IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 36, No. 3, June, 2022, pp. 489-495
Journal homepage: http://iieta.org/journals/ria

Word Recognition Method Using Convolution Deep Learning Approach Used in Smart Cities for Vehicle Identification

Srikanth Bhyrapuneni*, Anandan Rajendran

Department of Computer Science and Engineering, VELS Institute of Science, Technology and Advanced Studies, Chennai 600117, India

Corresponding Author Email: srikanthbt3010@gmail.com

https://doi.org/10.18280/ria.360318

Received: 7 April 2022
Accepted: 21 June 2022

Keywords:
smart cities, word recognition, number recognition, classification, clustering, image, text files

ABSTRACT

Recognizing words or numbers from an image or a text is a complex issue yet useful in different security applications. Dynamic coordinating between characters of a vocabulary passage and segment(s) of the information picture is utilized to rank the dictionary sections arranged by best match. In the proposed work words or numbers are recognized from images or text files or handwritten text, as this application can be applied to smart cities for improving security to recognize the new vehicles entering into the city. Traditional strategies frequently neglect to result acceptable outcomes in identification. Hence, in this proposed work, we propose to consolidate Key Pixel Locator (KPL) in an image and combine it with Convolutional Neural Network (CNN) to accomplish great recognition rate and identification rate. The characters are identified in a word from the vehicle number plate and the data extracted can be verified to recognize the vehicle and its owner details so that vehicle detection can be easy in case of theft or any criminal vehicle. Exploratory outcomes demonstrate that our methodology utilizing the variable length beats the strategy utilizing fixed span as far as both exactness and speed. Speed of the whole recognition procedure is around 200 ms and the recognition exactness is 97% is accomplished.

1. INTRODUCTION

A word is essentially an arrangement of characters; a common approach to word recognition is to divide the word into characters and recognise the individual characters using optical character recognizers (OCR). In many cases, it is reasonable to trust that a vocabulary is provided. A word recognition strategy based on word models (rather than character models) is proposed. The essential concept of this methodology is the early incorporation of the dictionary into the recognition procedure [1]. The word picture is contrasted with only words from the dictionary, eliminating the need for post processing. Since vocabularies are little in a larger part of utilizations this is an appealing methodology [2]. The procedure to recognize the words are clearly represented in Figure 1.

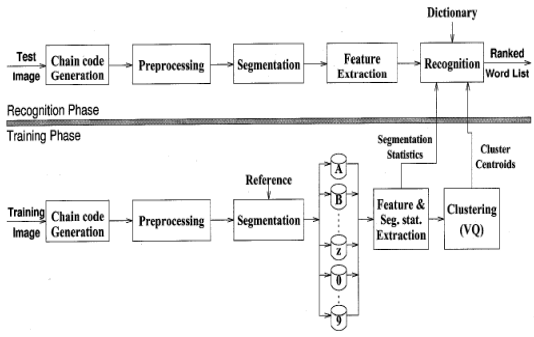

Figure 1. Word/Number Recognition Process

In Figure 1, the word recognition process is discussed. From the image, noise is removed after pre processing and then image segmentation is performed. From the segments, features are extracted and the word recognition is done based on the extracted features. Additionally, the idea of variable span, which is acquired from character division insights and utilized for deciding the size of coordinating window during the recognition, is presented. The variable span boosts the effectiveness of the vocabulary driven methodologies of the transcribed word recognition regarding both speed just as recognition precision [3]. An automatic recognition system's output could be cleaned up by detecting and correcting any word-level disfluencies as they occur. Contextual information is used in conjunction with a span-based training approach to identify disfluencies.

A linear mathematical model has a unique set of vectors called the eigenvectors. Eigenvectors have been employed extensively in the field of computer vision because of their unique properties. Information retrieval can be improved by applying the eigenvector to the image data corpus. The eigenvector can be used in word and character analysis, as demonstrated in this research.

A vigorous picture dealing with a wide range of content is proposed in this work. Chain code portrayal of shapes of word pictures is utilized for productive picture processing. In the proposed work, regular tests are performed to improve the recognition rate while not expanding the recognition time significantly. The primary parts of the recognition and check approach is demonstrated in the Figure 2. The hidden layers are represented in Figure 3.

