Algorithm for auto annotation of scanned documents based on subregion tiling and shallow networks

Komuravelli Prashanth

Department of Computer Science and Engineering
Indian Institute of Technology, Tirupati
Tirupati, India
Email: cs18s504@iittp.ac.in

Yeturu Kalidas*

Department of Computer Science and Engineering
Indian Institute of Technology, Tirupati
Tirupati, India
Email: ykalidas@iittp.ac.in

Abstract—There are millions of scanned documents worldwide in around 4 thousand languages. Searching for information in a scanned document requires a text layer to be available and indexed. Preparation of a text layer requires recognition of character and sub-region patterns and associating with a human interpretation. Developing an optical character recognition (OCR) system for each and every language is a very difficult task if not impossible. There is a strong need for systems that add on top of the existing OCR technologies by learning from them and *unifying* disparate multitude of many a system. In this regard, we propose an algorithm that leverages the fact that we are dealing with scanned documents of handwritten text regions from across diverse domains and language settings. We observe that the text regions have consistent bounding box sizes and any large font or tiny font scenarios can be handled in preprocessing or postprocessing phases. The image subregions are smaller in size in scanned text documents compared to subregions formed by common objects in general purpose images. We propose and validate the hypothesis that a much simpler convolution neural network (CNN) having very few layers and less number of filters can be used for detecting individual subregion classes. For detection of several hundreds of classes, multiple such simpler models can be pooled to operate simultaneously on a document. The advantage of going by pools of subregion specific models is the ability to deal with incremental addition of hundreds of newer classes over time, without disturbing the previous models in the *continual learning* scenario. Such an approach has distinctive advantage over using a single *monolithic model* where subregions classes share and interfere via a bulky common neural network. We report here an efficient algorithm for building a subregion specific lightweight CNN models. The training data for the CNN proposed, requires engineering synthetic data points that consider both pattern of interest and non-patterns as well. We propose and validate the hypothesis that an image canvas in which optimal amount of pattern and non-pattern can be formulated using a means squared error loss function to influence filter for training from the data. The CNN hence trained has the capability to identify the character-object in presence of several other objects on a generalized test image of a scanned document. In this setting some of the key observations are in a CNN, learning a filter depends not only on the abundance of patterns of interest but also on the presence of a non-pattern context. Our experiments have led to some of the key observations - (i) a pattern cannot be over-expressed in isolation, (ii) a pattern cannot be under-expressed as well, (iii) a non-pattern can be of salt and pepper type noise and finally (iv) it is sufficient to provide a non-pattern context to a modest representation of a pattern to result in strong individual sub-region class models. We have carried out studies and reported *mean average precision* scores on various data sets including (1) MNIST digits(95.77), (2) E-MNIST capital alphabet(81.26), (3)

EMNIST small alphabet(73.32) (4) Kannada digits(95.77), (5) Kannada letters(90.34), (6) Devanagari letters(100) (7) Telugu words(93.20) (8) Devanagari words(93.20) and also on medical prescriptions and observed high-performance metrics of mean average precision over 90%. The algorithm serves as a kernel in the automatic annotation of digital documents in diverse scenarios such as annotation of ancient manuscripts and handwritten health records.

I. INTRODUCTION

Building digital libraries involves computer vision based analysis of scanned documents. The characters or words or glyphs in a documents are peculiar to the language, culture, time period and several other domain specific factors. Building one single system for automatic annotation is challenging and is still an open area of research. There are several character level annotators known as optical character recognition (OCR) systems available worldwide today [1]. However each system is specific to its own domain of choice and interoperability across different document classes is limited.

There are around 50 popular OCR systems worldwide for different languages. Some of the famous examples include Tesseract, E-aksharayan, OCRopus, and OCRFeeder. For instance, *Tesseract* requires preparation of input in steps of - (i) preprocessing of the image, (ii) text localization, (iii) character segmentation, (iv) character recognition and (v) post processing. In case of other OCR systems as well, the steps are similar. Some of the reference works for word level segmentation include [2], [3] where settings are for Urdu and Hindi languages. Segmentation of subregions in hand written documents is challenging and the work by [4] addresses segmentation of text with cursive hand writing using Hidden Markov Models [5]. [6] proposed an OCR model for printed Urdu script using feed-forward neural network. [7] proposed a high-performance OCR model for printed English using long short term model networks [8]. There are works based on Markov random fields (MRF) [9] for Chinese character recognition and semi Markov fields for both Chinese and Japanese character recognition. In [10] proposes structural lattices for character and word patterns for Chinese texts. Segmentation is challenging problem in document scenarios when it comes to identifying touching characters and locating cutting points [11]. There are segmentation free character level annotation approaches

