Neurocomputing 243 (2017) 80–87

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Urdu Nastaliq recognition using convolutional–recursive deep learning

Saeeda Naz[a,b], Arif I. Umar[a], Riaz Ahmad[c], Imran Siddiqi[d], Saad B Ahmed[e], Muhammad I. Razzak[e,*], Faisal Shafait[f]

[a] Department of Information Technology, Hazara University, Mansehra, Pakistan
[b] GGPGC No.1, Abbottabad, Higher Education Department, Pakistan
[c] University of Kaiserslautern, Germany
[d] Bahria University, Islamabad, Pakistan
[e] King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
[f] National University of Sciences and Technology (NUST), Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 31 July 2016
Revised 16 January 2017
Accepted 27 February 2017
Available online 8 March 2017

Communicated by Ning Wang

Keywords:
RNN
CNN
Urdu OCR
BLSTM
MDLSTM
CTC

ABSTRACT

Recent developments in recognition of cursive scripts rely on implicit feature extraction methods that provide better results as compared to traditional hand-crafted feature extraction approaches. We present a hybrid approach based on explicit feature extraction by combining convolutional and recursive neural networks for feature learning and classification of cursive Urdu Nastaliq script. The first layer extracts low-level translational invariant features using Convolutional Neural Networks (CNN) which are then forwarded to Multi-dimensional Long Short-Term Memory Neural Networks (MDLSTM) for contextual feature extraction and learning. Experiments are carried out on the publicly available Urdu Printed Text-line Image (UPTI) dataset using the proposed hierarchical combination of CNN and MDLSTM. A recognition rate of up to 98.12% for 44-classes is achieved outperforming the state-of-the-art results on the UPTI dataset.

1. Introduction

Feature extraction is one of the most significant steps in any machine learning and pattern recognition task. In case the patterns under study are images, selection of salient features from raw image pixels not only enhances the performance of the learning algorithm but also reduces the dimensionality of the representation space and hence the computational complexity of the classification task. As a function of the problem under study, a variety of statistical and structural features computed at global or local levels have been proposed over the years [1,2]. Extraction of these manual features is expensive in the sense that it requires human expertise and domain knowledge so that the most pertinent and discriminative set of features could be selected. These limitations of manual features motivated researchers to extract and select automated and generalized features using machine learning models, especially, for problems involving visual patterns such as object detection [3], character recognition [4] and face detection [5].

A number of studies have shown that convolutional neural network (CNN), a special type of multi-layer neural network, realizes high recognition rates on a variety of classification problems. CNN represents a robust model that is able to recognize highly variable patterns [6] (such as varying shapes of handwritten characters) and is not affected by distortions or simple transformations of the geometry of patterns. In addition, the model does not require preprocessing to recognize visual patterns or objects as it is able to perform recognition from the raw pixels of images directly. Moreover, these visual patterns are easily detected regardless of their position in the image by observing CNN's shared weight property. In shared weights property, the CNN model uses replicated filters that have identical weight vectors and have local connectivity. This weight sharing eliminates the redundancy of learning visual patterns at each distinct location, consequently limiting each neuron in the model to have local connectivity to a local region of the entire image. Furthermore, weight sharing and local connectivity reduces over-fitting and computational complexity, giving rise to increased learning efficiency and improved generalizations for machine translation. Due to this robust weight sharing property of CNN architecture, it is sometimes known as shift invariant or shared weight neural network or space invariant artificial neural network. The general architecture of a CNN model illustrated in Fig. 1. The first layer, generally termed as the feature extractor part of the CNN, learns lower order specific features from the raw

* Corresponding author.
E-mail address: mirpak@gmail.com (M.I. Razzak).

http://dx.doi.org/10.1016/j.neucom.2017.02.081
0925-2312/

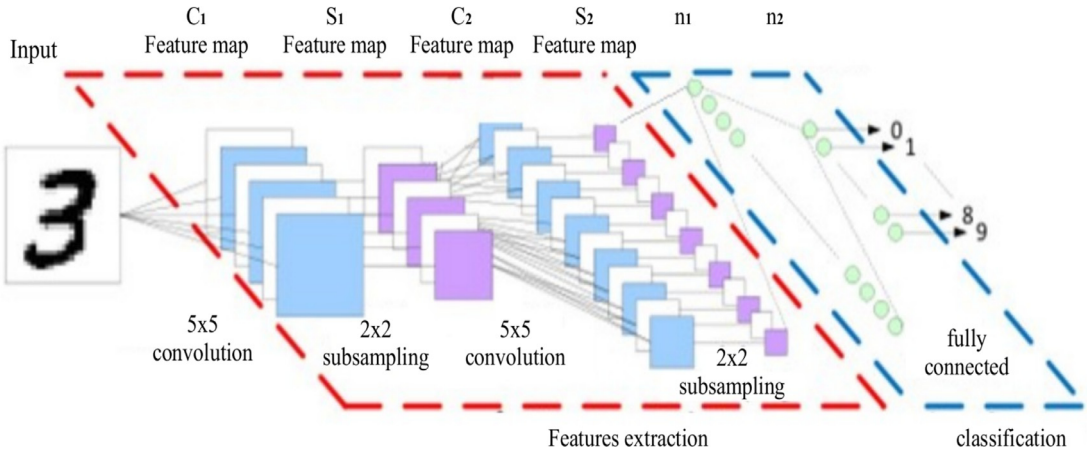

Fig. 1. General architecture of CNN [7].

image pixels [6]. The last layer is the trainable classifier which is used for classification. The feature extractor part also comprises two alternate operations of convolution filtering and sub-sampling. The illustrated model shows a convolution filtering (C) of size 5 × 5 pixels and a down sampling ratio (S) of 2, represented by $C1$, $S1$, $C2$ and $S2$ respectively.