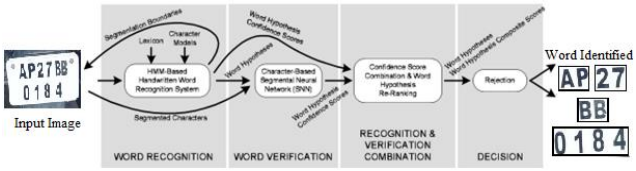

Figure 2. WORD recognition and Identification Framework

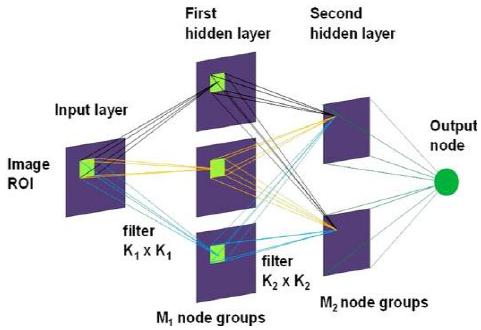

Figure 3. Hidden layers representation

We present a computational model that mimics the strong word recognition mechanism that is important and useful for numerous NLP tasks. Besides simple letter transpositions, additional sorts of noise were examined in further research on robust word processing mechanisms by humans. We conduct spelling correction experiments to see if the model can accurately identify the jumbled words. The proposed model may be able to decipher phrases with jumbled words if its powerful word-processing method is right.

The proposed approach utilizes the division theories given with KPL model identification framework, which is one of the qualities of the KPL approach, and returns to the information word to separate new features from the sections. A joined score of KPL and CNN is utilized to choose the best theory. This is another method for consolidating both KPL and CNN to accomplish great recognition rate and which varies from the current techniques.

2. RELATED WORK

Bai et al. [1] proposed the scene content getting issues. Numerous works in this period have been planned in power minimization systems. The arrival of testing open datasets, for example, SVT, IIIT-5K word assembled a great deal of consideration among the PC vision network.

Gomez et al. [2] proposed a convolutional neural system (CNN) for scene content acknowledgment. Their method is prepared on information assortment containing 8 million pictures. This technique altogether improved the best in class of scene content acknowledgment. Contemporary to this strategy,

He et al. [3] proposed a Rough Neural Network (RNN) structure for perceiving scene content. Here, authors speak to word pictures into a consecutive histogram of slope features, and a RNN is utilized to characterize these successive features.

Li et al. [4] proposed a strategy that build a binarization limit dependent on worldwide measurements of the picture, for example, intra-class change of content and border area, though nearby strategies register binarization edge dependent on neighbourhood insights of the picture for example, mean and fluctuation of pixel levels in patches. The peruser is urged to allude for more details of these techniques.

Zhan et al. [5] proposed four stages: denoising with a low-pass Wiener channel, good content and foundation guessing, using content and foundation gauge to predict nearby bounds, and present handling. Shi et al. [6] proposed a new administrator known as the stroke width change. The administrator records the stroke width at each pixel of the information image. At that moment, many heuristics are used to extract content.

Cheng et al. [7] offered a section-based tree-organized model for character differentiation proof, but it required point-by-point explained datasets to prepare. The dictionary-free techniques are somewhat bad performers, but appealing because they don't require a lexicon and are suitable for identifying characters from photos.

Toledo et al. [8] proposed a solo element learning technique for content and character order by mining important patches from the dataset and blending descriptors for the classifiers. The pruning technique based recognition strategy was improved with quick steady descriptors for character acknowledgment giving an ongoing start to finish framework. A single irregular classifier for location and acknowledgment was exhibited which depended on word reference search to address acknowledgment results.

There are two organizing stages: processing and recognition. Information pictures experience the means of generating chain code, pre-processing, clustering and feature extraction in the two stages.

1) Generating Chain code step changes over the paired picture contribution to a chain code portrayal by coding the limit shapes of parts in the picture while protecting the positional and directional data of neighbouring pixels [4]. The display is characterized for proficient portrayal and control of information. Single pixel segments are identified and expelled.