as well, such as [12] for Mongolian character recognition. Before research community started switching over to deep learning methods, [13] proposed a support vector machine based methodology for recognition of hand written digits. Most the methods are domain and language dependent due to heavy use of heuristics in preprocessing or postprocessing steps. There is a need for a system that simplifies subregion annotation mechanism and brings in domain agnosticism.

There are quite a few advances in image sub-region detection in the context of detection of common objects [14], [15], [16]. The algorithms are fundamentally a regression based formulation to predict the extent of bounding boxes and the object category. Moreover the methods employ much deeper convolutional neural networks and require enormous amount of training data. However general purpose image subregion detection approaches do not utilize the peculiarities of a scanned document scenario. The sizes of rectangles are smaller in a document image. The variability of pixel content for a given character class is lesser than the amount of variability in pixel content of a general purpose object category. There is a need for methods that leverage the specific aspects of document images and simplify the machine learning models.

Neural networks suffer from the problem of catastrophic forgetting [17]. When a new class arrives, the model needs to be retrained. However retraining may loose the performance of the model on prior data. This problem is mainly applicable to the case of multi class classification models. We need a system which is robust to adding of newer classes frequently and an incremental fashion. There are ensemble learning [18] based methods and other methods [19] to overcome these problems to certain extent. However in an ensemble, each constituent model is exposed to the same data distribution from the training data set. There is a need for a system where individual character subregion classes can be integrated in an effective way.

We report here a domain agnostic annotation algorithm that *unifies* many a disparate OCR system in a plug-n-play fashion through use of annotation as a service. We introduce a concept of *named model pools* where there are sets of subregion specific models allowing a user to add arbitrary number of newer classes over time. The idea of model pool is an *incremental learning* strategy where addition of a newer class does not disturb previous models thereby addressing the problem of catastrophic forgetting in a continual learning scenario. We introduce a concept of generating synthetic image data by tiling of subregions corresponding to pattern and anti-pattern categories. The tiled representation allows to build lightweight CNNs for recognition of individual character subregion classes. The algorithm builds very shallow networks sufficient enough to capture a single sub-region category with training time in only a couple of seconds. The algorithm has demonstrated high accuracies on datasets of documents having hand written text from Telugu, Kannada, Devanagari, English and clinician notes on medical prescriptions. The method has also been evaluated for mixed character content from different languages.

Fig. 1. Flow chart of the algorithm for sub-region annotation, model building and predictions.

II. METHODS

Manual or annotation service based, annotations are used from scanned documents for building a single class models. The models are then applied for auto annotation of new input bundles of scanned documents. The schematic of the algorithm is presented in (Figure 1). Key components of the proposed algorithm are - (i) general characteristics of subregion patterns in scanned documents and assumptions; (ii) subregion specific models; (iii) tiling based canvas representation for synthetic data generation; (iv) optimal amount of pattern and anti-pattern representation on canvas and (v) training methodology.

A. General characteristics of subregion patterns in scanned documents and assumptions

Visual appearances of a character subregion in several pages of a scanned document is expected to be consistent in terms of its pixel content and dimensions when scanned under controlled settings. There are cases where a character appears in larger or smaller font sizes. However such cases can be handled in preprocessing or postprocessing stages without affecting the general purpose algorithm presented here.

There are also cases to discuss and assume regarding orientation of scanned image. A CNN based model is sensitive to image orientation. Building a model that is rotation agnostic dilutes the major purpose of building a scalable document annotation system. It is expected to be handled in a preprocessing stage to detect right orientation of the input document or after application of subregion detectors based on too few elements detected.

The correctness of a recognized character in a given image subregion depends on the context of semantics of other characters around it. Such a semantic level analysis of character requires domain specific knowledge turned into heuristics and formulation of sequence based models such as recurrent neural networks [7]. We leave it to the postprocessing phase to deploy further text and semantics refinement models and focus mainly on character subregion detection from image.