In a number of studies, CNN model has been used to extract features while another model is applied for classification [8–10]. These include applications like emotion recognition [11], digit and character recognition [12–15] and visual image recognition [12]. Huang and LeCun [6] conclude that CNN learns optimal features from the raw images but it is not always optimal for classification. Therefore, the authors merged CNN with SVM, i.e. the features extracted by the CNN are fed to the SVM for classification of generic objects. These generic objects included animals, human figures, airplanes, cars, and trucks. The hybrid system realized a recognition rate of upto 94.1% as compared to 57% (only SVM) and 92.8% (only CNN).

In [8], Lauer et al. employed CNN to extract features without prior knowledge on the data for recognition of handwritten digits. Combining the features learned by the CNN with SVM, the authors report a recognition rate of 99.46% (after applying elastic distortions) on the MNIST database. In another similar study, Niu and Suen [9] employed CNN as a trainable feature extractor from raw images and used SVM as recognizer to classify the handwritten digits in the MNIST digit database. This hybrid systems realized a recognition rate of 94.40%.

Donahue et al. [10] investigated the combination of CNN and LSTM (Long-Short-Term-Memory network) for visual image recognition on UCF-101 database [16], Flickr30k database [17] and the COCO2014 database [18]. The combination reported promising classification results on these databases. In another interesting work [19], authors report the combination of convolution and recursive neural network for object recognition. CNN is used for extraction of lower features from images of RGB-D dataset followed by RNN forest for feature selection and classification. Similarly, Bezerra et al. [20] integrated a multi-dimensional recurrent neural network (MDRNN) with SVM classifiers to improve the character recognition rates. In [21], Chen et al. proposed T-RNN (transferred recurrent neural network). The authors extracted visual features using CNN and detected fetal ultrasound standard planes from ultrasound videos reporting very promising results. In a later study [22], the authors combined a fully convolutional network (FCN) and recurrent network for segmentation of 3D medical images. The proposed technique was evaluated on two databases and realized promising results.

Accurate sequence labeling and learning is one of the most important tasks in any recognition system. The sequence labeling needs not only to learn the long sequences but also to distinguish similar patterns from one another and assign labels accordingly. Hidden Markov models (HMM) [23], Conditional Random Field (CRF) [6], Recurrent Neural Network (RNN) and variants of RNN (BLSTM and MDLSTM) [4,24–26] have been effectively applied to different sequence learning based problems. A number of studies [27–30] have concluded that LSTM outperforms HMMs on such problems.

This paper presents a new convolutional–recursive deep learning model which is a combination of CNN and MDLSTM. The proposed model is mainly inspired from the one presented by Raina et al. [31] and is applied to solve character recognition problem on Urdu text in the Nastaliq script. The proposed system employs CNN for automatically extracting lower level features from a large MNIST dataset. The learned kernels are then convolved with text line images for extraction of features while the MDLSTM model is used as the classifier. Each (complete) text-line image is fed as a sequence of frames denoted by $X = (x_1, x_2, \ldots, x_i)$ with its corresponding target sequence denoted as $T = (t_1, t_2, \ldots, t_j)$. The input sequence of frames (X) is the set of all input character symbols from the text line images and target sequence is a set of all sequence of alphabets of labels (L) in ground truth or transcription file, i.e., $T = L^*$. The size of target sequence set (T) is less than or equal to input sequence set (X), i.e., $|T| \leq |X|$.

Let the data sample be composed of sequence pairs (X, T) taken from the training set (S) independently from the fixed distribution of both sequences $D_{X \times T}$. The training set (S) is used to train the sequence labeling algorithm $f: X \rightarrow T$ and then assign labels to the character sequence of the test set (S') having sample distribution ($S' \in D_{X \times T}$). The label error rate ($Error^{lbl}$) is computed as follows.

$$Error^{lbl} = \frac{1}{T} \sum_{(X,T) \in S'} ED(h(X), T) \tag{1}$$

where $ED(h(X), T)$ is the edit distance between the input character sequence (X) and the target sequence (T) and is employed to compute the error rates.

The main contributions of this study include:

- Demonstration of how convolutional–recursive architectures can be used to effectively recognize cursive text which forbids traditional feature learning due to the large number of classes/recognition units involved.

