Received August 21, 2021, accepted September 1, 2021, date of publication September 7, 2021,
date of current version September 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110858

Movie Title Extraction and Script Separation Using Shallow Convolution Neural Network

MRIDUL GHOSH[1,2], SAYAN SAHA ROY[3], HIMADRI MUKHERJEE[4], SK MD OBAIDULLAH[2], (Member, IEEE), XIAO-ZHI GAO[5], AND KAUSHIK ROY[6]

[1] Department of Computer Science, Shyampur Siddheswari Mahavidyalaya, Howrah 711312, India
[2] Department of Computer Science and Engineering, Aliah University, Kolkata 700160, India
[3] Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur 721302, India
[4] Department of Computer Science, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
[5] School of Computing, University of Eastern Finland, 70211 Kuopio, Finland
[6] Department of Computer Science, West Bengal State University, Kolkata 700126, India

Corresponding author: Kaushik Roy (kaushik.mrg@gmail.com)

ABSTRACT Graphical texts in natural images play an important role in portraying information in multitudinous fields such as communication, education, and entertainment to name a few. Recognizing text in scene images is challenging due to the inherent complexity of the images. Text recognition in natural images involves script identification, which requires text localization. This is not trivial for natural scene images due to the presence of disparate foreground/background components. For scene images like movie posters, the challenge is more dominant. The challenges aggravate due to the presence of composite characteristics of posters like complex graphics background and the presence of different texts like a movie title, names of actors, producers, directors, and tagline. These texts have miscellaneous fonts, variations in colors, size, orientation, and textures. In this work, an M-EAST (modified EAST) model is proposed, which is based on the EAST (efficient and accurate scene text detector) model for text localization. A novel movie title extraction is thereafter used for separating the title from the extracted text pool. Finally, the title script was identified using a shallow convolutional neural network (SCNN)-based architecture to ensure functionality in low-resource environments. Experiments were performed on a dataset of movie-poster images of Tollywood, Bollywood, and Hollywood industries, and a highest accuracy of 99.82% was obtained. The system performed better than the reported techniques.

INDEX TERMS Movie title extraction, multi-script, text localization, transfer learning.

I. INTRODUCTION

In human-computer interaction, the success of smartphones and broad demands for content-based image search/understanding has highly amplified the role of text recognition. Stable and efficient systems are required to deal with the text in natural scenes. This is because the text sometimes explicitly expresses the meaning in natural scenes. This attribute may act as a valuable source of knowledge for text in natural photographs and videos. Text in the image carries important information for many practical applications like text-based image search, road direction map, landmark detection, vehicle license-plate number recognition, etc.

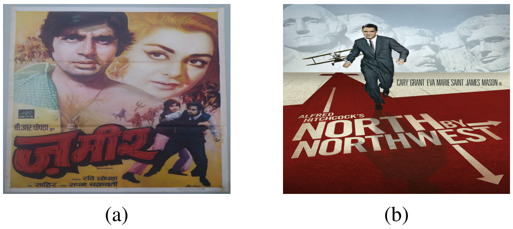
Automatic text localization and extraction [7], [8] is an important and demanding study topic in the field of image and pattern analysis. Extraction of texts from natural scene images is itself challenging, which aggravates in cases of low resolution, noise, blur, complex background images, etc. Text localization in scene images follows script identification and recognition of text. Automatic script recognition is essential to satisfy the increasing requirement for the electronic transmission of loads of documents composed of various scripts. This transmission is essential for day-to-day communications and thereafter understanding is essential. However, it is challenging to automatically understand such documents in a country like India that has several authorized state languages and scripts. A script can be used in one language or by more than one. For slight variations in graphical characters, they pose the same script class. For example, Hindi, Sanskrit,

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

Marathi, etc. can share the same script class of Devanagari in numerals.

The natural scene images are not graphically as rich as movie posters, which makes the latter a more challenging [1], [2] for text extraction and script identification. In movie posters, complex graphical texts and symbols, background arts, colors, textures, etc. are usually involved. Therefore, merely extracting texts from images is not enough. It is important to get the movie titles from the extracted text because the titles represent the genre, category, [16], [17] etc. This extraction is followed by script identification, which is a challenging task due to the presence of graphic-rich components. There are numerous types of graphical objects, such as images of actors, director(s), and catchy scenes in addition to several artistic texts related to the catchy tagline, movie title, production houses' names, actors' names, and directors' names, etc. The challenge is also because the words of the movie title are written in disparate fonts, orientations, and segments/syllables. Thus, the identification of the script of the movie title involves:

1) identification and extractions of the text regions;
2) extraction of movie titles amid different texts;
3) script identification of the extracted titles.



(a)

(b)

FIGURE 1. Samples of posters from: (a) Bollywood and (b) Hollywood movie.

Understanding a language is believed to be much simpler than reading it because the latter requires deep knowledge of the script [2], [3]. Quite often, audiences with only speaking proficiency in a particular language face problems in reading and understanding the title in a movie poster written in that language. This is especially true in a multi-script/lingual country like India, where movies are usually released in different languages in the same multiplex. Thus, the inability of reading a particular script might mislead the audience, who understand the corresponding language but are unable to recognize the title from the script. Moreover, for the visually impaired [9] or color-blind [10] persons, this application is a useful assisting system. That is, a script identification system can serve as a precursor, and help to decide the optical character recognition (OCR) [12] engine and recognize the title of the movie. Unfortunately, the existence of graphic-rich complex scenarios makes the task more challenging and has motivated us to work on such a topic.

For example, see Fig. 1 where two sample images, one from Bollywood (a) and another from Hollywood (b) movie posters are shown. These posters' visuals contain varied text syllables in addition to background visuals and foreground-background color similarities. In Fig. 1 (a), the movie title may also be oriented, written artistically, and the letters are not of the same size. On the other hand, the movie title in Fig. 1 (b) is slanted, multi-oriented, and parallel to the image's surface. The extraction of these movie titles from the two images is demanding.

In Fig. 2 the localized movie titles of two posters (Fig. 1 (a), (b)) are presented using the ViTSTR [4], EAST [33], and the proposed title localization methodology. In Figs. 2 (a), (b) the localized title box is parallel to the horizontal, but the title is tilt/oriented. For the Bollywood poster in Figs. 2 (a), and (b), more background graffiti was covered than (c). For both the ViTSTR and EAST methods, the rectangular bounding boxes of the Hollywood poster are not proper for slanted and oriented titles. These boxes do not cover the entire title, which eventually would produce wrong results from the OCR engine. The new method (in Figs. 2 (c)), could effectively detect the bounding boxes in both cases, even in the presence of multi-orientation in the title. The pipeline of the proposed work is depicted in Fig. 3.

In this work, a text localization method that is based on a pre-trained model and a process to extract the movie titles from localized text boxes was discussed. A shallow convolutional neural network framework was also developed to identify the scripts of movie titles, making it suitable for deployment in resource-constrained environments. The contributions of the present work are as follows:

- A modified EAST model (M-EAST) was developed to overcome the weakness of the EAST model for rotated and sheared scene images and at the same time ensure lower computational overhead which is evident from the increase in the FPS value.
- For detecting the predicted text boxes' coordinates for rotated and sheared images, novel rotation and shearing algorithms were proposed. Both positive and negative shearing were considered in the shearing process.
- The proposed title extraction procedure can extract the movie titles from the poster images in presence of multiple text blocks.
- A shallow convolutional neural network (SCNN) architecture was proposed for the separation of the scripts.

The paper is structured as follows: in section II and III the literature study and the proposed work is explained respectively; in section IV the experimental procedure is discussed and in section V, the conclusion of the article is conferred.

II. LITERATURE STUDY

Deep learning-based text identification systems have emerged as a significant advancement in image processing technologies in a few years. Girshick [5] introduced a fast region-based convolutional network (fast R-CNN) technique, which is nine times faster than the VGG-16 framework. The author extracted features from a series of the convolutional and pooling layers along with from the region of interest-pooling layer (ROIpool). The ROIpool features help to identify the objects in the image. Ren *et al.* [6]

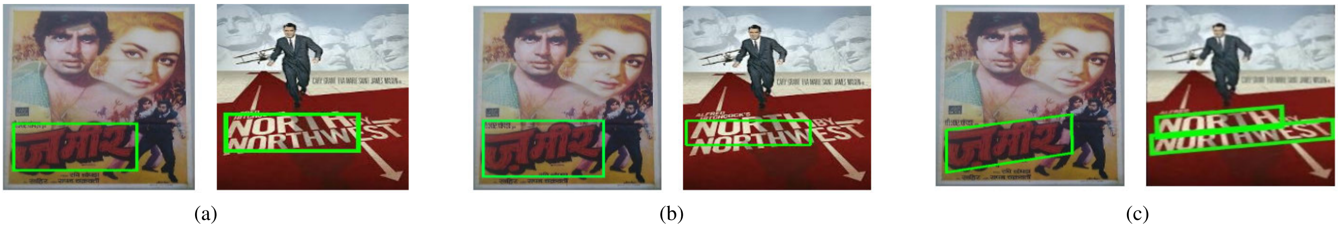


FIGURE 2. The localized movie titles marked in green using (a) ViTSTR method [4], (b) EAST method [33], and (c) the proposed method.

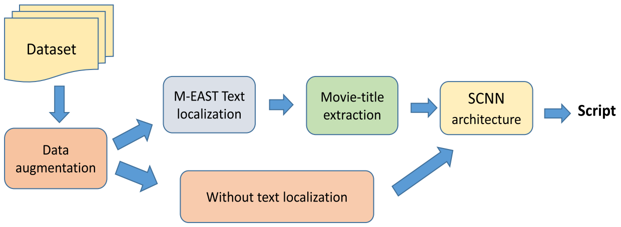


FIGURE 3. Detailed block diagram of the functioning of the proposed method.

proposed Region Proposal Network (RPN) for object detection. To enhance the accuracy, they joined the RPN and faster R-CNN network using the attention module to converge the bounding box into the text area. Zhong *et al.* [7] specified the weaknesses of Faster R-CNN. The accuracy of text box prediction drops for the text where long tail-like-shaped characters and for the cases where the text box itself is enclosed by texts. Also, for long text sequences, this method fails to detect all the text areas correctly. They used a two-way strategy to turn the multi-oriented text detection issue into a horizontal text detection issue, allowing the prediction of multi-oriented text cases while maintaining bounding box localization. They proposed a LocNet (localization network) [8] based framework, which was used to jointly work with faster R-CNN for text localization. They addressed the issue of tiny text and regular text backdrop and employed online-hard-example-mining and skip-pooling methods for the same. Naiemi *et al.* [11] proposed an accurate multi-oriented scene text localization (MOSTL) framework to gain accuracy in text localization in natural images. They tuned the inception network and ReLU (rectified linear activation unit) layer in their network.

Munjal *et al.* [12] proposed text localization with clustering of script (TeLCoS) framework, which is a fully trainable end-to-end system for concurrent text detection and script clustering. Their system minimizes the cost of distinct script identification components. They also introduced a pruning-based compression technique to deploy the system in resource-restricted gadgets. A multi-oriented scene text detector (MOST) framework was proposed by He *et al.* [13] where there are two modules. One is used for feature extraction and dynamically adjusted prediction text boxes, while the other one is concerned with the suppression of closely placed or inaccurate bounding boxes. Wang *et al.* [14] discussed a framework named R-YOLO (rotational you only look once) to recognize words in scene images that are positioned randomly. In their work, the text-bounding box is a rotatable reference rectangle that contains direction details that are used in multiple rotations.