2) Pre-processing step incorporates noise removal, remove unwanted data [5]. Noise presented by digitizing gadgets and transmission media, is wiped out by contrasting the size of associated parts and estimation of normal stroke width. Inclination point is evaluated by averaging direction edges of "vertical" strokes and moving the x-directions of segments in like manner.

3) Clustering step restores the division focuses to be utilized for gathering at least one cluster to shape important characters [6, 9]. The division focuses are resolved utilizing a mix of ligatures and concavity includes on the shape. Normal stroke width of a picture is evaluated and utilized in a versatile manner to decide the features.

4) Feature extraction step produces include vectors for combination of portions that are conjectured to be characters [7]. Worldwide and neighbourhood features are characterized and extricated from sub images of the cluster by their chain code.

5) Training stage utilizes a preparation set comprising of parallel word pictures unique in relation to the test set [8]. The references of character division focus in the info picture are resolved physically. Division measurements, for example, how regularly a specific character is part into what number of is accessed.

6) Recognition stage utilizes division measurements, character group centroids, a word reference, and feature vectors got from the test picture [10]. A dynamic coordinating plan is utilized to think about features of a portion or a blend of back to back sections with the group of centroids of a character in a vocabulary passage. This methodology is

utilized to rank the vocabulary sections

The division procedure at first recognizes associated segments. Some straightforward gathering and noise evacuation is performed [11]. The outcomes are alluded to as the underlying fragments. A component of an underlying division is commonly a noteworthy associated part in the word, or a gathering of associated segments. Those segments which are not "bars, (for example, the highest point of a "T" or the vertical bar in a "D") are sent to a part. The image and its segments are shown in Figure 4.

Figure 4. A word picture and its segments

Procedure which is intended to part associated segments comprising of various characters into basic fragments. The requirement for forceful part because of uncertainty of characters is very much archived somewhere else and won't be talked about here. The aftereffects of part are at that point used to shape the basic fragments [12]. A word picture and the subsequent basic sections. The division module portrayed varies from the division module depicted here in four different ways. The first is the output design. In the present methodology, dynamic writing computer programs are utilized to produce ideal divisions "on the fly" however in the referenced work, various divisions are framed in the division procedure and they stay static during the coordinating procedure [13]. The second real distinction is that a large number of the standards for gathering and part have been made less prohibitive so as to deal with unconstrained words instead of just handprinted words [14]. The previous division calculation incorporated a severe recognition module: an underlying section.

2.1 Handwriting & computer words recognition system (HCRS)

The pre processing stage takes out some inconstancy identified with the composition procedure and that isn't huge from the perspective of recognition, for example, the variableness because of the composition condition, composing style, procurement, and digitization of picture, and so on. The clustering technique plays out an explicit clustering of the words that intentionally proposes a high number of division focuses [15], offering, along these lines, a few division alternatives, the best ones to be approved during recognition.

This system may create accurately divided, under fragmented or over sectioned characters. Dissimilar to segregated character recognition, vocabulary driven word recognition methodologies don't expect features to be exceptionally separating at the character level in light of the fact that other data, for example, setting, word length, and so on for a character, or in excess of a character [16].

2.2 Identification method

The general issue of perceiving a manually written word w or, proportionally, a character succession obliged [17] to spellings in a dictionary L, is normally encircled from a measurable viewpoint where the objective is to discover the arrangement of names$e_1^L = (e_1 e_2 \dots e_L)$Lþ that is no doubt, given a grouping of T discrete perceptions $o_1^T = (e_1 e_2 \dots e_T)$: that are represented as:

$$\hat{w} \ni P(\hat{w}|o_1^T) = \max_{w \in L} P(w|o_1^T) \quad (1)$$

The word posterior probability is represented as:

$$P(\hat{w}|o_1^T) = \frac{P(w|o_1^T)P(w)}{P(o_1^T)} \quad (2)$$

where, $P(\hat{w}|o_1^T)$ is the probability of the word happening, which relies upon the terminology utilized and the recurrence includes in the preparation informational collection. The likelihood of information happening is obscure, however accepting that the word w is in the vocabulary L and that the decoder registers [18] the probabilities of the whole arrangement of potential speculations, at that point the probabilities must aggregate to one that is represented as:

$$\sum_{w \in L} P(w|o_1^T) = 1 \quad (3)$$

In such a way, evaluated back likelihood can be utilized as certainty gauges [19]. We acquire the back $P(w|o_1^T)$ for the word hypotheses equivalent.