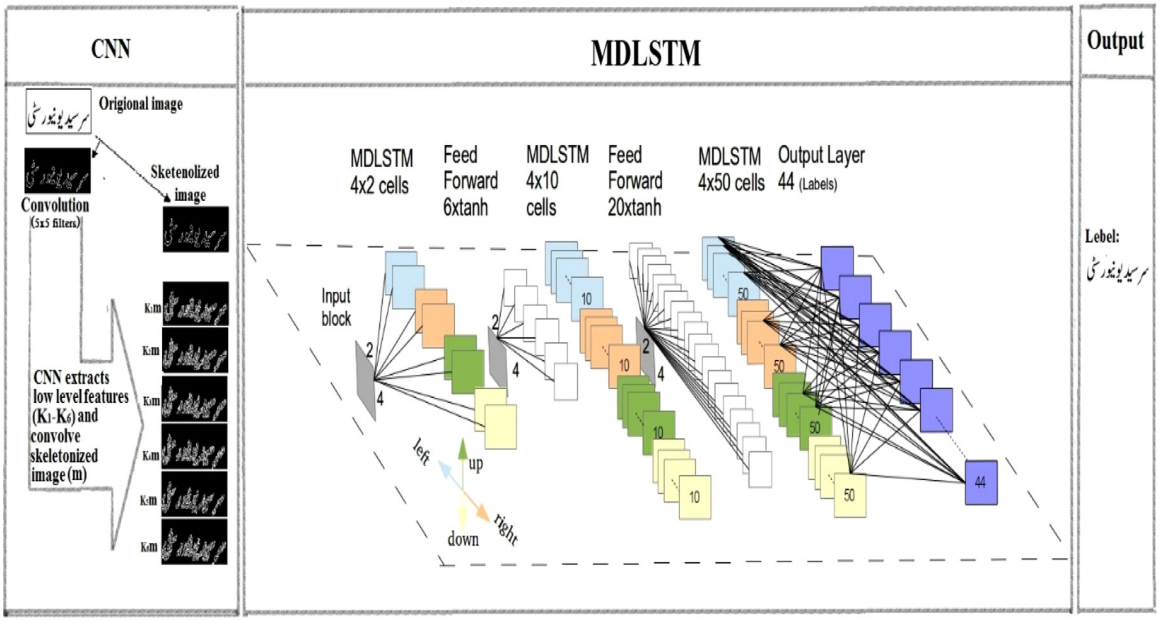

Fig. 2. An overview of convolutional–recursive deep learning model: a single CNN layer extracts low level features from Urdu textline. Six filters ($K1$–$K6$) taken form the first layer of CNN and filter with the contoured image. The convolutionalized images and contour representation of textline are given as input to a MDLSTM with random weights. Each of the neurons then recursively maps the features into a lower dimensional space. The concatenation of all the resulting vectors forms the final feature vector for a Connectionist Temporal Classification (CTC) output layer.

- Addressing the challenge of learning feature extraction from a huge number of ligature classes (over 20,000 in Urdu) by proposing a novel transfer learning mechanism in which representative features are learned from only a small set of classes.
- Showcasing the generalization of the feature extractor by training it on isolated handwritten English digits and then applying it for cursive Urdu machine printed text recognition.
- Evaluation performed on a benchmark UPTI dataset, thereby facilitating more informative future evaluations.

The rest of this paper is organized as follows. Section 2 details the proposed methodology of combining CNN and MDLSTM for character recognition. Experimental results along with a comparison with the existing systems are presented in Section 3 while Section 4 concludes the paper.

2. Convolutional–recursive MDLSTM based recognition system

In this section, we present the novel convolutional–recurisve deep learning technique proposed in this study. The proposed technique for recognition of Urdu text lines relies on machine learning based features extracted using the CNN. Features are learned using the MNIST digit database [32]. The first convolutional layer of the CNN learns generic features from images of digits. These features are then computed for Urdu text lines and are fed to the MDLSTM for learning higher level transient features and classification. Prior to feature extraction, the text line images are normalized in size by preserving the aspect ratio while the pixel values in the image are standardized using mean and standard deviation. The general idea of learning the features through CNN and performing classification using LSTM is illustrated in Fig. 2. The details on different key steps of the technique are presented in the following sections.

2.1. Dataset

We have realized the proposed system on Urdu Printed Text Image (UPTI) dataset [33]. The database comprises more than 10,000 Urdu text lines generated synthetically in Nastaliq font from a well-known Urdu newspaper (Jang).[1] This dataset covers a wide range of topics on political, social, and religious issues. The distribution of the database into training, validation and test sets is summarized in Table 1. In supervised classification, class labels are required to be generated for data elements in the input space. This is known as ground truth or transcription. LSTM being a supervised learning model, also requires the ground truth values for each image in the input space. In our study, the shape variations of a character including beginning, middle, ending and isolated forms (of a basic character such as "ب") are grouped into a single class and are assigned one label. This produces a total of 44 unique labels at character level transcription. Among these labels, 38 labels represent basic characters, 4 labels represent the commonly occurring secondary characters (noonghuna, wawohamza, haai, and yeahamza), 1 label for SPACE and 1 extra label for the blank. The ground truth transcription of each text line is provided as an input to the network along with the sequence of feature vectors. An example text line and its ground truth transcription are illustrated in Fig. 3.

Table 1
Distribution of UPTI dataset in training, validation and test sets.

Sets	Text lines	Characters
Training set	6800	506,569
Validation set	1600	137,785
Test set	1600	126,985

2.2. Normalization and standardization

Data normalization, in general, refers to fit the data within unity and is mostly realized using the following equation.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

[1] http://jang.com.pk

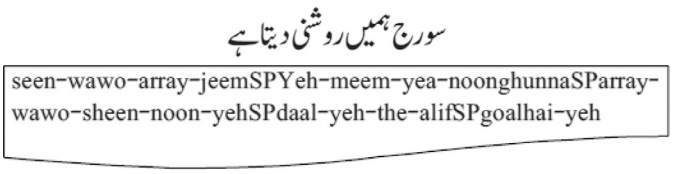

Fig. 3. A sentence in Urdu: (a) Text line image. (b) Ground truth or transcription.

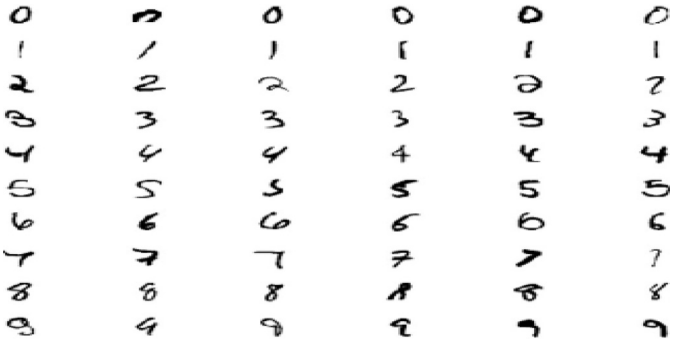

Fig. 4. Sample images of digits (0–9) from the MNIST dataset.

In our case, we deal with 8-bit grayscale images having pixel values in the interval [0–255]. We normalize the pixel values by dividing each value by 255 hence ensuring that the normalized pixel values are in the interval [0–1]. Likewise, we also carry out standardization of the pixel values. Standardization provides meaningful information about each data point and gives a general idea about the outliers (values above or below a *z*-score). Standardization is carried out by subtracting the mean intensity from each pixel value of the image and dividing by the standard deviation of the pixel values as summarized in the following equation.

$$X_{new} = \frac{X - \mu}{\sigma} \tag{3}$$

Where
X represents a data point
μ The average of all the sample data points
σ The sample standard deviation
The $\overline{X}_s$ (average) and $\sigma_{x,\,s}$ (standard deviation) are later reused in normalizing the test and validation data.