The proper text extraction can contribute to script identification with high accuracy. Script identification methods can be categorized in two ways: based on handcrafted features and deep learning-based architectures. The handcrafted-based features require machine learning classifiers to predict output classes while deep learning-based techniques do not require other classifiers for class separation. Ghosh *et al.* [19] proposed an iterative dilation model to join the broken character's components and extracted features considering the shape and texture-based methods to identify the multi-character scripts. In [20] they used an extreme learning-based classifier to identify the same and obtained 97.95% accuracy. Shi *et al.* [21] designed a model that considers both deep and mid-level features and feeds them together in a deep network to efficiently identify the scripts, which have a close resemblance to each other. Mei *et al.* [22] designed a network that consisted of CNN and RNN for script identification from natural images. Gomez *et al.* [23] proposed a novel technique of script identification by combining the convolutional features and the characteristics of the Naive Bayes classifier. Bhunia *et al.* [24] proposed a method to identify scene images and video scripts considering local and global features using the CNN-LSTM network. They also used an attention-based patch weight structure in the CNN layer for local feature extraction and global features are extracted from LSTM.

Lei *et al.* [25] proposed a shallow deep learning model to classify images. In this network, they used seven layers: CPCPDD (C, P, D denotes convolution, pooling, and dense layers respectively). They considered MNIST and Fashion-MNIST databases for their experiment. An attention-based deep neural network was proposed by Ma *et al.* [26] for script identification of scene text images where Res2Net was deployed to extract multiscale features. A flexible channel-wise attention framework was also used to increase the feature map's spatial sensitivity. Also used global max-pooling to increase the separation performance. Khalil *et al.* [27] discussed a method for script identification in scene text images by augmenting the EAST model using a fully connected network module. To retrieve information such as genre, category, rating, etc., movies necessitate automatic data analysis. Khan *et al.* [15] discussed a deep learning-based approach to separate the movie tags from frames. Simões *et al.* [16] proposed a CNN-based architecture to classify movie genres. A hybrid model was proposed by Battu *et al.* [17] to identify the genre, and rating based on the summary of the movie. The occurrence of beating scenes in blockbusters action movies was identified using CNN-based framework [18].

III. PROPOSED METHOD

The proposed method is divided into two phases. In the first phase, a pre-processing technique was discussed, which is followed by the proposed M-EAST model and title box extraction technique. In the second phase, a convolutional neural network-based architecture was considered for script identification.

A. PRE-PROCESSING

To test the robustness of the system, the augmented dataset was prepared. For this process, Affine transformation [28] was considered for rotation, blur, and shear. The motivation for using this transformation is due to its co-linear properties and the fact that it remains in the Affine space after the transformation. Also, a wavelet-based single level decomposition [29] as a low-pass filter and Gaussian noise [30] for noise contamination in the images were considered for augmentation.

1) ROTATION

The image can be rotated in different degrees. Based on Affine transformation, the effect of rotation at an angle θ was determined in the augmentation.

2) SHEAR

The shape of the image is skewed by applying shear. Shearing can be done both in horizontal and vertical directions. Here, only horizontal shearing was considered. In this work, 0.45 shearing value was considered.

3) NOISE

Gaussian noise [30] was considered to contaminate the poster images to represent noisy scenarios at the time of image capture or image acquisition. Poor lighting, elevated heat, unpredictable variations, etc., are all factors that contribute to this acquisition process of signals.

4) BLUR

Pixel value varies quickly at the edges. To allow the low frequency to join while stopping high frequency, a blurring filter which is known as a low pass filter is required. In this process, for a filter dimension of p × p, the pixel having similar values is divided with p^2.

5) WAVELET DECOMPOSITION

Here, a single level decomposition of the discrete Haar wavelet [29] was considered for decomposing the images and retained with approximation level. The sample images from the developed dataset are presented in Fig. 4.

B. M-EAST

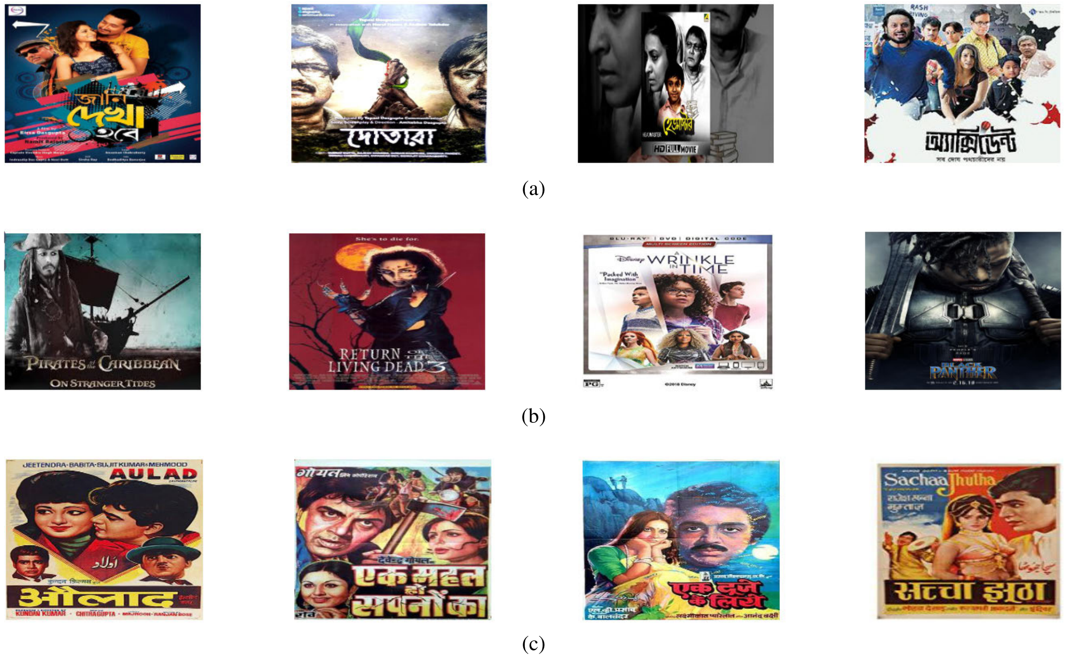
Deep learning [31] is a component of a wider community of machine learning techniques founded on artificial neural networks with representation learning. To train deep learning-based models necessitate a large amount of training data and computation power. In this work, the transfer-learning [32] technique was used for text localization. A pre-trained model is being used in the transfer-learning process that does a similar task as required in the new problem and removes training in a large amount of data from scratch. In this study, an M-EAST model was developed, which is based on the EAST [33] (considered in the transfer-learning process) text detector model. The M-EAST model comprises the proposed lightweight EAST module, coordinate extraction block, and rotation and shearing module. The M-EAST model is shown in Fig. 5.

In EAST, there are three stages: feature extraction, feature merging, and output stage. In the first stage, there are four convolutional and corresponding pooling layers. This network was trained on the ImageNet dataset. The feature maps f1, f2, f3, f4 (say) of sizes 1/32, 1/16, 1/8, and 1/4 respectively of the input image, are produced from this stage. In the second stage, the unpooling layer is used to double the feature map size. The unpooling layer acts as an upsampler that remembers and uses the positions of maximum elements in the original feature map to retain the size. Two other convolutional layers are used for feature extraction having filter dimensions 1 × 1 and 3 × 3, respectively. The output from the unpooling layer and the convolutional layers are concatenated to generate the feature map.

1) LIGHTWEIGHT EAST

Let the height and width of the original images be H and W, respectively. After normalization, the height and width become nH and nW (i.e., image dimension becomes nH × nW) respectively. The ratio of H, nH and the ratio of W, nW is rH and rW respectively. The values rH and rW are required in the title extraction phase to retain the coordinates of the text boxes according to the original dimension of the image. In this study, poster images were normalized to 320 × 320 based on trial runs.

Here, the EAST module was modified by removing the stage 4 convolutional layer from the feature extraction stage followed by an entire convolutional-concatenation-unpooling block and a 1 × 1 convolutional layer from the second convolutional-concatenation-unpooling block in the feature-merging stage. The quadrangle coordinate block from the output stage was also removed. Instead of a quadrangle block, a coordinate extraction block was created here. It was observed that after the modification of this module, the FPS (frames per second i.e., number of frames processed per second to get the localized texts) value improved by 4.17, which is discussed in the results and analysis section.



(a)

(b)

(c)

FIGURE 4. A snapshot of the dataset. The posters were collected from (a), Tollywood; (b) Hollywood; and (c) Bollywood movies.

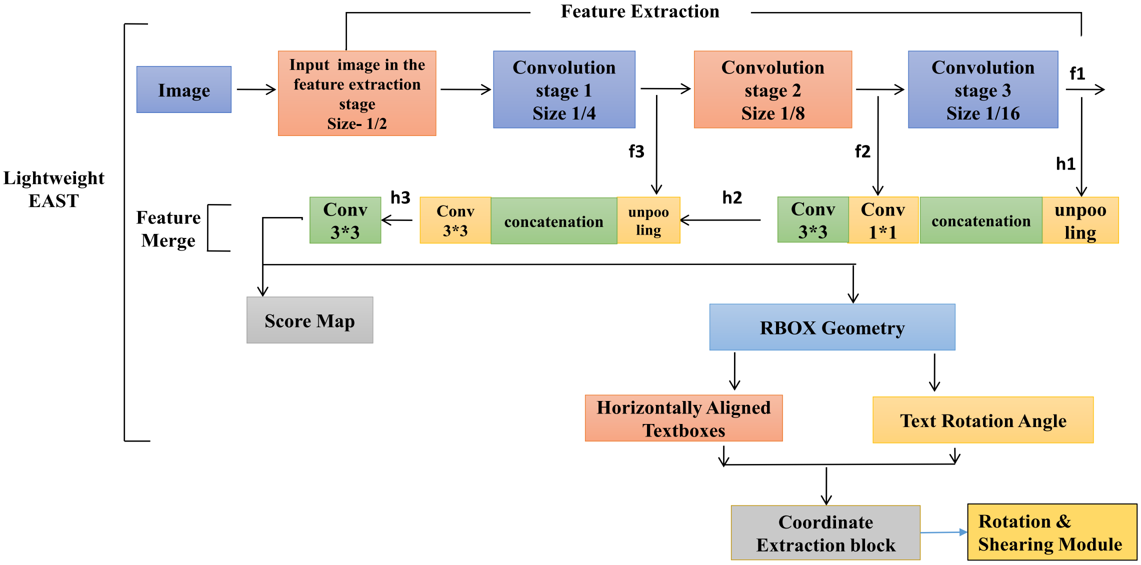

Feature Extraction
Lightweight EAST
Image
Input image in the feature extraction stage Size- 1/2
Convolution stage 1 Size 1/4
Convolution stage 2 Size 1/8
Convolution stage 3 Size 1/16
f1
f2
f3
h1
h2
h3
Feature Merge
Conv 3*3
Conv 3*3
concatenation
unpooling
Conv 3*3
Conv 1*1
concatenation
unpooling
Score Map
RBOX Geometry
Horizontally Aligned Textboxes
Text Rotation Angle
Coordinate Extraction block
Rotation & Shearing Module

FIGURE 5. A schematic diagram of the proposed M-EAST technique.

