Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7699815

SCENE TEXT RECOGNITION USING ENERGY ENRICHED SELF ORGANIZING MAP

Dr. T. Beula Bell
Associate Professor, Department of Computer Application, Nesamony Memorial Christian College, Marthandam, Email :beulamaglin@gmail.com

Dr. S. Immaculate Shyla
Assistant Professor, Department of Computer Science, St. Alphonsa College of Arts and Science, Karinkal, Email: immaculatejudit@gmail.com

Dr.C. Jaspin Jeba Sheela
Assistant Professor, Department of PG Computer Science, Nesamony Memorial Christian College, Marthandam, Email :jaspinjebasheela@gmail.com

Dr. J. R. Maglin
Professor, Department of Electrical and Electronics Engineering, C.S.I. Institute of Engineering, Thovalai, Email: maglincsiit@gmail.com

Abstract

The identification and recognition of scene-text remains an open and challenging subject in the field of computer vision. This is due to the fact that text often covers just a tiny portion of an image and has a backdrop that is not uniform. The scene-text detection and identification problem are made more difficult by factors such as noise, blur, occlusions, and reflections. As a result, insufficient accuracy and excessive time complexity are produced as a result of this challenge. In order to locate the accentuated edges in the scene-text image, the Sobel and Gaussian Edge Detectors are converted into Energetic Edge Detectors. In order to accurately localise the scene-text, the updated Local Directional Number descriptor on the Energetic Edge Information drives a course of action forward. The innovative algorithm that identifies the text characters of localised scene-text is able to account for these phenomena. The recognised output is supplied in the form of characters from the English language. The Energy enriched SOM that is empowered by the Histogram of Oriented Gradients and the Symlet Transform is the primary factor responsible for the success of the procedure. The output of this highly accurate scene-text recognition system may be coupled to simple activities that include image processing.

Keywords: *scene-text recognition, Neural Networks, self-organising map, Hog Descriptor*

1. INTRODUCTION

One of the most ubiquitous types of visual content in natural situations is the scene's language. The ability for humans to recognise scenes and the context in which they occur is very necessary. In a wide variety of applications, including multimedia retrieval, industrial automation, language translation, navigation, and many others, the scene-text is the primary

aspect that determines success. It does so by converting the text in the image to an ASCII format, and the text string that is produced as a consequence is referred to as the recognized-text. Scene-text recognition is much closer to the traditional OCR process than traditional OCR, but scene-text recognition is much more difficult due to variations in foreground texts and background objects, aspect ratios, arbitrary orientations, and uncontrollable lighting conditions. Scene-text recognition is much more difficult than traditional OCR. Therefore, the accuracy of scene-text recognition is much lower in comparison to that of document-text recognition. This technique attempts to build a novel scene-text-recognition system in order to overcome the constraints of current methods that are considered to be state of the art at this time. This study intends to design a new algorithm for onscene-text recognition and will be termed "Scene text recognition using Energy Enriched self-organizing map." The goal of this research is to develop the algorithm. This new plan is broken out in great detail in the next chapter. The image with the scene text indicated on it is sent as input to the module that does scene-text recognition, and the approach that is being suggested does scene-text recognition from that image. The ASCII characters that were taken from the scene's text are what the recognition output consists of. The scene-text recognition approach that has been suggested may be broken down into the following four sub-modules.

- A training method based on the self-organizing map.
- Character extraction from query texts as well as feature extraction
- Testing with energy-enhanced SOM. Post-processing.

The schematic of the general architecture may be seen in figure 1. In order to complete the training, the EESOM method is given the information in a font-oriented manner. After the characters from the query text have been retrieved, the energy characteristics may be determined by making use of those text-images. The testing procedure for Energy Enriched SOM is able to identify each individual character of the scene-text. In order to achieve accurate text recognition, the last step is to re-join the results of the text recognition.

2. RELATED WORK

The identification of scene text is still a difficult task because of the wide variety of possible text curvatures, orientations, and aspect ratios. The challenge of accurately representing text instances that take on random forms is one of the most difficult aspects of this undertaking. In spite of the fact that several approaches have been suggested, modelling irregular texts in a flexible way, the majority of these approaches sacrifice simplicity and robustness. The detection performance and generalisation ability are both hindered by the extensive post-processing steps they need, as well as by the regression caused by the Dirac delta distribution. In this article, we suggest an effective representation of text instances that we call CentripetalText (CT). This representation breaks down text instances into a mix of text kernels and centripetal shifts, and it is called CentripetalText. To be more specific, we make use of the centripetal shifts to accomplish pixel aggregation, which directs the pixels of the exterior text to the kernels of the inside text. Because the relaxation operation is included into the dense regression for centripetal shifts, it is now possible to provide an accurate forecast within a range rather than a single number. Our technique ensures a high detection accuracy as well as a quick inference speed because to the straightforward reconstruction of text outlines as well as its tolerance of prediction mistakes. In addition to this, we have shrunk our text

detector into a proposal generating module known as CentripetalText Proposal Network (CPN). This module has replaced Segmentation Proposal Network (SPN) in Mask TextSpotter v3 and is responsible for delivering more accurate suggestions [17].SwinTextSpotter is the name of the framework for finding text in scenes from beginning to finish. With the use of a new Recognition Conversion mechanism, we were able to integrate the two jobs by using a transformer encoder with a dynamic head as the detector. This allowed us to actively lead text localisation through the process of recognition loss. The uncomplicated design produces a simple framework that does not need the installation of a rectification module or character-level annotation for the arbitrarily-shaped text[18].The process, which is referred to as SVTR, begins by disassembling an image text into a series of smaller patches that are referred to as character components. After then, hierarchical steps are repeatedly carried out by the mixing, merging, and/or combining of components at various levels. In order to perceive the inter-character and intra-character patterns, global and local mixing blocks have been developed. This has resulted in a multi-grained perception of character components. Therefore, characters may be identified with the use of a straightforward linear prediction. The efficiency of SVTR was tested using English and Chinese scene text recognition tasks, and the findings showed that it performed well in both languages. SVTR-L (Large) produces accuracy that is very competitive in English and surpasses other approaches in Chinese by a significant margin, all while operating at a higher speed. In addition, the SVTR-T (Tiny) model is efficient despite its little size, and it exhibits a speed at inference that is quite attractive [19].We present a brand new operation called Strip Convolution, which is a convolution that has been developed specifically for the purpose of extracting characteristics from thin strokes. In addition to this, we make use of a Hierarchical Correlation technique, which entails the use of multi-level attention mechanisms for the purpose of capturing common text properties. On the basis of this, we construct a new framework for scene text recognition that we call a Hierarchical Correlated Strip Convolutonal network. Extensive trials show that the HCSC network that was presented is better, significantly boosting the accuracy of text recognition[20].A Cross-Attention Network with Two Branching Branches (DBCAN). Unlike the previous methods, which heavily relied on semantic information, DBCAN can improve position clues and learn semantic relations with two separate branches and fuse them by a tailored Cross-Attention Module. This is a significant departure from the previous methods, which heavily relied on semantic information (CAM). In addition, a Convolution-Based 2D Positional Embedding (CBPE) is presented as a method for describing the spatial interdependence of characters in 2 dimensions. Extensive trials show that our DBCAN is more accurate and reliable than the approaches that came before it, and it achieves state-of-the-art performance on a number of benchmarks, most notably CUTE (93.4%) [21].A linear structure that is used in the processing of the model's embedding layers as well as its prediction layers. When compared to ResNet, the AutoMLPMixer is capable of achieving a 16.4% reduction in effective parameter size. We are able to further decrease the number of parameters used in scene text recognition by 14.89M by using a grouped linear structure as a solution to the issue of an excessive dictionary. The speed at which the model processes images improved by 29% when both AutoMLPMixer and clustered linear structure were allowed to coexist in the same environment[22].The process of extracting text from photographs of natural scenes is referred to as scene text recognition (STR). Recent research has begun to add contrastive learning methodologies for the STR