In a nutshell, the algorithm proposed here focuses on identification and recognition of character subregions in scanned text images. The algorithm includes in its scope, high speed training of subregion classes, incremental addition of newer classes over time and plug-n-play of annotation engines as services. The algorithm provides for convenient features of named model pools that simultaneously operate on a given image. It is assumed here that computational power is that of a standard server grade machine with large memory and parallel processing capabilities as commonly available today.

Fig. 2. A two-layer shallow neural network showing schematic of the steps for input, convolutions, fusion and regression against given label.

B. Subregion specific models

Model building is carried out on a per class basis as against building a single monolithic multi class model. Multiple exemplars of sub-regions of a specific class are determined through the annotation phase. These sub-regions are variable in size in terms of their dimensions of width and height. The widths and heights may not be identical, however they fall within a band of similar values for a given character. With this intuition, we have devised rectangular convolutional filters whose dimensions approximate average width and height of a set of rectangles of a given sub-region type.

The input canvas having multiple instances of pattern of interest along with non-patterns is passed to a shallow CNN model. The first layer of convolutions have filters whose dimensions match with the average width and average height of the input sub-regions. The second layer corresponds to fusion of convolved images into a single image. The predicted output is then regressed over the actual label. The schematic of the process and neural network architecture is shown in (Figure 2).

C. Synthetic training data - tiling based representation of canvas

A training document is represented as a tiling of sub-regions of interest. The number of repetitions of a sub-region in a synthetic training document image directly influence the filter values in CNN. The more the repetitions the more the influence and the lesser the repetition of subregion patterns, the lesser is the influence.

For instance, given a blank document with all pixels 0, with much less repetition of pattern, results in CNN filters that predict blank regions on any new document. Too high repetition of pattern in the synthetic canvas, causes to learn filters that make any input document just white. In addition, only pattern is not sufficient in the synthetic document. In order to force filters to learn subregion patterns, anti-pattern are to be provided in the training document for the loss function to force change the filter contents to learn pixel patterns of required pattern.

The number of times anti-patterns appear in a given document also influence the quality of the model. The anti-pattern can be sub-regions from other class or it can be pepper kind of noise. The results and insights are discussed in (Section III-A). A brief schematic of the steps are shown in (Algorithm 1). An

Fig. 3. Sample train document image and its label.

Fig. 4. Example of pixel wise false positives and false negatives in prediction.

example of training image and its corresponding label created using tiling based formulation with pepper noise as anti-pattern is shown in (Figure 3).

Algorithm 1 Schematic for tiling based representation

- 1: Select a sub-region of interest for a given class
- 2: Repeat the sub-region on random locations in a given canvas
- 3: Randomly select patches from the data set as non-pattern
- 4: Implant non-pattern patches in the same document image in random locations
- 5: Adjust the number of times a pattern repeats
- 6: Adjust the number of times a non-pattern repeats

D. Optimal amount of pattern representation on canvas

There is a need of optimal representation of pattern on canvas created using tiling based representation. A pattern cannot be over-expressed or under-expressed as well. It is sufficient to provide a non-pattern context to modest representation of pattern to result in strong individual sub-region class model. An example of test image, true label and predicted label with pixel wise false positives and false negatives is shown in (Figure 4).

Optimal amount of pattern gives less number of false positives and false negatives. The pictures, (Figure 5) and (Figure 6) depict plots with percentage of pattern from *MNIST digit class 4* in each training canvas on X-axis and number of false positives and false negatives on Y-axis respectively. Our empirical studies indicate that having 40 to 60 percent of

Fig. 5. Amount of pattern vs number of false positives.

Fig. 6. Amount of pattern vs number of false negatives.

pattern on each training canvas will results in less number of false positives as well as false negatives.

E. Training methodology

The loss function is formulated as a multi-layer convolutional pattern mapping problem from input tensor to output tensor. Input tensor corresponds to canvas tiled with optimal amount of pattern where as output tensor corresponds to another canvas of same size determines presence of pattern at middle pixel level. A two layer shallow CNN as shown in (Figure 7) is trained. We have used Adam optimizer with learning rate of 0.001 for our experiments.

Fig. 7. Two layer shallow network architecture.

