's histograms [26]. In other's ways, the flattening layer only transfers the data to the classifier layer in the form of statistical metrics such as skewness, kurtosis, moments, entropy, co-occurrence matrix, etc. We will discuss all these metrics in the case of real datasets for comparison purposes.

All experiments are conducted using a Python-Programming tool and tested on Intel CORE-i5 CPU with 3GB RAM on Linux Mint operating system (OS). The experimental results on popular benchmarks datasets using our proposed algorithm achieved encouraging results. We have used the Kimia's 99 [30], the Myth dataset, the Tool's dataset, the Modified National Institute of Standards and Technology (MNIST) database [5] and the MPEG-7 Core Experiment CE-Shape-1 [31],[32] dataset for the experiments.

A. Kimia's dataset

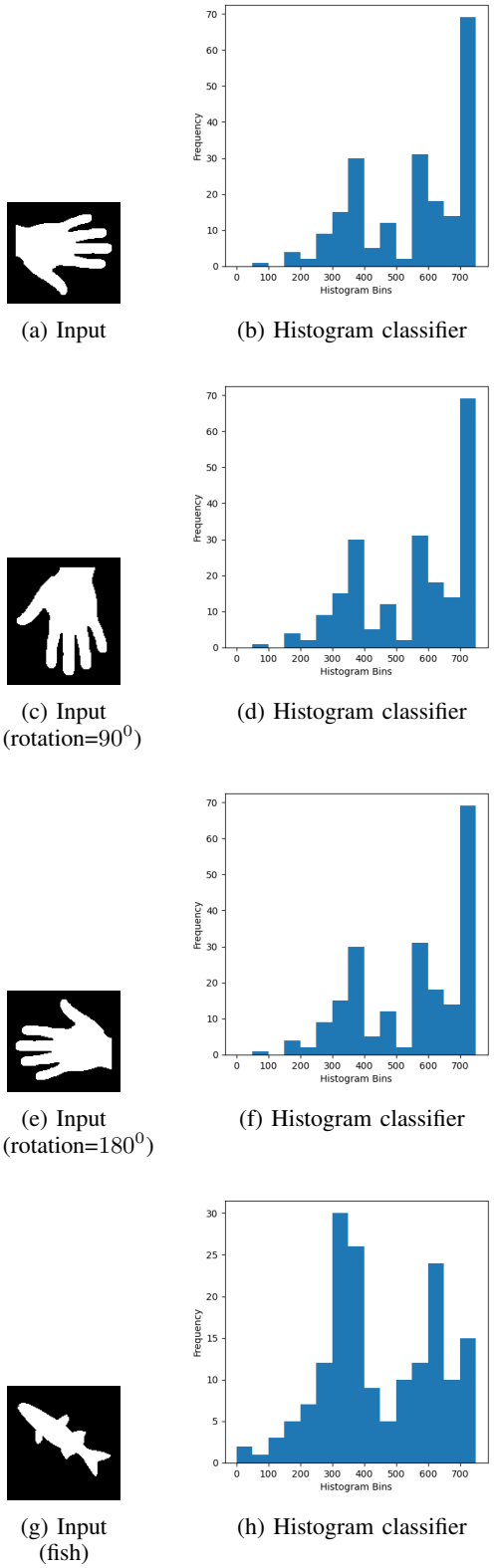
The kimia's 99 [30] data set is widely used for testing the performances of shape contour preserving descriptors in the recent era of shape matching and classification. It contains 99 images from nine categories, each category contains eleven images (as shown in figure 12). In the experiment, every binary object in the data set is considered a query, and the retrieval result is summarized as the number of tops 1 to top 10 closest matches in the same class (excluding the query object). Therefore, the best possible result for each of the rankings is 99. Table I lists the results of our proposed method and some other recent methods. The performance of our approach is comparably better than recent approaches.

Figure 12. The kimia's 99 data set.

TABLE I. Retrieval results on kimia's 99 data set.

