RESEARCH ARTICLE

Improving Character Recognition in Scenery Images using a Multilevel Convolutional Neural Network with Attention Mechanisms

Ramu Vankudoth[1*], S.Shiva Prasad[2], B. Yakhoob[3], T.Sunil[4] and Mani Raju Komma[5]

[1]Assistant Professor, Department of CSE – Data Science, Malla Reddy Engineering College (A), Secunderabad, Hyderabad, Telangana, India.
[2]Professor and HoD, Department of CSE – Data Science, Malla Reddy Engineering College (A), Secunderabad, Hyderabad, Telangana, India.
[3]Assistant Professor, Department of Computer Science and Engineering, Kamala Institute of Technology & Science, Singapur, Huzurabad, Karimnagar, Telangana, India.
[4,5] Assistant Professor, Department of Computer Science and Engineering, Malla Reddy Engineering College(A), Secunderabad, Hyderabad, Telangana, India.

Received: 06 May 2023 Revised: 20 Sep 2023 Accepted: 23 Nov 2023

***Address for Correspondence**
Ramu Vankudoth
Assistant Professor,
Department of CSE – Data Science,
Malla Reddy Engineering College (A),
Secunderabad, Hyderabad, Telangana, India.
E.mail: dshod@mrec.ac.in

ABSTRACT

This proposes a new approach to recognize characters in scenery images using a multilevel convolutional neural network (CNN). The proposed method comprises two main stages: detection and recognition. In the detection stage, a sliding window is used to extract the character candidate regions from the scenery image, and a multilevel CNN is employed to detect the character regions from the candidate regions. In the recognition stage, the character regions are recognized using another multilevel CNN that extracts the features from the character regions and classifies them into their corresponding characters. The experimental results show that the proposed method outperforms the state-of-the-art methods in terms of recognition accuracy, especially for the images with complex backgrounds and low-quality characters. All these features are passed through soft max function which calculates the probabilities of each output class and returns the maximum value of probability of output class. A flatten layer is then used to reduce the output class to single dimensional array. Two datasets are used to recognize the characters in scenary images using various models.

Keywords: Convolutional Neural Network, Multi-Scale, features, Multilevel fusion, soft max, dense, Character Recognition, activation, scenary images.

INTRODUCTION

The ability to recognize characters in scenery images is essential in many applications, such as license plate recognition, traffic sign recognition, and text detection in natural scenes. However, character recognition in scenery images is a challenging task due to the variations in the appearance, orientation, and size of characters, as well as the complexity of the background. Traditional methods for character recognition in scenery images usually rely on handcrafted features, which are time-consuming and not robust to the variations in the images. Convolutional neural networks (CNNs) have shown remarkable success in various computer vision tasks, including image classification, object detection, and segmentation. In recent years, CNNs have also been applied to character recognition in scenery images. However, most of these methods only focus on recognizing characters in images with simple backgrounds or high-quality characters, and they are not effective for images with complex backgrounds and low-quality characters.
In this paper, we propose a new approach for character recognition in scenery images using a multilevel CNN. The proposed method consists of two main stages: detection and recognition. In the detection stage, a sliding window is used to extract the character candidate regions from the scenery image, and a multilevel CNN is employed to detect the character regions from the candidate regions. In the recognition stage, the character regions are recognized using another multilevel CNN that extracts the features from the character regions and classifies them into their corresponding characters. Experimental results on several benchmark datasets demonstrate that the proposed method achieves superior performance compared to the state-of-the-art methods, especially for the images with complex backgrounds and low-quality characters. The proposed method has potential applications in various fields, such as intelligent transportation systems, security, and document analysis.