$$P(\hat{w}|o_1^T) = \frac{P(w|o_1^T)P(w)}{\sum_{w \in L} P(o_1^T|w)P(w)}, \quad (4)$$

3. PROPOSED METHOD

3.1 Pre-processing

Before processing our models with the dataset, pre-processing and information increase procedures are applied on the dataset so as to make our information progressively good with the models and to make our dataset increasingly strong to genuine circumstances [20]. The dimensionality reduction method is used in the proposed technique to eliminate the noisy data and consider only useful data and hierarchy generation model is used for arranging the data in a order.

3.2 Word recognition using key pixel locator

In the Key Pixel Locator, from the input image provided only the key pixels which form a word are extracted and maintained as a record. From an image ‘IM’, the words ‘w’ are extracted as:

$$IM=(W_1, W_2, \dots\dots,) \quad (5)$$

where, Wε DS(IM).

From the words identified, the pixels are extracted from the machine by using trained dataset as a reference and then the order of the pixels 'O' are identified and weights are assigned. The process is depicted as:

$$PE\left(W,T_{set}\right)=\frac{1}{2}\sum_{(\mathbf{x}_i,y_i)\in T_{set}} IMO_k\left(\mathbf{x}_i\right)-y_i^{\,2} \tag{6}$$

The training set is represented as:

$$\mathrm{L(D|\ W)}=-\sum_{\left(G,\mathrm{y}^*\right)\in\mathcal{D}}\log \mathrm{P}\left(\mathrm{y}^*|\ G\right) \tag{7}$$

where, *L* is the optimizing objective, *D* is the dataset, and *W* is the word, *G* is the set of words to be trained.

The edges of the images are eliminated and unwanted pixels are removed using the equation by considering vectors *v*.

$$\begin{aligned} E(G,\mathbf{y}) &= \sum_{v\in V}\mathbf{w}^{VT}\cdot\Phi_v\left(y_v,G\right) \\ &+\sum_{(u,v)\in E}\mathbf{w}^{ET}\Phi_{(u,v)}\left(y_u,y_v,G\right) \end{aligned} \tag{8}$$

The boundary limits of the word size are calculated as:

$$\partial\mathcal{R}=\{p\in\mathcal{D}\setminus\mathcal{R}:\exists q\in\mathcal{R}:pAq\} \tag{9}$$

3.3 Grouping of recognition and verification scores

In this system, inputs are N-best rundown score from the individual classifiers and the score of both classifiers are generally made standardized before combination.

$$CS(B_n)=\frac{b_n}{\sum_{n=}^{N} b_n} \tag{10}$$

where, $CS(B_n)$ represents assurance score of n^{th} hypothesis and b_n corresponds score of entity one (here N=10).

3.4 Padding images

The pictures are of various sizes in light of the fact that various words are of various lengths and sizes. For example, the picture of the word 'machine' has a lower width than the picture of the word 'methodologies' due to the length of the words. So also, the picture sizes contrasted among pictures because of the height of the characters. Most importantly, the angle proportion of each picture is extraordinary. This implies on the off chance that we needed to set the width of a picture to a particular worth, the other measurement would have been diverse for the various pictures with various perspective proportions.

Also, in the event that we resized the picture on the height and width measurements freely, at that point the picture would get misshaped, and imagine that this again would adversely influence proposed model since the natural qualities of the image are being lost. In this manner, cushion the pictures with whitespace to the most extreme width and height present in our dataset are performed. At the same time, the blank area was included uniformly the two sides of the stature and the width measurements. The rotation of images for accurate identification is shown in Figure 5.

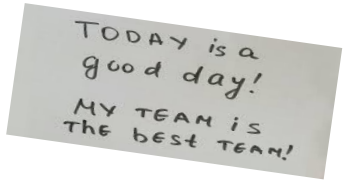

Figure 5. Rotating of images for Identification

3.5 Rotation of pictures

This event happens all around as often as possible, all things considered, regardless of whether the page has lines, to make information increasingly vigorous to this issue by pivoting a picture towards the privilege by an exceptionally little point with irregular likelihood and adding that picture to our preparation set. This information enlargement method helped us to make the model progressively powerful to some minor yet so regular refinement that may come up in the test set. The word recognition indexing process is shown in Figure 6.