2.3. Feature extraction using CNN

We employed a five layered CNN model (Fig. 1) for extraction of generic and abstract features from 60,000 handwritten digits images of the MNIST database. The major motivation of using this database for learning of features is that segmentation of text into words or sub-words is a challenging problem in cursive scripts like Nastaliq. Since CNNs require labeled training data in a large amount, manually creating segmented data from Nastaliq ligatures is not feasible. Our hypothesis is that the isolated digits consist of strokes (horizontal, vertical, diagonal, circular and oval etc.) which also make the foundation of any other writing style such as Urdu Nastaliq script – essentially writing is stroke-based in all scripts and languages. Sample digit images of the database are illustrated in Fig. 4. On the training set, we realized an error rate of 0.11% (classification rate of 99.89%) on the MNIST dataset as illustrated in Fig. 5.

The first convolution layer $C1$ of the CNN extracts abstract and generic features such as lines, edges and corner information from the raw pixels of the image. The inner layers are known to extract relatively low level features. We, therefore, selected features from the first convolutional layer $C1$ in the form of convolution filters or kernels ($K1$–$K6$) as shown in Fig. 6. These kernels are then used to convolve the Urdu text line images (m) and result in convolutionalized text line images $mK_1 = m * K_1$, $mK_2 = m * K_2$, ...$mK_6 = m * K_6$ for training the MDLSTM as discussed in the next section.

2.4. Learning and training using MDLSTM

As discussed earlier, the system is trained using a multi-dimensional LSTM. LSTM represents a variant of the recurrent neural networks (RNN) [34]. Recurrent neural networks are artificial neural networks with cyclic paths or loops. The loops not only allow dynamic temporal behavior of the network but also enable the network to process arbitrary sequences of inputs through internal memory. These networks, however, cannot learn long term dependencies. The problem was addressed by introduction of LSTM–RNN [35] which are capable of retaining and correlating information for longer delays. The basic unit of LSTM architecture is a memory block with memory cells and three gates (input, forget and output). The standard one dimensional LSTM network can also be extended to multiple dimensions by using n self connections with n forget gates [36].

To train the LSTM on Urdu text lines, we first find the skeletonized image of each line. The six kernels ($K_1 - K_6$) extracted through CNN are then used to convolve the skeletonized images of text lines. The skeletonized image of a text line (Fig. 7(b)) and the six convolved images (Fig. 7(c)–(h)) are used as features and are fed to the MDLSTM for training as outlined in Fig. 2. As discussed earlier, the kernels are extracted using the MNIST database as the digit images share many common strokes with the Urdu text and are already segmented.

The values of different parameters for MDLSTM classifier are shown in Table 2. The extracted feature vector is divided into 4 $\times$ 1 small patches having a height of 4 rows and width of 1 column and fed to the MDLSTM with the corresponding ground truth. The MDLSTM model scans the input patch in all four directions. The network comprises 3 hidden layers of LSTM cells having sizes of 2, 10 and 50 respectively. All these hidden layers are fully connected and each of them is further divided into two sub-sampling layers having sizes of 6 and 20 respectively. The sub-sampling layers are feed-forward tanh layers. The features are collected into 4 $\times$ 2 hidden blocks and these blocks are then fed to the feed forward layer which employs tanh summation units for the cell activation. The MDLSTM activation finally collapses into a one dimensional sequence. The Connectionist Temporal Classification (CTC) layer [37] then labels the contents of the one dimensional sequence. The CTC output layer has the same number of labels (L) of target sequences (T) with one additional label for blank/null, hence the total labels (L'^*) are $L \cup \{blank/null\}$. Each element of L'^* is known to be a path for each input character sequence x and is denoted as η. The CTC output layer computes the conditional probabilities for η against each input sequence x as shown in the following.

$$p(\eta|x) = \prod_{n=1}^{N} Y_{\eta t}^{t} \tag{4}$$

Where $Y_{\eta t}^{t}$ is output activation against input unit at time t.

We have used gradient descent optimizer to reduce the loss. The loss is obtained by Connectionist Temporal Classification (CTC) loss function. Assuming S to be a training set containing pairs of input and target sequences (X, T), provided $|T| \leq |X|$, the objective function Ω for CTC is the negative log probability of the network correctly labelings all of S.

$$\Omega = -\sum_{(X,T)\in S} ln\ p(T/X) \tag{5}$$

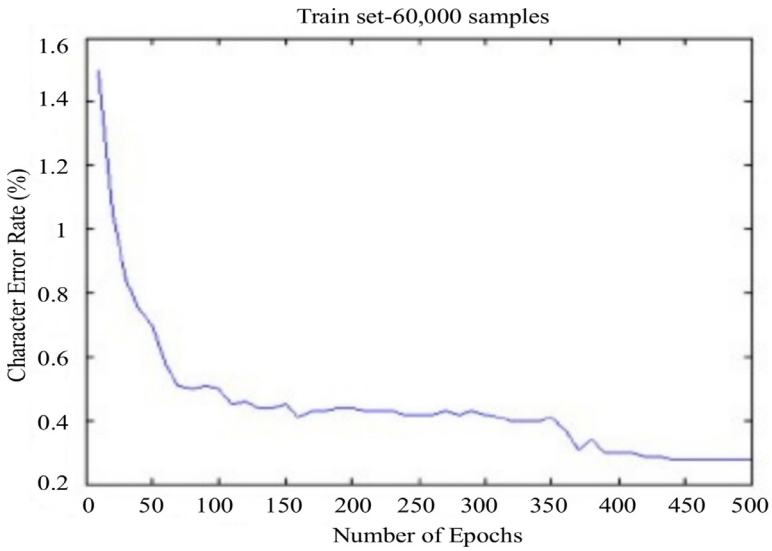

Fig. 5. Error rate of CNN on 60,000 samples images from MNIST dataset on different number of epochs.