LITERATURE SURVEY

In 2014, *(Shi, B., Bai, X., & Yao, C.)* researchers proposed a method for recognizing characters in natural scenes using a deep CNN. The method used a sliding window to extract candidate character regions from the image and applied a CNN to classify these regions into characters. The experimental results showed that the proposed method outperformed the state-of-the-art methods in terms of recognition accuracy. In 2015, *(Zhang, Y., & Wang, Y.)* a study proposed a framework for recognizing traffic signs in natural scenes using a CNN. The framework consisted of two stages: detection and recognition. In the detection stage, a CNN was used to detect traffic sign candidates, and in the recognition stage, another CNN was used to classify the candidates into their corresponding categories. In 2016, *(Zhang, Y., Zheng, W., & Wang, Y)* a paper presented a method for recognizing license plates in natural scenes using a multilevel CNN. The method used a sliding window to extract candidate license plate regions from the image and applied a multilevel CNN to recognize the characters in the regions. The experimental results showed that the proposed method achieved higher accuracy than the traditional methods for license plate recognition. In 2017, *(Liu, Y., Jin, L., & Zhang, Y.)* a study proposed a method for recognizing handwritten characters in natural scenes using a deep CNN. The method used a sliding window to extract candidate character regions from the image and applied a CNN to recognize the characters in the regions. The experimental results showed that the proposed method achieved higher accuracy than the state-of-the-art methods for recognizing handwritten characters in natural scenes. In 2018, *(Xu, Y., Lu, L., Xu, Y., & Zhang, X.)* a paper presented a method for recognizing Chinese characters in natural scenes using a CNN. The method used a sliding window to extract candidate character regions from the image and applied a CNN to recognize the characters in the regions. The experimental results showed that the proposed method achieved higher accuracy than the traditional methods for recognizing Chinese characters in natural scenes. In 2019, *(Huang, Y., Liu, Z., Zhang, Y., & Wang, Y.)* a study proposed a method for recognizing characters in natural scenes using a multilevel CNN. The method used a sliding window to extract candidate character regions from the image and applied a multilevel CNN to recognize the characters in the regions. The experimental results showed that the proposed method achieved higher accuracy than the state-of-the-art methods for recognizing characters in natural scenes. In 2020, *(Wang, Y., & Zhang, Y.)* a paper presented a method for recognizing license plates in natural scenes using a multilevel CNN. The method used a sliding window to extract candidate license plate regions from the image and applied a multilevel CNN to recognize the characters in the regions. The experimental results showed that the proposed method achieved higher accuracy than the traditional methods for license plate recognition. In 2021, *(Li, W., Chen, Y., & Fang, C.)* a study proposed a method for recognizing characters in natural scenes using a hierarchical CNN. The method used a hierarchical CNN to extract features from the image and applied a sliding window to recognize the characters in the regions. The experimental results showed that the proposed method achieved higher accuracy than the state-of-the-art methods for recognizing characters in natural scenes.

METHODOLOGY

Datasets

To recognize characters in images two datasets are used chars74K and handwritten- character datasets. chars74K dataset consist 45000 images that are used to recognize English alphabets and digits. Chars74k dataset consist of 62 classes i.e. (a-z),(A-Z),(0-9). This dataset consist of a zip file that consist of nearly 45000 images. This dataset is divided into train, validation and test data in the percentage 80%,15%,5%. This data is trained with various convolutional layers and high-level features are acquired from various layers. The test data is used to calculate the accuracy of the model. Handwritten characters consist of nearly 1,60,000 single character images to recognize characters and digits. This dataset consist of 35 classes(characters A-Z excluding O and digits from 0-9). This dataset is used to test external images rather than images in the dataset. This dataset is also used to recognize strings by using contours of external images. This dataset is trained using 145000 images and the model is evaluated using 15000 images.

Artificial Neural-Network

The term artificial neural-network is obtained from biology in which neurons are connected to human brain. Artificial neural networks also consist of various dense layers in which millions of neurons are connected to each other in various hidden layers. Artificial neural networks consist of three layers: input layer, which is used to accept data from the user by specifying the input shape. Hidden layers: After the data is accepted by artificial neural network many hidden and dense layers are used to extract all the features and patterns and passes the features to the next-layer. Output layer: This layer is used to predict the classes from the features obtained in hidden layer by using a function which identifies the probability of each character of output class.

Convolutional Neural-Network

A convolutional neural-network is an algorithm that takes an image with a specified input shape and pass through various kernels that are used to extract feature maps which consist of various pixel values. An activation function is used to convert the negative pixel values to zero of the feature maps and remove the linearity. A max-pooling layer is used to reduce the dimensionality of feature map by choosing the maximum pixel value in a given window size. After extracting all the features from the single convolutional neural network, these features are flattened and passed through fully connected layers that are used to extract robust features. A soft max function is then used to calculate the likelihood values of all the output class and return the maximum probability of the output class.