From this pre-trained model, the blobs in the images were detected as the areas of the image in which the colors, contrasts, and brightness significantly differ from the surrounding area and it consists of a similar type of pixels in the boundary. The probability/score of being a blob enclosing the text/non-text region is obtained from the feature extraction and feature-mapping property of the CNN architecture. Depending on the score of the blobs i.e., the higher the score of the blobs, the higher is the probability of the blobs containing text region, the text and non-text regions are segregated. In this study, the score value of the blobs was considered as or above 0.8 as the text-blob. After blob/bounding box selection, the coordinates of the blobs were determined. The process of blob detection is based on parameters such as channel-scaling factor and average pixel intensity. In the case of the scaling factor for the channel normalization process, the resultant channel values were obtained based on the mean pixel intensity values. Let, μ_x, μ_y, μ_z represent the mean pixel intensity values of red, green, and blue channels, respectively. Subtracting these mean values from the original input channels (r, g, b), the resultant channels were obtained as follows: $R = r - \mu_x, G = g - \mu_y, B = b - \mu_z$. Using the channel scaling factor γ, the normalized channels were obtained as $r' = \frac{R}{\gamma}, g' = \frac{G}{\gamma}, and\ b' = \frac{G}{\gamma}$

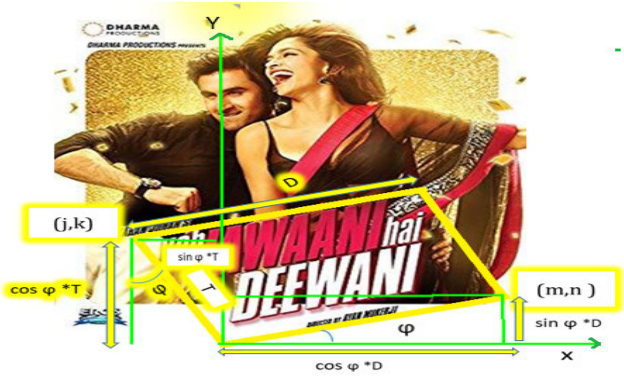


FIGURE 6. Coordinate derivation for the bounding box allocation of a Bollywood poster where the movie title is orientated.

2) COORDINATE EXTRACTION BLOCK

It was seen that the text areas in poster images are not always horizontally aligned rather they are oriented. In these cases, the horizontal rectangular box can enclose more areas and thus more background information can get confined. To avoid this or to enclose the minimum area as in the texts reside, the coordinates surrounding the texts need to be found to get a bounding box with the orientation of the texts. Let, the rectangle surrounding the text creates an angle ϕ (obtained from the text rotation angle block of lightweight EAST) with the horizontal axis.

To draw a rectangle, at least two coordinate points were needed. Let, (j, k) be the initial coordinate or an offset and find another coordinate say (m, n) (see Fig. 6). Using the following equations the other coordinate (m, n) can be determined considering T and D as the height and width of the two sides of the oriented rectangle.

$$m = j + D * cos\phi + T * sin\phi \tag{1}$$

$$n = k - D * sin\phi + T * cos\phi \tag{2}$$

The description of the above process can be understood by Fig. 6. After bounding box selection, it was found that there were many closely placed bounding boxes to enclose a single text, as shown in Fig. 7. To have a proper i.e., minimum enclosing bounding box NMS (non-maximum suppression) [34] method was adopted. In this method, the best-scored bounding box i.e., whose IoU (intersection over union) value is highest, is retained and the rest of the boxes are suppressed. The images after applying NMS are depicted in Fig. 8.

3) ROTATION AND SHEARING MODULE

To get the highest predicted scorer text boxes for rotated and sheared images, further rotation, and shearing are needed.

In the case of rotated movie poster images, to get the maximum score-valued text boxes (leads to maximum precision and recall for the text box), the image was further rotated by $d°$ in each step and checked for its score value. If further $d°$ rotation gives a better score value than the previous then that value along with the orientation (θ), predicted text boxes' coordinates are stored. But, if the value is less than the previous then previous values were retained. The process continues until the rotation reaches 360°. This process of rotation is explained in the algorithm 1. Here, d = 1 was set which was obtained from the ablation study (section IV-C1).

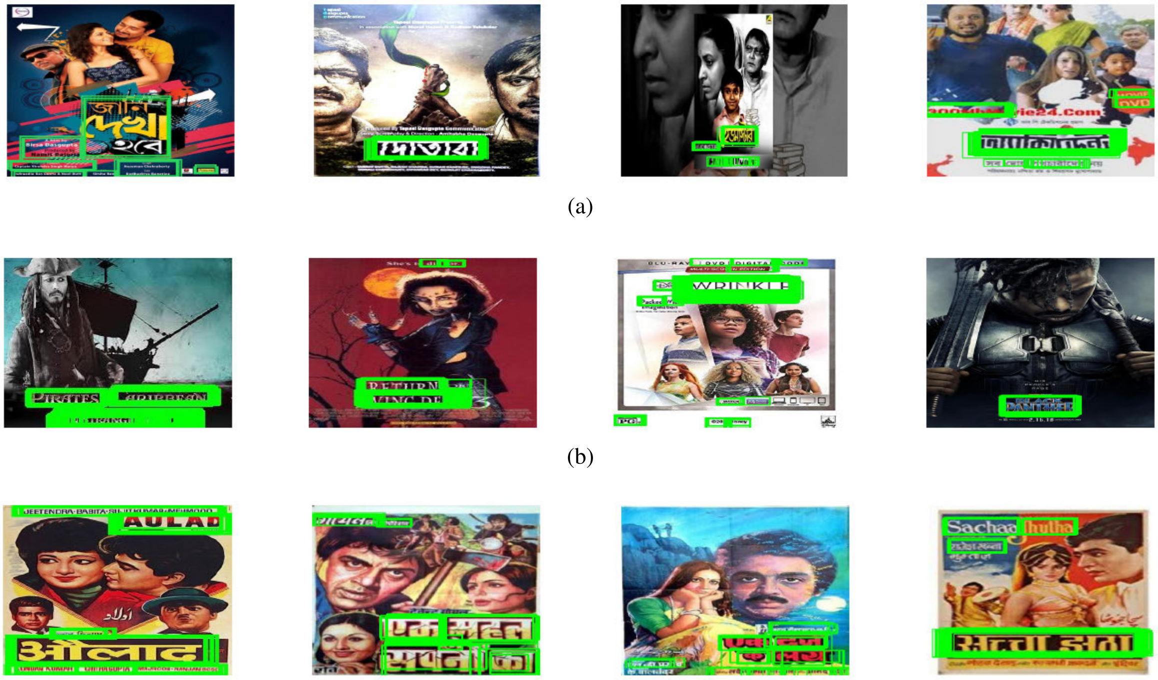
Algorithm 1 Rotation Algorithm

```
Input   : Rotated image
Output: ($\theta$), predicted text boxes' coordinates
Initialization;
I = input image; /* Input an rotated poster image*/
$\theta = 0°$; /* Angle of rotation*/
nW = 320; nH = 320; /*nH, nW = new height and new width of the input image*/
rW = W / float(nW);
rH = H / float(nH); /*rH and rW are the ratios of original and normalized images*/
C1 = 0;
while $\theta \leq 360°$ do
    /*rotate image by $d°$ */
    $\theta = \theta + d°$;
    call Lightweight_EAST
    C = confidence_score /* obtained from the score map of lightweight EAST */
    if $C \geq C1$ then
        Cmax = C;
        $(startX, startY, endX, endY)$ = Text_box(C);
    else
        Cmax = C1;
        $(startX, startY, endX, endY)$ = Text_box(C1);
    end
end
```

For any sheared image, there are two situations: positively or negatively sheared. Here, both cases are considered simultaneously. According to this algorithm 2, two iterations were considered, so that either of the value of the shearing variable is zero (i.e, (sh1 || sh2) == 0). Since the shearing status (+ve or -ve) is unknown to us, the value of sh1 and sh2 will be updated accordingly. The shearing value (sh1) is decreasing by 0.1 (update sh1) until the condition is satisfied i.e, sh1>= −1. Similarly, sh2 is increasing by 0.1 (update sh2) until the condition holds i.e, sh2<=1 in each step to get the maximum score-valued text boxes (leads to maximum precision and recall for the text box). The image was further sheared by 0.1 or -0.1 in each step and checked the score. If the latter one produces a better score value than the previous then that score value along with the sheared value (sh1 or sh2), and the predicted text boxes' coordinates are stored. For script identification, the stored rotation angle θ and sheared value will be required to use in the localized text boxes to use as of the original image's rotated or sheared position.

C. TITLE BOX EXTRACTION

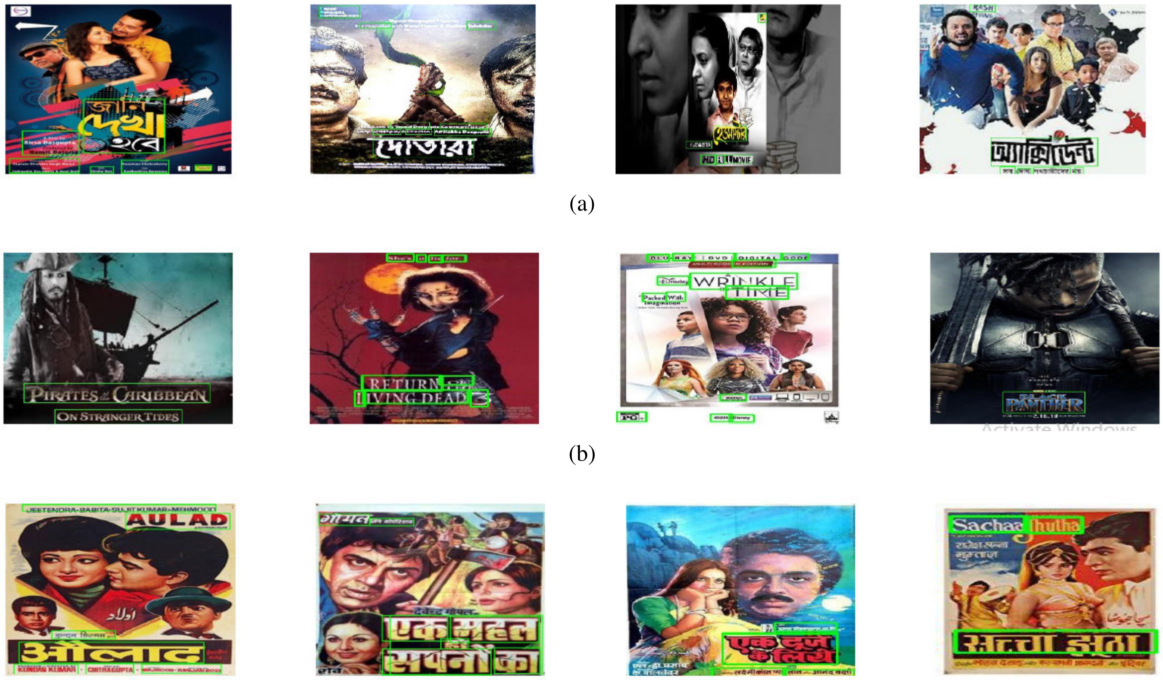
It was observed that the textboxes containing the movie titles are bigger compared to the other supplementary boxes in


(a)

(b)

(c)

FIGURE 7. Presence of initial multiple bounding boxes for the images shown in Fig. 4 (a-c).


(a)

(b)

(c)

FIGURE 8. Localized text box's (shown in (a)-(c)) after applying non-maximum suppression on Fig. 7.

general. But, in all cases, it is not always true. Based on the rectangle parameters, the wrong text box could be selected as a title box. Keeping this into account the boxes must be chosen intelligently in such a way that the entire movie title is extracted properly.