problem. This is being done since researchers have discovered that self-supervised contrastive learning has several benefits. These studies, on the other hand, concentrate almost entirely on the data arguments of pictures from a visual point of view, neglecting the fact that scene text often consists of a substantial amount of noise and has the qualities of variety. In this research, we propose a text-level contrastive learning technique to learn a better representation of text in scene text pictures in order to successfully enhance the prediction performance of the STR task. This will be done in order to solve the problem that was raised before. The success of our strategy is shown by the results of our comprehensive tests, which we run on a number of publicly available benchmark datasets and then compare with baseline models[23]. the basic semantic feature is determined with the help of Semantic GAN, and then the Balanced Attention Module is used to determine the scene text. The goal of the Semantic GAN is to bring the semantic feature distribution of the support domain and the target domain into alignment with one another. In contrast to the traditional image-to-image translation methods, which operate at the image level, the Semantic GAN generates and discriminates information at the semantic level using the Semantic Generator Module (SGM) and the Semantic Discriminator Module. These two modules are responsible for the generation and discrimination, respectively (SDM). The Semantic Generator Module is responsible for generating basic semantic features for target pictures, which are scene text images. These features share the same feature distribution as support images (clear text images). The Semantic Discriminator Module is put to use in order to differentiate the semantic characteristics of the target domain from those of the support domain. In addition, an issue with attention drift is addressed with a Balanced Attention Module that was developed for this purpose. The Balanced Attention Module will first learn a balancing parameter based on the visual glance vector and the semantic glimpse vector before carrying out the balancing operation in order to achieve a balanced glimpse vector. This will be done in order to obtain a balanced glimpse vector[24].

3. SELF ORGANIZING MAP BASED TRAINING

The binarized text regions of database image which contains font's information are undergone the *character separation process.* After character segmentation, each font's character prototypes are undergone training process based on *Energy Enriched Self Organizing Map*. Herein, *multiple instances* of specific text-character are used for training purpose. The training process is sub divided into the following models.

(i) Extraction of Font-images from database andSample vector generation
(ii) SOM based training

2.1. Extraction of Font Images from Database and Sample Vector Generation

This module receives as its input the database image of the typeface that includes all of its character forms in their respective positions. This font-specific image has been binarized for your viewing convenience. The horizontal projection of the histogram is used to separate each line of text, while the vertical projection of the histogram is used to separate each character image.Each character's *height* and *width* are stored as $DB_H^k$and DB_W^krespectively. The character pairs which are producing uncertainty in recognition process are flagged as *uncertaintycharacters.* Herein, *uncertainty means* the *shape-wise-similar characters*. In

English language, that character pairs are given below, I and l, 0, o and O, s and S, c and C, v and V, w and W, x and X, z and Z

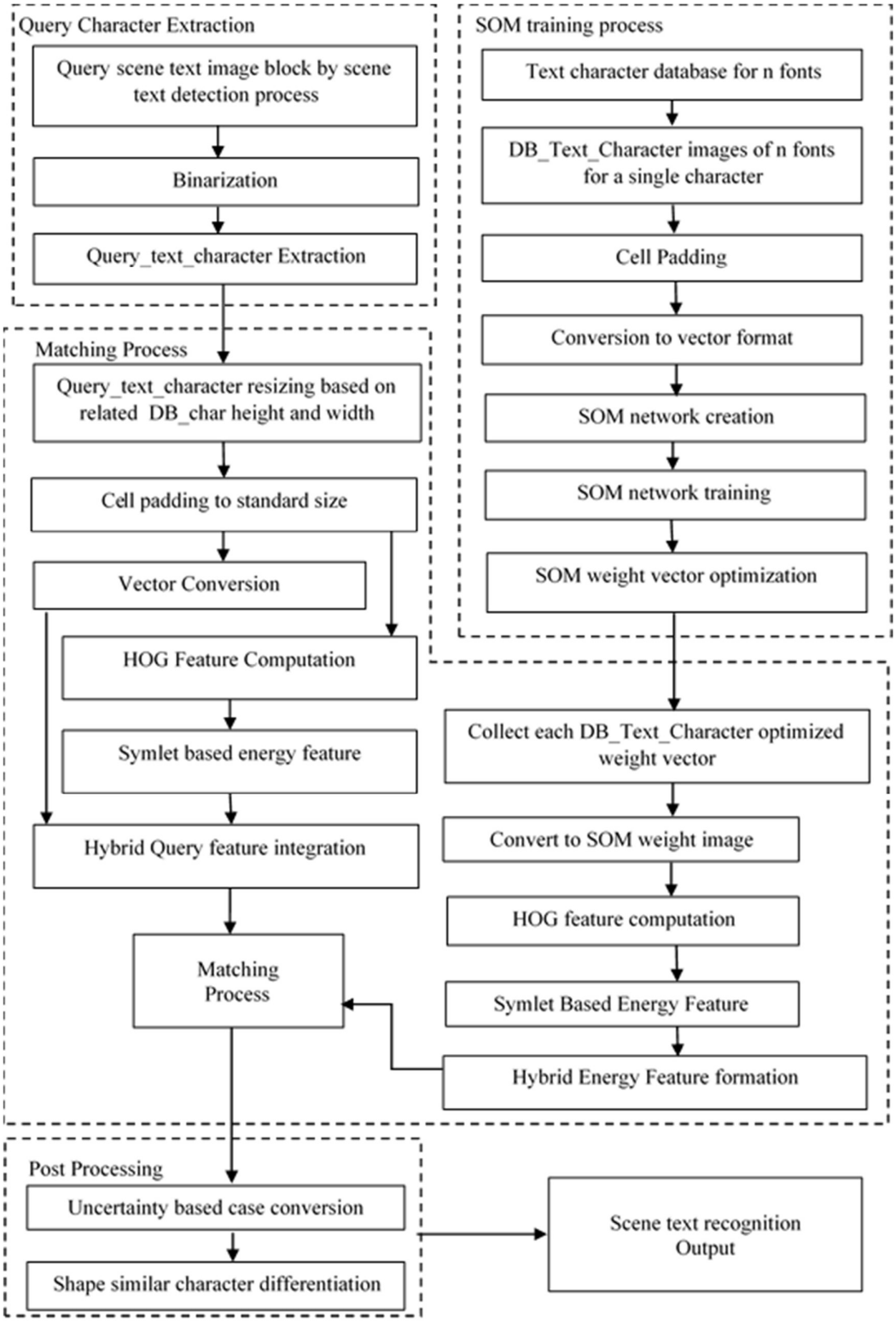


Fig.1. Architecture diagram of the scene-text recognition method.