Fig. 8. Randomly curated document with Devanagari letters

Fig. 9. Randomly curated document with Kannada letters

F. Benchmark data sets

The Devanagari script consists of 47 primary characters. The data set¹ contains only 36 character classes and 40 annotations for each class. We have evaluated on all the 36 character classes containing a total sum of 1431 images. Figure 8 is the example of randomly curated document with annotations of Devanagari letters.

Kannada language has 53 primary characters. The data set² is a collection of 25 image annotations for each of the 53 character classes. We have evaluated on 1315 images belonging to all the 53 character classes. Figure 9 is the example of randomly curated document with annotations of Kannada letters.

We have evaluated our algorithm over the data set([20]) contains around 10000 annotations for each of the 26 English capital alphabets. Figure 10 is the example of randomly curated document with annotations of English capital letters.

Kannada handwritten digits data set([21]) consists 6000 annotations for each of the 10 Kannada digit classes. We have carried out our experiments on all the 60000 present in this data set. Figure 11 is the example of randomly curated document with annotations of Kannada digits.

¹http://www.iapr-tc11.org/mediawiki/index.php?title=Devanagari_Character_Dataset

²<http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/>

Fig. 10. Randomly curated document with English capital letters

Fig. 11. Randomly curated document with Kannada digits

MNIST handwritten digits data set³ consists 6000 annotations each for 10 digit classes. We have done our experiments on all the 60000 annotations. Figure 12 is the example of randomly curated document with annotations of MNIST digits.

The Devanagari words dataset [22] consists of 90000

³<http://yann.lecun.com/exdb/mnist/>

Fig. 12. Randomly curated document with MNIST digits

Fig. 13. Schematic representation of platform.

annotations for different word classes, we have built models and carried out validations on 10 annotation classes.

The Telugu words dataset [23] contains 120000 annotations for several words, we have built models and carried out validations on 10 annotation classes.

G. Annotation platform

We have also developed a platform for for large-scale, domain agnostic and crowd enabled document content annotation which supports 2D annotations by humans or other OCR like softwares on document images and provides automatic annotation support by building AI models from human annotations. To build the models (section II-E) we tiling based method (section II-C) to generate synthetic documents by placing annotations by users or other softwares optimally (section II-D). The schematic representation of platform is as shown in Figure 13

a) *System overview*: The platform provides an interface for users to make 2D bounding box annotations, and also an auto-annotation engine which runs model building algorithm. For giving desirable annotation experience to the users we integrated a web based tool LabelMe[24] to our platform.

b) *Model Pooling*: Users can group 2 or more sub-region models to create a model pool. A user can create n number of such *model pools*. The platform employs multiple models or multiple model pools on a given document to perform automatic annotation.

III. RESULTS

We report here results and evaluations to substantiate the conclusions drawn from the work. The algorithm has been evaluated on aspects for (i) optimal expression of pattern and anti-pattern subregions; (ii) recognition of subregions in complex settings (iii) performance of model pool over monolithic approach and (iv) learning of Tesseract word level annotations.

A. Optimal expression of pattern and anti-pattern subregions

In order to train a shallow network model for a single class, tiling based representation is used (Algorithm 1). However we have the some of critical observations on the amount of representation of pattern and non-pattern context in connection

xi	yi	Pattern	Non-pattern	Prediction
W,B	W,B	Not there or Single or Multiple	Not there or Single or Multiple	All White
B,W	B,W	Single	Not there	All Black
B,W	B,W	Multiple	Not there	All White
B,W	B,W	Optimal*	Optimal*	Optimal

TABLE I

OBSERVATIONS ON PREDICTIONS BASED ON BACKGROUND, CHARACTER COLOURS, AMOUNT OF PATTERN AND NON PATTERN. (W,B) OR (B,W) INDICATE BACKGROUND AND CHARACTER COLOURS RESPECTIVELY

with the ability to learn patterns, summarized in (Table I). The xi and yi variables correspond to input image and output prediction respectively. For the yi, central pixel of the bounding box is determined. However, the bounding box size is same as the size of the kernels used in the first convolution layer (Figure 2).