Algorithms	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}
Salient Points [34]	99	99	98	96	95	93	93	90	84	77
IDSC+DP [25]	99	99	99	98	98	97	97	98	94	79
Height function [33]	99	99	99	99	98	99	99	96	95	88
LBP of Segments [23]	99	99	99	99	98	99	98	96	97	94
Our Proposed CNN Model	**99**	**99**	**99**	**99**	**98**	**99**	**98**	**97**	**97**	**95**

According to an experimental setting, Table I shows the outcomes of the suggested methodology, which outperforms the current strategy. In the experimental setting, we have designed the shape in terms of the pixel's structure [23]. Here, we have used the hand image from Kimia's 99 dataset for the experimental study of our model. The histogram of



(a) Input (b) Histogram classifier

(c) Input (rotation=90^0) (d) Histogram classifier

(e) Input (rotation=180^0) (f) Histogram classifier

(g) Input (fish) (h) Histogram classifier

Figure 13. The proposed CNN model's output histogram for various rotations.

the model output for different rotations of the hand image is shown in figure 13. The hand image having dimensions 116 x 130 contains 5270 object pixel's. The images in figure 13 (a), (c), and (e) are first inverted into the bright foreground and then rotated by 90^0 successively before going into the input of the model. Herein, a 3x3 box filter has been used to convolve the hand object pixels. For each object cell, we get an integer from the range of 1 to 9. Now the main task is to make the model's scale invariant. For this purpose, we have applied ReLUs on convolved output 17 times. Here, we have reduced the object pixels from, 5270 to 212 by preserving the structure of shape boundaries. Thereafter, convolutional layers 2 and 3 have been applied for stretching the region of interest between 1-to-729. The histogram of outputs shown in figure 13 (b), (d) and (f) for the respective input images are strongly supporting the accuracy of our proposed CNN model. The flattening layer constructs the histogram using bins, which have been defined between 1-750 with a bin interval of 50. The image in figure 13 (g) comes from the same dataset but is from a different class, and figure 13 (h) displays their histogram. The fish image has 2583 object pixels and applied 8-times ReLUs operations to get 171 non-zero integer values in their reduced ROI. The interclass dissimilarity and intraclass similarity of the histogram represent the accuracy of our model.

B. The Myth and Tools Dataset

The myth dataset used by [25][33], contains 15 samples of binary images having dimensions 540 X 530, consists of 3 classes (humans, horses, centaurs) and each class contains 5 sample images. The myth dataset has been shown in figure 14, and table II represents its corresponding average shape feature statistics, which have been extracted using the outputs of the proposed model for three different classes.

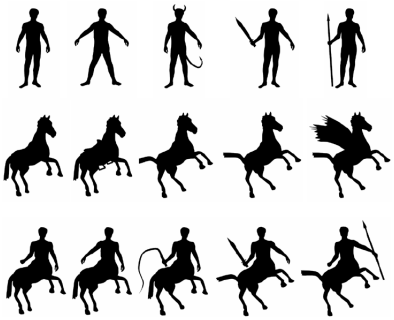


Figure 14. The Myth data set.

The average number of object pixels for humans, horses, and centaurs in the myth dataset is 3200, 4500, and 5000, respectively. Before being image's used as model input, the dataset is first inverted into an object with bright pixels. The model has performed ReLUs operations an average of 35, 42, and 40 times when the number of points in reduced ROI is taken to be 250, resulting in an average of 156, 180, and 191 non-zero integers between 1 and 729 for humans, horses, and centaurs, respectively. The shape statistics of the objects in table II show that the statistics of humans are more separated from the statistics of horses and centaurs. And the statistics of horses and centaurs are approximately close but differentiable. The model visualizes the average statistics for the myth shapes dataset as the same as the human visualization. Human vision saw the similarity between the horses and centaurs, and both are less differentiable than human shapes.

TABLE II. The average shape statistics of the Myth dataset.