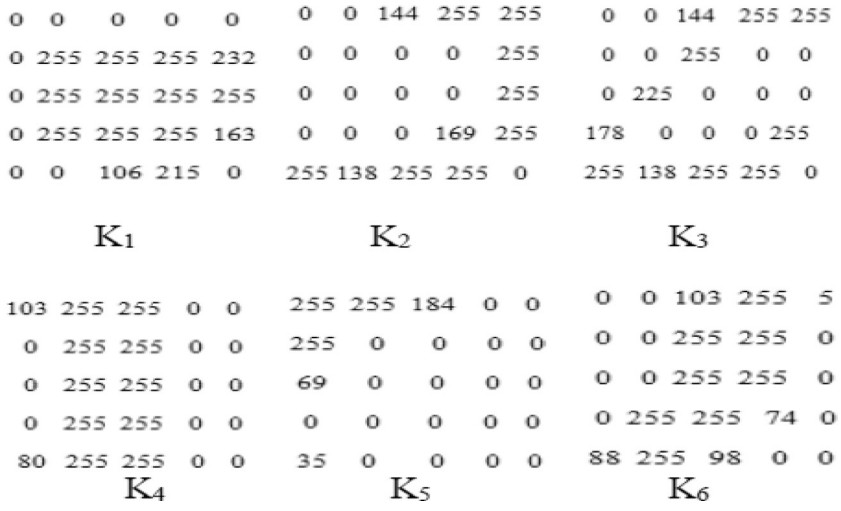

Fig. 6. Selected feature kernels *K*1, *K*2, *K*3, *K*4, *K*5 and *K*6.

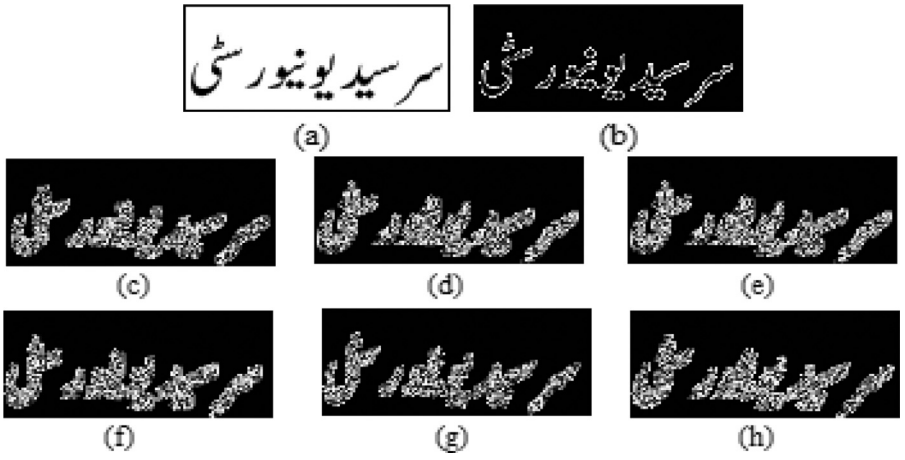

Fig. 7. Urdu text line (a) Original image (b) Skeletonized image (c)–(h) Six convolutionalized images representing results of filtering the skeletonized text lines image (*m*) with each of the kernels (*K*1–*K*6) extracted by CNN.

Table 2
Parameters values for training the MDLSTM network using automatic features extracted by CNN.

Parameters	Values	Horizontal sampling	Vertical sampling
Input block size	4 × 1	1	4
Hidden block size	4 × 2 and 4 × 2	2	4
Subsample sizes	6 and 20	–	–
Hidden sizes	2, 10 and 50	–	–
Learning rate	1 × 10–4	–	–
Momentum	0.9	–	–
Total network weights	143,581	–	–

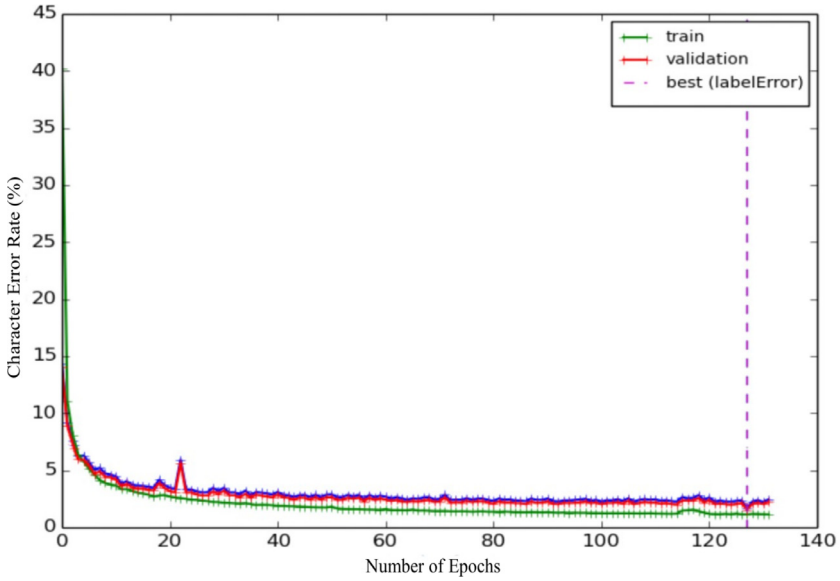

Fig. 8. Training of MDLSTM on different number of epochs using CNN features.

Table 3
Accuracies achieved by hybrid Urdu recognition system.