Critical observations on tiling based formulation of pattern and non-pattern patches for canvas representation. The columns xi and yi correspond to input and predicted image representation. The letters W and B correspond to White and Black colours. For each of the xi and yi, the tuples, (W,B) or (B,W) indicate background and character colours respectively. Pattern and Non-pattern repetitions are indicated whether zero times i.e. not there or single time or multiple times. The optimal values are determined after

B. Recognition of sub-regions in complex settings

Sub-regions in scanned documents of diverse language and domain categories are evaluated. In order to test recognition capability, a mixed document is given and tested for domain agnosticism. The second test here is change of position of character subregions and recognize.

1) *Domain agnostic recognition:* (Figure 14) is the example of predictions on randomly curated document using character level annotations of Devanagari characters. Similarly the same algorithm will work on document containing regions from multiple languages, to support that a single document is created with sample characters from different language - Telugu, Kannada, Devanagari and English. The algorithm is able to recognize these characters simultaneously as show in (Figure 15).

Sub-regions may correspond to character or word level or any region of interest and a test document may have mixture of sub-regions from diverse categories. Figure 16 is a single document showing the recognition of words from Telugu, English, Devanagari languages.

2) *Segmentation free recognition:* The algorithm does not require sub-region level segmentation as prerequisite step. The sub-regions may be spaced anywhere in the document in a random manner as shown in the (Figure 17).

C. Performance of model pool over monolithic approach

In our approach a pool of models is used for sub-region annotation rather than a single monolithic model (Section II-G) (we assume model means deep neural network based model in all of our discussions). A monolithic model requires complete

Fig. 14. Sample Devanagari document and its predictions.

Fig. 15. Recognized characters are shown in rectangular boxes. The characters or word level sub-regions correspond to Telugu, Kannada, Devanagari and English all in one document.

retraining as new character classes are added over time. However prior performance is not guaranteed on older classes of data points due to change of weights of the neural network. This phenomenon is called catastrophic forgetting. To evaluate this behaviour and present advantage of a pool of multiple single class models, we have demonstrated the proof of concept on MNIST handwritten digits. The digits arrive over time in the order of 0 to 9 classes. Each time a new digit class arrives, the entire model is retrained and accuracies are computed on the past digit classes. We can clearly see in (Table II) the accuracies fluctuate over time. In case of our proposed model

Fig. 16. Recognition of words from Telugu, English, Devanagari languages in single document.

Fig. 17. The recognized sub-regions are scattered across the document in a random manner without requiring segmentation.

Fig. 18. Training data point for a drug to get detected in medical prescriptions.

pool based approach, the accuracies on prior classes remain high as those models are not touched (Table III).

D. Performance of algorithm on diverse data sets

The algorithm has been tested on datasets (Section II-F) of Telugu, Devanagari, Kannada, EMNIST, handwritten medical prescriptions and diverse font categories and reported accuracies in Table IV. The accuracies can further increased if we increase the number of layers and kernels in neural network. We performed the experiments on some datasets with bit deeper networks(4 layers) and reported accuracies in Table V

Figure 19 is the example prescription on which detection of a drug happend with the help of very sparse network trained for for that specific drug using training images like Figure 18

E. Tesseract OCR annotations

We built models by taking annotations made by tesseract as training data to verify the level of mimicking the behaviour of Tesseract. We observed an *mAP* score of 90 for the models trained on Tesseract annotations of 10 word level classes. Figure 20 is document having predictions from the models trained on tesseract annotations.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

We report here an algorithm to recognize sub-regions in images of scanned documents in a domain agnostic and segmentation free setting. We also report critical observations on the importance of pattern and non-pattern context resulting

DD FORM 1289 1 NOV 71 DOD PRESCRIPTION	
FOR (Full name, address, & phone number) (If under 12, give age) John R. Doe, HMA, USA	
U.S.S. Never forgotten (DD 178)	
MEDICAL FACILITY U.S.S. Never forgotten (DD 178)	DATE 23 Jan 99
R (Superscription) gm or ml. (Inscription) In Belladonna 15 ml Amphogel good 120 ml (Subscription) M & JT Solution (Signa) Sig: 5ml b.i.d. a.c.	
MFGR: Worth LOT NO: P39K106	EXP DATE: 12/02 FILLED BY: RMT
R NUMBER 10072	Signature: Jacob R. Frost LCPLR M.D. USNR SIGNATURE RANK AND DEGREE
EDITION OF 1 JAN 60 MAY BE USED FOR S/N 0102 LF 012 6201	

Fig. 19. Recognition of drug names on medical prescription.

