IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 29, 2020, accepted June 7, 2020, date of publication June 11, 2020, date of current version June 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001605

Cursive Character Recognition in Natural Scene Images Using a Multilevel Convolutional Neural Network Fusion

ASGHAR ALI CHANDIO[1,2], MD. ASIKUZZAMAN[1], (Member, IEEE), AND MARK R. PICKERING[1], (Member, IEEE)

[1]School of Engineering and Information Technology, University of New South Wales, Canberra, BC 2610, Australia
[2]Department of Information Technology, Quaid-e-Awam University, Nawabshah 67450, Pakistan

Corresponding author: Asghar Ali Chandio (a.chandio@student.adfa.edu.au)

ABSTRACT The accuracy of current natural scene text recognition algorithms is limited by the poor performance of character recognition methods for these images. The complex backgrounds, variations in the writing, text size, orientations, low resolution and multi-language text make recognition of text in natural images a complex and challenging task. Conventional machine learning and deep learning-based methods have been developed that have achieved satisfactory results, but character recognition for cursive text such as Arabic and Urdu scripts in natural images is still an open research problem. The characters in the cursive text are connected and are difficult to segment for recognition. Variations in the shape of a character due to its different positions within a word make the recognition task more challenging than non-cursive text. Optical character recognition (OCR) techniques proposed for Arabic and Urdu scanned documents perform very poorly when applied to character recognition in natural images. In this paper, we propose a multi-scale feature aggregation (MSFA) and a multi-level feature fusion (MLFF) network architecture to recognize isolated Urdu characters in natural images. The network first aggregates multi-scale features of the convolutional layers by up-sampling and addition operations and then combines them with the high-level features. Finally, the outputs of the MSFA and MLFF networks are fused together to create more robust and powerful features. A comprehensive dataset of segmented Urdu characters is developed for the evaluation of the proposed network models. Synthetic text on the patches of images with real natural scene backgrounds is generated to increase the samples of infrequently used characters. The proposed model is evaluated on the Chars74K and ICDAR03 datasets. To validate the proposed model on the new Urdu character image dataset, we compare its performance with the histogram of oriented gradients (HoG) method. The experimental results show that the aggregation of multi-scale and multilevel features and their fusion is more effective, and outperforms other methods on the Urdu character image and Chars74K datasets.

INDEX TERMS Cursive text recognition, natural scene Urdu character recognition, multi-scale feature aggregation, multi-level feature fusion, convolutional neural network (CNN).

I. INTRODUCTION

Rapid developments in camera-based portable devices have facilitated the acquisition of a large number of images every day. These images not only contain scenery information but also plentiful textual information, which is one of the most informative sources of information. The text extracted from natural scene images also provides a rich source of semantic information, which is helpful in many real-world applications, such as unmanned vehicle navigation [1], robot navigation [2], intelligent transport systems [3], content-based image retrieval [4], assistance for visually impaired people [5], language translation for foreign tourists [6] and many more [7], [8]. Traditionally, optical character recognition (OCR) techniques have been applied to recognize text from scanned documents, where the text is generally well-formatted and captured in a well-controlled environment. Though OCR methods have shown very good accuracy in

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

scanned document images [9], when applied for character recognition in natural scene images, the performance of OCR is limited [8]. This is generally due to the specific features associated with the text in natural scene images. For example, characters in natural scene images have variations in their font sizes, writing styles, aspect ratios, colors, orientations and complex backgrounds, whereas characters in scanned documents mostly have a unique text color, size, orientation and have clean/plain backgrounds. Figure 1 compares Urdu text in scanned document[1] and natural scene images. In addition, image degradation due to environmental constraints, such as un-even lighting conditions, multi-orientated text and blur, as shown in Figure 2, create major problems in the recognition of natural scene text.

FIGURE 1. A comparison of Urdu text in scanned and natural scene image documents. (a) scanned document (b) natural scene image.

A series of robust reading competitions (RRC) and workshops organized by the International Conference on Document Analysis and Recognition (ICDAR) [10]–[12] have been the catalyst for recent developments in text detection and recognition in natural scene images. In these competitions, a set of standard data-sets was published to facilitate research into the problem. The ICDAR dataset published in [12] was the first dataset where the text was not focused while capturing natural scene images. Hence, text detection and recognition in this dataset is more complex and challenging than other datasets [10], [11]. Althogh most of the datasets published were for English text, recently the RRC competitions and workshops for Chinese [13] and other multi-language text have also been organized [14]. The ICDAR has published a multi-language natural scene image dataset that includes Arabic and eight other languages [14], whereas the datasets, techniques, evaluation protocols and the results achieved for Chinese text detection and end-to-end recognition are reported in [13]. In these ICDAR robust reading competitions, the problem of text extraction is generally divided into four sub-tasks: (i) text detection, (ii) isolated character recognition, (iii) cropped word recognition and (iv) end-to-end text recognition.

The purpose of text detection is to put bounding boxes around all word instances present in an image. This is one of the important requirements for enhancing the accuracy of text recognition. A large number of novel methods and ideas have been proposed to address the problem of text localization [15], [16]. However, locating text in unconstrained natural scene images is still an open problem. The purpose

[1] http://cle.org.pk/clestore/urduocr.htm

FIGURE 2. Scene text images with challenging text.

of isolated character recognition is to convert each character image into an editable label. Initially, traditional OCR techniques were used to address this problem [17], [18], but the performance of these techniques is poor due to the inadequate binarization of natural images with complex backgrounds, the low resolution of the input images, as well as variations in layout and other distortions. Recently, a significant number of methods have been developed that eliminate the binarization process and extract the features directly from the input images and feed to the classifiers [19], [20]. The cropped word recognition problem is considered to be a sequence-to-sequence problem and has been addressed using CNNs [21], conditional random fields [22], hidden markov models [23] and RNNs [24]. This method is more suited to situations where the segmentation of characters is more difficult, i.e., cursive text, degraded text, dot-matrix text or text in low-quality images. Finally, an end-to-end system detects image locations where text exist and then recognizes the text [25], [26]. Though proposed powerful deep CNN techniques have shown good results, it is still an open challenge to achieve satisfactory accuracy for scene text recognition in complex images.

In the literature, most of the work on natural scene character recognition has been performed for English, Chinese

and Indic scripts [7], [8] with very little attention devoted to cursive scripts such as Arabic and Persian [27], [28]. Urdu is the national language of Pakistan. It is a type of cursive language, written in the right-to-left direction, and is derived from the Arabic and Persian scripts. To the best of our knowledge, no research has been conducted for the Urdu language, except our previous works [29]–[31].

The development of an Urdu text detection and recognition system in natural scene images is an important task as more than 500 million speakers in the world speak Urdu/Hindi [32] and a majority of the signs, shop names and advertisement banners in Pakistan and some parts of India carry Urdu text. A scene text recognition system will assist non-Urdu speakers by translating Urdu text into other languages. It can also be helpful for visually impaired people to understand product labels and automated teller machine instructions. A significant number of visual images with Urdu text can be found on the Internet and a scene text recognition system will be helpful for content-based search and retrieval of these images. In addition, an Urdu scene text detection and recognition system will be helpful for the automation of vehicle number plate recognition in Pakistan.

FIGURE 3. Shapes of the Urdu characters. First row: Urdu text characters in isolated, initial, medial and final shapes. Second row: isolated Urdu characters with the same shape as a segment of another character. Third row: Urdu characters with similar shape and basic structure, but a different number or position of dots around the shape.

In Urdu text, characters usually have four shapes: isolated, initial, medial and final as shown in the first row in Figure 3. Depending upon the shape of a character in a ligature, it can appear in any of the above forms. The text may also contain different dialects and writing styles. Therefore, the dimensions of the feature vector of a character in Urdu text are higher than for Latin text. In cursive scripts like Urdu, a word is made by joining two or more characters with each other and may contain one or more ligatures as shown in Figure 4. Each sub-word in Figure 4(b) represents a ligature and the two ligatures form a single word هوٹل pronounced as ''Hotel''. Hence, automatic segmentation of individual characters from a word is more challenging and complex, which may not provide the desired segmentation results. In the case of natural scene images, Urdu text may be written in different writing styles and may contain various widths and heights as shown

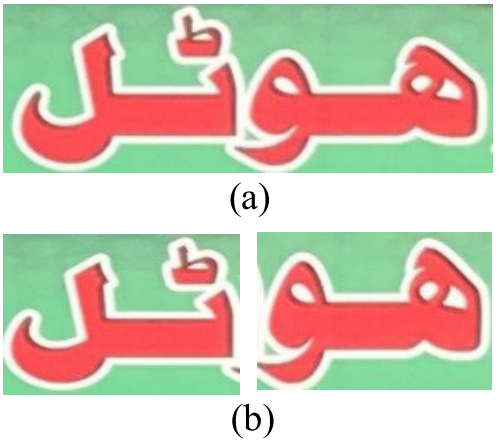

FIGURE 4. Ligature representation of an Urdu word "Hotel": (a) original word (b) two ligatures representing the word in (a).