After that, the following font-oriented database image, which should include character prototypes, is used as input, and each character is then retrieved from it. Because of this, it is

possible to process an n-thousandth of the typefaces, and as a result, each character will have n samples based on distinct fonts. The padding method is used to repack these character-image samples such that they have a defined width MW and fixed height MH. The term "padding method" refers to the process of filling up blank spaces with zeros. In this context, the terms "maximum height of the typeface" (MH) and "maximum width of the font" (MW) refer to the same thing.

It is possible to determine the MH by measuring the height of n font samples, and the MW may be determined by measuring the width of n font samples. Generally speaking, the letter 'Q' has the greatest possible height in English typefaces, while the character 'W' has the greatest possible width in English fonts. Therefore, the average height of the letter 'Q' is shown by the notation MH, while the average width of the character 'W' is indicated by the notation MW. After being repacked and shrunk, the characters are then transformed into a one-dimensional vector representation. This article takes into consideration a total of 62 different characters. It has 10 numbers in addition to the 26 capital letters and 26 lowercase letters that make up the alphabet. Concerned in this discussion are n sets of 62 different characters each. The vector representation of the pth sample's qth alphabet for the kth typeface may be specified as V if desired (k,p,q). In this study, the values for k, p, and q are as follows: k = [0, 62-1], p = [0, n-1], and q = [0, (MH*MW) -1].

Algorithm 1: Text Detection

Input: Scene text image

Output: Segment image with character bounding

BEGIN

Step 1: PerformBinarization for local minima $c(T) = 1/p$

For dark object on a light background$c(T) = 1 - {}^{1}/_{p}$

points =[F.spatial_transformer_grid(localization, self.target_shape) for localization in localizations]

rois = [F. spatial_transformer_sampler (images, point) for point in points]

Step 2: Perform Matching Process

Apply HOG Feature Computation

$$V_{HOG}=\left[F_1,F_2,\ldots\ldots\ldots\ldots,F_i,\ldots\ldots\ldots F_{36}\right]$$

Apply Symlet based energy feature

$$\Sigma_n w(n)w(n-2k) = \delta(k)$$

Step 3: Similarly Apply SOM Training Process

Step 4: Perform Post-processing

Step 5: Scene text recognition

End

2.2. SOM Based Training

Kohonen network is the well-known name for the structure that SOM network has. A feed-forward structure is used to design this network. It consists of a single computational layer that is organised into rows and columns. Within this particular instance, every single neuron is completely linked to the whole of the source nodes that make up the input layer.

Initialization is a procedure in which all of the linked weights are given random values with a narrow range, and this process gives the weights their starting values. The process for the competition is specified as follows: for each input pattern, the neurons compute their corresponding values using a discriminant function, and the neuron specific with the least value of the discriminant function is proclaimed to be the winner of the competition. The discriminant function in this context is the squared Euclidean distance between each neuron's input vector x and its weight vector w, and it is explained by *equation (1).*

$$d_j(x) = \sum_{i=1}^{D}(x_i - w_{ji})^2 \quad (1)$$

Where

$d_j(x) - distance\ metrice$

w_{ji} – *weight vector for* j^{th} *neuron*

Therefore, it is possible to map the continuous input space onto the discrete output space associated with neurons. The cooperation step is characterised by the fact that the cooperation between neighbouring neurons affects the spatial position of the winning neuron in relation to other stimulated neurons in the neighbourhood. The topological neighbourhood of the neurons is determined by computing it based on characteristics such as the lateral distance between the neurons, the index of the winning neuron I(x), and the size of the neighbourhood. The adaptation phase is constructed in such a way that the stimulated neurons lessen their responsibility in the linked weights. This makes it possible for the winning neurons to have maximal matching.

Take into consideration the kth character for the aim of training. It is made up of n sets of vectors in total. A training procedure has been carried out on these n sample vectors. The n sample vectors that are associated with the kth character are trained using the SOM neural network. Each character results in the creation of its own unique SOM [6][7][8]. The SOM network is constructed by using the function with the following parameters by using the equation (2),

$$Net_{SOM}.Epoches=100$$
$$Net_{SOM}.Goal=0.0$$
$$Net_{SOM}.Time = inf$$
$$Net_{SOM}.\sigma = 3$$
$$Net_{SOM} = Net_{SOM}.FuncCreateSom(V^K) \quad (2)$$

The term *Epoch* provides the time info to determine the neighbour count info. The term goal means the learning quality by specifying the convergence error. The term Time points out the number of iterations to be proceeded. The parameter σ means the initial number of neighbour count. The $FuncCreateSom(\ )$ is the SOM network creation function. The Net_{SOM} is the indicator of SOM network.The SOM network training [9][10][11] is proposed by *equation (3)*.

$$Net_{SOM} = Net_{SOM}.Train(V^k) \quad (3)$$

The trained SOM network contains n weight vectors, i.e.$W^{i,j}$, where i ϵ [0, n-1], j ϵ [0, MH*MW-1]. The weight vectors naturally contain the fractional values. But there is the need of '1s and 0s' for the further calculation. Hence, the weight matrix is converted to '0s and 1s' model. This phenomenon is described in *equation (4).*

$$W_B^{i,j} = \begin{cases} 1 & , if\ W^{i,j} > 0.1 \\ 0 & otherwise \end{cases} \quad (4)$$

$i \in [0, n-1]$

$j \in [0, MH*MW-1]$

Where

$W_B^{i,j} - Binarized\ weight\ matrix$

Finally the *centre-positioned weight vector* is chosen as the *final trained vector* related to k^{th} *font.* This phenomenon is explained in *equation (5).*

$$W_{SOM}^{k,q} = W_B^{fix\left(\frac{n}{2}\right),q} \quad (5)$$

$q \in [0, MH*MW-1]$

Where

$W_{SOM} - Final\ SOM\ weight\ vector.$

In this way, the entire k fonts are processed to generate the final SOM weight vector W_{SOM}.