Fig. 20. Mimicking the behaviour of Tesseract.

in a novel tiling based formulation for canvas representation. The algorithm is scalable due to employing a pool of multiple sub-region specific models. Each model itself is a very shallow network having just two layers apart from input layer. Training time for each model is in few seconds and all the models in minutes compared to several hours of any monolithic model. Our algorithm is incremental in nature to allow for plug and play of newer classes of data over time and avoids the phenomenon of catastrophic forgetting. Sub-regions may correspond to character or word level or any region of interest and a test document may have mixture of sub-regions from diverse categories. The algorithm has been tested on data sets of Telugu, Devanagari, Kannada, EMNIST, handwritten medical prescriptions and diverse font categories where it reported on average high mean average precision of more than 90%. The algorithm can be extended for general purpose images however after addressing the problem of diversity of subregion content for a given category. The method has potential for deployment on edge devices and distributed computing environment to provision model pools. It is also possible to extend the work to build person specific models and effectively use in conjunction with plug-n-play annotation services.

n/c	0	1	2	3	4	5	6	7	8	9
6	85.9	90.8	81.7	83.3	81.8	78.4	-	-	-	-
7	86.5	89.7	73.6	84.6	79.4	73.1	85.2	-	-	-
8	84.9	80.6	77.6	79.2	80.3	73.3	86.6	83.7	-	-
9	83.5	82.8	81.4	83.3	79.7	76.7	84.5	97.9	91.2	-
10	82.5	83.6	82.9	87.9	79.8	76.4	81.1	96.9	91.1	92.5

TABLE II

ACCURACIES FLUCTUATE AS NEW CLASSES ARRIVE IN A MONOLITHIC MODEL SETTING. n IS NUMBER OF CLASSES c IS CLASS LABEL.

n/c	0	1	2	3	4	5	6	7	8	9
6	90.3	91.4	67.6	87.1	80.4	73.9	-	-	-	-
7	90.3	91.4	67.6	87.1	80.4	73.9	89.1	-	-	-
8	90.3	91.4	67.6	87.1	80.4	73.9	89.1	85.2	-	-
9	90.3	91.4	67.6	87.1	80.4	73.9	89.1	85.2	73.0	-
10	90.3	91.4	67.6	87.1	80.4	73.9	89.1	85.2	73.0	85.2

TABLE III

ACCURACIES REMAIN HIGH WITHOUT ANY FLUCTUATIONS AS NEW CLASSES ARRIVE IN A MODEL POOL SETTING. n IS NUMBER OF CLASSES c IS CLASS LABEL.

Dataset	classes	Max	Avg	Min
EMNIST CAPITAL LETTERS	26	92.34	81.26	70.33
EMNIST SMALL LETTERS	26	88.18	73.32	42.29
MNIST	10	97.94	95.99	94.40
DEVANAGARI LETTERS	36	100	100	100
KANNADA DIGITS	10	98.37	95.77	92.02
KANNADA LETTERS	53	97	90.34	82
DEVANAGARI WORDS	10	95	93.20	90
TELUGU WORDS	10	100	96.18	91

TABLE IV

FOR EACH OF THE DATA SETS, THE NUMBER OF SUB-REGION CLASSES AND AGGREGATE METRICS OF MAXIMUM, AVERAGE AND MINIMUM SCORES OF MEAN AVERAGE PRECISIONS ARE INDICATED.

Dataset	classes	Max	Avg	Min
EMNIST CAPITAL LETTERS	26	99	93.09	83
EMNIST SMALL LETTERS	26	98	87.85	50
KANNADA DIGITS	10	100	99.90	99.56
KANNADA LETTERS	53	100	99.83	95.27

TABLE V

FOR EACH OF THE DATA SETS, THE NUMBER OF SUB-REGION CLASSES AND AGGREGATE METRICS OF MAXIMUM, AVERAGE AND MINIMUM SCORES OF MEAN AVERAGE PRECISIONS ARE INDICATED.

V. DISCUSSIONS

The platform can be accessed at <https://services.iittp.ac.in/annotator>.