FIGURE 5. Variations in text width, height and writing styles in Urdu text.

in Figure 5. Furthermore, a segment of an isolated character may have similar shape and structure to another character, as shown in the second row in Figure 3. Many of the Urdu characters have a similar basic shape and structure, and can only be differentiated either by the number of dots or by the positions of these dots around the shape, as shown in the third row in Figure 3. Correct classification and recognition of such characters is a challenging task. Researchers, therefore, have not investigated these languages thoroughly and this area of research is still in its infancy.

Another limitation for the development and evaluation of state-of-the-art methods to address this problem is the unavailability of standard datasets. A large dataset is a preliminary requirement for the training and testing of classifiers in cursive text recognition. Therefore, in this research study, a new dataset is created which contains images of isolated characters that are manually segmented from natural scene images containing Urdu text. Before passing this dataset to the CNN classifier for classification and recognition, preprocessing operations are performed to give the dataset a uniform and standard representation. All the labels for the Urdu characters are encoded using a standard label encoding technique.

To achieve high text recognition accuracy in natural scenes with traditional feature extraction techniques, it is necessary to design a robust feature engineering method which can

select appropriate features. However, deep learning-based methods have the capability to extract high-level spatial features from natural scenes automatically. In our previous work, the model proposed in [29] implemented a sequential CNN without analyzing the effect of multi-scale and multi-level feature fusion. Similarly, the model proposed in [30] implemented a histogram of oriented gradients (HoG) method. Both models were trained on an augmented dataset. However, in this paper,a multi-scale feature aggregation and a multi-level feature fusion network are proposed to extract low-level and high-level features and fuse them to recognize Urdu characters in natural scene images. The architecture for the proposed network is created by aggregating multi-scale convolutional features by up-sampling and element-wise addition operations and then combining with the multi-level feature fusion network. Finally, the feature maps of both networks are fused and passed to the classifier to generate character class probabilities. To add multi-scale convolutional features, a nearest-neighbour interpolation method is proposed that widens the spatial dimensions of smaller feature maps. The performance of the proposed network is compared with sequential CNN and HoG networks. The experimental results show that the proposed model outperforms the methods proposed in [29] and [30] without data augmentation. Furthermore, the proposed model is evaluated on the Chars74K [33] and ICDAR03 [34] datasets, where our method outperforms on the Chars74K dataset and achieves competitive results on the ICDAR03 dataset.

The key contributions presented in this paper are summarized below:

1) The first benchmark dataset for Urdu text detection and recognition in natural scene images is developed.
2) The first large annotated dataset of Urdu characters cropped from natural scene images is developed.
3) A new CNN architecture is proposed that integrates multi-scale feature aggregation and multi-level feature fusion networks to make a CNN fusion by concatenating multi-scale and multi-level CNN features for Urdu character recognition in natural scenes.
4) The proposed model is evaluated on two natural scene character datasets, where it outperforms on the Urdu Chars and Chars74K datasets.

The rest of this paper is organized as follows: Section II presents related work. Section III highlights the characteristics of the proposed Urdu natural scene text dataset and the segmentation process. Section IV presents the proposed methodology and feature extraction techniques. Section V describes the experimental setup and results. Finally, Section VI presents conclusions and future work.

II. RELATED WORK

The problem of character recognition for cursive and non-cursive scripts in scanned and handwritten documents has been studied for at least the last three decades [35]–[39]. Several methods have been successfully developed to address the problem of OCR for cursive scripts such as Arabic and Urdu [36], [40]. However, research into the problem of text recognition in natural scene images for these scripts is still at a preliminary stage. The background complexities and other challenges associated with text in natural scene images make its recognition more complex and challenging than for printed or handwritten text. Some earlier works proposed for character recognition in natural images used local or global feature representations [33], [41]–[46], while recent works use deep neural network techniques [21], [23], [47], [48].

A. LATIN AND INDIC SCENE TEXT RECOGNITION

In [33] an annotated dataset of English and Kannada characters (Chars74k) was presented. In this dataset, all the individual characters were manually segmented from the captured images. The dataset was used to evaluate six different local feature extraction techniques including shape, edge-based methods, and texture representation methods using their commonly adopted parameter values. Most of the experiments were performed on English characters and some preliminary experiments were performed on Kannada characters. The results of this evaluation showed that the Nearest Neighbor classifier performed better than a support vector machines (SVM) classifier on shape context and geometric blur feature descriptors. Local feature descriptors were found to be more stable and to be capable of extracting more information from the image patch when the characters were degraded. In [41], a comparison of different feature descriptors, dictionary sizes, SVM kernels and sampling methods was analyzed on the feature representation and classification for natural scene characters. They evaluated the model on ICDAR2003 and Chars74k dataset and obtained the best results on the HOG descriptor with SVM.

In [42], a group of local features was used to describe the classes of the characters and the mutual positions of these local features were used to represent the structure of the character. Learning was performed in a weakly-supervised fashion and the parameter estimation was achieved using an expectation-maximization optimization approach. A probabilistic model proposed in [43] integrated a mixture of character properties including character similarity, character appearance, bi-gram appearance and lexicon decisions to recognize characters in natural images. Similar to [43], a character similarity method proposed in [44] trained a similarity expert to classify either each pair of the characters is same or not. An integer program was then formulated to find the optimum features of each character. A neural network-based technique proposed in [45] used histogram of gradient (HoG) features as the input to a neural network for recognizing cropped words in natural scenes. The results reported showed better accuracy. An extended form of the HoG descriptor, proposed in [46] extracted features at multiple scales and evaluated the performance of the method on the Chars74k [33] and ICDAR03-CH datasets. This system achieved promising results using the same evaluation framework as in [33].

In recent years, competitive results have been achieved with deep neural networks. In [47], the HoG feature extraction process was extended and two new feature descriptor approaches: Co-occurrence HoG and Convolutional Co-HoG were proposed for multi-lingual character recognition in natural images. The Co-occurrence HoG feature captures the co-occurrence of gradient orientation pairs of neighboring pixels and captures more spatial as well as contextual information. The Convolutional Co-HoG feature extracts co-occurrence features from all the possible image patches in a character image for deep spatial information. Both these techniques were evaluated on five different datasets of characters including Chinese and Bengali. An unsupervised learning approach was proposed in [23] where the features were directly extracted from unlabelled data and were input into a 62-way CNN character classifier. The method was evaluated on the cropped ICDAR03-CH dataset, and state-of-the-art results were reported. Most of the recently proposed methods do not use segmentation step and apply sliding window techniques for candidate character detection. A CNN classifier was used in [48] to first detect the candidate characters. Then a graph method was applied to examine the best series of characters. A CNN model was proposed in [21] for text detection as well as character recognition modules. A beam search and non-maximal suppression along with lexicons were used to recognize cropped words in the ICDAR03 dataset.

The methods that recognize individual characters and then combine them into words or text lines are usually prone to errors due to the complex backgrounds of natural images and text size variations. To overcome this problem, several methods have been developed that recognize the whole word or a text line without localizing or recognizing each character individually [49], [50]. The methods proposed in [49], [50] have focused on recognizeing arbitrary oriented whole word text in natural images. However, in the case of cursive scripts like Urdu, the recognition of a whole word or a text line from a natural scene image is more complex due to the different shapes of each character and other challenges as discussed in Sections I and III.