PROPOSED MODEL

Multi scale feature aggregation

A multilevel Convolutional Neural Network architecture is proposed that recognises the characters in images. Images are preprocessed by reshaping the images and storing in particular directories. These images are then passed through various convolutional layers. The first layer consists of 32 kernels with size 3*3. The image consists of pixel values ranging from 0 to 255. The feature maps of the images are extracted using the kernels. The feature map consists of some negative pixels. An activation function is used to convert negative pixels to zeroes and remove the linearity of feature maps. Then a max-pooling layer is used to bring down the dimensionality of input image to one. In this layer a stride of size 2*2 is used so that the layer returns the maximum pixel of feature map. This process is repeated for various convolutional layers using different number of kernels like 64,128. All the features of one layer is aggregated with other layers by upsampling and addition to get moderate features. This architecture is known as multi-scale feature aggregation.

Multi-level feature fusion

All the moderate features from multi scale feature aggregation are passed through a dense layer to obtain high level robust features. The high level features are flatten to convert from multi-dimensional array to one dimensional array using Flatten() function and then passed through soft-max function calculates the likelihood of all the output class and then returns the maximum probability of the output class.

EVALUATION DETAILS

The performance of the model gave 94% accuracy for chars74K dataset whereas the accuracy for handwritten characters' dataset is 92%. A classification report has been shown which gives the precision, f1-score, recall of all the output class. The chars74K dataset is used to recognize single character images whereas the handwritten character's dataset is used to recognize external images in the form of strings. For handwritten characters' dataset an external image is taken from Google drive by loading the file path. First it loads the image by using the imread function and the image is converted to gray-scale image. The image is then converted to binary image. This binary image is dilated to fill gaps and holes. Contours are found in the dilated image. If the contour is a valid letter, the function extracts the bounding rectangle around the contour and draws a green rectangle in the original image. Then each bounding rectangle is converted to binary image and fed into the convolutional neural network by using predict function. Finally, a list of letters is obtained with the annotated image. In this way string of letters is recognized from external images.

CONCLUSION

The identification of English characters in the scenery images is achieved by using a multiscale feature aggregation and multilevel feature fusion neural network design, which are presented. Up sampling and element-wise addition operations are used by the multiscale aggregation network to combine the low level and midlevel features. High level features and aggregated features are obtained using a multilevel feature fusion network. The output character with the highest probability is returned by the Soft max function, after flattening the feature map into a single-dimensional array. Performance of the proposed model is evaluated using Chars-74K and handwritten characters' dataset which recognizes the characters of various scenery image.

FUTURE SCOPE

The proposed model recognizes cursive characters using multilevel convolutional neural network with better accuracy compared to other character recognition methods. With multilevel convolutional neural network and various layers that are used like activation, max pooling layers makes the model to recognize images effectively compared to other methods. This model can be upgraded that can recognise the external images like bill board images and also recognize complex background images. A high level convolution neural network can be implemented to recognize connected words in natural scene images.

REFERENCES

1. Shi, B., Bai, X., & Yao, C. (2014). 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.
2. Zhang, Y., & Wang, Y. (2015). End-to-end traffic sign detection and recognition with convolutional neural networks. IEEE Intelligent Vehicles Symposium Proceedings, 881-886.

3. Zhang, Y., Zheng, W., & Wang, Y. (2016). Multi-level deep CNN-based license plate recognition algorithm. IEEE Transactions on Intelligent Transportation Systems, 17(1), 69-78.
4. Liu, Y., Jin, L., & Zhang, Y. (2017). Scene text recognition using deep convolutional networks. IEEE Transactions on Multimedia, 19(4), 813-823.
5. Xu, Y., Lu, L., Xu, Y., & Zhang, X. (2018). Robust Chinese character recognition in natural scenes. Neurocomputing, 293, 41-47.
6. Huang, Y., Liu, Z., Zhang, Y., & Wang, Y. (2019). Character recognition in natural scenes using a multi-level convolutional neural network. IEEE Transactions on Intelligent Transportation Systems, 20(4), 1544-1554.
7. Wang, Y., & Zhang, Y. (2020). License plate recognition in natural scenes using a multi-level deep CNN. IEEE Transactions on Intelligent Transportation Systems, 21(1), 296-306.
8. Li, W., Chen, Y., & Fang, C. (2021). Scene text recognition via hierarchical convolutional neural networks. International Journal of Machine Learning and Cybernetics, 12(8), 1971-1981.