Training data sets, test images and predictions, per class average precision for all the data sets can be found at <https://tinyurl.com/xk6hasmn>.

REFERENCES

- [1] M. Jamshed, S. Maira, R. A. Khan, and U. Mueen, "Handwritten optical character recognition : A comprehensive systematic literature review," *IEEE Access*, vol. 8, pp. 142 642–142 668, 2020.
- [2] B. Z. Haris, A. R. Agha, and A. Awais, "Urdu word segmentation using conditional random fields (CRFs)," *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2562–2569, 2018.
- [3] S. Bikash, K. P. Swapan, and M. Shridhar, "Offline handwritten devanagari word recognition: A segmentation based approach," *Proceedings of the 19th International Conference on Pattern Recognition*, pp. 1–4, 2008.
- [4] N. Arica and F. T. Yarman-Vural, "Optical character recognition for cursive handwriting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 801–813, 2002.
- [5] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE Signal Processing Magazine*, vol. 3, no. 1, pp. 4–16, 1986.
- [6] I. Shamsher, A. Zaheer, J. K. Orakzai, and A. Awais, "OCR for printed urdu script using feed forward neural network," *Proceedings of the World Academy of Science, Engineering and Technology*, vol. 23, 2007.
- [7] M. B. Thomas, U. H. Adnan, A. A. Mayce, and S. Faisal, "High-Performance OCR for Printed English and Fraktur using lstm networks," *Proceedings of the 12th International Conference on Document Analysis and Recognition*, vol. 1, pp. 683–687, 2013.
- [8] H. Sepp and S. Jürgen, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, 1997.
- [9] Z. Jia and L. Zhi-Qiang, "Markov random field-based statistical character structure modeling for handwritten chinese character recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 767–780, 2008.
- [10] W. Qiu-Feng, Y. Fei, and L. Cheng-Lin, "Handwritten chinese text recognition by integrating multiple contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1469–1481, 2012.
- [11] M. Simone, G. Marco, and S. Giovanni, "Artificial neural networks for document analysis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 23–35, 2005.
- [12] Z. Hui, W. Hongxi, B. Feilong, and G. Guanglai, "Segmentation-free printed traditional mongolian OCR using sequence to sequence with attention model," *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, vol. 01, pp. 585–590, 2017.
- [13] A. Boukharouba and A. Bennia, "Novel feature extraction technique for the recognition of handwritten digits," *Applied Computing and Informatics*, vol. 13, no. 1, pp. 19 – 26, 2017.
- [14] R. Joseph, D. Santhosh, G. Ross, and F. Ali, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.
- [15] W. Keze, L. Liang, Y. Xiaopeng, C. Ziliang, Z. Dongyu, and Z. Lei, "Cost-effective object detection: Active sample mining with switchable selection criteria," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 834–850, 2019.

- [16] Z. Zhong-Qiu, Z. Peng, X. Shou-Tao, and W. Xindong, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
- [17] P. Inyoung, O. Sangjun, K. Taeyeong, and K. Injung, "Overcoming catastrophic forgetting by neuron-level plasticity control," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5339–5346, 2020.
- [18] C. Robert, M. Aaron, and A. Itamar, "Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1623–1634, 2013.
- [19] P. Gyeong-Moon, Y. Sahng-Min, and K. Jong-Hwan, "Convolutional neural network with developmental memory for continual learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2020.
- [20] C. Gregory, A. Saeed, T. Jonathan, and V. S. Andre, "EMNIST: Extending MNIST to handwritten letters," *Proceedings of the International Joint Conference on Neural Networks*, pp. 2921–2926, 2017.
- [21] P. V. Uday, "Kannada-MNIST: A new handwritten digits dataset for the Kannada language," *arXiv e-prints*, p. abs/1908.01242, 2019.
- [22] D. Kartik, K. Praveen, M. Minesh, and C. V. Jawahar, "Offline handwriting recognition on devanagari using a new benchmark dataset," *Proceedings of the International Workshop on Document Analysis Systems*, pp. 25–30, 2018.
- [23] —, "Towards spotting and recognition of handwritten words in indic scripts," *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pp. 32–37, 2018.
- [24] C.R.Bryan, T.Antonio, P.M.Kevin, and T.F.William, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.