B. CURSIVE TEXT RECOGNITION IN VIDEO AND NATURAL SCENE IMAGES

In recent years, several novel works for cursive text detection and recognition in video images have been developed [51]–[54], while a limited work is presented for cursive text recognition in natural scenes [55]–[57]. Ahmed *et al.* [55], modified the maximally stable extremal region method to extract the scale-invariant features and passed to the multi-dimensional long short term memory (MDLSTM) classifier. They evaluated their method on two datasets i.e., ICDAR15 and their proposed English-Arabic scene text recognition dataset. In another work [56], they used a sequential CNN to recognize isolated Arabic characters in natural images. They developed a dataset of manually segmented Arabic character images. The dataset contained a limited number of isolated Arabic letters and represented each character class of Arabic letters with 20 samples. To further increase the size of the training set, each character was oriented at five different angles. After performing this augmentation, the dataset contained 2700 samples, which were divided into 2450 training and 250 test samples. The lowest error rate achieved was 14.57%. In [57], a modified bag of features technique, based on spatial pyramid matching and sparse coding, was used to represent local features for Arabic character recognition in natural scene images. Each character image was divided into some fine sub-regions and the histogram of local features was calculated for each of the sub-region. A multi-scale max pooling method was then used to sum up all the histograms of local features, which were then passed to an SVM classifier to perform character classification. Experiments were performed on a custom developed Arabic scene character dataset and accuracy of 60.40% was achieved.

Jain *et al.* [51] used a hybrid CNN-RNN network to recognize Arabic text in videos and natural images. To train the network, they created a large-scale synthetic dataset. They evaluated their method on two Arabic news video text datasets and a synthetic natural scene text dataset. They reported a character recognition rate of 98.17%, 97.44% for the Arabic news video text and 75.05% for the synthetic Arabic natural scene character image datasets. A multi-language end-to-end scene text recognition system was presented in [58]. The authors used the ResNet50 [59] network for text localization, VGG16 [60] pre-trained on ImageNet [61] data for script identification and an OCR module proposed in [62] for multi-language text recognition. They evaluated their model on six different scripts including Arabic and reported an accuracy of 26.80% for the Arabic scene text recognition.

Several approaches have been proposed for typewritten Urdu ligature recognition [63], [64]. This type of text is usually horizontally aligned, written with one font type, font size and on a plain/clean background. Moreover, some results have been reported for artificial Urdu text detection from video frames and still images [65], [66]. In these systems, the artificial text is only written horizontally and has a fixed font size and similar color. Compared to typewritten and artificial text, natural scene text detection or recognition is a more complex and challenging task. To the best of our knowledge, no research results have yet been reported for detection or recognition of Urdu text in natural scene images.

C. FEATURE FUSION IN CNN

Several studies in the computer vision domain have combined multi-layer convolutional features to detect and recognize objects [67]–[72]. In a CNN, the low-level layers represent the edge and corner information, while the high-level features help to classify the objects. In [67] convolutional features at multi-levels were integrated to classify electroencephalograph (EEG) images. In [68], a multi-layer convolutional feature fusion network was proposed by combining

CNN features to classify high spatial resolution images. To integrate the multi-layer convolutional features, they used the VGG16 [60] model pre-trained on ImageNet [61] data. In [69], the spectral-spatial features of the hyperspectral images were fused with multi-level and multi-scale CNNs. They used a 3D CNN to extract spectral features and a 2D CNN to extract the spatial features and combined them using a residual block. To prevent the problem of model over-fitting and reduce the network parameters, they used global average pooling instead of fully connected layers. In [70] the visual relationship between different objects and attention mechanisms was used to create a multi-modal feature fusion algorithm for a question answering system. A gated recurrent unit was implemented to model the question, then a bi-linear attention method was applied to determine the weights of all regions of an image. The model achieved state-of-the-art results on the two visual question-answer datasets. In [71], three deep CNN frameworks were used to extract features separately from remote scene images. These features were fused together to create a single feature vector for the classification stage. The model achieved better classification results on the three remote scene classification datasets. In another study [72], feature fusion of hyperspectral and multispectral images was implemented using network-in-network, batch normalization and skip connections to fuse the network features at multi-scales. The model outperformed other methods on four remote scene datasets. Multi-level and multi-scale CNN feature fusion techniques have also been used in many other techniques, where competitive results have been achieved in the specific domains [73]–[76]. Cursive text in natural scene images has variations in aspect ratio, font sizes, writing styles, orientations and the characters have various shapes for similar base structures. These challenges have motivated us to use novel multi-level feature integration and feature fusion techniques to recognize cursive text in natural scene images.

III. THE URDU NATURAL SCENE TEXT DATASET

Synthetic datasets [77]–[79] have been commonly used to train deep learning models. However, the synthetic text data is not complex like the real text and lacks variations in writing styles, aspect ratio, background complexities and other environmental factors. Hence, deep learning models trained on synthetic data may not perform well on the real natural scene text.

Therefore, a set of 2000 natural scene images containing Urdu text were photographed with a mobile phone and digital camera to develop a new Urdu natural scene text dataset. The captured images mostly contain scenes with signboards, advertisement banners, shop names, hoardings, and road signs. The images captured have different resolutions and were captured in varying lighting, weather and perspective conditions. Some samples of the originally captured images are shown in Figure 6. Some images contain tri-lingual text including English, Urdu and Sindhi are shown in the first image in Figure 6. The dataset has a large number of images

FIGURE 6. Some samples of the natural scene images with Urdu text from the proposed dataset: images with variations in font size, blur, illumination, multi-language and handwritten text on the walls.

with handwritten or painted text as illustrated in the second and third images in Figure 6. This handwritten text on walls or signboards is more challenging to detect and recognize.

As explained in Section I, the words in cursive scripts such as Urdu are formed by joining two or more characters and many of the words overlap with each other. Therefore, it is quite complex to segment individual characters from the words. Hence, automatic segmentation algorithms, developed for Latin text, have very low accuracy for Urdu text. For this reason, all of the character images were manually segmented from the original images. The cropped character images were scaled to form images with a dimension of $48 \times 48 \times 3$ pixels. The process of manually segmenting characters from a given word is shown in Figure 7. Using this segmenting and scaling procedure, a dataset of 18500 Urdu character images was created.

The Urdu alphabet contains 58 symbols out of which 39 form the basic character set while 18 are digraphs. When all the forms of the characters are included, there are more than 150 visually distinct characters. The Urdu language has no lower or upper-case letters (small or capital letters).

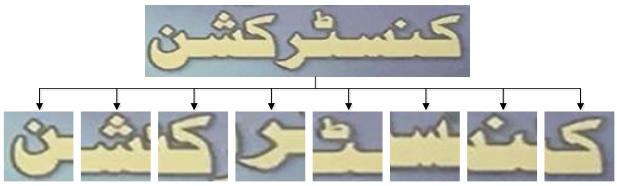

FIGURE 7. Manual segmentation of a word into characters.

TABLE 1. The statistics of the Urdu dataset.

Data	Number of Samples
Number of cropped words	13880
Number of segmented characters	18500
Number of samples after synthetic images	19000
Number of classes	42
Training set size	13300
Test set size	5700

The new Urdu dataset contains 42 character classes consisting of 39 basic Urdu characters in their different forms. However, some of these classes such as ث, ژ and ظ, rarely occur in everyday Urdu text. Therefore, to make the number of samples of these classes equal to 100, a synthetic Urdu character dataset was created using the background from real natural images and varying font types, styles, and sizes for the text. Figure 8 shows some sample images from the synthetically generated Urdu character dataset. The statistics of the dataset are summarized in Table 1.

FIGURE 8. Some examples of the synthetically generated characters.

The new dataset contains a huge variety of Urdu text with variations in font size, shape, illumination, background complexities, color and hand-writing styles. In order to cover all varieties of character shapes, the character images were segmented into their different positions within a word. In general, the number of Urdu text images captured in unconstrained environments is sufficient for the purposes of training deep learning networks. Figure 9 shows some examples from the new Urdu character dataset. The development of such a dataset requires a huge amount of time and effort. Therefore, this new Urdu natural scene text dataset will be extremely useful to the document analysis and recognition community in the following ways:

FIGURE 9. Some examples of the segmented character images in the new Urdu natural scene text dataset.

1) The dataset will be made freely available so that researchers can further investigate this challenging problem and propose new solutions.
2) It will be helpful to researchers working on multilingual text detection and recognition in natural scenes as well as language-independent OCR algorithms.
3) It will be helpful to develop end-to-end natural scene text detection and recognition systems.

IV. METHODOLOGY

To handle the challenging problem of Urdu text recognition in natural scene images, we propose a new CNN architecture that integrates convolutional features of the network at different layers and then combines them with the high-level layers to create a fused feature. Cropped character images of Urdu text in natural scene images are fed into the proposed model to extract discriminative features with different convolutional filters. In further experiments, we extract HoG features from the same-cropped character images and compare their character recognition performance with the proposed method. A brief description of the CNN and HoG methods is explained in Section IV-A and Section IV-B respectively, and the proposed method is explained in Section IV-C.

A. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Networks are widely used in computer vision, pattern recognition, document analysis and many other areas. Generally, a CNN architecture is divided into two parts. The first part consists of various convolutional layers, pooling layers and non-linear activation functions, and the second part consists of fully connected layers and the classification layer. This CNN architecture makes the network able to learn features at multiple scales. Initial layers of the CNN extract low-level features such as edges and corners, while the later layers extract features that are more abstract. Each convolutional layer has a receptive field with which the image is convolved. The output of the convolutional layer creates an activation map, which is passed as the input to the next layer. For Urdu character recognition in natural scenes, a fixed size image of $W_i \times H_i$ pixels is given to the input layer.

The output of this layer is passed to the convolutional layer to perform the convolutional operation. The convolutional layer calculates the output feature map as:

$$F_m^s = \sum_{k=1}^{D} W_k * I^k + b \tag{1}$$

where I is an input image to the convolutional layer or a feature map produced by the previous layer with dimensions $W_i \times H_i \times D$. W is the filter bank of size $w \times h \times D$ which connects the input and output feature maps of two layers. $*$ is a 2D discrete convolutional operator, b represents the trainable bias parameter and F_m is an output feature map of the convolutional layer.

An activation function is then used to add the non-linearity in the output of the convolutional layer. Initially, the sigmoid function was used as an activation function, but it reduces the gradient to almost zero if the local gradient is very low. This causes the problem of vanishing gradients. To handle this problem, the rectified linear unit (ReLU) has become a common choice for the activation function. This non-linear operation is applied on the feature map F_m of the convolutional layer. The non-linear operation is calculated as (2):

$$F_{RL} = f(F_m) \tag{2}$$

where $f(\cdot)$ is commonly selected as the ReLU function and is defined as $f(x) = \max(0, x)$. The ReLU sets the output value to zero when the input value is zero or negative, otherwise, the output value remains the same as x.

In the CNN architectures, a pooling layer is commonly used in between convolutional layers to down-sample the feature map by combining a collection of the filter responses. This down-sampling of the features reduces network computations. For this purpose, max pooling is a common strategy used in CNNs. In max pooling, a maximum value is taken from the small spatial region G of an active feature map F_m as shown in (3):

$$F_{PL} = \max F_{RL_i} \tag{3}$$

After convolutional and pooling layers the next part of the CNN is based on fully connected layers where two or more than two dense layers are stacked to form a standard multi-layer perceptron. The Softmax function is mostly used to perform classification at the last fully connected layer as shown in (4):

$$\text{sm}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_j}} = \hat{y}_j \tag{4}$$

where z is the vector of inputs to the last output layer of the CNN, which generates the output vector of $\hat{y}$ categorical probability distributions of all classes. The cross-entropy loss l between the one-hot encoded label y and $\hat{y}$ for adjusting the parameters of the data is calculated as shown in (5):

$$l(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{N} \sum_{j=0}^{J} [y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j)] \tag{5}$$

As Urdu script contains multiple character classes, a sparse categorical cross-entropy cost function is utilized in our architecture.

B. HISTOGRAM OF ORIENTED GRADIENTS

The HoG features [80] are extracted by taking the histograms of edge orientations in an image patch. The orientation information is more robust to uneven lighting conditions and taking the histograms provides translational in-variance. The gradients for the gray scale pixels (x, y) in an image I are computed by convolving $I(x, y)$ in the x-direction with the kernel $K_x = [-1, 0, 1]$ and in the y-direction with the kernel $K_y = [1, 0, -1]^T$ as shown in (6) and (7):

$$G_x(x, y) = K_x * I(x, y) \tag{6}$$

$$G_y(x, y) = K_y * I(x, y) \tag{7}$$

At each pixel location (x, y), the magnitude $M(x, y)$ and orientation $\theta(x, y)$ of the gradient are calculated as follows:

$$M(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{8}$$

$$\theta(x, y) = \frac{G_y(x, y)}{G_x(x, y)} \tag{9}$$

Next, the image is split into cells of size $n \times n$. For every pixel in a cell, a magnitude is computed and then added to an appropriate orientation bin. A HoG feature vector, equal to the total magnitude for each orientation bin, is then extracted from each cell.

For Urdu character recognition, various values of n were investigated and a value of $n = 8$ produced optimal results. When the value of n was increased, the number of extracted features in an image patch decreased. When the size of n was greater than 16, there was a loss of extracted feature details and when the value of n was small noise was introduced. A visualization of the HoG features generated for the character 'ب', when different cell sizes were used, is shown in Figure 10.

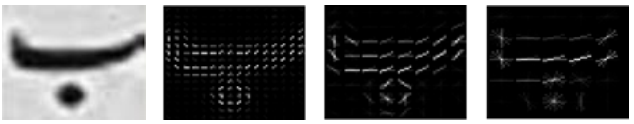

FIGURE 10. (a) The original image of the character "bay"; visualization of the HoG features generated from (a) with: (b) a cell size of 4 × 4 and feature vector length of 14112, (c) a cell size of 8 × 8 and feature vector length of 2592 and, (d) a cell size of 12 × 12 and feature vector length of 648.

The orientation histogram bins selected were 8 so that a trade-off between orientation details and the feature vector's size be achieved. The size of pixels-per-cell was selected to be 4. This small size of pixels-per-cell is helpful in capturing the significance of local pixels and handles illumination variations of the HoG features. Various experiments were performed by changing the value of cells-per-block and value of 2×2 was selected. Although, more information is captured

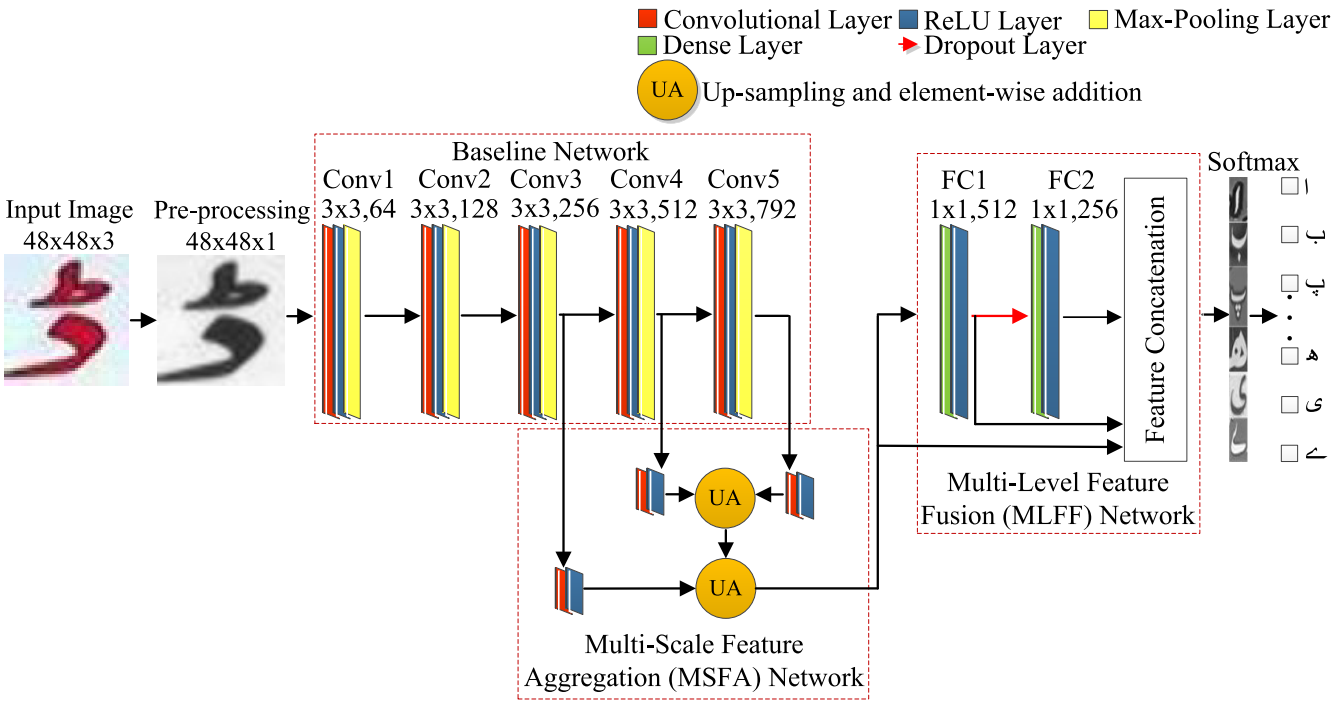

FIGURE 11. The framework of the proposed cursive character recognition model. The baseline network uses five convolutional layers. The multi-scale feature aggregation network is used to integrate low-level and mid-level features, which are then processed by the multi-level feature fusion network to obtain robust context information.

by keeping the large value of cells-per-block, it also makes the size of feature vector larger. The HoG features are generated for each key point of the image and the histogram entries of all cells around every key point make the feature of that key point. The HoG features are then fed to the SVM classifier for classification.