3. QUERY TEXT CHARACTER AND FEATURE EXTRACTION

This module takes as its input the query scene-text designated image I ST, from which it extracts individual Scene-text blocks. Through the use of the vertical projection of the histogram, the retrieved binary blocks are broken down into individual characters. With the help of the padding procedure, the image of a single letter that is of the rectangle type is resized to the standard height of MH and the standard width of MW. Both the original query character image and the scaled version of the character image are stored separately as independent memories. The format of the scaled image is then changed to a vector format, which has its size set to [1, (MH*MW)-1], and the name of the query vector is 〖V〗_(Q)^i, with I falling within the range [0, (MH*MW)-1].

4. ENERGY ENRICHED SOM TESTING

The Histogram of Oriented Gradients, sometimes known as HOG, is a prominent feature descriptor that is frequently used in many image processing applications. Finding the frequencies of different gradient orientations in a particular area of an image is the goal of the HOG approach. This method is comparable to the other approaches, which are referred to as scale-invariant feature transform descriptors, edge orientation histograms [12], and shape contexts. The primary distinction between its working idea and traditional methods is that calculations are performed on a dense grid that is consistently spaced.

The key idea behind HOG is that the local appearance and form of objects inside an image may be characterised by the distribution of intensity gradients or edge directions. This is the basis of the HOG algorithm.

The HOG feature descriptor [13][14][15] divides the image into a number of tiny linked sections that are referred to as cells, and then computes a histogram of gradient directions for the pixels that are included inside each cell. HOG descriptors are given names based on the composition of these histograms. The HOG descriptor's strength lies in the fact that it acts on local cells rather than global areas; hence, it is not affected by changes in geometry or photometry [3, 5]. This is a significant benefit. The calculation of HOG is carried out in five stages, which are the computation of the gradient, the creation of bins associated with orientation, the construction of blocks, the normalising of blocks, and the generation of HOG

feature vectors. In most cases, the output of HOG is kept in the form of a one-dimensional vector.

The wavelet transform is used in order to extract characteristics from one-dimensional data or signals. The temporal and frequency information of a signal or an image may be generated concurrently via the use of the wavelet transform. A complicated vector may be broken down into its component parts by the use of wavelet transform. Ingrid Daubechies is the person responsible for developing a family of orthogonal wavelets known as the Daubechies Wavelets. These wavelets have vanishing moments.

The vanishing moments of the wavelets function are the sole thing that sets Symlet wavelets from fromDaubechies wavelets [4]. Symlet wavelets are quite comparable to Daubechies wavelets.

Symlets are Daubechies least symmetric wavelets, and their support is highly densely concentrated. Although the construction of symlets is quite comparable to that of daubechies, the symmetry of symlets is far more robust than that of daubechies. Wavelets with a Daubechies structure have the maximum phase, whereas Symlet structures have the minimum phase [1].

With the use of the HOG feature vector and the Symlet transform, the trained weight vector W SOMk of the SOM is transformed into an image format so that it can support the energised form. HOG is implemented with the help of a built-in function in MATLAB that is associated with the extraction of HOG features. In order to carry out the Symlet Transformation, this article makes use of a built-in function of MATLAB that is associated with the Discrete Wavelet Transformation. The Symlet 2 type configuration is utilised wherever possible in the Symlet implementation. In this case, the HOG supplies the 1D data, which is then subjected to the Symlet transform for additional processing. The output of the Symlet transform is likewise preserved in its original 1D form. These energised features, together with the weight vector W_SOM^k, are concatenated to produce the hybrid feature, which is then used to transform the SOM into an energy-enriched version of itself. When this mode is selected, the whole of the training vectors are transformed into an energised form. This outcome is informed by the function 〖W〗_EESOM^(k,i), where k ϵ where k may range from 0 to n-1 and I can range from 0 to (MH*MW)-1.

Cubic interpolation is used to scale the single character rectangle type image of the query so that it corresponds to the height and width of the associated database character's DB_H and DB_W values. Both the original query character image and the enlarged image of the character are stored separately in their respective memory. After being scaled, the query character image is repacked into its original size, which is measured in MH x MW. This image format is then translated into vector format, which also has its size set to [1, (MH*MW)], and the query vector is given the name 〖V〗_(Q)^i, with I ranging from [0, (MH*MW-1)], where I is the index number. In order to calculate the energised SOM model with the HOG feature and the Symlet transform, the query feature vector 〖V〗_(Q)^I is first transformed into image format. The hybrid feature vector V EESOM is formed by concatenating the resulting vector with the appropriate query feature vector V (Q)i in order to provide the desired result. The equation that will be used for the Energy Enriched SOM-based testing may be found here (6).

$$\alpha^k = \sum_{i=0}^{l-1}(W_{EESOM}^{k,i} - V_{EESOM}^{i})^2 \qquad (6)$$

$$k \in [0,62\text{-}1]$$

Where

α – matching score corresponding to k^{th} database character related with query image

l – Length of hybrid query vector feature.

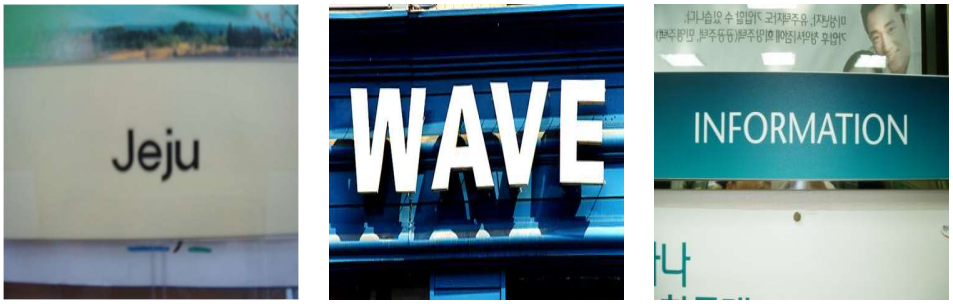
The least scoring database character is declared as the intermediately matching character.