	Humans	Horses	Centaurs
Mean	190.241	499.50	520.69
MAX	729	729	729
MIN	50	80	90
Skewness	0.9	-0.60	-0.46
Kurtosis	-0.46	-0.96	-1.06
Second moment	22707.89	31964.45	331078.74
Third moment	4183860.35	-1520847.63	-2765168.6
Standard deviation	151	182	177

The Tool's dataset, shown in figure 15, used by [25][33], consisting of 7 classes of 35 sample objects of different types of instruments, each class contains 5 images. Each image has dimensions of 476 X 635, containing on average, 15000 object pixels. The shape retrieval results of the Tool's dataset corresponding to figure 15 are represented in Table III, and its retrieval outcomes have been displayed in terms of shape statistics. To evaluate the recognition result for each image, the four most similar matches are chosen from other images in the dataset. The retrieval result is summarized as the number of first, second, third, and fourth most similar matches that come from the correct object.

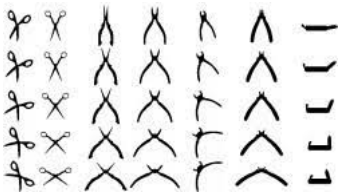


Figure 15. Tools dataset.

TABLE III. The retrieval results of the Tool's dataset.

Algorithms	First	Second	Third	Fourth
IDSC+DP [25]	35	29	30	22
Our Methodology	35	35	33	31

C. The MPEG-7 data set

The other widely tested data set is MPEG-7 CE-Shape-1 [31][32], that consists of 1400 silhouette images from 70 classes. Each class has 20 different binary objects, some typical objects are shown in figure 16. The recognition rate is measured by the Bullseye test used by several authors in literature [23],[25],[33]. The Bull's eyes score for every query image in the datasets is described by hit ratio. It is matched with all other images in the dataset and the top 40 most similar images are counted. These 40 images, at most

20 images are from the query image class that is correctly hit. The score of the test is the ratio of the number of correct hits of all images to the highest possible number of hits. In this case, the highest possible number of hits is $20 * 1,400 = 28,000$. Table IV shows the result of our proposed algorithm and comparison with some other existing context. In this table, we have calculated the retrieval rate on the MPEG-7 data set in terms of the percentage of Bull's eyes score.

Figure 16. The MPEG-7 CE-Shape-1 data set.

The analysis of the accuracy of our proposed model for the MPEG-7 CE-Shape-1 dataset is based on the probability of the occurrence of the digits in the bin. In this data set, we have reduced the region of interest to between 300-to-500 object pixels by applying the Generalization of Rectified linear units (G-ReLUs). We have used the normalizing function on the output of the convolutional layer 3 to make the reduced ROI scale invariant. Scale, translation, and rotation invariant characteristics are represented by the probability of occurrence of non-zero integers 1-to-729 with bins of size 10. As a result, in the interval of 10, there will be 73 bins for the data 1-to-729. Therefore, for each bin, the probability of occurrence is the sum of the integer in a given interval divided by the total sum of the integer in a region of interest. Thereafter, we have been applying the dissimilarity measure technique (Euclidean distance) between two vectors.

TABLE IV. Retrieval Rate (Bullseye Score) of Different Algorithms for the MPEG-7 CE-Shape-1 Data Set.

Algorithms	Score
IDSC+DP [5]	85.40%
Salient Points [17]	87.54%
A bioinformatics approach [18]	77.24%
Height functions [2]	89.66%
Height functions + shape complexity [2]	90.35%
Height functions+LCDP [2]	96.45%
LBP of Segments [23]	96.89%
Our Method	**97.05%**

D. The MNIST data set

The Modified National Institute of Standards and Technology (MNIST) database is a large database of handwritten digits that is commonly used for training various image processing systems. The MNIST database contains 60,000 training images and 10,000 testing images. Figure 17 provides a few examples of sample images that were created from comma-separated values. In their original paper, they use a support-vector machine to get an error rate of 0.8% [5]. A variety of machine learning classifiers are used for the MNIST dataset. These machine learning techniques are organized into six broad categories: linear classifiers, k-nearest neighbors, boosted stumps, nonlinear classifiers, support vector machines (SVMs), neural nets (with no convolutional structure), and the convolutional nets. In this work, our developed methodology saves unnecessary efforts on data preprocessing and formatting [35].