The maximum area and length of the text box were stored after comparing it with all text boxes. As $p = max(area)$ and $c = max(length)$, the width (d) is calculated by using the ratio between p and c as d $= \frac{p}{c}$. In Fig. 9 M denotes the box, which encloses the movie title and T represents the

Algorithm 2 Shearing Algorithm

```
Input  : Sheared image
Output: Sheared value, predicted text boxes'
        coordinates
Initialization;
I = input a sheared poster image
nW = 320; nH = 320;
/*nH, nW = new height and new width of the input
image */
rW = W / float(nW);
rH = H / float(nH);
/*rH and rW are the ratios of original and normalized
images*/
sh1 = 0; sh2 = 0;/*sh1, sh2 = Shearing variables*/
C1 = 0;
while $sh1 \geq -1$ do
    sh1 = sh1- 0.1;/* considering input image has a
    positive sheared value*/
    I1= $shear(I, sh1)$; /*I1 is sheared image of I by a
    value sh1*/
    call $Lightweight\_EAST$
    C = $confidence\_score$
    if $C \geq C1$ then
        Cmax = C;
        $(startX, startY, endX, endY) = Text\_box(C)$
    else
        Cmax = C1;
        $(startX, startY, endX, endY) = Text\_box(C1)$
    end
    while $sh2 \leq 1$ do
        sh2 = sh2 + 0.1;/* considering input image has
        a negative sheared value*/
        I2 = $shear(I, sh2)$ /*I2 is sheared image of I by
        a value of sh2*/
        call $Lightweight\_EAST$
        C = $confidence\_score$
        if $C \geq C1$ then
            Cmax = C;
            $(startX, startY, endX, endY) =$
            $Text\_box(C)$
        else
            Cmax = C1;
            $(startX, startY, endX, endY) =$
            $Text\_box(C1)$
        end
    end
end
```

box containing other supplementary text. L, D, and A denote the length, width, and area of the title box M, respectively. L1, D1, and A1 reflect other supplementary text's (T) length, width, and area, respectively.

The following cases can occur:

Case 1: If L > L1 but D < D1 (L, D ∈ M & L1, D1 ∈ T) i.e., the length of the movie title box is longer, the width is narrower, and another text box's length T is greater. But, there is a small difference in their widths. Here, confusion will arise in choosing the correct movie title box (shown in Fig. 9 a).

Case 2: If L < L1 and D ≫ D1 i.e., the length of movie title box M is less compared to T but the width of the movie title box is much greater, confusion will arise in the selection of box as movie title box (shown in Fig. 9 b).

Case 3: If L > L1 and D < D1, but A < A1, i.e., the length of the title box is greater and width is less than Box T, but the area of the title box is less than the area of T, confusion will arise in choosing the correct title box (shown in Fig. 9 c).

For case 1 and case 2, the solution can be expressed by considering the area metric only i.e., the box having greater area will be considered the title box but, in case 3 it is seen that considering the only area can lead us to the wrong title-text. For the title box extraction process, the text box coordinates and the number of predicted text boxes obtained after non-maximum suppression were taken as input. To get the rectangle box according to the original image's dimension, the coordinates of the rectangles were rescaled as $a = sX \times rH$, $b = sY \times rW$, $X = eX \times rH$, $Y = eY \times rW$. Here, sX, sY, eX, eY represents the coordinates of the rectangle box obtained after text detection from the normalized images (i.e., 320 × 320). X, Y, a, and b represent the coordinates of the rectangle box according to the scale of the original input image. The length and width of the text box corresponding to the input image size can be represented as $c = X - a$, $d = Y - b$. Thus, the area of the rectangle becomes p = c × d. To select the title boxes from all localized text boxes, the height, width, and area parameters of the text boxes were rescaled by considering equations 3-5. The area was rescaled as

$$p1 = \frac{1}{\alpha} \times max(p, p1) \tag{3}$$

The length was rescaled as

$$c1 = \frac{1}{\beta} \times max(c, c1) \tag{4}$$

The width was rescaled as

$$d1 = \frac{1}{\beta} \times max(d, d1) \tag{5}$$

Initially, p1 = c1 = d1 = 0. If (p ≥ p1 or c ≥ c1 and d ≥ d1) holds, then the rectangle can be cropped using the coordinate ((a, b), (X, b), (a, Y), X, Y)). The process of title box extraction was repeated for all the predicted text boxes in the image.

The process of text localization and movie title extraction was initiated with scaling factor $\gamma = 1$. It was observed that with $\gamma = 1$, the titles were correctly extracted in 80% images. This whole process is repeated for the rest of the images using different γ values. With $\gamma = 1.3$ about 15%, with $\gamma = 1.8$ the rest 5% titles were correctly extracted. By ablation study (section IV-C1), the parameters α and β were set.

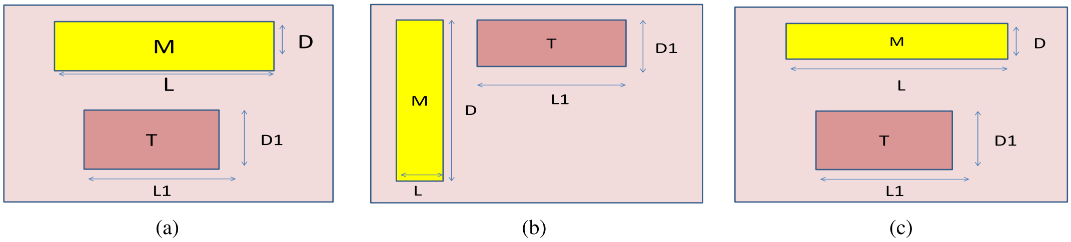


FIGURE 9. Scenarios that may lead to improper information of title box, shown as (a), (b), and (c) representing cases 1-3, respectively.

Using $\alpha = 2$ and $\beta = 1.2$, titles in the above cases, as well as the titles in dissimilar sized multiple boxes, were well extracted.

From the algorithm 1 and 2 the rotation angle θ and the shear value obtained were used to get the original augmented position of the movie titles for script identification of the corresponding movie titles. In Fig. 10 and 11 the bounding boxes of the movie titles and extracted titles are shown respectively. In Fig. 12 localized original and augmented movie title images of a Tollywood poster are shown.

In Table 1 the localized title and the metric values of a rotated (45° and 135°) and sheared Tollywood poster image using EAST and the proposed approach are presented. In the proposed methodology the images were rotated (θ) until the maximum-predicted score was obtained. Thus, the recall value (R) using the proposed method is a little higher than in EAST. The improvement of 8.39, 8.28, and 4.79 in F-score (F) was observed for rotated and sheared images, respectively using the new approach over EAST.

D. SHALLOW CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

Here, a shallow convolutional neural network (SCNN) was presented for the identification of the scripts of the extracted movie titles. A CNN [35] has generally three layers: the convolution, pooling, and fully connected or dense layer. However, there are architectures [36] which do not employ pooling. In the convolution layer, the input image is convoluted with a small kernel element-wise and by sliding process, the features are generated, which can be fed to the immediate next layer i.e. Pooling layer. It acts as a dimensionality reducer of the feature map. But, information loss occurs due to the dimensionality reduction which leads to a loss in accuracy of the system. The generation of feature map from this layer can be expressed as

$$F_{q,r} = \sum_{q=1}^{q}\sum_{r=1}^{r} I_{(q-a,r-b)} * k^{c}_{a,b} \tag{6}$$

where, $k^{c}_{q,r}$ represents the kernel for the a^{th}, b^{th} pixel in c^{th} layer over the instance $I_{q,r}$ and $*$ denotes the convolution operator.

$$F_{q,r} = \varsigma(\sum_{q=1}^{q}\sum_{r=1}^{r} I_{(q-a,r-b)} * k^{c}_{a,b}) \tag{7}$$

TABLE 1. Title box localization based on EAST and the proposed method, respectively are shown for rotated and sheared images.

	Augmentation	Localize	P	R	F
EAST	45°		91.98	71.87	80.69
	135°		91.94	72.06	80.80
	Shear		91.01	78.49	84.29
Proposed	45°		90.30	87.91	89.08
	135°		90.30	87.91	89.08
	Shear		90.30	87.91	89.08

here, ς denotes ReLU (rectified linear unit) activation function which can be expressed as

$$\varsigma(z) = max(0; z) \tag{8}$$

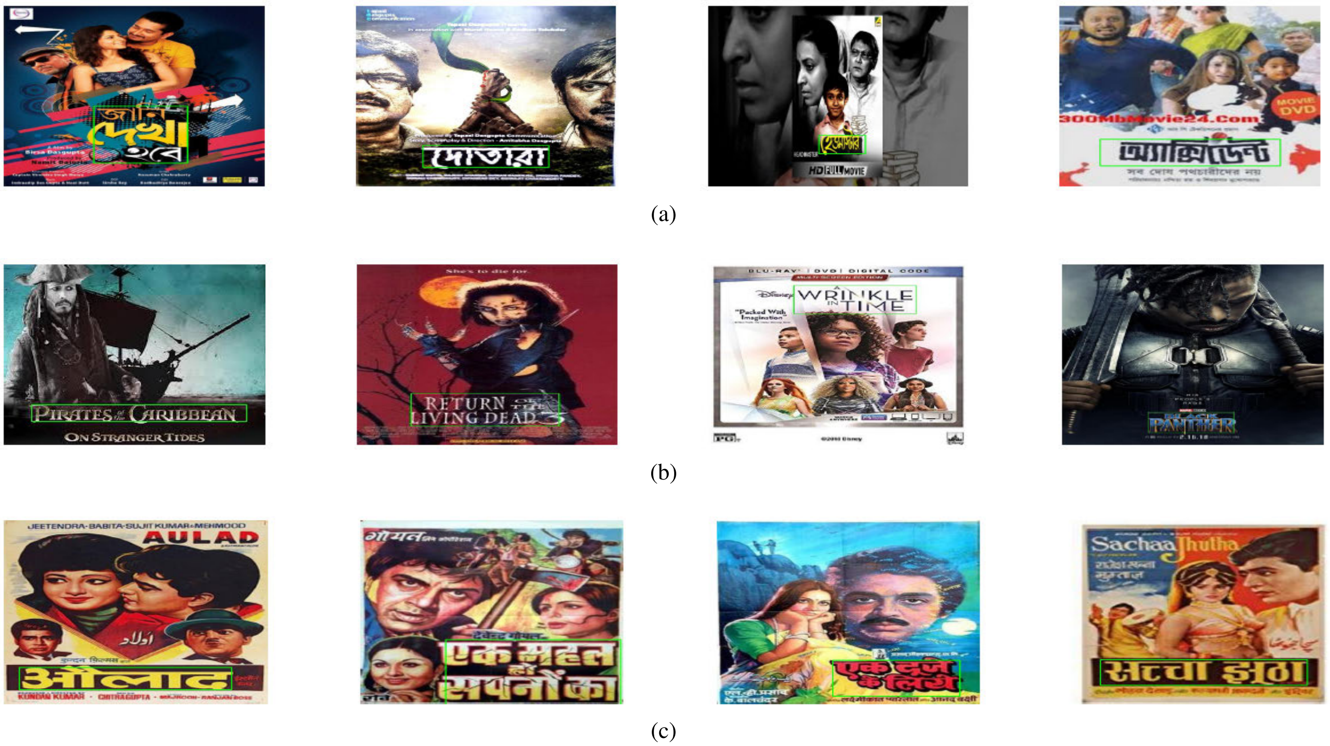
i.e., $if(z < 0); Re(z) = 0, otherwise, Re(z) = z$, where, z denotes the neuron input.

The output of pooling is fed to a dense layer/fully connected layer which has an all-to-all connection to the neurons to every neuron to the previous layer. Another activation function, named Softmax, is used in the final dense layer of the network, which can be expressed as

$$\chi(Y) = \frac{e^{Y}}{\sum_{1}^{L} e^{Y}} \tag{9}$$

here, Y denotes the input vector in this layer whose size is L.