5. POST PROCESSING

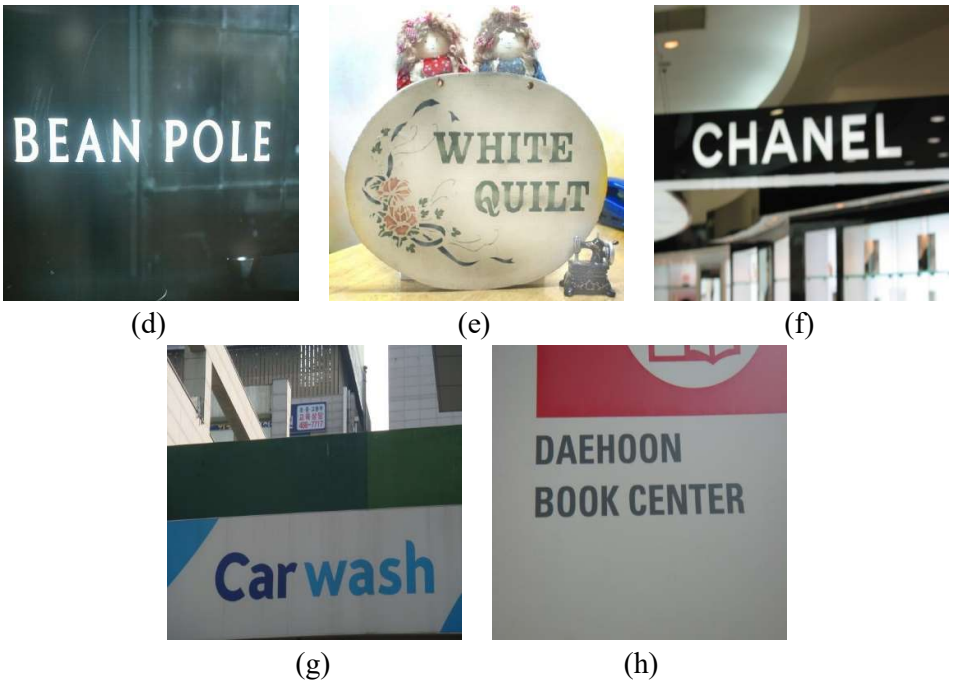
The result of the uncertainty-flagged-type matching is handled in order to produce the redefined-result by using the font-height attribute. This post processing phase takes care of the minor variations in characters with shapes that are quite close to one another by taking into account the extension-specific alterations in the font structure. Therefore, each character inside the query block is identified, and then, thereafter, these individual recognitions are concatenated to form the result of the scene-text recognition. The same task is done over and over again for each text block that is included in the scene-text-detected image 〖I〗_ST. The results of the scene-text recognition based on the proposed paper are shown, with the recognised text characters serving as the output.This paper uses a scene-text database:

- KAIST scene-text database[16]
  (https://sites.google.com/site/pedestrianbenchmark/download)

The KAIST scene-text database consists of 3000 scene-text images with a variety of situations, such as indoor and outdoor scenes with different lighting criteria, such as clear day, well artificial lighting, night environment, etc. The database also includes a variety of scene-text images with different lighting criteria. In order to capture photos, a high resolution camera and a cell phone with a modest resolution are both used as image acquisition devices. Within this database, the typical dimensions of an image are 640 pixels wide by 480 pixels tall. For the purpose of producing an effective paper on scene-text management, the database supports the Korean language in addition to the English language and other languages. In order to assess how well the suggested strategy works, this research employs 200 test photos taken from the aforementioned database. This database will be referred to as KAIST DB throughout this article.



(a) (b) (c)



(d) (e) (f)

(g) (h)

Fig 2. Sample test images form *KAIST_DB* database.

The names of the sample images of KAIST_DB (given in Fig.3) are given below.*KAIST_DB_img0, KAIST_DB_img1, KAIST_DB_img2, KAIST_DB_img3, KAIST_DB_img4, KAIST_DB_img5, KAIST_DB_img6, KAIST_DB_img7*

6.ANALYSIS ON SCENE-TEXT DENOISING

The analysis on scene-text denoising is progressed by comparing the proposed method with the performance of the following three state-of–the art existing denoising methods.

- *Modified Decision Based Unsymmetrical Trimmed Median Filter (MDBUTMF Filter)*
- *Geometric MultiscaleRidgelt Support Vector Transform Filter (GMRSVTF Filter)*
- *Effective and Adaptive Algorithm for Pepper and Salt noise removal (EAA Filter)*

An algorithm for the restoration of grayscale and colour pictures that have been severely contaminated by salt and pepper noise that is based on a modified decision and uses an asymmetrical trimmed median filter. During the capture and transmission of images, impulsive noise often causes the images to become damaged. As a result, an effective noise suppression strategy is necessary before continuing with following image processing processes. Due to its power of denoising and its effectiveness in computing terms, the median filter, often known as MF, is commonly employed in noise reduction techniques. However, it is only useful in environments with a low noise density.The development of effective image denoising technology has been made feasible by recent advances in machine learning. In order to generate a multiscale, multidirectional, undecimated, dyadic, aliasing, and shift-invariant geometric multiscaleridgelet support vector transform, the multiscaleridgelet support vector filter (MRSVF) must first be inferred from the ridgelet support vector machine (RSVM)

(GMRSVT). After that, multiscale dictionaries are learnt from examples in order to cut down on the amount of noise that is present in GMRSVT coefficients. The MRSVF has the ability to extract prominent features from pictures that are related with linear singularities. As a result, GMRSVT is capable of accurately approximating image edges, contours, and textures, and it is able to prevent the ringing effects that are caused by sampling during the process of multiscale decomposition of image data.

We can accurately evaluate whether a pixel has noise by counting the number of closed grey-level and noise-free pixels in the vicinity of a pixel that is suspected of having noise. This allows us to establish whether the pixel has noise or not. Noise filtering does not apply to pixels that are devoid of noise. An adaptive filtering technique that uses a weighted mean that is based on the Euler distance is able to produce great noise reduction while maintaining a reasonable level of feature retention for the noisy pixels. Because the algorithm is able to deal with varying degrees of noise, it is not necessary for us to manually alter the parameters and thresholds.

6.1 Mean Square Error (MSE) Analysis for Denoising

MSE analysis is an analytic metric which is used to measure the performance of image denoising methods. The *MSE* computation is involved with the two parameters such as original image and noise-free image. The computation of *MSE* is described in *equation (7).*

$$MSE = \frac{1}{H*W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}(I_{ORIG}^{i,j} - I_{NF}^{i,j})^2 \quad (7)$$

Where

H- Height of the original image

W- Width of the original image

I_{ORIG}*- Original image*

I_{NF}*- Noise-free image*

The less-*MSE* score indicates the better denoising power of a specific paper method and higher-*MSE* score indicates the poor denoising power of that method.

6.1.1. *MSE* Analysis on Denoising for *KAIST_DB* Database

The four methods such as *MDBUTMF* method, *GMRSVTF* method, E*AA* method and proposed method for 60% noise corruption and this analysis is performed over the scene-text images taken from *KAIST_DB* database.

Table 1: MSE analysis for denoising for 60% noise corruption on *KAIST_DB*

Database Name	Image Name	MSE			
		MDBUTMF Method	GMRSVTF method	EAA method	Proposed STD-PELE method
KAIST_DB database	KAIST_DB-img0	208.487	173.412	152.433	57.422
	KAIST_DB-img1	264.899	223.913	186.243	59.990
	KAIST_DB-img2	316.286	246.649	203.741	67.310
	KAIST_DB-img3	287.131	235.548	191.902	66.386
	KAIST_DB-img4	209.932	283.844	238.276	93.128
	KAIST_DB-img5	255.905	210.416	200.484	55.986
	KAIST_DB-img6	284.498	257.087	218.816	72.124