Figure 17. Sample images from MNIST dataset.

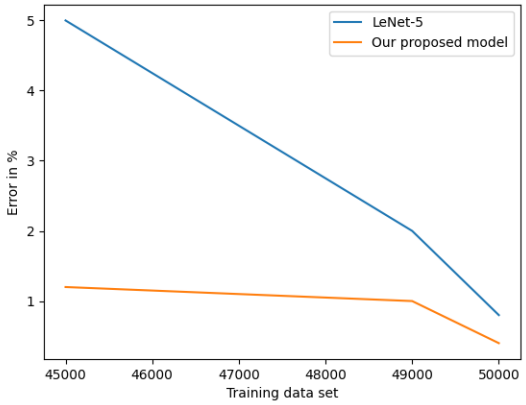


Figure 18. The error rate in %.

Here, we have used the radial basis function (RBF) classifier on the extracted features. The convolutional kernel used for feature extraction is shown in figure 6. In figure 18, we have analysed the error rate when the model is trained with less training data. The justification offered above firmly backs the model we have suggested for feature extraction.

V. Conclusion

This paper presents a model that describes the structure of pixels as features based on the neighbouring pixels over a shrunk region of interest (ROI) for object's shape recognition. The proposed model yield a structural erosion layer tool that can be applied to many situations where the domain knowledge or the state of information is propagated by the layer's in a given architecture. The basic objective of the domain-specific procedural knowledge is to integrate the unsupervised and supervised image-processing techniques into a comprehensive model. This can be achieved by propagating the information from one layer to another without getting pieces of information loss. Unfortunately, the majority of object recognizer's have used different types of pooling or subsampling methodology, which carries inefficient knowledge for the next layers and does not reduce

the dimensions of ROI. Here, we have solved the above-described issues by using our proposed model and yield a better recognition and classification rate compared to the existing model on standard shape datasets.

REFERENCES

[1] P. Maragos, "Pattern Spectrum and Multiscale Shape Representation," IEEE Transaction on pattern analysis and machine intelligence, vol II, no. 7, July 1989, pp.(701-716).

[2] ROBERT M. HARALICK, R. STERNBERG, and X. ZHUANG, "Image Analysis Using Mathematical Morphology," IEEE Transaction on pattern analysis and machine intelligence, vol. PAMI-9, no. 4, July 1987, pp.(532-550).

[3] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, "Convolutional neural networks: an overview and application in radiology" insight into imaging, (2018)9:611-629.

[4] Bart M. N. Smets, Jim Portegies, Erik J. Bekkers, Remco Duits, "PDE-Based Group Equivariant Convolutional Neural Networks" Journal of Mathematical Imaging and Vision (2022), https://doi.org/10.1007/s10851-022-01114-x.

[5] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[6] Hubel D. H., Wiesel T. N., "Receptive fields and functional architecture of monkey striate cortex," J Physiol. 1968 Mar;195(1):215-43. doi: 10.1113/jphysiol.1968.sp008455. PMID: 4966457; PMCID: PMC1557912.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[8] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. ILSVRC-2012, 2012. URL http://www.image-net.org/challenges/LSVRC/2012/.

[9] F. Sultana, A. Sufian and P. Dutta, "Advancements in Image Classification using Convolutional Neural Network," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2018, pp. 122-129, doi: 10.1109/ICRCICN.2018.8718718.

[10] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," Proceedings of the 27th International Conference on International Conference on Machine Learning, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104425

[11] D. Cire¸san, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. Arxiv preprint arXiv:1202.2745, 2012.

[12] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In International Conference on Computer Vision, pages 2146–2153. IEEE, 2009.

[13] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, pages 253–256. IEEE, 2010.

[14] R. Kumar and K. Mali, "Fragment of Binary Object Contour Points on the Basis of Energy Level for Shape Classification," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), 2021, pp. 609-615, doi: 10.1109/SPIN52536.2021.9566080.