Set	Accuracy (%)
Training	99.4
Validation	98.73
Testing	98.12

The network is trained by using gradient descent optimizer with a learning rate of 1 x 10^{-4} and a momentum of 0.9. First, Ω is differentiated with respect to the outputs. Backpropagation is then used through time to find the derivatives with respect to the weights.

The total number of weights of the network cells are 143,581. The training was stopped when there was no improvement in the error rate of validation set for 30 consecutive epochs.

The curves for character error rates on different number of epochs for training and validation sets are illustrated in Fig. 8. The classification rates read at 99.40% and 98.73% on training and validation sets respectively on epoch 128. Table 3 summarizes the character error rates on training set and validation set for best network.

3. Results and comparative analysis

Table 4 compares the performance of the proposed technique with the existing systems evaluated on the UPTI database. These include implicit segmentation based approaches [38–41] and the segmentation free approach using context shape matching technique presented in [33].

The meaningful comparisons of our system are possible with the work of Ul-Hassan et al. [38] and Ahmed et al. [39] where the authors employed BLSTM on raw pixels. Ul-Hasan et al. [38] reported an error rate of 5.15% while Ahmed et al. [39] achieved an error rate of 11.06%. BLSTM scans images in only horizontal dimension hence it is likely to make errors in the presence of excessive dots or diacritics or vertically overlapped ligatures. It should, however, be noted that in [38], authors employ 10,064 text lines with 46% in the training set, 34% in the validation set and 20% in the test set. In [39], authors employ the extended version of the UPTI database where different degradations are applied to the original text lines to increase the database size. A total of 27,195 text lines are employed in [39] with 45.6% in training set, 43.9% in validation set and 10.4% in the test set. Further comparison is possible with our previous works [40,41] where we extracted manual features and employed MDLSTM using the same UPTI dataset. Recognition rates of 94.97% and 96.4% are reported in [40,41] respectively. The experimental protocol in [40,41] is exactly the same as in the present study. Our proposed technique realizes better performances reporting an error rate of 1.88% using CNN based features as compared to 3.6% and 5.25% in the work of Naz et al. [40,41], representing an over 50% reduction in the error rate. The authors in [33] employed segmentation free approach to extract contour features and then applied context shape matching technique. Recognition rates of upto 91% are reported in this study.

Fig. 9 shows the recognition results of different systems [38–41] on two sample text-line images from the UPTI dataset. It can be noticed that the BLSTM could not learn some complex ligatures as compared to the MDLSTM network, though it is more efficient with respect to the execution time. The character "noon" (ںـ) in the

Table 4
Comparison of Urdu recognition system on UPTI dataset.

Systems	Segmentation	Features	Classifier	Accur. (%)
Ul-Hassan et al. [38]	Implicit	Pixels	BLSTM	94.85
Ahmed et al. [39]	Implicit	Pixels	BLSTM	88.94
Naz et al. [40]	Implicit	Statistical	MDLSTM	94.97
Naz et al. [41]	Implicit	Statistical	MDLSTM	96.4
Sabbour and Shafait [33]	Holistic	Contour	BLSTM	91
Proposed	Implicit	Convolutional	MDLSTM	98.12

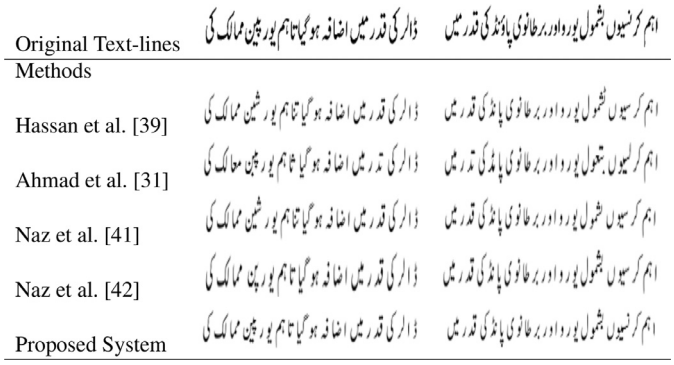

Fig. 9. Recognition results of different systems on sample Urdu text-lines from UPTI dataset.

second word (کرنسوں) is deleted. In the third word (بشمول), "bay" (ب) is replaced with "teh" (ٹ). In word (پاؤند), the character "hamzawawo" (ؤ) is missed in the recognition step in Ul-Hasan et al.'s network [38] as shown in Fig. 9(b). The proposed system recognized the lines correctly and there is just one error in first sentence that is the deletion of the character "hamzawawo" (ؤ) in word (پاؤند) as shown in Fig. 9(f) while the second text-line is perfectly recognized.

4. Conclusion

We proposed a convolutional–recursive deep learning model based on a combination of CNN and MDLSTM for recognition of Urdu Nastaliq characters. The CNN is used to extract low level translational invariant features and the extracted features are fed to MDLSTM. The MDLSTM extracts high order features and recognizes the given Urdu text line image.The combination of CNN and MDLSTM proved to be an effective feature extraction method and outperformed the state of the art systems on a public dataset. Without extracting traditional features, convolutional–recursive deep learning (CNN–MDLSTM) based system achieved accuracy of 98.12% on UPTI dataset.

While the present study employs CNN for feature extraction and MDLSTM for classification, it would also be interesting to train the complete framework (CNN+LSTM) and compare the performances with other models. It is also worth investigating to extend the proposed combination of CNN and MDLSTM model to other applications. The application of this work is easy to extend to the sub-set of Urdu like printed/synthetic scripts such as Arabic and Persian. We can also apply this model to handwritten Urdu, Arabic or Persian language after studying the different handwriting styles of characters by writers in these languages.