	KAIST_DB-img7	266.735	233.389	194.572	61.955

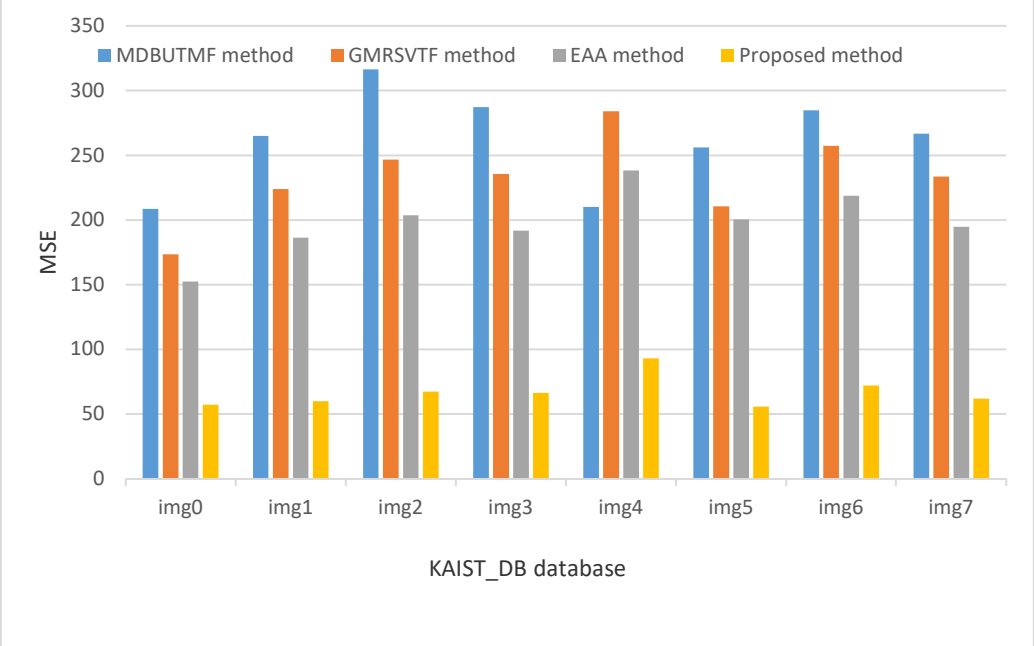


Fig.3.*MSE* analysis chart on denoising for 60% noise corruption on KAIST_DB.

Table 1 and Fig.3 prove that the proposed method gains less-*MSE* than the other existing methods. Hence, it proves the best performance of the proposed method. The second-best-method according to this analysis is *EAA* method. The least-performance-method in denoising based on *MSE* is *MDBUTMF* method. The least-MSE gained by the proposed method is 55.986 and the corresponding image is KAIST_DB-img5. This analysis proves the best performance of the proposed method than the existing methods corresponding with *KAIST_DB* database.

6.2 Peak Signal to Noise Ratio (PSNR) Analysis on Denoising

The PSNR analysis is a dominant analytic measure to evaluate the denoising performance of the image denoising algorithms. The PSNR computation is done by using the MSE value corresponding to *original image* and *noise-free image*. The PSNR is computed based on *equation (8)*.

$$PSNR = 10 * log_{10}(\frac{255^2}{MSE}) \quad (8)$$

The numeric 255 means the highest value of intensity in an image. The high PSNR denotes the best method and vice versa.

6.2.1 *PSNR* Analysis on Denoising for *KAIST_DB*

The PSNR values of the four methods, such as *MDBUTMF*, *GMRSVTF*, *EAA* and the proposed, for *KAIST_DB* database are shown in the following table.

Table 2: PSNR analysis for denoising for 60% noise corruption on *KAIST_DB*

Database Name	Image Name	PSNR (in db)			
		MDBUTMF	GMRSVTF	EAA method	Proposed STDR-PELE

		method	method		method
KAIST_DB database	KAIST_DB-img0	24.94	25.74	26.30	30.54
	KAIST_DB-img1	23.90	24.63	25.43	30.35
	KAIST_DB-img2	23.13	24.21	25.04	29.85
	KAIST_DB-img3	23.55	24.41	25.30	29.91
	KAIST_DB-img4	24.91	23.60	24.30	28.44
	KAIST_DB-img5	24.05	24.90	25.11	30.65
	KAIST_DB-img6	23.59	24.03	24.73	29.55
	KAIST_DB-img7	23.87	24.05	25.24	30.21

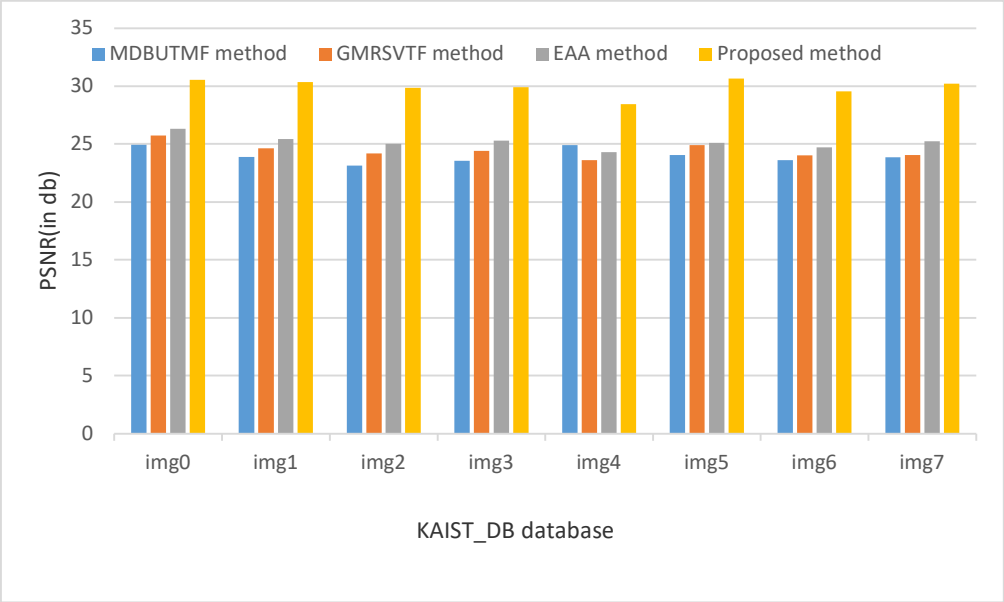


Fig.4.PSNR analysis chart on denoising for*KAIST_DB* for 60% noise corruption.

Table 2 and Fig.4express that the proposed method achieves the first grade denoising performance because it attains highest PSNR than the other existing methods. The *EAA* method receives second-grade denoising performance because it gets second-level PSNR values. The poor output producer in denoising is the *MDBUTMF* method because it does not contains the switching median approach. The high power of the proposed method shows the power of the combination of '*decision based filter'*, '*trimmed median'*, '*switching median'*, '*fuzzy logic'* and '*pattern based median*'. The higher PSNR obtained by the proposed method is 30.65. This analysis produces the higher performance on denoising for the proposed method over the *KAIST_DB* database.