[15] R. Kumar and K. Mali, "Local Binary Pattern for Binary Object Classification using Coordination Number (CN) and Hu's Moments," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-7, doi: 10.1109/ICRITO51393.2021.9596458.

[16] R. Kumar and K. Mali, "Bond Dissociation Energy and Pattern Spectrum for Shape Classification," 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2021, pp. 788-793, doi: 10.1109/ICEECCOT52851.2021.9707936.

[17] R. Kumar and K. Mali, "Shape Classification Via Contour Matching Using the Perpendicular Distance Functions", International Journal of Engineering and Applied Physics, vol. 1, no. 2, pp. 192–198, May 2021.

[18] R. Kumar, K. Mali. "Binary Shape Segmentation and Classification using Coordination Number (CN)*," Research and Reviews: Discrete Mathematical Structures. 2021; 8(1): 22–30p.

[19] R. Kumar, K. Mali. "A Novel Approach to Edge Detection and Performance Measure Based on the Theory of "Range" and "Bowley's Measure of Skewness" in a Noisy Environment," Journal of Image Processing and Pattern Recognition Progress. 2021; 8(1): 31–38p.

[20] R. Kumar, K. Mali, "Concurrent Lines Perpendicular Distance Functions for Contour Points Analysis," Available at SSRN: https://ssrn.com/abstract=3998737 or http://dx.doi.org/10.2139/ssrn.3998737.

[21] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," J. Phyiiol. (1962), 160, pp. 106-154, https://doi.org/10.1113/jphysiol.1962.sp006837.

[22] Yee Whye Teh and Geoffrey E. Hinton, "Rate-coded Restricted Boltzmann Machines for Face Recognition," Advances in Neural Information Processing Systems 13 (NIPS 2000).

[23] R. Kumar, K. Mali, "Local Binary Patterns of Segments of a Binary Object for Shape Analysis," J Math Imaging Vis (2022). https://doi.org/10.1007/s10851-022-01130-x.

[24] S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape contexts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 509-522, April 2002, doi: 10.1109/34.993558.

[25] H. Ling and D. W. Jacobs, "Shape Classification Using the Inner-Distance," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 2, pp. 286-299, Feb. 2007, doi: 10.1109/TPAMI.2007.41.

[26] Sung-Hyuk Cha, S. N. Srihari, "On measuring the distance between histograms" in Pattern Recognition 35 (2002) 1355-1370.

[27] YOSSI RUBNER, CARLO TOMASI AND LEONIDAS J. GUIBAS, "The Earth Mover's Distance as a Metric for Image Retrieval" in International Journal of Computer Vision 40(2), 99–121, 2000.

[28] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," in IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[29] Zhihu Huang and Jinsong Leng, "Analysis of Hu's moment invariants on image scaling and rotation," 2010 2nd International Conference on Computer Engineering and Technology, 2010, pp. V7-476-V7-480, doi: 10.1109/ICCET.2010.5485542.

[30] Sebastian, T.B., Klein, P.N., Kimia, B.B.: Recognition of shapes by editing their shock graphs. IEEE Trans. Patterns Anal. Mach. Intell. 25, 116–125 (2004)

[31] Latecki, L. J., Lakamper, R., Eckhardt, U.: Shape descriptors for non-rigid shape with a single closed contour. In: CVPR International Conference on Computer Vision and Pattern Recognition. pp 424–429 (2000)

[32] L. J. Latecki, R. Lakamper and T. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), 2000, pp. 424-429 vol.1, doi: 10.1109/CVPR.2000.855850.

[33] Wang, J., Bai, X., You, X., Liu, W., Latecki, L.J.: Shape matching and classification using height functions. Pattern Recogn. Lett. 33, 134–143 (2012)

[34] Shen, Y., Yang, J., Li, Y.: Finding salient points of shape contour for object recognition. In: Proceedings of the 2015 IEEE Conference on Robotics and Biomimetics Zhuhai. China. (2015)

[35] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 141-142, Nov. 2012, doi: 10.1109/MSP.2012.2211477.