Different from the HoG method proposed in [30], we fine-tune the hyper-parameters of the HoG feature vector block using the $L2$ normalization method.

C. ARCHITECTURE OF THE PROPOSED MODEL

The proposed model includes three components: 1) the baseline network, 2) a multi-scale feature aggregation (MSFA) network and 3) a multi-level feature fusion (MLFF) network. The overall framework of the proposed model is shown in Figure 11.

1) BASELINE NETWORK

In the baseline network, five convolutional layers were used. The number of filters and their sizes for each of the convolutional layers were: {number of filters: filter size}, {64: 3, 128:3, 256: 3, 512: 3, and 792: 3}. To preserve the spatial size of the output feature maps, the input to every convolutional layer was padded with the values at the border of the image. To add non-linearity, a ReLU activation layer was used after every convolutional layer. A max-pooling layer with a kernel size of [2, 2] and a stride value of [2, 2] was applied after every convolutional layer. Kernel filter sizes of 3×3, 5×5 and 7×7 were trialled and the best accuracy was achieved with a size of 3×3. Generally, text in natural scene images is placed at different locations and its size is relatively small comared with the size of the entire image. Keeping the minimum filter size helps to extract more pixel details of the Urdu character images.

2) MULTI-SCALE FEATURE AGGREGATION NETWORK

This component of the network aggregates diverse features of the convolutional layers at multiple scales. Conventional CNNs hierarchically represent the features and share the weights across different spatial locations to reduce the network complexity. Aggregating different hierarchical features is useful in representing different features of the input images. In the baseline network, the output size of the feature maps for Conv3, Conv4 and Conv5 is $[w/3, h/3, c_3]$, $[w/4, h/4, c_4]$ and $[w/5, h/5, c_5]$, respectively, where w and h are the width and height of the input image and $c_x = \{256, 512, 792\}$ for $x = 3,\ 4$ and 5 are the number of channels of the feature maps. In the MSFA network, we use three convolutional layers with 1×1 convolution kernels to reduce the dimensions of the feature maps and 256 output channels. This module aggregates the low-level and mid-level features of the convolutional layers with a different number of channels by applying an "up-sampling and addition" operation. This operation first widens the feature maps with smaller dimensions to the same size as the larger ones by using nearest-neighbour interpolation, and then performs element-wise addition to produce

a feature map. Finally, the two feature maps in the MSFA network are aggregated together by applying the ''up-sampling and addition'' operation to obtain multi-scale features. These features have more rich information than the individual low-level to mid-level features. Different from the sequential CNN model proposed in [29], we aggregate low-level and mid-level convolutional features at multi-scales. To increase the spatial dimensions of the feature maps, we up-sample the latter layers and perform an element-wise operation to integrate feature maps from different layers.

3) MULTI-LEVEL FEATURE FUSION NETWORK

This component of the network concatenates the feature map of the MSFA network with the high-level features. The input to the MLFF network is the aggregated feature map from the MSFA network. The MLFF network consists of two fully connected layers with 512 and 256 output units. The output of the MSFA is concatenated with FC1 and FC2 and passed to the Softmax function in the classification layer. The final outputs represent the probabilities of each of the character classes. This multilevel feature fusion technique represents each class of the Urdu character set with more robust features.

$$\text{MFCNN} = \text{Concat}(\text{MSFA}, \text{FC1}, \text{FC2}) \tag{10}$$

where the features in MSFA, FC1 and FC2 are concatenated to create a single feature vector that is passed to the classification layer, as shown in Figure 11.

Different from the CNN model proposed in [29], we concatenate up-sampled aggregated features of different layers with the high-level layers. This multi-level feature fusion captures more robust spatial information that helps to boost the recognition accuracy for Urdu characters.

V. EXPERIMENTAL SETUP AND RESULTS ANALYSIS

A. DATASETS

We evaluated the proposed model on a custom-developed Urdu character image dataset, as described in Section III, and two publicly available datasets of English characters called Chars74K [33] and ICDAR03 [34]. Chars74K dataset consists of synthetically generated single character images and natural scene images of English and Kannada letters. This dataset includes 12402 natural scene images of 62 English characters (A-Z, a-z, 0-9). As our work is focused on natural scene text, we used only natural scene character images of English letters and ignored synthetically generated characters to train the model. The performance of the proposed method were compared with state-of-the-art natural scene character recognition methods.

B. IMPLEMENTATION DETAILS AND NETWORK TRAINING

The proposed framework was implemented using the Keras [81] python-based open-source deep learning library, while the HoG based model was implemented with Scikit-learn [82]. Images from the cropped character datasets were shuffled and randomly divided into training (70%) and testing (30%) sets. All the images were resized to $48 \times 48 \times 3$ pixels. To further boost the accuracy of the proposed model, a Keras library-based real-time data augmentation method was used for the Urdu character images where only the training samples were augmented. We did not use data augmentation for Chars74K dataset. The accuracy of the model was compared with and without data augmentation methods. To measure the robustness of the proposed models, we repeated all the experiments for five times and the best classification results achieved are taken.

The proposed model was trained by back-propagation using a sparse categorical cross-entropy loss function [83] and stochastic gradient descent optimization with momentum [84]. Different batch sizes were trialled and a batch size of 64 was found to be the most efficient. Different learning rates were trialled and a learning rate of 0.005 was found to be optimal. The number of epochs for training were chosen to be 80. The network contains approximately 10 million parameters.

C. EVALUATION METRICS

The performance of the proposed model and HoG-based classification method was evaluated using the standard evaluation metrics for assessing the performance of multilingual text recognition in natural scene images [85]. We first calculate the precision and recall scores. Precision corresponds to the number of true character class predictions that actually belong to the positive class, while recall corresponds to the number of true character class predictions that are made out of all positive samples in the dataset.

After precision and recall calculations, the overall accuracy of the models are calculated through the F-score as:

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

where F-score is the weighted harmonic mean of precision and recall.

D. CLASSIFICATION RESULTS ON URDU CHARACTER DATASET

Classification results for Urdu characters in natural scenes with the proposed model and HoG method are shown in Table 2. The results are compared with the previous results [29], [30]. In [29], a data augmentation method of rotating images at three different angle values was used to enhance the training samples. Several images in the Urdu script have similar shapes and the augmented images created with angle rotation were matched with other character shapes. This increased the ratio of false positives. In this paper, we do not apply the augmentation method for HoG based experiments and create synthetic character images to increase the number of infrequently used characters. As shown in Table 2, the precision, recall and F-score for the proposed feature fusion method is 90%, 91% and 91%, respectively, whereas for the HoG model using the SVM classifier these scores are 82%, 83% and 83%, respectively. The precision, recall and F-score for the sequential CNN and the machine learning

TABLE 2. The performance of the proposed model on the Urdu character dataset and comparison with previously published results.

Model	Precision (%)	Recall (%)	F-score (%)
Ours [29]	84	85	85
Ours [30]	73	73	73
HoG	82	83	83
Proposed Model	90	91	91

TABLE 3. The performance of the proposed model on the Urdu character dataset with data augmentation and comparison with previously published results.

Model	Precision (%)	Recall (%)	F-score (%)
Ours [30]	88	89	89
Proposed Model	93	93	93

methods are 84%, 85%, 85% and 73%, 73%, 73%, respectively. These results show that the data augmentation method with angle rotation is not effective for cursive character recognition. The results in Table 2 also show that the aggregation of multi-scale features with high-level features outperforms the sequential CNN and conventional feature extraction methods such as HoG.

To further investigate the problem of the limited size of the datasets, a real-time data augmentation method was used, where the virtual images were created on the fly with zooming, scaling and translation of the original images. The proposed model was trained with real-time augmented data. Table 3 shows the precision, recall and F-score of the proposed method and compares with the previous results. It can be seen that the proposed model improves the recognition results by 4% than the previous augmented method [29]. The model with data augmentation also improves 2% recognition accuracy than the model without data augmentation result (Table 2). This shows that the deep convolutional neural networks perform better when using larger training datasets.