6.3. Image Enhancement Factor (*IEF*) Analysis on Denoising

The Image Enhancement Factor (IEF) is used to evaluate the image-denoising performance of a specific noise-reduction algorithm. In this metric, three parameters are participated to obtain the IEF value. The IEF computation uses the *original image, Noisy-image* and *Noise-free image* to find the goodness of a denoising-algorithm. IEF is computed based on *equation (9) to (11).*

$$SE_{NOISY} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} (I_{ORIG}^{i,j} - i_{NOISY}^{i,j})^2 \quad (9)$$

$$SE_{NFREE} = \sum_{i=0}^{H-1} \sum_{j=0}^{w-1} (I_{NF}^{i,j} - I_{ORIG}^{i,j})^2 \quad (10)$$

$$IEF = \frac{SE_{NOISY}}{SE_{NF}} \quad (11)$$

Where

SE_{NOISY} *- Square error between noisy image and original image*

SE_{NF} *- Square error between noise-free image and original image*

The higher-IEF means better method and lower-IEF means poor method in image denoising.

6.3.1. *IEF* Analysis on Image Denoising for *KAIST_DB*

The findings of the analyses using the four different approaches, namely MDBUTMF, GMRSVTF, EAA, and the suggested. This investigation is carried out at a noise corruption level of 60% on scene-text taken from the KAIST DB database. The analysis in Table 3 and Fig.5 explain about the effectiveness of the suggested technique on scene-text denoising by investigating the greatest IEF value produced by the proposed method. Based on this examination, the MDBUTMF method is considered to be a subpar denoising technique due to the fact that it achieves the lowest IEF values. The innovative PDTM filter that is included into the suggested approach is one of the primary reasons why this method is able to attain a higher IEF than other methods. The performance of the suggested technique is improved because to the addition of the efficient components that PDTM filter provides. The suggested approach for the KAIST DB database has been shown to have first-grade denoising capabilities as a result of this investigation.

Table 3: IEF analysis for 60% noise corruption on KAIST_DB database

Database Name	Image Name	IEF			
		MDBUTMF method	GMRSVTF method	EAA method	Proposed STD-PELE method
KAIST_DB database	KAIST_DB-img0	36	52	48	325
	KAIST_DB-img1	31	51	60	330
	KAIST_DB-img2	37	51	61	337
	KAIST_DB-img3	32	48	54	347
	KAIST_DB-img4	37	48	54	341
	KAIST_DB-img5	41	50	67	340
	KAIST_DB-img6	42	49	48	317
	KAIST_DB-img7	37	47	53	327

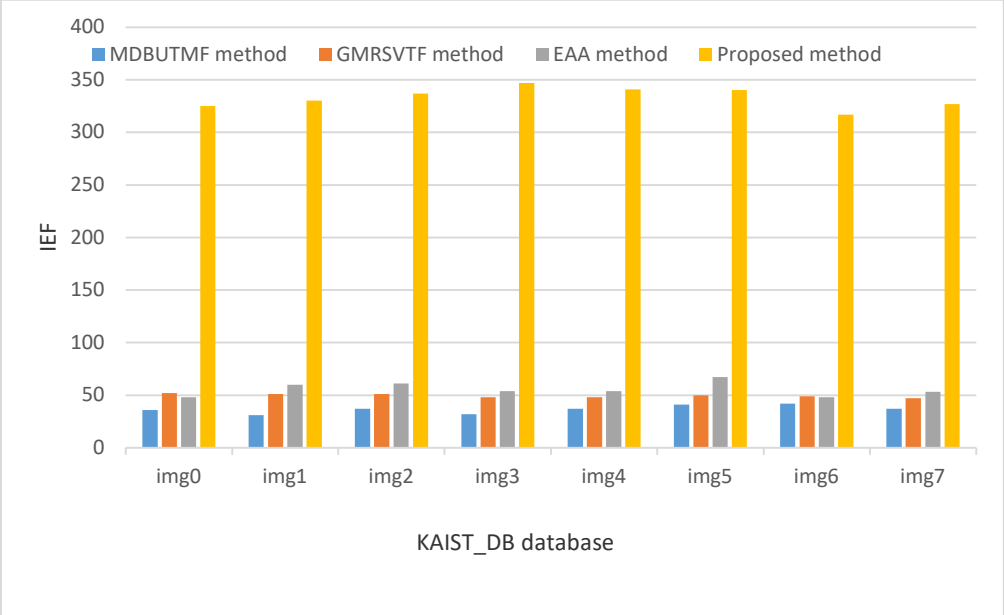


Fig.5. IEF analysis chart on denoising for*KAIST_DB* for 60% noise corruption.

6.4. Analysis on Scene-text Recognition

Recall analysis is an analytic metric to assessment the proposed method against the existing methods. It can be used to measure the scene-text recognition performance. The Recall computation is performed based on *equation (12).*

$$Recall = \frac{TP}{TP+FN} (12)$$

Where

TP- True positive

FN- False negative

The positive class is referred by all the text in the recognized output file (whether it may be correctly or incorrectly recognized). The term true positive refers the correctly recognized items in the positive class. The false recognition term indicates the items that are not labelled as belonging to the positive class, but should have been recognized. The higher recall percentage means better scene-text recognition method and vice-versa.

7. SAMPLE SCREEN SHOTS

The outputs generated after scene text recognition using energy enriched self-organizing map is
shown in 7.1

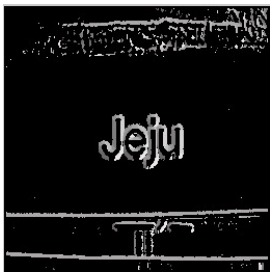
7.1 Sample screen shots for the proposed methodOn KAIST_DB Database

Input onscene image Gray image Noisy image

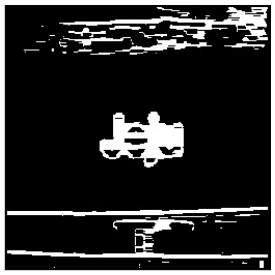
Noise reduced image

Sobel fused edge image

Gaussian fused edge image

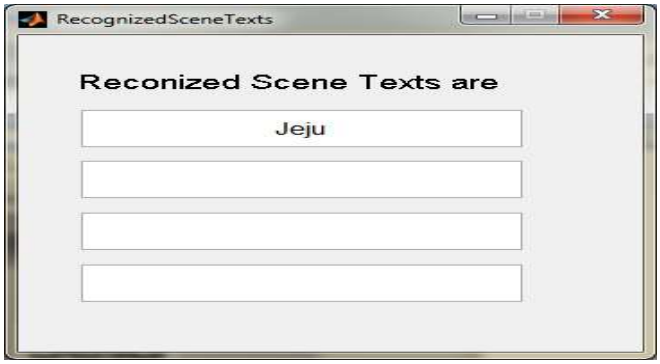


LDN image



Linked map

Scene Text marked image



Recognized Scene Text

Fig.6Sample screenshots for the proposed method on KAIST_DB database.