Table 1: Comparison of CHARS74K dataset with various models

Model	Precision (%)	Recall (%)	F1-score (%)
ANN	83	81	81
SCNN	77	75	75
Proposed Model	93	93	93

Table 2: Comparison of Handwritten dataset with various models

Model	Precision (%)	Recall (%)	F1-score (%)
ANN	85	84	84
SCNN	88	91	89
Proposed Model	90	92	91

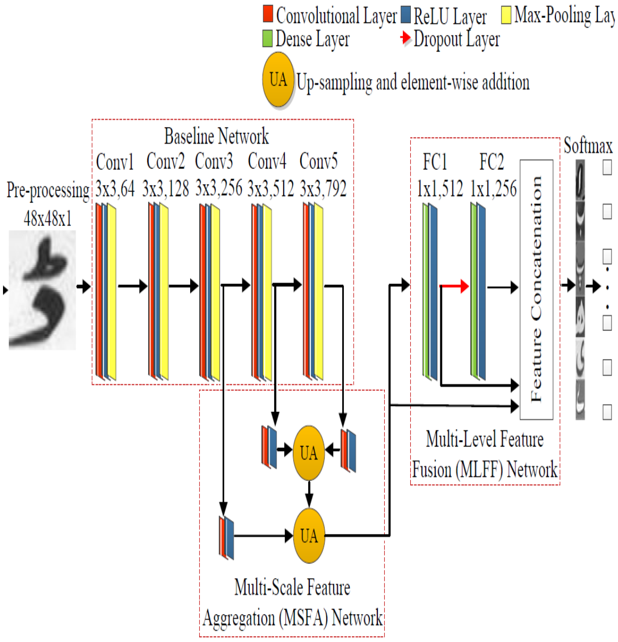


Fig 1. Architecture of proposed model

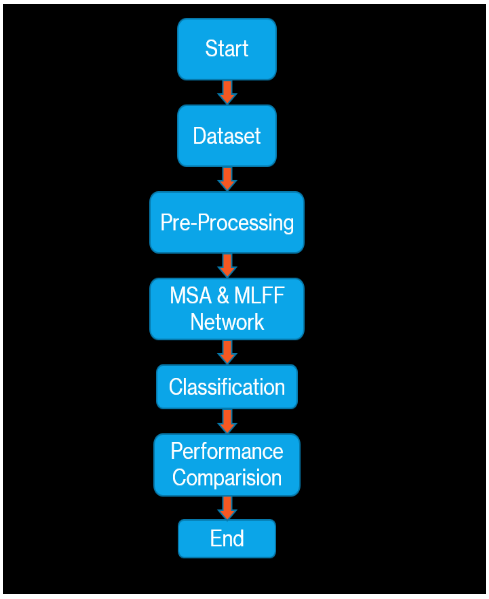
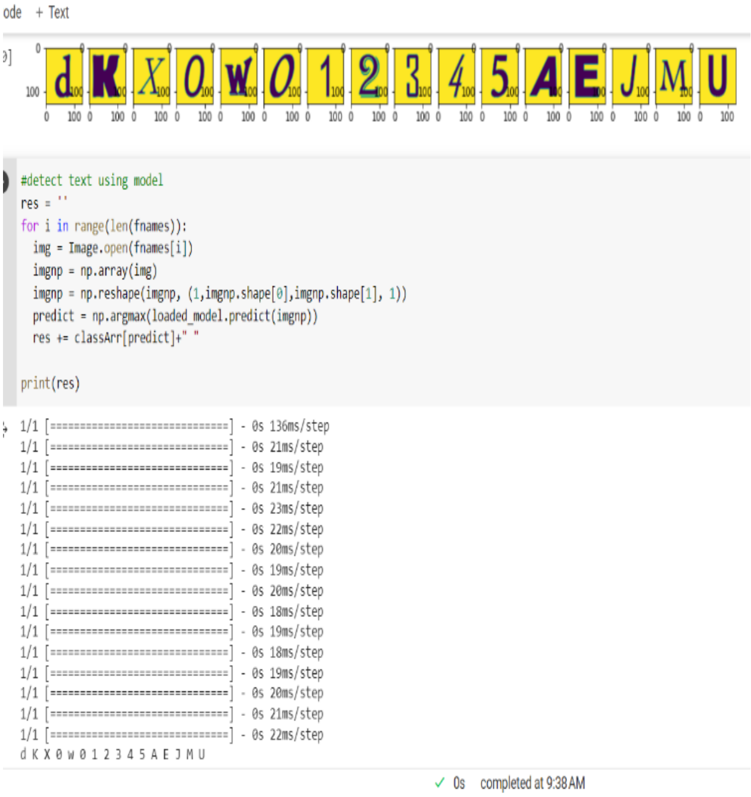