References

[1] D. Trier, A. Jain, T. Taxt, Feature extraction methods for character recognition-a survey, Pattern Recognit. 29 (4) (1996) 641–662.
[2] S. Naz, K. Hayat, M.I. Razzak, M.W. Anwar, S.A. Madani, S.U. Khan, The optical character recognition of urdu-like cursive scripts, Pattern Recognit. 47 (3) (2014) 1229–1248.
[3] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 2, IEEE, 1999, pp. 1150–1157.
[4] S. Naz, A.I. Umar, R. Ahmad, M.I. Razzak, S.F. Rashid, F. Shafiat, Urdu Nastaliq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks, SpringerPlus 5 (1) (2016) 1–16.
[5] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Trans. Image Process. 19 (6) (2010) 1635–1650.
[6] F.J. Huang, Y. LeCun, Large-scale learning with SVM and convolutional nets for generic object categorization, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, IEEE, 2006, pp. 284–291.
[7] M. Peemen, B. Mesman, H. Corporaal, Efficiency optimization of trainable feature extractors for a consumer platform, in: Proceedings of the Thirteenth International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, 2011, pp. 293–304.
[8] F. Lauer, C.Y. Suen, G. Bloch, A trainable feature extractor for handwritten digit recognition, Pattern Recognit. 40 (6) (2007) 1816–1824.
[9] X.X. Niu, C.Y. Suen, A novel hybrid CNN-SVM classifier for recognizing handwritten digits, Pattern Recognit. 45 (4) (2012) 1318–1325.
[10] J. Donahue, K. Saenko, T. Darrell, U.T. Austin, U. Lowell, U.C. Berkeley, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2625–2634.
[11] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimed. 16 (8) (2014) 2203–2213.
[12] Q.A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, Curran Associates, Inc., 2012, pp. 1097–1105.
[13] P. Sermanet, S. Chintala., Y. LeCun, Convolutional neural networks applied to house numbers digit classification, in: Proceedings of the 2012 IEEE International Conference on Pattern Recognition (ICPR), 2012, pp. 3288–3291.
[14] S. Pan, Y. Wang, C. Liu, X. Ding, A discriminative cascade CNN model for offline handwritten digit recognition, in: Proceedings of the 2015 IEEE IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 501–504.
[15] D.C. Ciresan, U. Meier, L.M. Gambardella, J. Schmidhuber, Convolutional neural network committees for handwritten character classification, in: Proceedings of the 2011 IEEE International Conference on Document Analysis and Recognition (ICDAR), 2011, pp. 1135–1139.
[16] K. Soomro, A.R. Zamir, M. Shah, in: UCF101: A dataset of 101 human actions classes from videos in the wild, 2012. arXiv preprint: arXiv:1212.0402.
[17] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, TACL 2 (2014) 67–68.
[18] P.D.T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, C.L.Z. Ar, Microsoft COCO: common objects in context, in: Proceedings of the 2014 European Conference on Computer Vision (ECCV), in: Lecture Notes in Computer Science, 8693, 2014, pp. 740–755.
[19] R. Socher, B. Huval, B. Bath, C.D. Manning, A.Y. Ng, Convolutional–recursive deep learning for 3d object classification, Advances in Neural Information Processing Systems, Curran Associates, Inc., 2012, pp. 665–673.
[20] B.L.D. Bezerra, C. Zanchettin, V.B.D. Andrade, A MDRNN-SVM hybrid model for cursive offline handwriting recognition, Artificial Neural Networks and Machine Learning (ICANN), 2012, pp. 246–254.
[21] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, P.-A. Heng, Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks, Medical Image Computing and Computer-Assisted Intervention (MICCAI-2015), Lecture Notes in Computer Science, 9349, 2015, pp. 507–514.
[22] J. Chen, L. Yang, Y. Zhang, M. Alber, D. Chen, Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation, in: Proceedings of the 2016 Neural Information Processing Systems (NIPS), 2016.
[23] H.K. Al-Omari, M.S. Khorsheed, System and methods for Arabic text recognition based on effective Arabic text feature extraction. U.S. Patent 8,369,612, issued February 5, 2013.
[24] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (5) (2009) 855–868.

[25] S.B. Ahmed, S. Naz, S. Swati, M.I. Razzak, Ucom offline dataset an urdu handwritten dataset generation, Int. Arab J. Inf. Technol. 14 (2017) 228–241.
[26] S. Naz, S.B. Ahmed, R. Ahmad, M.I. Razzak, Zoning features and 2D LSTM for urdu text-line recognition, Proc. Comput. Sci. 96 (1) (2016) 16–22.
[27] M. Liwicki, A. Graves, H. Bunke, J. Schmidhuber, A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks, in: Proceedings of the Ninth International Conference on Document Analysis and Recognition, 1, IEEE, 2007, pp. 367–371.
[28] A. Graves, Supervised sequence labelling, in: Supervised Sequence Labelling with Recurrent Neural Networks, Springer Berlin Heidelberg, 2012, pp. 5–13.
[29] R. Ahmad, S. Naz, M.Z. Afzal, H.S. Amin, T. Breuel, Robust optical recognition of cursive Pashto script using scale, rotation and location invariant approach, PLoS One 10 (9) (2015a) 1–16.
[30] R. Ahmad, M.Z. Afzal, S.F. Rashid, M. Liwicki, T. Breuel, Scale and rotation invariant OCR for Pashto cursive script using MDLSTM network, in: Proceedings of the Thirteenth International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015b, pp. 1101–1105.
[31] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, in: Proceedings of the Twenty-fourth International Conference on Machine Learning, 2007, pp. 759–766.
[32] Y. LeCun, C. Cortes, C.J. Burges, in: The MNIST database of handwritten digits, 1998.
[33] N. Sabbour, F. Shafait, A segmentation-free approach to Arabic and Urdu OCR, in: Proceedings of the 2013 SPIE International Society for Optics and Photonics, 86580, 2013.
[34] H. Jaeger, Tutorial on Training Recurrent Neural Networks, Covering BPPT, RTRL, EKF and the "Echo State Network" Approach, GMD-Forschungszentrum Informationstechnik, 2002.
[35] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[36] A. Graves, J. Schmidhuber, Offline handwriting recognition with multidimensional recurrent neural networks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2009, pp. 545–552.
[37] A. Graves, S. Fernndez, F.J. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 2006 International Conference on Machine Learning (ICML), 2, 2006, p. 369â;;376.
[38] A. Ul-Hasan, S.B. Ahmed, F. Rashid, F. Shafait, T.M. Breuel, Offline printed urdu Nastaleeq script recognition with bidirectional LSTM networks, in: Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2013, pp. 1061–1065.
[39] S.B. Ahmed, S. Naz, M.I. Razzak, S.F. Rashid, M.Z. Afzal, T.M. Breuel, Evaluation of cursive and non-cursive scripts using recurrent neural networks, Neural Comput. Appl. 27 (3) (2016) 603–613.
[40] S. Naz, A.I. Umar, R. Ahmad, S.B. Ahmed, S.H. Shirazi, M.I. Razzak, Urdu Nastaliq text recognition system based on multi-dimensional recurrent neural network and statistical features, Neural Comput. Appl. 26 (8) (2015) 1–13.
[41] S. Naz, A.I. Umar, R. Ahmad, S.B. Ahmed, I. Siddiqi, M.I. Razzak, Offline cursive Nastaliq script recognition using multidimensional recurrent neural networks with statistical features, Neurocomputing 177 (2016) 228–241.