The principle of designing a shallow convolutional neural network is to reduce the computational time, and space using fewer parameters so that it can be deployed in mobile devices that have low resource constraint issues. Also, to avoid the over-fitting issue that happens in the case of deep architecture, for longer inference time and use of a huge number of parameters. Here only two successive convolution layers followed by a single dense layer/ output layer were considered, which can be termed as CCD where C, D represents the convolution and dense, respectively. The kernel size of the first

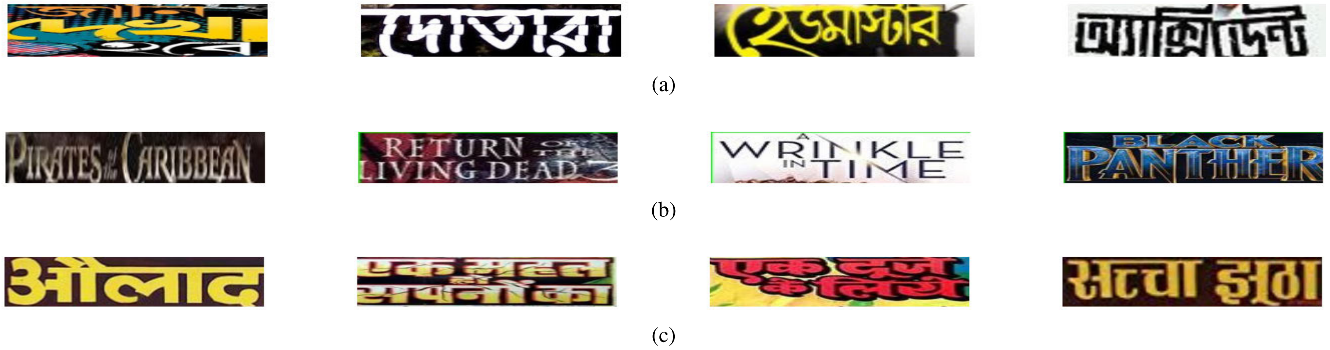


(a)

(b)

(c)

FIGURE 10. Localized movie title boxes from Fig. 8.

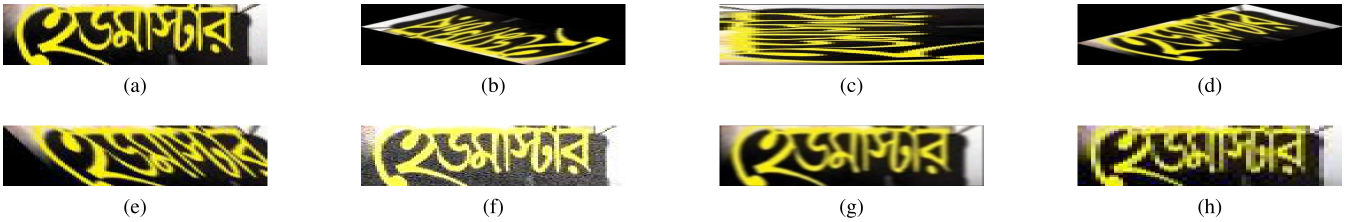


(a)

(b)

(c)

FIGURE 11. Extracted movie titles from Fig. 10. In the first row (a) the Bangla movie titles from Tollywood posters; in the second row (b) the Roman titles from Hollywood posters and the Devanagari titles from Bollywood posters are presented in the third row (c).


(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 12. Movie title (a) of 3^{rd} image of Fig. 11 (c) extracted from the 3^{rd} image of Fig. 4 (c) and its augmented forms are shown in (b-h).

convolution layer was 5 × 5 and for the second convolution, was 3 × 3. There were 32 and 16 filters used in the first and second convolutional layers. The features generated by the second layer were fed to a dense layer of size 3 (for three classes). The stride value for the convolutions was kept one. Table 2 shows the number of parameters used in the training phase. There were 1813587 trainable parameters used in total.

The production of the number of parameters in the convolutional layer and dense layer along with their dimensions are expressed by equations 10, 11, and 12. For an input image having dimension Q × R × 3, using the kernel of dimension

TABLE 2. The number of parameters used in the proposed SCNN.

Layer Type	Dimension	# Param
Convolution1	196x 196x 32	2432
Convolution2	194x 194x 16	4624
Dense	3	1806531
Total parameters	-	1813587

$k_i \times l_i$ in the i_{th} convolution layer, and the number of filters χ_i, ψ_i in the i_{th} and in the previous layer respectively, it can be written as

$$Dimension_i = (Q - k_i + 1) \times (R - l_i + 1) \times \chi_i, \quad (10)$$

$$Parameter_{cov_i} = ((k_i \times l_i \times \psi_i) + 1) \times \chi_i. \quad (11)$$

The parameters generated in the dense layer depends on the input sizes τ_{in} received from the previous layer (i.e., convolutional layer2) and the output size τ_{out} of the dense layer, can be written by

$$Parameter_{Dense} = \tau_{out} \times (\tau_{in} + 1) \quad (12)$$

Table 2 can be comprehended by the above analogy through the equations 10, 11, and 12. For input image dimensions 200 × 200 × 3, layerwise feature dimensions and corresponding number of parameters were calculated as

$Dimension_1 = (200 - 5 + 1) \times (200 - 5 + 1) \times 32 = 196 \times 196 \times 32$,

$Parameter_{conv1} = ((5 \times 5 \times 3) + 1) \times 32 = 2432$,

$Dimension_2 = (196 - 3 + 1) \times (196 - 3 + 1) \times 16 = 194 \times 194 \times 16$,

$Parameter_{conv2} = ((3 \times 3 \times 32) + 1) \times 16 = 4624$,

$Parameter_{Dense} = (3 \times ((194 \times 194 \times 16) + 1)) = 1806531$

It is obvious that $(Parameter_{conv1} + Parameter_{conv2} + Parameter_{Dense}) = 1813587$ is the total number of trainable parameters used in the proposed architecture. In Fig. 13 the activation map generated from the proposed shallow convolution neural network is depicted.

IV. EXPERIMENTS

A. DATASET

Working on the movie-poster image experiment requires a dataset to examine the efficiency of the method suggested. But, the unavailability of such a dataset leads us to create a dataset. The dataset[1] used in this paper was collected from various sources such as IMDB,[2] Pinterest.[3] Here, multi-script images of Bangla, Roman, and Devanagari were considered. Thus, the sets were composed of poster images of Tollywood, Hollywood, and Bollywood having dissimilar graphical styles, font styles, and diversity in colors, contrast, etc. This dataset consists of 1431 poster images. Among these, there are 484 images where the names of the films are printed/ written in Bangla script, 607 in Roman, and 340 in Devanagari. For experimentation, voluminous, as well as a diverse set of images, was required. For this purpose, the augmentation procedure was adopted, which is discussed in section III-A. Seven augmentation methods were considered: three types of rotation (45°, 90°, and 135°), shearing, blurring, single-level wavelet decomposition, and noise. The dataset was divided into two main sets: non-localized and localized. In the non-localized set, the normalized raw images along with their augmentations were considered. In the localized set, only localized movie titles along with their augmentations were regarded.

In Table 3 the details of the number of scripts along with their augmentations are shown. To deal with the real-world scenario, the augmented data were mixed with the original data in both the non-localized and localized cases. In the non-localized set, the original non-localized images were mixed with their corresponding augmented images. Also, a set was built by considering seven augmentations along with the original set. In the same way, the localized set was made. The volume of the non-localized and the localized sets are 32,913 each. The total dataset size is 65,826 considering the localized, non-localized, original, and seven-way augmentation.

B. EVALUATION PROTOCOL

1) LOCALIZATION PROTOCOL

After the text box extraction, the text boxes were compared with their corresponding ground truth. The following formulae were considered to compare with the ground truth and dynamically created coordinates of each box. The union over intersection on the generated box (ρ) and the ground truth (σ) provide the degree of matching with each other, which can be written as

$$M(\rho;\sigma) = max[I(\rho;\varphi|\varphi \in \sigma)] \quad (13)$$

here, $I(\rho;\varphi)$ denotes the area of intersection of the created box and the ground truth. The precision (P), recall (R), and F-score (F) were measured based on Equation 13, can be defined as

$$Precision = \frac{\sum_{x_i \in E}, \mathrm{M}(x_i; a_i)}{|E|}, \quad (1 \le i \le n) \quad (14)$$

$$Recall = \frac{\sum_{y_i \in T} \mathrm{M}(y_i; x_i)}{|T|}, \quad (1 \le i \le n) \quad (15)$$

here, x_i and y_i denote each generated, and the ground truth bounding boxes, respectively. E and T denote the set of generated and ground truth bounding boxes, respectively. The F-score can be evaluated from Equations 14 and15 as

$$Fscore = \frac{2 * (Precision * Recall)}{Precision + Recall}. \quad (16)$$

2) SEPARATION PROTOCOL & TRAINING REGIME

The n-fold cross-validation system was adopted for the script separation. The entire dataset was split into n subsets using such approach. The one subset was a test set among them. The training was regarded for the residual (n-1) subset.

[1] This dataset is available upon request for research purposes by email: mridulxyz@gmail.com.

[2] https://www.imdb.com, visited on 14.06.2021.

[3] https://www.pinterest.co.uk, visited on 14.06.2021.

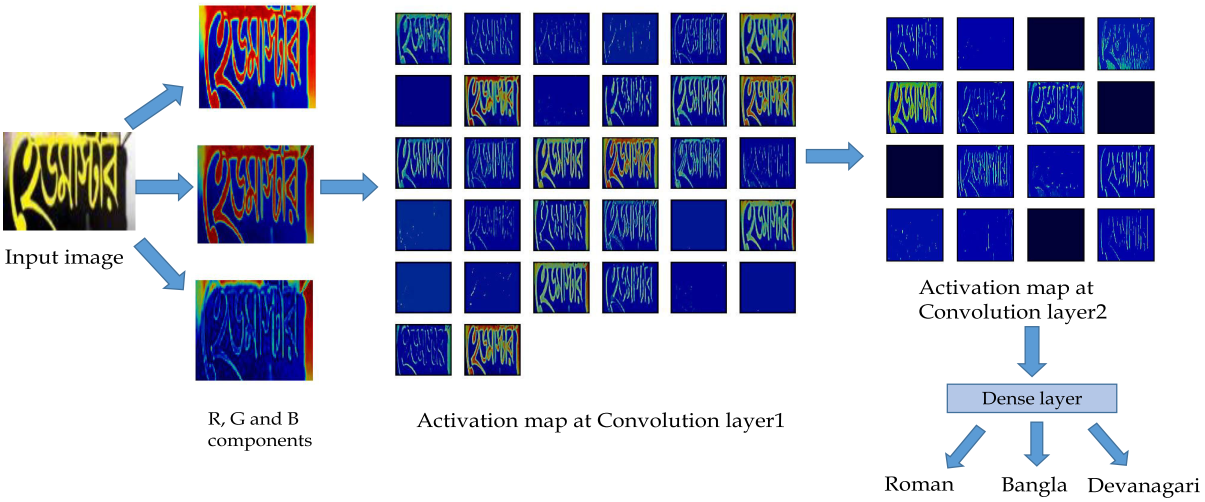

Input image
R, G and B components
Activation map at Convolution layer1
Activation map at Convolution layer2
Dense layer
Roman
Bangla
Devanagari

FIGURE 13. Layerwise activation map produced by the proposed SCNN model for a movie title from Tollywood as shown in Fig 12 (a).