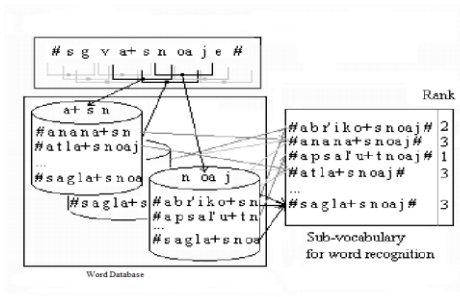

Figure 6. Word recognition indexing process

We characterize oa as component vector comparing to l^{th} word portion and ve as character class of l^{th} word fragment given by HRS. The word recognition framework is shown in Figure 7.

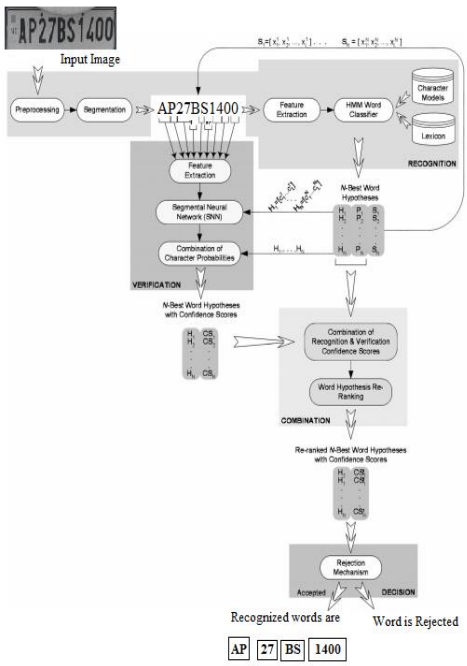

Figure 7. Coordination of word recognition framework and choice stages

In the proposed methodology multiple vehicle details are extracted simultaneously and then the vehicles data are sent to central station for verification and checking miscellaneous vehicles entering into the city. The proposed method multiple image extraction and word conversion is depicted in the Figure 8.

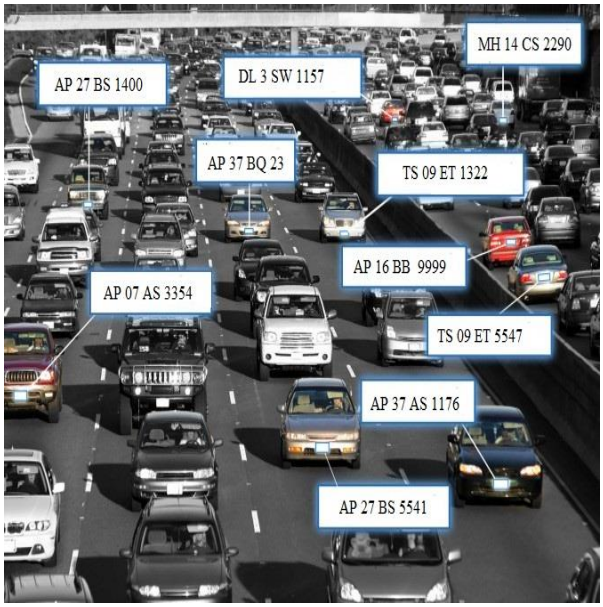

Figure 8. Multiple vehicle numbers recognition

3.6 Optimization function

In the proposed work for identification of hidden layers and the pixels, optimization activation functions are used for improving the accuracy of the system. In the proposed work, Leaky ReLU function is used which is better in improving the accuracy rate.

Unstructured text data is processed and analysed using a new field termed 'text mining'. Classifying a vast volume of text data is the most critical step in the evaluation process. As a result, we place a high value on text classification, one of the most widely used text mining techniques. The text data is sorted into groups based on its similarity. Thousands or even millions of features are required for each training instance in the text classification task, which uses a number of features to represent a document. That is, the dataset's dimensionality is determined by the number of features. Having so many input options slows down and complicates training. The curse of dimensionality is the term used to describe this issue. Solving this problem is the focus of a broad scientific topic known as dimensionality reduction.