Saeeda Naz an Assistant Professor by designation and Head of Computer Science Department at GGPGC No.1, Abbottabad, Higher Education Department of Government of Khyber-Pakhtunkhwa, Pakistan, since 2008. She did her Ph.D. in Computer Science from Hazara University, Department of Information Technology, Mansehra, Pakistan. She has published two book chapters and more than 30 papers in peer reviewed national and international conferences and journals. Her areas of interest are Optical Character Recognition, Pattern Recognition, Machine Learning, Medical Imaging and Natural Language Processing.

Arif Iqbal Umar was born at district Haripur Pakistan. He obtained his M.Sc. (Computer Science) degree from University of Peshawar, Peshawar, Pakistan and Ph.D. (Computer Science) degree from BeiHang University (BUAA), Beijing PR China. His research interests include Data Mining, Machine Learning, Information Retrieval, Digital Image Processing, Computer Networks Security and Sensor Networks. He has at his credit 22 years' experience of teaching, research, planning and academic management. Currently he is working as Assistant Professor (Computer Science) at Hazara University Mansehra Pakistan.

Riaz Ahmad is a Ph.D. student in Technical University at Kaiserslautern, Germany. He is also a member of Multimedia Analysis and Data Mining (MADM) research group at German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His Ph.D. study is sponsored by Higher Education Commission of Pakistan under Faculty Development Program. Before this, he has served as a faculty member at Shaheed Benazir Bhutto University, Sheringal, Pakistan. His areas of research include document image analysis, image processing and Optical Character Recognition. More specifically, his work examines the invariant approaches against scale and rotation variation in Pashto cursive text.

Imran Siddiqi is received his Ph.D.in Computer Science from Paris Descartes University, Paris, France in 2009. Presently, he is working as an Associate Professor at the department of Computer Science at Bahria University, Islamabad, Pakistan. His research interests include image analysis and pattern classification with applications to handwriting recognition, document indexing and retrieval, writer identification and verification and, content based image and video retrieval.

Saad Bin Ahmed is serving as Lecturer at King Saud bin Abdulaziz University for Health Sciences, Saudi Arabia. He is completed his Master of computer sciences in intelligent systems from University of Technology, Kaiserslautern, Germany and has been served as research assistant at Image Understanding and Pattern Recognition (IUPR) research group at University of Technology, Kasierslautern, Germany. He had served as Lecturer at COMSATS institute of information technology, Abottabad, Pakistan and Iqra University, Islamabad, Pakistan. He has also performed his duties as project supervisor at Allama Iqbal Open University, Islamabad, (AIOU) Pakistan. His area of interests is document image analysis, medical image processing and optical character recognition. He is in field of image analysis since 10 years and has been involved in various pioneer research like handwritten Urdu character recognition.

Imran Razzak is working as Associate Professor, Health Informatics, College of Public Health and Health Informatics, King Saud bin Abdulaziz University for Health Sciences, National Guard Health Affair, Riyadh Saudi Arabia. Besides, is associate editor in chief of International Journal of Intelligent Information Processing (IJIIP) and member of editorial board of PLOS One, International Journal of Biometrics Indersciences, International Journal of Computer Vision and Image Processing and Computer Science Journal, as well as scientific committee of several conferences. He is a writer of one US/PCT patent and more than 80 research publications in well reputed journals and conferences. His research area/field of expertize includes health informatics, image processing and intelligent system.

Dr. Faisal Shafait is working as the Director of TUKL-NUST Research & Development Center and as an Associate Professor in the School of Electrical Engineering & Computer Science at the National University of Sciences and Technology, Pakistan. He has worked for a number of years as an Assistant Research Professor at The University of Western Australia, Australia, a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), Germany and a visiting researcher at Google, CA, USA. He received his Ph.D. in Computer Engineering with the highest distinction from TU Kaiserslautern, Germany in 2008. His research interests include machine learning and computer vision with a special emphasis on applications in document image analysis and recognition. He has co-authored over 100 publications in international peer reviewed conferences and journals in this area. He is an Editorial Board member of the International Journal on Document Analysis and Recognition (IJDAR), and a Program Committee member of leading document analysis conferences including ICDAR, DAS, and ICFHR. He is also serving on the Leadership Board of IAPRs Technical Committee on Computational Forensics (TC-6) and as the President of Pakistani Pattern Recognition Society (PPRS).