TABLE 3. Details of the number of scripts along with their augmentations. Scripts with their original, mixed augmented (i.e., localized/non-localized original data were mixed with their corresponding augmented data).

		Mixed							
	Original	Rotation(°)			Shear	Noise	Blur	Decomposed	All aug
		45	90	135					
Bangla	484	968	968	968	968	968	968	968	3872
Roman	607	1214	1214	1214	1214	1214	1214	1214	4856
Devanagari	340	680	680	680	680	680	680	680	2720
Total	1431	2862	2862	2862	2862	2862	2862	2862	11448

Second time, another subset was employed for the test set, with the remainder (n-1) subset being reserved for training. This procedure was done n times more. The experiment was conducted with 5 and 10-fold cross-validation. The images were normalized to 200 × 200 pixels for script separation purposes. The script separation accuracy can be defined as

$$Accuracy = \frac{(\#correctly\ separated\ images)}{(\#total\ images)} * 100\% \quad (17)$$

Initially, 200 epochs were used with a batch size of 100. Here a dropout value of 0.5 was considered. The learning rate was considered 0.001. The exponential decay rate for the first momentum and the second momentum was kept at 0.9 and 0.999, respectively. The epsilon value was 0.0000001. The value of the Boolean optimization parameter variable named AMSGrad was made false.

C. RESULTS & ANALYSIS

1) ABLATION STUDY

Extensive experiments were conducted for proper tuning of various parameters used in this work.

It is noted from Figure 14 that when the image is rotated by a step of 2°, 3°, 4°, and 5°, the precision drops by 1.5%, 2.4%, 5.4%, and 6.2% respectively as compared to the 1° step. The execution time was only 1.12, 1.28, 1.5, and 1.6 times higher than 2°, 3°, 4°, 5° rotation intervals respectively. Hence, the experiment was performed with the 1° step.

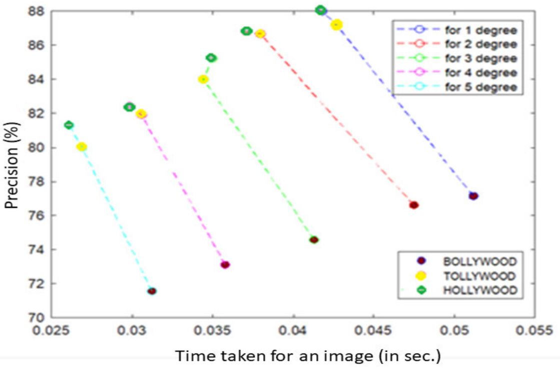


FIGURE 14. Precision vs. time taken to rotate an image (in sec.) considering 1 to 5° step of rotation for Tollywood, Bollywood, and Hollywood images while experimenting on core i5, and 4GB RAM machine.

The scaling factor (γ) for text localization is taken as 1, 1.3, and 1.8 based on the trial run on the repetitive experiments. Similarly, for title localization, the parameters affecting the probability of proper title box localization viz α and β (used by equations 3, 4, and 5 for scaling rectangle parameters) are also chosen from a set of experiments, as depicted in Table 4. It is observed that for $\alpha = 2$ and $\beta = 1.2$ highest probability of 0.915 was obtained. The ablation study was also carried out for the script separation. Using a batch size and epochs of 100, this experiment was conducted. The evaluation parameters of script separation are tabulated in Table 5. It is observed that for 10 fold cross-validation the highest accuracy was obtained for both the original and mixed decomposed set.

2) LOCALIZATION PERFORMANCE

The performance of the M-EAST model was assessed using the evaluation metrics like precision, recall, f-measure, and FPS, which is depicted in Table 6.

TABLE 4. The probability (Pr) of extracting correct movie title box(s) from text boxes for different value of α and β.

α	1	1	0.8	1	1.2	1.4	1.6	1.8	**2**	2.2	2	2
β	1	0.8	1.2	1.4	1.2	1.2	1.2	1.2	**1.2**	1.2	1.4	0.8
Pr	0.76	0.65	0.77	0.76	0.79	0.80	0.83	0.87	**0.91**	0.86	0.87	0.70

TABLE 5. Different evaluation metrics for different folds for the original and mixed decomposed set are presented.

Set	#Fold	Accuracy(%)	Precision	Recall	F-score
Original	03	97.02	0.976	0.969	0.972
	05	98.81	0.988	0.988	0.988
	07	98.31	0.986	0.988	0.987
	09	98.84	0.989	0.981	0.985
	10	**98.95**	**0.992**	**0.989**	**0.990**
	12	98.12	0.987	0.979	0.983
Decomposed	03	98.15	0.984	0.979	0.981
	05	99.23	0.994	0.991	0.992
	07	99.10	0.995	0.988	0.991
	09	99.33	0.994	0.992	0.993
	10	**99.65**	**0.998**	**0.995**	**0.996**
	12	98.89	0.992	0.986	0.989

TABLE 6. The evaluation of the proposed M-EAST localization method is tabulated.

	Precision	Recall	F-Score	FPS
Tollywood	87.16	83.11	85.09	23.4
Bollywood	77.12	77.67	77.39	19.5
Hollywood	**87.98**	**83.77**	**85.82**	**23.9**
Average	84.09	81.52	82.77	22.27

The ROC (receiver operating characteristic) plot with the true positive rate and the false positive rate are shown in Fig. 15. The AUC (area under the curve) values obtained were 0.8463 for Bangla (Tollywood), 0.8727 for Roman (Hollywood), and 0.7012 for Devanagari (Bollywood), respectively.

3) SCRIPT SEPARATION PERFORMANCE

The separation was accomplished using 5-fold and 10-fold cross-validation techniques. The set that provided us the high accuracy was used for the detailed testing. The experimentation regarding the separation of scripts was conducted in two phases. In the first phase, the non-localized set, and in the second phase the localized movie title set was processed into the SCNN framework. Table 7 describes the script separation accuracy of non-localized and localized scripts where the original, as well as a combination of original and augmented images, were regarded, respectively.

In Table 7 it is seen that for original poster images, the accuracy of 91.33% and 96.15% were obtained for 5 and 10-fold cross-validation respectively making 8.67% and 3.85% error. It is observed that for mixed (original + augmented) 45°, 90°, 135° rotations, sheared, noisy, blurred, and decomposed the improvement in accuracy for 10-fold compared to 5-fold increased by 5.74%, 8.07%, 6.08%, 4.85%, 4.93%, 0.52%, and 0.24%, respectively.

For localized movie titles, it is seen from the table that 0.14% less error was generated for 10-fold cross-validation than 5 fold. For mixed data i.e., 45°, 90°, 135° rotations, sheared, noisy, and blurred the improvement in accuracy for 10-fold compared to 5-fold increased by 4.36%, 5.62%, 4.79%, 4.82%, 0.53%, and 0.5% respectively. It is observed that using mixed decomposed data, the highest accuracy of 99.65% was obtained using 10-fold cross-validation, making an error of 0.35%. The confusion matrix of the highest accuracy is explored and presented in Table 11. The experimentation was also conducted separately to test the efficiency of the proposed network by mixing the original and augmented data (all seven kinds) for the localized and the set that wasn't localized. The results are depicted in Table 8.

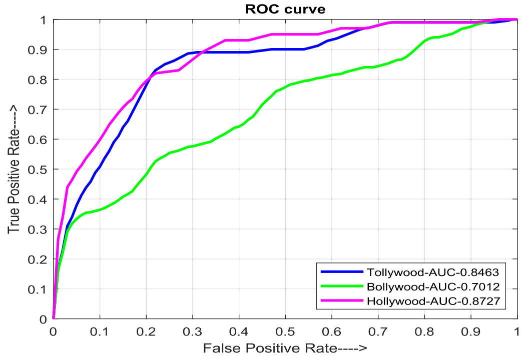


FIGURE 15. ROC curves for Tollywood (Bangla), Bollywood (Devanagari), and Hollywood (Roman) posters, which are depicted in blue, green, and magenta lines, respectively.

The experiment was further conducted with the set having the highest accuracy i.e., mixed decomposed to get increased accuracy. The epoch number was increased from 200 to 400 with an interval of 50 epochs and found out the accuracy which is tabulated in Table 9. From this Table, it is seen that at 300 epoch 99.82% accuracy was obtained, which is the highest among other epochs. Keeping this epoch (300) fixed, the batch size was increased from 100 to 300 with a batch size interval of 50 and tabulated the result in Table 10. It is realized from this Table is that for 100 and 200 batch size 99.82% accuracy was obtained i.e., there was no improvement in accuracy after 100 batch size.

To test the robustness of the proposed scheme, the system was also trained with original data and tested with the augmented set. The results are shown in Table 12.

4) ERROR ANALYSIS

From Tables 7, and 8, it is observed that there are 7.38% and 2.8% improvements in accuracy for 5 and 10-fold cross-validation using original localized title images over non-localized images. For mixed augmented data, mixed decomposed images provided the highest accuracy and compared to non-localized decomposed images 0.88% improvement on localized decomposed sets in 10-fold cross-validation. For localized all-mixed-data 0.07% error was generated compared to localized mixed decomposed data. After increasing the number of epochs to 300 the error was decreased to 0.18%.

In Fig. 16 few instances of improper localized images are shown. Images (a) and (e) suffered from improper localization. The possible reason is small letter issues and

TABLE 7. The accuracy of the original and mixed non-localized and localized images for 5 and 10-fold cross-validation.

Set	Folds	Original (%)	Original+ Augmented(Mixed)						
			45° (%)	90° (%)	135° (%)	Sheared (%)	Noisy (%)	Blur (%)	Decomposed (%)
Non-localized	5	91.33	88.81	86.51	88.36	90.32	93.32	97.97	98.53
	10	96.15	94.54	94.58	94.44	95.17	98.25	98.49	**98.77**
Localized	5	98.81	90.46	89.69	89.93	91.61	98.91	98.31	99.23
	10	98.95	94.82	95.31	94.72	96.43	99.44	98.81	**99.65**

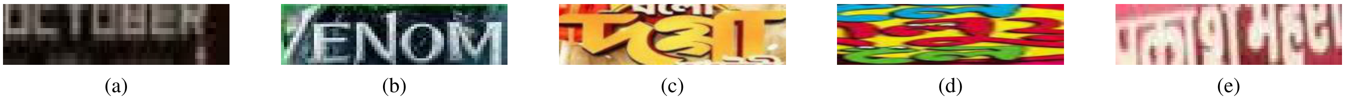

(a) (b) (c) (d) (e)

FIGURE 16. Some of the cases where the proposed localization scheme could not properly localize the movie titles. (a-b) from Hollywood posters; (c-d) from Tollywood and (e) from a Bollywood poster.

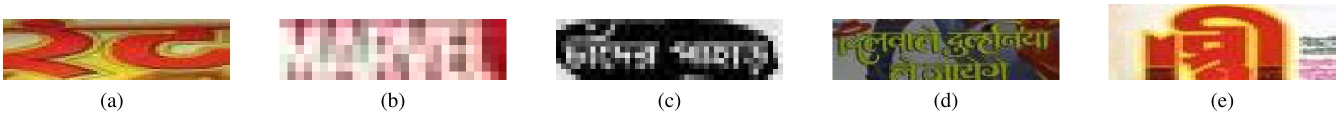

(a) (b) (c) (d) (e)

FIGURE 17. Script separation error: two Devanagari title images (from Bollywood poster): (a) and (d) were wrongly identified as Bangla script; a Roman title image (from Hollywood poster) (b) was wrongly identified as Bangla; two Bangla title images (from Tollywood poster), (c) and (e) were also wrongly identified to be in Devanagari class.