Leaky ReLU function is an enhanced version of ReLU function. In ReLU function, the gradient is 0 for x<0, which made the neurons die for activations in that region. Leaky ReLU is defined to address this problem. Instead of defining the ReLu function as 0 for x less than 0, we define it as a small linear component of x. It can be defined as:

$$f(x)=ax,\ x<0$$
$$=x,\ x>=0 \qquad (11)$$

We have replaced the horizontal line with a non-zero, non-horizontal line. Here a is a small value like 0.01 or so. It can be represented on the graph as shown in Figure 9.

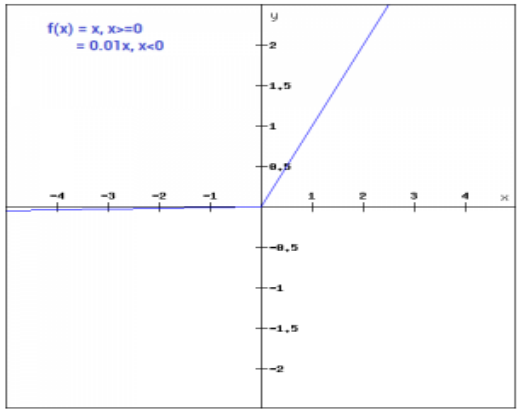

Figure 9. ReLu representation

The main advantage of replacing the horizontal line is to remove the zero gradient. Here the gradient of the left side of the graph is non zero and so we would no longer encounter dead neurons in that region. The gradient of the graph is represented as in Figure 10.

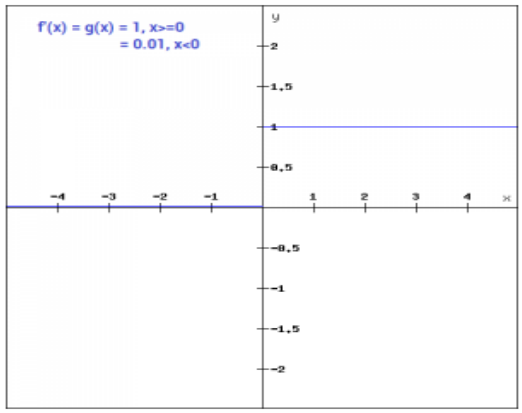

Figure 10. Leaky ReLu representation

4. RESULTS

The proposed word or number recognition method is implemented in Python and executed in ANACONDA platform. The proposed method deals with word or number recognition on the vehicles entering into a city for providing security by identifying the danger vehicles, where content pictures are of various colours, sizes, textual styles, alongside variation sorts of damage.

A point by point investigation of execution on datasets is seen as far as state and Gaussian number for each character models. From the examination, it is noticed that, sliding window width of 6 pixels with covering proportion half gives better outcome.

We have done four distinct kinds of trials: recognition of recorded words, recognition of separated manually written words, mix of identification and checking to streamline the general identification rate of word speculations to advance the general dependability of the identification procedure. In any case, in this proposed work, the tests identified with the recognition, confirmation, and notice of written by hand words are done. The Figure 11 depicts the KPL and CNN based recognition model.

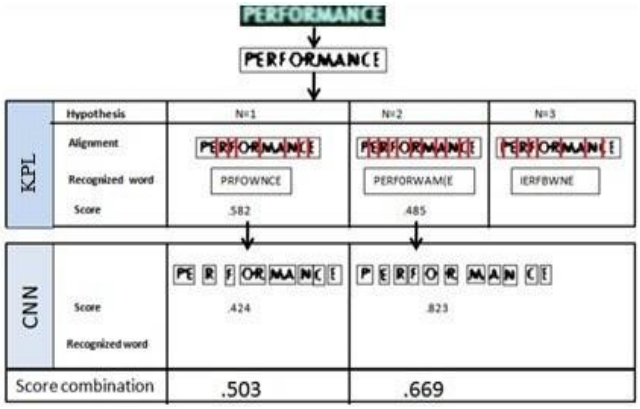

Figure 11. Score grouping from KPL and CNN based recognition

The investigations identified with the recognition of disengaged manually written words are portrayed in this work. Table 1 shows the word recognition rate and recognition time. To assess the outcomes, the following measures are utilized: recognition rate, error rate, identification rate, and unwavering quality, which are characterized as:

$$Recognition\ Rate = \frac{N_{RG}}{N_{TD}} \times 100$$
$$Error\ Rate = \frac{N_{ERR}}{N_{TD}} \times 100 \quad (12)$$