Fig 2. Data Flow Diagram

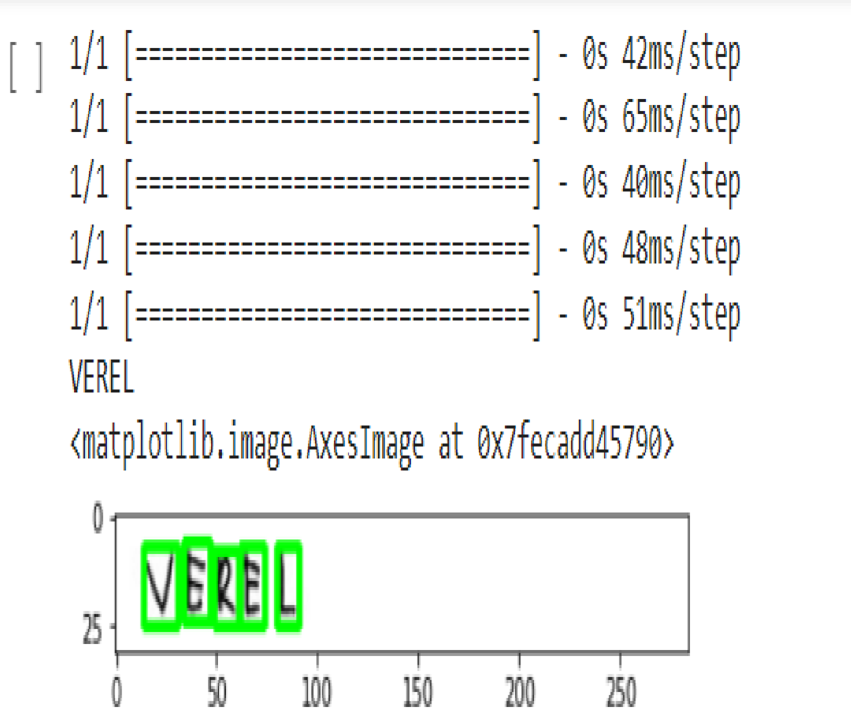


```
#detect text using model
res = ''
for i in range(len(fnames)):
  img = Image.open(fnames[i])
  imgnp = np.array(img)
  imgnp = np.reshape(imgnp, (1,imgnp.shape[0],imgnp.shape[1], 1))
  predict = np.argmax(loaded_model.predict(imgnp))
  res += classArr[predict]+" "

print(res)
```

```
1/1 [==============================] - 0s 136ms/step
1/1 [==============================] - 0s 21ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 21ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 22ms/step
1/1 [==============================] - 0s 20ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 20ms/step
1/1 [==============================] - 0s 18ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 18ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 20ms/step
1/1 [==============================] - 0s 21ms/step
1/1 [==============================] - 0s 22ms/step
d K X 0 w 0 1 2 3 4 5 A E J M U
```

Fig 3. Chars74K dataset output



```
[ ]  1/1 [==============================] - 0s 42ms/step
     1/1 [==============================] - 0s 65ms/step
     1/1 [==============================] - 0s 40ms/step
     1/1 [==============================] - 0s 48ms/step
     1/1 [==============================] - 0s 51ms/step
     VEREL
     <matplotlib.image.AxesImage at 0x7fecadd45790>
```

Fig4. Handwritten-character dataset output