TABLE 8. The accuracy of the mixed (original and all seven types) augmented images for non-localized and localized sets.

Folds	Non-local-original+augmented (%)	local-original+augmented (%)
5	88.18	**98.31**
10	93.86	**99.58**

TABLE 9. The accuracies corresponding to the different epochs of the localized mixed-decomposed set keeping the fixed batch size of 100.

Batch	100	150	200	250	300
Accuracy(%)	**99.82**	99.72	**99.82**	99.72	99.54

blurriness and noisiness and low illuminations. Image (b) suffers from the foreground and background similarity. In (c) the movie title was written in two lines and the word in the first line is too small compared to the second. Image (d) was written very artistically, and there is background similarity.

It is observed from the confusion matrix of Table 11, that there is a misunderstanding between Bangla and Devanagari scripts. Since both Bangla and Devanagari were derived from the Bramhi script, there are similarities in certain letters. In Fig. 17 the wrongly separated title images are presented. Images (a) and (e), Devanagari and Bangla script wrongly separated from each other. This wrong separation was due to the similarity in cursive characters of both these classes and improper localization issues. Images (b) and (c), written in Roman and Bangla scripts, were wrongly separated into Bangla and Devanagari, respectively. They suffered from small letters and low illumination as well as noisiness and blurriness issues. Image (d), written in Devanagari, was wrongly separated into Bangla class due to foreground-background similarity and similarity in cursive letters in both classes.

5) COMPARISON

The performance of the original EAST model on the developed dataset is depicted in Table 13.

TABLE 10. The accuracies corresponding to the different batch sizes of the localized mixed-decomposed set.

Epoch	200	250	300	350	400
Accuracy(%)	99.65	99.79	**99.82**	99.72	99.61

TABLE 11. The confusion matrix corresponding to the highest accuracy of 99.82% for the mixed single-level decomposed set.

	Bangla	Roman	Devanagari
Bangla	967	0	2
Roman	1	**1210**	0
Devanagari	2	0	680

TABLE 12. The script separation performance (%) for 7:3 and 8:2 train-test ratio using original localized set as training and augmented as testing.

Ratio	45°	90°	135°	Sheared	Noisy	Blur	Decomposed
7:3	87.81	86.51	84.36	85.32	87.32	86.97	93.52
8:2	89.54	88.58	86.44	87.17	93.02	92.49	**95.21**

TABLE 13. The performance using EAST method.

	Precision	Recall	F-Score	FPS
Tollywood	82.32	79.82	81.06	18.7
Bollywood	77.40	73.61	75.46	16.2
Hollywood	**84.27**	**80.99**	**82.60**	**19.4**
Average	81.34	78.14	79.71	18.10

Comparing the performance of the EAST and M-EAST models (Tables 6 and 13), it was observed that the F-score value got improved by 4.03%, 1.93%, and 2.70% for Tollywood, Bollywood, and Hollywood poster images, respectively. The major improvement was found in the case of frame per second (FPS) value. The average FPS value was increased by 4.07 using the proposed modified EAST method over EAST.

The performance of the proposed localization technique was also compared with the state-of-the-art methods of public datasets, ICDAR 2017 [53] and ICDAR 2019 [54] RRC MLT (robust reading challenge multilingual

TABLE 14. **The proposed text localization technique's performance was evaluated and compared with state-of-the-art using the ICDAR 2017 RRC MLT dataset [55].**

Method	Precision	Recall	F-Score
EAST [33]	72.90	67.40	70.10
Dasgupta et al. [52]	88.60	73.9	80.50
TeLCos [12]	78.70	64.90	71.13
FCOS+FPN [53]	80.12	76.42	78.23
FCOS+BiFPN[53]	83.41	78.26	80.75
Proposed M-EAST	**90.01**	**81.07**	**84.50**

TABLE 15. **The proposed text localization technique's performance was evaluated and compared with state of the art using the ICDAR 2019 RRC MLT dataset [54].**

Method	Precision	Recall	F-Score
EAST [33]	67.14	61.60	64.25
Tencent-DPPR Team [54]	**87.52**	**80.05**	**83.61**
CRAFTS [56]	81.42	62.73	70.86
Multiplexed TextSpotter [56]	85.53	63.16	72.66
Proposed M-EAST	86.45	79.98	83.08

TABLE 16. **The performance of state-of-the-art methods evaluated on the developed dataset in this work.**

Method	Precision	Recall	F-Score
EAST [33]	81.34	78.14	79.71
TexRNet [46]	84.60	78.09	81.21
BiFPN [53]	85.14	80.00	82.49
ViTSTR [56]	**85.23**	79.87	82.46
Proposed M-EAST	84.09	**81.52**	**82.77**

scene text) (Task-1). The quantitative results are presented in Table 14 and 15 respectively. In Table 14 it was observed that an increase of 1.41%, 7.17%, and 4% in precision, recall, and F-score respectively was obtained respectively as compared to [52]. In Table 15 it is seen that the proposed method lacked behind Tencent-DPPR Team by only 1.07%, 0.07%, and 0.53% in terms of precision, recall, and F-score respectively.

The text localization performance (average precision, recall, and F-score) was also compared with the other methods, which are shown in Table 16. It is seen that the precision of the proposed method was 1.14% less as compared to ViTSTR. However, 0.52% and 0.28% higher results were obtained for recall and F-score respectively as compared to BiFPN.

The script separation performance of the proposed SCNN model was compared with the handcrafted feature-based methods like WLD (Weber's local descriptor) [37], Gabor wavelet transforms (GWT) based feature [38], and LBP (local binary pattern) [39]. For this experiment, here a localized mixed decomposed set was used since it gave us the highest accuracy in the proposed shallow convolutional network. Using the Gabor wavelet transform, for 3 scales and 5 orientations, 9375 features were extracted. For WLD, using parameters radius as 3 and 8 neighbors, 960 features were extracted. Working with such a large number of features, the performance of the system would fall. Thus, the feature dimensions were reduced using a dimension reduction technique named PCA (principal component analysis method) [40] and retained the feature dimension of WLD and GWT to 100, respectively. From LBP, 59 features were obtained. Using standard machine learning classifiers, Random forest (RF) [41], Naive Bayes (NB) [42], Multilayer perceptron (MLP) [43], Radial basis function (RBF) [44] the evaluation metrics such as accuracy (Acc), precision (P), recall (R), F-score (F) were calculated, which are tabulated in Table 17.

Experiments were also conducted with standard deep neural network architectures and compared them with the proposed architecture using the evaluation matrices. Here VGG19 [45], Xception [47], InceptionV3 [48], ResNet50 [49], and EfficientNetB0 [50] networks were used to compare with the proposed network. The accuracy (%), precision, recall, F-score, number of trainable parameters, inference time (in millisecond), and model size (in Mb) of the models are tabulated in Table 18.

The number of layers deployed in VGG19, Xception, InceptionV3, ResNet50, and EfficientNetB0 is 19, 71, 48, 50, and 237 respectively. Here, better accuracy, precision, recall, and F-score values were achieved using the proposed SCNN that is only 3 layered networks compared to these deep networks. It is also observed that the number of trainable parameters used in the network is 11.04, 9.38, 6.98, 8.32, and 1.78 times less compared to VGG19, Xception, InceptionV3, ResNet50, and EfficientNetB0 networks, respectively. The inference time and model size of the SCNN was also less compared to the standard networks. The experimentation was conducted in a system having 32 GB RAM. It was equipped with the NVIDIA Quadro RTX 5000, 16 GB GPU.

The proposed M-EAST was also compared with the localization technique using the dataset of [51]. It was found that in the proposed M-EAST, there is a 1.8% improvement in precision for Hollywood posters whereas 2.3%, 0.7% improvement in recall, 1.5%, 1.4% improvement in the F-score for Hollywood and Bollywood posters. The improvement in FPS values observed were 0.7, 0.6, and 0.5 for Tollywood, Bollywood, and Hollywood poster images. For script separation, it was also seen that using the proposed SCNN architecture (3 layers), only 0.04% loss in accuracy was obtained over the CPCPDDD (7 layers) architecture.

TABLE 17. **The performance using handcrafted feature based methods which were separated with machine learning classifiers.**

	WLD				GWT				LBP			
	Acc(%)	P	R	F	Acc(%)	P	R	F	Acc(%)	P	R	F
RF	67.85	0.688	0.679	0.666	89.58	0.901	0.896	0.895	69.32	0.696	0.693	0.672
NB	47.37	0.519	0.474	0.472	67.74	0.674	0.677	0.674	54.22	0.537	0.542	0.528
MLP	61.74	0.616	0.617	0.617	91.54	0.916	0.915	0.916	65.51	0.654	0.655	0.655
RBF	52.86	0.536	0.529	0.532	67.15	0.669	0.672	0.669	53.00	0.523	0.53	0.524

TABLE 18. Comparison with standard deep learning architecture by using different evaluation metrics.

	Acc (%)	P	R	F	#Trainable param	Inference time(ms)	Model size (Mb)
VGG19	98.39	0.985	0.983	0.984	20025923	22	152
Xception	82.35	0.856	0.824	0.840	17016379	26	144
InceptionV3	71.56	0.760	0.716	0.737	12670403	4	132
ResNet50	78.55	0.790	0.781	0.785	15090915	10	122
EfficientNetB0	96.50	0.970	0.960	0.965	3235071	4	42.6
Proposed	**99.82**	**0.998**	**0.998**	**0.998**	**1813587**	**2**	**4.9**

TABLE 19. Different metrics of script identification were compared with available methods in the literature for ICDAR 2017 and ICDAR 2019 RRC MLT dataset (Task2).

Dataset	Method	Acc(%)	P	R	F
ICDAR 2017[55]	SCUT-DLVCla	87.69	-	-	-
	CNN based method	88.09	-	-	-
	Proposed	**94.73**	0.947	0.947	0.947
ICDAR 2019 [54]	SOT	91.81	-	-	-
	Tencent-DPPR Team	94.03	-	-	-
	Proposed	**95.18**	0.951	0.951	0.951

The script identification performance of the proposed SCNN framework was also compared with the state-of-the-art methods on ICDAR 2017 and ICDAR 2019 RRC MLT dataset (Task2). The results are tabulated in Table 19. It is seen that better performance was achieved using the proposed shallow network.

V. CONCLUSION

Movie posters contain complex graphical texts and compound background graphics. These texts are written artistically to attract audiences and to give information about the movie. The extraction of these texts is a challenge when there are resemblances in the foreground and background graffiti. In this work, an automatic text extraction method, M-EAST, was proposed, which is based on the transfer-learning process. The rotated and sheared images often create hindrances in proper text localization. Thus, properly aligned images were required for text localization well. This was achieved using the rotation and shearing module of M-EAST. An automatic movie title extraction algorithm was also proposed.

To work with diverse data and to increase the volume of the dataset, seven types of data augmentation techniques were adopted. The highest accuracy of 99.82% was obtained for mixed decomposed sets in case of script separation. For the decomposition, a single level 'Haar' wavelet was considered. By decomposing into a single level, the low-frequency components were arrested. Using seven-way augmentation, the accuracy dropped off by 0.07% only in 200-epoch and for a batch size of 100, compared with the highest accuracy in 10-fold cross-validation.

The proposed M-EAST works very efficiently in the horizontal and for the tilt/oriented/ sheared texts. But, for long, blurry, noisy text, and close similarity in foreground-background in the image, the localization is often not accurate. In the future, these localization issues along with the minimization of the class separation error rate will also be addressed. To deploy the proposed technique in resource-constrained devices like mobile platforms are being planned. Also, the development of a uniform architecture for text localization as well as script identification in scene images is planned soon.