8.CONCLUSION

The noisy environment, the variable intensity, and the variable text size are the primary obstacles associated with this activity. Inaccuracy is a problem for the present approaches that have been created for scene text recognition. This is a problem in text localization as well as text character recognition. This situation motivates people in the real world to create new scene text recognition algorithms in order to address the shortcomings of the approaches that are already in use. Using the Jerod et al. technique, the Marouna et al. method, and the Sezer et al. method as comparisons and the three benchmarked databases from KAIST DB as aids, the suggested approach is evaluated in comparison to the current methods. The technique incorporates procedures for detecting noise at the pixel level and filtering out noise. The technique of noise detection produces high levels of proper detection while simultaneously maintaining low levels of false alarms. This is accomplished by counting the number of pixels in the surrounding area that have a closed grey level. An examination is carried out on each of the three subdivisions, which include scene text denoising, detection, and identification. The performance that the suggested approach achieves is superior to that of the other methods in

each of these three areas. This study was carried out using a wide range of analytic parameters, and it is a compelling demonstration of the superiority of the approach that was provided for comprehending scene-text relationships. In this work, the process of extracting scene-text from scene images is broken down and explained in detail. With the assistance of the SOM neural network trainer, the database training may be accomplished quite well. It teaches the text content in a manner that is effective for the feature selection. When attempting to recognise the text included inside an image, certain characteristics, such as HOG and Symlet, are used.

References

[1].Ayush, D, Bhawna, G& Sunil, A 2016, 'Performance Comparison of Different Wavelet Families Based on Bone Vessel Fusion', Asian Journal of Pharmaceutics, vol. 10, no. 4, pp. S791-S795.

[2].Baoguang, S, Xiang, B & Cong, Y 2017, 'An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298 – 2304.

[3].Boran, Y &Hongjie, W 2017, 'Chinese Text Detection and Recognition in Natural Scene Using HOG and SVM', Proceedings of the International Conference on Information Technology for Manufacturing Systems, pp.148-152.

[4].Masumdar, RE &Karandikar, RG 2016, 'Comparative Study of Different Wavelet Transforms in Fusion of Multimodal Medical Images', International Journal of Computer Applications, vol. 146, no.11, pp. 18-24.

[5].Shiva, RP &Giri, PMN 2016, 'Extracting text from natural scene images by HOG Character Descriptor', International Journal of Advanced Paper in Electronics and Communication Engineering, vol. 5, no. 12, pp. 2518-2523.

[6].Strouthopoulos, C &Papamarkos, N 1998, 'Text identification for document image analysis using a neural network', Journal of Image and Vision Computing, vol. 16, no. 13, pp. 879-896.

[7].Sung, HC, Jong, PY, Seung& Sang, WK 2010, 'Edge-based Text Localization and Character Segmentation Algorithms for Automatic Slab Information Recognition', Proceedings of the IEEE International Conference on Image Analysis and Signal Processing, pp. 387–392.

[8].Teofilo, EC, Bodla, RB &Manik, V 2009, 'Character Recognition in Natural Images', Proceedings of the International Conference on Computer Vision Theory and Applications, pp. 1-

[9].Thynzar, S & Pike, T 2005, 'Recognition and Translation of the Myanmar Printed Text Based on Hopfield Neural Network', Proceedings of the Sixth IEEE Asia-Pacific Symposium on Information and Telecommunication Technologies, pp.99-104.

[10].Toan, ND, Jonghyun, P &Guee, SL 2010, 'Voting Based Text Line Segmentation in Handwritten Document Images', Proceedings of the Tenth IEEE International Conference on Computer and Information Technology, pp. 529 – 535.

[11].Weijuan, W, Xianglin, H, Lifang, Y, Zhao, Y &Pengju, Z 2009, 'An Efficient Method for Text Location and Segmentation', IEEE WRI World Congress on Software Engineering, pp. 3-7.

[12].Xiaohang, R, Zhou, Y, Zheng, H, Jun, S, Xiaokang, Y & Kai, C 2017, 'A Novel Text Structure Feature Extractor for Chinese Scene Text Detection and Recognition', IEEE Journals & Magazines, vol. 5, pp. 3193 – 3204

[13].Yuanping, Z, Jun, S & Satoshi, N 2011, 'Recognizing Natural Scene Characters by Convolutional Neural Network and Bimodal Image Enhancement', International Workshop on Camera-Based Document Analysis and Recognition CBDAR 2011: Camera-Based Document Analysis and Recognition, pp. 69-82.

[14].Zaidah, I, Dino, I &Rajprasad, R 2008, 'Text and Non-text Segmentation and Classification From Document Images', Proceedings of the IEEE Conference on Computer Science and Software Engineering, pp. 973-976.

[15].Julinda, G, Ralph, E, & Bernd, F 2004, 'A Text Detection, Localization and Segmentation System for OCR in Images', Proceedings of the Sixth IEEE International Symposium on Multimedia Software Engineering, pp. 310-317

[16].KAIST Multispectral Pedestrian Dataset Available from https://sites.google.com/site/pedestrianbenchmark/download [09 November 2015].

[17] Sheng, Tao, Jie Chen and ZhouhuiLian. "CentripetalText: An Efficient Text Instance Representation for Scene Text Detection." *Neural Information Processing Systems* (2021).35th Conference on Neural Information Processing Systems (NeurIPS 2021) pp 1-17

[18]Huang, Mingxin, Yuliang Liu, ZhenghaoPeng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Jing Yuan, Kai Ding and Lianwen Jin. "SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022): 4583-4593.

[19] Yongkun Du, Zhineng Chen, CaiyanJia, Xiaoting Yin, TianlunZheng, Chenxia Li, Yuning Du, Yu-Gang Jiang, "**SVTR: Scene Text Recognition with a Single Visual Model**" Computer Vision and Pattern Recognition pp:1-7, 2022

[20] Liu, Chang and Junjie Lai. "Pattern Matters: Hierarchical Correlated Strip Convolutional Network for Scene Text Recognition." *2022 IEEE International Conference on Multimedia and Expo (ICME)* (2022): 1-6.

[21] Gao, Xinjian, Ye Pang, Yuyu Liu, Jun Yu, Maokun Han, Kai Hou and W. Wang. "DBCAN: Dual-Branch Cross-Attention Network for Scene Text Recognition." *2022 IEEE International Conference on Multimedia and Expo (ICME)* (2022): 1-6.

[22]Lan, Tianxiang and Dong Yin. "A Lightweight Backbone Used for Scene Text Recognition." *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)* (2022): 261-265.

[23] Zhuang, Junbin, YixuanRen, Xia Li and Zhanpeng Liang. "Text-Level Contrastive Learning for Scene Text Recognition." *2022 International Conference on Asian Language Processing (IALP)* (2022): 231-236.

[24]Zhong, Dajian, ShujingLyu, PalaiahnakoteShivakumara, Bing Yin, Jiajia Wu, Umapada Pal and Yue Lu. "SGBANet: Semantic GAN and Balanced Attention Network for Arbitrarily Oriented Scene Text Recognition." *European Conference on Computer Vision* (2022).

.