In Figure 12, we present some qualitative results for correctly recognized characters with the proposed convolutional feature fusion method (Figure 12(a)) and the HoG method (Figure 12(b)). For each of the character images, the labels on the bottom-left and the bottom-right are the ground truth and network predictions, respectively. Figure 13 shows some incorrectly recognized Urdu characters with the proposed convolutional feature fusion model (Figure 13(a)) and HoG method (Figure 13(b)). For each of the images, labels at the bottom-left are ground truths while at bottom-right with red color are the network predictions. Depending upon the position of a character within a word, it may appear in one of four different shapes. Some characters may have similar shapes, which can only be differentiated by the number of dots or their positions (i.e., above, below or between the characters). These characteristics of cursive scripts make the text recognition problem more challenging than for non-cursive scripts. As shown in image 1 of Figure 13(a), the initial shape of the letter ٹ matches with the initial shape of the letter ٹ. Similarly, the initial shapes of letters ب and ٹ as shown in image 3 and 16 in Figure 13(b), match with the initial shapes of ب and ٹ. However, they are only differentiated by the number of dots above or below the shapes. An isolated shape of the letter د in image 10 matches with the initial shape of the letter ح, while the initial shape of the latter letter in image 20 matches with the isolated shape of the earlier letter in image 10. Furthermore, due to different writing styles, several characters look visually similar to other characters, which also increases the false recognition rate. As shown in image 9 and 21 in Figure 13(a), final shapes of the letters ل and ع look similar to the final shapes of the letters ر and ح. The environmental factors such as low resolution, blur, image degradation etc., as shown in images 3, 4, 8 and 19 in Figure 13(a), and images 6, 17 and 20 in Figure 13(b) make the problem of scene text recognition more complex.

E. PERFORMANCE COMPARISON

This paper introduces the baseline investigation for Urdu character recognition in natural scene images and no prior work has been reported for this problem. The dataset of Urdu text in natural scene images used in this paper is the first to be published. Therefore, the validity of the proposed model is evaluated on Chars74K and ICDAR03 datasets. Furthermore, some research works for isolated Arabic character recognition in natural images have been reported. However, the datasets developed for Arabic scene text recognition are not publicly available yet. Hence, the performance of the proposed method is compared with the existing results of Arabic natural scene character recognition.

1) ENGLISH NATURAL SCENE CHARACTER DATASET

To analyze the quality of the proposed method, we evaluated our method on the Chars74K [33] and ICDAR03 [34] datasets, and compared its performance in terms of F-score with a number of state-of-the-art character recognition methods. The recognition results of each of the methods are presented in Table 4. As shown in Table 4, the proposed multi-level feature fusion method outperforms than the state-of-the-art methods on the Chars74K dataset with an F-score of 81.45%. Several lower and upper-case letters of English script (e.g., 'C', 'c', 'K', 'k', 'O', 'o', 'W', 'w', 'X', 'x', 'Z', 'z') have similar structure and shape, which increase the false recognition rate. Moreover, sometimes these letters are very difficult for human beings to recognize correctly.

The ICDAR03 [34] dataset consists of 6185 character images in the training set and 5430 in the test set. The deep learning models need a large amount of data samples

FIGURE 12. Qualitative results of correctly recognized Urdu characters in natural images: (a) proposed convolutional feature fusion method and (b) HoG method. Labels on the bottom-right are the ground truths, while on the bottom-left are the network predictions.

for training. The F-score of our proposed method on the ICDAR03 test set is 76.49%, which is approximately 6% less than the state-of-the-art method as reported in [87]. However when we combined the ICDAR03 and Chars74K dataset together and trained the model, the F-score of the model on the ICDAR03 test set was 82.07%, which is 0.63% lower than the method reported in [87]. This shows that the proposed method can be used to further improve recognition accuracy when trained on larger datasets. The results in Table 4 show the superiority of the proposed method as it works comparatively well on different datasets.

FIGURE 13. Some examples of incorrectly recognized Urdu characters in natural scene images: (a) proposed convolutional feature fusion method and (b) HoG method. Both methods fail to recognize some blurred, low resolution or characters which have similar shapes.

2) ARABIC NATURAL SCENE CHARACTER RECOGNITION METHODS

As explained in Section III, Urdu is a derived language of Arabic and Persian scripts, it has 39 basic alphabets, whereas Arabic and Persian scripts have 28 and 32 basic alphabets respectively. It inherits many characters from both scripts, however, Urdu is written in a Nataliq style whereas Arabic text is written in Naskh style. Urdu text is written more compactly from the top right to bottom left with taller and tighter letters than in Arabic and Persian text. Therefore, in this paper, the recognition performance of the HoG and proposed convolutional feature fusion models was compared with the performance of the algorithms proposed in [56] and [57]. In [56], only 20 samples for each of the Arabic

TABLE 4. Comparison of the proposed method and a number of state-of-the-art methods for the Chars74K dataset in terms of the F-score (%).

Method	Chars74K	ICDAR03
GHoG + SVM [41]	62.00	76.00
LHOG + SVM [41]	58.00	75.00
HOG + NN [23]	58.00	52.00
SIFT+ SVM [33]	21.40	–
Strokelet [86]	62.00	69.00
Co-Strokelets [87]	67.50	82.70
SIFT + Autoencoder + SVM [88]	75.40	79.30
SIFT + Sparse RBM + SVM [88]	73.90	78.10
SIFT + Sparse Coding + SVM [57]	73.10	75.30
PCNet [89]	64.00	75.00
Rank 1 Tensor [90]	79.00	74.00
CoHOG + Linear SVM [91]	—	79.04
Proposed Model	**81.45**	**76.49**

TABLE 5. Performance comparison of the proposed model with state-of-the-art techniques.

Model	Script	Feature Extraction Method	Error Rate (%)
Ahmed *et. al.* [56]	Arabic	CNN	15
Moalla *et. al.* [57]	Arabic	SIFT	24
Ours [29]	Urdu	CNN	11
HoG	Urdu	HoG	17
Proposed Model	Urdu	Multilevel CNN Fusion	09
Proposed Model with Data Augmentation	Urdu	Multilevel CNN Fusion	07

characters were manually segmented from natural scenes and then each character of the sample was oriented at five different values to increase the size of the dataset. The total number of samples after orientation augmentation was 2700, where 2450 samples were used for training and the remaining 250 samples were taken to test the CNN network. The authors evaluated their method in terms of error rate and have reported a 15% error rate. In [57], only 260 images were captured and a total of 100 classes with 30 to 40 samples, for each of the Arabic characters in different positions within a word, were manually segmented. The characters were represented by sparse coding of scale invariant-feature transform (SIFT) features. Table 5 shows a performance comparison of the HoG, proposed convolutional feature fusion model and the methods proposed in [29], [56] and [57]. This table shows that using HoG features, the model outperforms than the SIFT features [57], while, the performance of the techniques using a CNN in [56] is better than the HoG features. The error rate of our previous sequential CNN model [29] is 11%, which is less than the CNN model proposed in [56].

TABLE 6. Performance comparison of the proposed network and its variants.

Network	F-score (%)
Baseline	86
Baseline + MSFA	89
Baseline + MLFF	88
Baseline + MSFA + MLFF	91

TABLE 7. Average running time (seconds) of the feature fusion method on the three datasets.

Dataset	Training	Testing
ICDAR03	0.032	0.009
Chars74K	0.035	0.010
Urdu chars	0.037	0.011

While the error rate of the proposed model with and without data augmentation is 7% and 9%, respectively. These results show the effectiveness of convolutional features when fused together at multi-scales and multi-levels.

F. ABLATION STUDY

The proposed model consists of MSFA and MLFF network modules. To measure the effectiveness of each of the network modules, we modified the proposed model into four variants. The first variant denoted as baseline did not use the MSFA and MLFF networks. The second variant used the MLFF network with the baseline network, the third variant used the MSFA network, while the fourth variant used both the MSFA and MLFF networks. All the network variants were trained with the same hyper-parameters. Table 6 presents a quantitative comparison of the F-scores for all network variants. As shown in Table 6, the fourth network variant that used both the MSFA and MLFF modules outperformed the other three variants. By aggregating low-level and mid-level features at different CNN layers, and then concatenating with high-level features, our proposed model recognizes Urdu characters more accurately.

G. COMPUTATIONAL EFFICIENCY

All the experiments were performed on an Intel Core i7 3.60 GHz CPU with 16GB of random access memory. The average training and testing times for the ICDAR03, Chars74K and Urdu character datasets are given in Table 7.