Table 1. Word recognition rate and recognition time

Lexicon Size	Word Recognition Rate (%) TOP 1	TOP 5	TOP 10	Recognition Time (sec/word)
10	97.84	98.96	100	0.010
1,000	92.01	97.32	96.71	0.372
10,000	84.06	93.58	94.36	1.996
40,000	72.23	83.86	85.63	7.392
80,000	66.65	80.32	83.10	13.25

$$Rejection\ Rate = \frac{N_{RE}}{N_{TES}} \times 100$$
$$Reliability = \frac{N_{RG}}{N_{RG} + N_{ERR}} \times 100 \quad (13)$$

where, N_{RE} is count of words exactly classify, N_{ERR} is count of words misclassified, N_{REJ} is count of words rejected, and N_{TD} is count of words tested.

4.1 Performance of the convolutional neural network (CNN)

The presentation of the CNN was assessed on a database of disengaged written by hand characters got from the SRTP database. The preparation set contains 25874 disengaged words with an equivalent circulation among the classes. For those classes with low example recurrence, engineered tests were created by a stroke twisting system. An approval set of 24152 words was additionally utilized during the preparation of the CNN to screen the speculation and a test set of 15985 words was utilized to assess the exhibition of the CNN.

Feature vectors made out of three component types were produced from the character tests. The CNN was actualized by the engineering methods and it was prepared utilizing the back propagation calculation.

Table 2. Character recognition rate

Dataset	SRTP Number of Samples	SRTP Recognition Rate (%)	Proposed Method Number of Samples	Proposed Method Recognition Rate (%)
Training	84,760	77.54	74,880	94.76
Validation	36,170	72.47	23.670	90.45
Test	46,700	75.68	23,941	86.45

Table 2 demonstrates the word recognition rates accomplished on the preparation, approval, and the SRTP database when the CNN was utilized as standard classifier.

4.2 Performance of the verification stage (VS)

A few diverse test sets were created from the N-best word speculation records to survey the exhibition of the VS. Having the division limits of each word theory, we come back to the word picture to remove new features from each portion speaking to a character. The new component vectors are utilized in the check procedure.

Table 3. Word recognition rates

Lexicon Size	Proposed Method Word Recognition Rate (%) Average Level 1	Average Level 5	Average Level 10	Product Level 1	Product Level 5	Product Level 10
10	92.04	97.64	100	81.34	84.47	100
1,000	86.35	96.58	96.45	77.54	80.76	84.79
10,000	85.76	93.47	93.75	75.45	74.46	79.45
40,000	77.56	84.89	85.37	68.57	68.95	75.93
80,000	74.87	82.47	87.98	65.35	72.64	70.67

The results in Table 3 show that the word recognition rate of the proposed method is high than the traditional methods. The word recognition rate is shown in Figure 12.

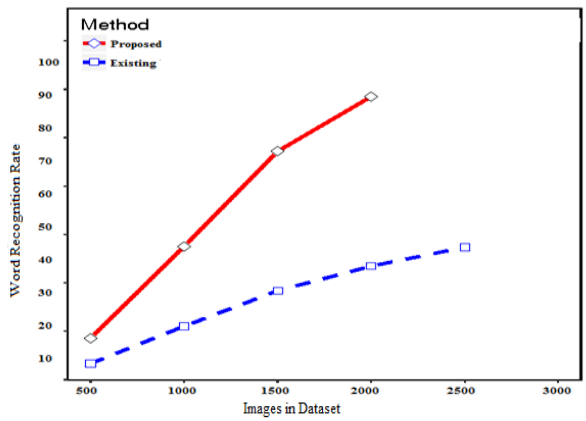

Figure 12. Word recognition rate

5. CONCLUSIONS

The combination of KPL and CNN is presented as a word-recognition technique. According to our evaluation, the combination technique outperforms other autonomous systems. The results confirm that the framework proposed is promising for true application when material appears in handwritten or normal graphics. It is recommended that two different methods of order be combined, each of which can be used in a different depiction space. While the deferral in the

overall process is negligible, the recognition rate achieved by combining the recognition and confirmation approaches is fundamentally superior than that achieved by the KPL framework alone. The combination that has been proposed is both workable and efficient in terms of processing power. The Proposed method is effective in identification of words or numbers from the images provided with different styles and fonts and this application can be implemented in smart cities for identifying different kinds of vehicles entering into the city to improve safety measures in the city. The proposed method results in 97% accuracy rate in identification of the numbers or words from the provided input.

REFERENCES

[1] Bai, F., Cheng, Z., Niu, Y., Pu, S., Zhou, S. (2018). Edit probability for scene text recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 1508-1516. https://doi.org/10.1109/CVPR.2018.00163

[2] Gómez, L., Karatzas, D. (2017). Textproposals: A text-specific selective search algorithm for word spotting in the wild. Pattern Recognition, 70: 60-74. https://doi.org/10.1016/j.patcog.2017.04.027

[3] He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., Li, X. (2017). In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 3047-3055. https://doi.org/10.1109/ICCV.2017.331

[4] Li, H., Wang, P., Shen, C. (2017). Towards end-to-end text spotting with convolutional recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 5238-5246. https://doi.org/10.1109/ICCV.2017.560

[5] Zhan, H., Wang, Q., Lu, Y. (2017). Handwritten digit string recognition by combination of residual network and RNN-CTC. In International Conference on Neural Information Processing, pp. 583-591. https://doi.org/10.1007/978-3-319-70136-3_62

[6] Shi, B., Bai, X., Yao, C. (2016). 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, 39(11): 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371

[7] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S. (2017). Focusing attention: Towards accurate text recognition in natural images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 5076-5084. https://doi.org/10.1109/ICCV.2017.543

[8] Toledo, J.I., Dey, S., Fornés, A., Lladós, J. (2017). Handwriting recognition by attribute embedding and recurrent neural networks. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, pp. 1038-1043. https://doi.org/10.1109/ICDAR.2017.172

[9] Bai, X., Yao, C., Liu, W. (2016). Strokelets: A learned multi-scale mid-level representation for scene text recognition. IEEE Transactions on Image Processing, 25(6): 2789-2802. https://doi.org/10.1109/TIP.2016.2555080

[10] He, T., Huang, W., Qiao, Y., Yao, J. (2016). Text-attentional convolutional neural network for scene text detection. IEEE Transactions on Image Processing, 25(6): 2529-2541. https://doi.org/10.1109/TIP.2016.2547588

[11] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A. (2014). Deep structured output learning for unconstrained text recognition. arXiv preprint arXiv: 1412.5903. https://arxiv.53yu.com/abs/1412.5903

[12] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. International Journal of Computer Vision, 116(1): 1-20. https://doi.org/10.1007/s11263-015-0823-z

[13] Johnson, J., Karpathy, A., Li, F.F. (2016). Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565-4574.

[14] Gordo, A. (2015). Supervised mid-level features for word image representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2956-2964.

[15] Almazfian, J., Gordo, A., Fornfies, A., Valveny, E. (2014). Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(12):2552-2566. https://doi.org/10.1109/TPAMI.2014.2339814

[16] Li, H., Wang, P., Shen, C. (2017). Towards end-to-end text spotting with convolutional recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5248-5256. https://doi.org/10.1109/ICCV.2017.560

[17] Gupta, A., Vedaldi, A., Zisserman, A. (2016). Synthetic data for text localisation in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315-2324. https://doi.org/10.48550/arXiv.1604.06646

[18] Liao, M., Shi, B., Bai, X., Wang, X., Liu, W. (2017). Textboxes: A fast text detector with a single deep neural network. In Thirty-first AAAI conference on artificial intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.11196

[19] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. In European Conference on Computer Vision, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[20] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell., 39(4): 640-651. https://doi.org/10.1109/TPAMI.2016.2572683