VI. CONCLUSION AND FUTURE WORK

This paper presents a convolutional feature fusion method with multi-scale feature aggregation and multi-level feature fusion networks for Urdu character recognition in natural scenes. The multi-scale feature aggregation network integrates low-level and mid-level convolutional features of different layers by up-sampling and element-wise addition operations. The aggregated features are then given

to the multi-level feature fusion network to be combined with high-level features. Finally, the aggregated and multi-level features are fused together and passed to the Softmax classifier to generate the predictions. The performance of the proposed network is evaluated on three datasets including the Chars74K, ICDAR03 and a custom developed Urdu character image dataset that consists of 18500 manually segmented character images. This dataset contains a mixture of natural scene images, which cover a variety of text in terms of font size, font style, orientations and handwriting styles. To the best of the authors' knowledge, this is the first benchmark dataset for Urdu text in natural scene images. A real-time data augmentation technique was used to further enhance the training set samples in the dataset. Several experiments were performed by fine-tuning the network hyperparameters and combining different mid-level and high-level layers. Experiments show that our proposed convolutional feature fusion model outperforms on the new Urdu character image and Chars74K datasets and achieves competitive results on the ICDAR03 dataset. Currently, only isolated Urdu characters in natural scenes are investigated. In the future, a method for recognition of cropped words and an end-to-end system for text recognition will be proposed.

DATASET AVAILABILITY

The dataset used in this research study is available from the corresponding author upon request.

ACKNOWLEDGEMENT

The author Asghar Ali Chandio thanks the University of New South Wales, Australia, for supporting his Ph.D. candidature with a scholarship.

REFERENCES

[1] A. Broggi, A. Zelinsky, Ü. Özgüner, and C. Laugier, "Intelligent vehicles," in *Springer Handbook of Robotics* (Springer Handbooks), B. Siciliano and O. Khatib, Eds. Cham, Switzerland: Springer, 2016, pp. 1627–1656.
[2] G. N. DeSouza and A. C. Kak, "Vision for mobile robot navigation: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 237–267, Feb. 2002.
[3] M. Alam, J. Ferreira, and J. Fonseca, "Introduction to intelligent transportation systems," in *Intelligent Transportation Systems* (Studies in Systems, Decision and Control), vol. 52, M. Alam, J. Ferreira, and J. Fonseca, Eds. Cham, Switzerland: Springer, 2016, pp. 1–17.
[4] Q. Ye, Q. Huang, W. Gao, and D. Zhao, "Fast and robust text detection in images and video frames," *Image Vis. Comput.*, vol. 23, no. 6, pp. 565–576, Jun. 2005.
[5] X. Liu, "A camera phone based currency reader for the visually impaired," in *Proc. Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2008, pp. 305–306.
[6] X. Chen, J. Yang, J. Zhang, and A. Waibel, "Automatic detection and recognition of signs from natural scenes," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 87–99, Jan. 2004.
[7] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 19–36, Feb. 2016.
[8] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.
[9] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," *J. Inf. Commun. Technol.*, vol. 10, no. 2, pp. 1–4, 2016.
[10] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1491–1496.
[11] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.
[12] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.
[13] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "ICDAR2017 competition on reading chinese text in the wild (RCTW-17)," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 1429–1434.
[14] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J. C. Burie, C. L. Liu, and J. M. Ogier, "ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019," in *Proc. IEEE Int. Conf. Document Anal. Recognit.*, Sep. 2019, pp. 1582–1587.
[15] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," 2018, *arXiv:1811.04256*. [Online]. Available: http://arxiv.org/abs/1811.04256
[16] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "TextField: Learning a deep direction field for irregular scene text detection," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, Nov. 2019.
[17] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3538–3545.
[18] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2004, pp. 366–373.
[19] J. J. Weinman, Z. Butler, D. Knoll, and J. Feild, "Toward integrated scene text reading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 375–387, Feb. 2014.
[20] J. L. Feild and E. G. Learned-Miller, "Improving open-vocabulary scene text recognition," in *Proc. IEEE Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 604–608.
[21] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, Jan. 2016.
[22] H. Zhang, C. Liu, C. Yang, X. Ding, and K. Wang, "An improved scene text extraction method using conditional random field and optical character recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 708–712.
[23] S. Roy, P. P. Roy, P. Shivakumara, G. Louloudis, C. L. Tan, and U. Pal, "HMM-based multi oriented text recognition in natural scene image," in *Proc. 2nd IAPR Asian Conf. Pattern Recognit.*, Naha, Nov. 2013, pp. 288–292.
[24] B. Su and S. Lu, "Accurate recognition of words in scenes without character segmentation using recurrent neural network," *Pattern Recognit.*, vol. 63, pp. 397–405, Mar. 2017.
[25] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1457–1464.
[26] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 3304–3308.
[27] S. Ahmed, S. Naz, M. Razzak, and R. Yusof, "Arabic cursive text recognition from natural scene images," *Appl. Sci.*, vol. 9, no. 2, p. 236, Jan. 2019.
[28] M. Darab and M. Rahmati, "A hybrid approach to localize farsi text in natural scene images," *Procedia Comput. Sci.*, vol. 1, no. 13, pp. 171–184, Jan. 2012.
[29] A. Ali, M. Pickering, and K. Shafi, "Urdu natural scene character recognition using convolutional neural networks," in *Proc. IEEE 2nd Int. Workshop Arabic Derived Script Anal. Recognit. (ASAR)*, Mar. 2018, pp. 29–34.
[30] A. A. Chandio, M. Pickering, and K. Shafi, "Character classification and recognition for urdu texts in natural scene images," in *Proc. Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Mar. 2018, pp. 1–6.
[31] A. Ali and M. Pickering, "Feature-level fusion using convolutional neural network for multi-language synthetic character recognition in natual images," in *Proc. Digit. Image Comput. Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–6.
[32] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 22nd ed. Dallas, TX, USA: SIL International, Feb. 2019. Accessed: Jul. 10, 2019. [Online]. Available: http://www.ethnologue.com

[33] T. D. Campos, B. R. Babu, and M. Varma, "Character recognition in natural images," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2009, p. 273–280.

[34] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. 7th Int. Conf. Document Anal. Recognit.*, 2003, pp. 682–687.

[35] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," 2017, *arXiv:1710.05703*. [Online]. Available: http://arxiv.org/abs/1710.05703

[36] N. H. Khan and A. Adnan, "Urdu optical character recognition systems: Present contributions and future directions," *IEEE Access*, vol. 6, pp. 46019–46046, 2018.

[37] H. Althobaiti and C. Lu, "A survey on arabic optical character recognition and an isolated handwritten arabic character recognition algorithm using encoded freeman chain code," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2017, pp. 1–6.

[38] P. V. Bhagyasree, A. James, and C. Saravanan, "A proposed framework for recognition of handwritten cursive english characters using DAG-CNN," in *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. (ICIICT)*, Apr. 2019, pp. 1–4.

[39] M. Rajalakshmi, P. Saranya, and P. Shanmugavadivu, "Pattern recognition-recognition of handwritten document using convolutional neural networks," in *Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS)*, Apr. 2019, pp. 1–7.

[40] C. Boufenar, A. Kerboua, and M. Batouche, "Investigation on deep learning for off-line handwritten arabic character recognition," *Cognit. Syst. Res.*, vol. 50, pp. 180–195, Aug. 2018.

[41] C. Yi, X. Yang, and Y. Tian, "Feature representations for scene text character recognition: A comparative study," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Washington, DC, USA, Aug. 2013, pp. 907–911.

[42] B. Zhang, W. Zhao, J. Liu, R. Wu, and X. Tang, "Character recognition in natural scene images using local description," in *Proc. Intell. Sci. Intell. Data Eng.-2nd Sino-Foreign-Interchange Workshop*, 2011, pp. 193–200.

[43] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, Oct. 2009.

[44] D. L. Smith, J. Field, and E. Learned-Miller, "Enforcing similarity constraints with integer programming for better scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 73–80.

[45] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 591–604.

[46] A. J. Newell and L. D. Griffin, "Multiscale histogram of oriented gradient descriptors for robust character recognition," in *Proc. IEEE Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1085–1089.

[47] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C. L. Tan, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognit.*, vol. 51, pp. 125–134, Mar. 2016.

[48] K. Elagouni, C. Garcia, F. Mamalet, and P. Sébillot, "Combining multi-scale character recognition and linguistic knowledge for natural scene text OCR," in *Proc. 10th IAPR Int. Workshop Document Anal. Syst.*, Mar. 2012, p. 120.

[49] F. Zhan and S. Lu, "ESIR: End-to-end scene text recognition via iterative image rectification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2059–2068.

[50] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards arbitrarily-oriented text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5571–5579.

[51] M. Jain, M. Mathew, and C. V. Jawahar, "Unconstrained scene text and video text recognition for arabic script," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, Nancy, France, Apr. 2017, pp. 26–30.

[52] O. Zayene, S. Masmoudi Touj, J. Hennebert, R. Ingold, and N. Essoukri Ben Amara, "Open datasets and tools for arabic text detection and recognition in news video frames," *J. Imag.*, vol. 4, no. 2, p. 32, Jan. 2018.

[53] S. Yousfi, S.-A. Berrani, and C. Garcia, "Contribution of recurrent connectionist language models in improving LSTM-based arabic text recognition in videos," *Pattern Recognit.*, vol. 64, pp. 245–254, Apr. 2017.

[54] O. Zayene, S. M. Touj, J. Hennebert, R. Ingold, and N. E. B. Amara, "Multi-dimensional long short-term memory networks for artificial arabic text recognition in news video," *IET Comput. Vis.*, vol. 12, no. 5, pp. 710–719, Aug. 2018.

[55] S. B. Ahmed, S. Naz, M. I. Razzak, and R. B. Yusof, "A novel dataset for English-arabic scene text recognition (EASTR)-42K and its evaluation using invariant feature extraction on detected extremal regions," *IEEE Access*, vol. 7, pp. 19801–19820, 2019.

[56] S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated arabic scene character recognition," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, Apr. 2017, pp. 46–51.

[57] M. Tounsi, I. Moalla, A. M. Alimi, and F. Lebouregois, "Arabic characters recognition in natural scenes using sparse coding for feature representations," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1036–1040.

[58] M. Bušta, Y. Patel, and J. Matas, "E2E-MLT—An unconstrained end-to-end method for multi-language scene text," in *Proc. Springer Asian Conf. Comput. Vis.*, Dec. 2018, pp. 127–143.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[62] M. Busta, L. Neumann, and J. Matas, "Deep TextSpotter: An end-to-end trainable scene text localization and recognition framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2204–2212.

[63] K. U. U. Rehman and Y. D. Khan, "A scale and rotation invariant urdu nastalique ligature recognition using cascade forward backpropagation neural network," *IEEE Access*, vol. 7, pp. 120648–120669, 2019.

[64] S. Y. Arafat and M. J. Iqbal, "Two stream deep neural network for sequence-based urdu ligature recognition," *IEEE Access*, vol. 7, pp. 159090–159099, 2019.

[65] A. Raza, I. Siddiqi, C. Djeddi, and A. Ennaji, "Multilingual artificial text detection using a cascade of transforms," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 309–313.

[66] A. Jamil, I. Siddiqi, F. Arif, and A. Raza, "Edge-based features for localization of artificial urdu text in video images," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1120–1124.

[67] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE Access*, vol. 7, pp. 18940–18950, 2019.

[68] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, early access, Feb. 2020, doi: 10.1109/LGRS.2019.2960026.

[69] Mu, Guo, and Liu, "A multi-scale and multi-level spectral-spatial feature fusion network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 1, p. 125, Jan. 2020.

[70] W. Zhang, J. Yu, H. Hu, H. Hu, and Z. Qin, "Multimodal feature fusion by relational reasoning and attention for visual question answering," *Inf. Fusion*, vol. 55, pp. 116–126, Mar. 2020.

[71] W. Xue, X. Dai, and L. Liu, "Remote sensing scene classification based on multi-structure deep features fusion," *IEEE Access*, vol. 8, pp. 28746–28755, 2020.

[72] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, and G. Li, "HAM-MFN: Hyperspectral and multispectral image multiscale fusion network with RAP loss," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 2020, doi: 10.1109/TGRS.2020.2964777.

[73] J. Qu, Y. Li, Q. Du, and H. Xia, "Hyperspectral and panchromatic image fusion via adaptive tensor and multi-scale retinex algorithm," *IEEE Access*, vol. 8, pp. 30522–30532, 2020.

[74] J. Kang, W. Lu, and W. Zhang, "Fusion of brain PET and MRI images using tissue-aware conditional generative adversarial network with joint loss," *IEEE Access*, vol. 8, pp. 6368–6378, 2020.

[75] R. Dong, M. Liu, and F. Li, "Multilayer convolutional feature aggregation algorithm for image retrieval," *Math. Problems Eng.*, vol. 2019, Jun. 2019, Art. no. 9794202.

[76] E. Y. Huan and G. H. Wen, "Multilevel and multiscale feature aggregation in deep networks for facial constitution classification," *Comput. Math. Methods Med.*, vol. 2019, Dec. 2019, Art. no. 1258782.

[77] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," 2014, *arXiv:1406.2227*. [Online]. Available: http://arxiv.org/abs/1406.2227

[78] F. Zhan, S. Lu, and C. Xue, "Verisimilar image synthesis for accurate detection and recognition of texts in scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 249–266.
[79] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion GAN for image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3653–3662.
[80] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.
[81] F. Chollet. Keras. Github, 2015. [Online]. Available: https://github.com/fchollet/keras
[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[83] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4148–4158.
[84] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognit. Model.*, vol. 5, no. 3, p. 1, 1988.
[85] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, W. Khlif, M. M. Luqman, J.-C. Burie, C.-L. Liu, and J.-M. Ogier, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification—RRC-MLT," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 1454–1459.
[86] X. Bai, C. Yao, and W. Liu, "Strokelets: A learned multi-scale mid-level representation for scene text recognition," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2789–2802, Jun. 2016.
[87] S. Gao, C. Wang, B. Xiao, C. Shi, W. Zhou, and Z. Zhang, "Learning co-occurrence strokes for scene character recognition based on spatiality embedded dictionary," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5956–5960.
[88] M. Tounsi, I. Moalla, and A. M. Alimi, "Supervised dictionary learning in BoF framework for scene character recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3987–3992.
[89] C. Chen, D. H. Wang, and H. Wang, "Scene character recognition using PCANet," in *Proc. 7th Int. Conf. Internet Multimedia Comput. Service*, 2015, pp. 1–4.
[90] M. Ali and H. Foroosh, "Character recognition in natural scene images using rank-1 tensor decomposition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2891–2895.
[91] S. Tian, S. Lu, B. Su, and C. L. Tan, "Scene text recognition using co-occurrence of histogram of oriented gradients," in *Proc. 12th IEEE Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 912–916.

ASGHAR ALI CHANDIO received the B.S. degree in information technology from the Institute of Information Technology, University of Sindh, Pakistan, in 2008, and the M.S. degree in information technology from the Quaid-e-Awam University of Engineering, Science and Technology (QUEST), Pakistan, in 2014. From 2010 to 2015, he was a Lecturer with the Department of Information Technology, QUEST, where he has been an Assistant Professor, since 2016. He is currently a Ph.D. Research Scholar with the School of Engineering and Information Technology, University of New South Wales, Canberra, Australia. His major research interests include machine learning, deep learning, handwritten text recognition, text extraction in natural scene images, document analysis, and semantic text similarity matching.

MD. ASIKUZZAMAN (Member, IEEE) received the B.Sc. degree in electronics and telecommunication engineering from the Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, in 2010, and the Ph.D. degree in electrical engineering from the University of New South Wales, Canberra, Australia, in 2015, under a very competitive University International Postgraduate Award Scholarship. From 2015 to 2019, he was a Research Associate with the School of Engineering and Information Technology, University of New South Wales, where he is currently a Senior Research Associate. His current research interests include 2D and 3D video watermarking, privacy preservation, deep learning, medical imaging, and video coding. He was the Technical Program Chair for the 2018 International Conference on Digital Image Computing: Techniques and Applications. He is also serving as an Associate Editor for IEEE Access.

MARK R. PICKERING (Member, IEEE) was born in Biloela, Australia, in 1966. He received the B.Eng. degree in electrical engineering from the Capricornia Institute of Advanced Education, Rockhampton, Australia, in 1988, and the M.Eng. and Ph.D. degrees in electrical engineering from the University of New South Wales, Canberra, Australia, in 1991 and 1995, respectively. He was a Lecturer, from 1996 to 1999, a Senior Lecturer, from 2000 to 2009, and an Associate Professor, from 2010 to 2017, with the School of Electrical Engineering and Information Technology, University of New South Wales, where he is currently a Professor. His research interests include video and audio coding, medical imaging, data compression, information security, data networks, and error-resilient data transmission.