REFERENCES

[1] J. A. Wi, S. Jang, and Y. Kim, "Poster-based multiple movie genre classification using inter-channel features," *IEEE Access*, vol. 8, pp. 66615–66624, 2020.

[2] Y.-F. Huang and M.-C. Hsieh, "Text extraction and recognition from posters for movie title retrieval," in *Proc. 19th Int. Database Eng. Appl. Symp. (IDEAS)*, 2014, pp. 180–185.

[3] Z. Guo, Y. Li, Y. Wang, S. Liu, T. Lei, and Y. Fan, "A method of effective text extraction for complex video scene," *Math. Problems Eng.*, vol. 2016, pp. 1–11, Jul. 2016.

[4] R. Atienza, "Vision transformer for fast and efficient scene text recognition," 2021, *arXiv:2105.08582*. [Online]. Available: http://arxiv.org/abs/2105.08582

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[7] Z. Zhong, L. Sun, and Q. Huo, "Improved localization accuracy by LocNet for faster R-CNN based text detection in natural scene images," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106986.

[8] S. Gidaris and N. Komodakis, "LocNet: Improving localization accuracy for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 789–798.

[9] S. P. F. Joan and S. Valli, "An enhanced text detection technique for the visually impaired to read text," *Inf. Syst. Frontiers*, vol. 19, no. 5, pp. 1039–1056, Oct. 2017.

[10] S. Ganguly, H. Mukherjee, and K. Roy, "Towards automatic detection of colour blindness," in *Proc. Int. Conf. Emerg. Technol. Sustain. Develop. (ICETSD)*, 2019, pp. 399–402.

[11] F. Naiemi, V. Ghods, and H. Khalesi, "MOSTL: An accurate multi-oriented scene text localization," *Circuits, Syst., Signal Process.*, vol. 40, pp. 1–22, Feb. 2021.

[12] R. S Munjal, M. Goyal, R. Moharir, and S. Moharana, "TeLCoS: OnDevice text localization with clustering of script," Apr. 2021, *arXiv:2104.08045*. [Online]. Available: http://arxiv.org/abs/2104.08045

[13] M. He, M. Liao, Z. Yang, H. Zhong, J. Tang, W. Cheng, C. Yao, Y. Wang, and X. Bai, "MOST: A multi-oriented scene text detector with localization refinement," Apr. 2021, *arXiv:2104.01070*. [Online]. Available: http://arxiv.org/abs/2104.01070

[14] X. Wang, S. Zheng, C. Zhang, R. Li, and L. Gui, "R-YOLO: A real-time text detector for natural scenes with arbitrary rotation," *Sensors*, vol. 21, no. 3, p. 888, Jan. 2021.

[15] U. A. Khan, M. A. Martinez-Del-Amor, S. M. Altowaijri, A. Ahmed, A. U. Rahman, N. U. Sama, K. Haseeb, and N. Islam, "Movie tags prediction and segmentation using deep learning," *IEEE Access*, vol. 8, pp. 6071–6086, 2020.

[16] G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz, "Movie genre classification with convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 259–266.

[17] V. Battu, V. Batchu, R. R. R. Gangula, M. M. K. R. Dakannagari, and R. Mamidi, "Predicting the genre and rating of a movie based on its synopsis," in *Proc. 32nd Pacific Asia Conf. Lang., Inf. Comput.*, Dec. 2018, pp. 1–11.

[18] N. Ejaz, U. A. Khan, M. Á. M. del Amor, and H. Sparenberg, "Deep learning based beat event detection in action movie franchises," in *Proc. 10th Int. Conf. Mach. Vis. (ICMV)*, vol. 10696, Apr. 2018, Art. no. 1069608.

[19] M. Ghosh, S. M. Obaidullah, K. C. Santosh, N. Das, and K. Roy, "Artistic multi-character script identification using iterative isotropic dilation algorithm," in *Proc. Int. Conf. Recent Trends Image Process. Pattern Recognit.*, vol. 1037. Singapore: Springer, Jul. 2019, pp. 49–62.

[20] M. Ghosh, H. Mukherjee, S. M. Obaidullah, K. C. Santosh, N. Das, and K. Roy, "Artistic multi-script identification at character level with extreme learning machine," *Proc. Comput. Sci.*, vol. 167, pp. 496–505, Jan. 2020.

[21] B. Shi, X. Bai, and C. Yao, "Script identification in the wild via discriminative convolutional neural network," *Pattern Recognit.*, vol. 52, pp. 448–458, Apr. 2016.

[22] J. Mei, L. Dai, B. Shi, and X. Bai, "Scene text script identification with convolutional recurrent neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 4053–4058.

[23] L. Gómez and D. Karatzas, "A fine-grained approach to scene text script identification," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Apr. 2016, pp. 192–197.

[24] A. K. Bhunia, A. Konwer, A. K. Bhunia, A. Bhowmick, P. P. Roy, and U. Pal, "Script identification in natural scene image and video frames using an attention based convolutional-LSTM network," *Pattern Recognit.*, vol. 85, pp. 172–184, Jan. 2019.

[25] F. Lei, X. Liu, Q. Dai, and B. W.-K. Ling, "Shallow convolutional neural network for image classification," *Social Netw. Appl. Sci.*, vol. 2, no. 1, pp. 1–8, Jan. 2020.

[26] M. Ma, Q.-F. Wang, S. Huang, S. Huang, Y. Goulermas, and K. Huang, "Residual attention-based multi-scale script identification in scene text images," *Neurocomputing*, vol. 421, pp. 222–233, Jan. 2021.

[27] A. Khalil, M. Jarrah, M. Al-Ayyoub, and Y. Jararweh, "Text detection and script identification in natural scene images using deep learning," *Comput. Electr. Eng.*, vol. 91, May 2021, Art. no. 107043.

[28] Q. Lin, C. Luo, L. Jin, and S. Lai, "STAN: A sequential transformation attention-based network for scene text recognition," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107692.

[29] H. Kanagaraj and V. Muneeswaran, "Image compression using Haar discrete wavelet transform," in *Proc. 5th Int. Conf. Devices, Circuits Syst. (ICDCS)*, Mar. 2020, pp. 271–274.

[30] F. Luisier, T. Blu, and M. Unser, "Image denoising in mixed Poisson–Gaussian noise," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 696–708, Mar. 2010.

[31] A. Mosavi, S. Ardabili, and A. R. Varkonyi-Koczy, "List of deep learning models," in *Proc. Int. Conf. Global Res. Educ.* Cham, Switzerland: Springer, Jan. 2019, pp. 202–214.

[32] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jul. 2020.

[33] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5551–5560.

[34] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 745–753.

[35] M. Ghosh, H. Mukherjee, S. M. Obaidullah, K. C. Santosh, N. Das, and K. Roy, "Identifying the presence of graphical texts in scene images using CNN," in *Proc. Int. Conf. Document Anal. Recognit. Workshops (ICDARW)*, vol. 1, Sep. 2019, pp. 86–91.

[36] E. Walach and L. Wolf, "Learning to count with cnn boosting," in *Proc. 14th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 660–676.

[37] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, Sep. 2010.

[38] A. Ahmadian and A. Mostafa, "An efficient texture classification algorithm using Gabor wavelet," in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 1, Apr. 2004, pp. 930–933.

[39] D. Fronitasari and D. Gunawan, "Palm vein recognition by using modified of local binary pattern (LBP) for extraction feature," in *Proc. 15th Int. Conf. Qual. Res. (QiR), Int. Symp. Electr. Comput. Eng.*, Jul. 2017, pp. 18–22.

[40] M. Ghosh, H. Mukherjee, S. M. Obaidullah, and K. Roy, "STDNet: A CNN-based approach to single-/mixed-script detection," *Innov. Syst. Softw. Eng.*, vol. 17, pp. 1–12, Apr. 2021.

[41] D. Denisko and M. M. Hoffman, "Classification and interaction in random forests," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 8, pp. 1690–1692, Feb. 2018.

[42] S. L. Ting, W. H. Ip, and A. H. Tsang, "Is Naive Bayes a good classifier for document classification," *Int. J. Softw. Eng. Appl.*, vol. 5, no. 3, pp. 37–46, Jul. 2011.

[43] S. Raghu and N. Sriraam, "Optimal configuration of multilayer perceptron neural network classifier for recognition of intracranial epileptic seizures," *Expert Syst. Appl.*, vol. 89, pp. 205–221, Dec. 2017.

[44] D. Giveki and M. Karami, "Scene classification using a new radial basis function classifier and integrated SIFT–LBP features," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1071–1084, Feb. 2020.

[45] 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

[46] X. Xu, Z. Zhang, Z. Wang, B. Price, Z. Wang, and H. Shi, "Rethinking text segmentation: A novel dataset and a text-specific refinement approach," 2020, *arXiv:2011.14021*. [Online]. Available: http://arxiv.org/abs/2011.14021

[47] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[49] 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.

[50] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[51] M. Ghosh, S. S. Roy, H. Mukherjee, S. M. Obaidullah, K. C. Santosh, and K. Roy, "Understanding movie poster: Transfer-deep learning approach for graphic-rich text recognition," *Vis. Comput.*, pp. 1–20, Mar. 2021, doi: 10.1007/s00371-021-02094-6.

[52] K. Dasgupta, S. Das, and U. Bhattacharya, "Scale-invariant multi-oriented text detection in wild scene image," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2041–2045.

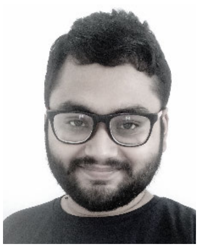
[53] D. Cao, J. Dang, and Y. Zhong, "Towards accurate scene text detection with bidirectional feature pyramid network," *Symmetry*, vol. 13, no. 3, p. 486, Mar. 2021.

[54] 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. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1582–1587.

[55] 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)*, vol. 1, Nov. 2017, pp. 1454–1459.

[56] J. Huang, G. Pang, R. Kovvuri, M. Toh, K. J Liang, P. Krishnan, X. Yin, and T. Hassner, "A multiplexed network for end-to-end, multilingual OCR," Mar. 2021, *arXiv:2103.15992*. [Online]. Available: http://arxiv.org/abs/2103.15992

MRIDUL GHOSH received the B.Tech. and M.Tech. degrees in computer science and engineering from the University of Calcutta. He is currently pursuing the Ph.D. degree with Aliah University, Kolkata, India. He was a Faculty Member with the Seacom Engineering College, and the Supreme Knowledge Foundation Group of Institutions, West Bengal, India. He is currently an Assistant Professor with Shyampur Siddheswari Mahavidyalaya, Howrah, India. He has published a book, a book chapter, and 16 research papers in journals and conferences. His research interests include image processing, pattern recognition, and machine intelligence.



SAYAN SAHA ROY received the B.Tech. degree in electronics and communication engineering from the University of Calcutta, India, in 2020. He is currently working as a Junior Research Fellow with IIT Kharagpur, West Bengal, India. His research interests include image processing, artificial intelligence, machine learning, and deep learning. He has published two conference papers and one article.

HIMADRI MUKHERJEE received the B.Sc. degree in computer science from APC College, in 2013, and the M.Sc. and Ph.D. degrees in computer science from West Bengal State University, in 2015 and 2020, respectively. He is currently working as a Postdoctoral Fellow with New York University Abu Dhabi. He has published more than 50 research papers in reputed conferences and journals. His research interests include audio signal processing, music processing, image processing, natural language processing, pattern recognition, and machine intelligence.

SK MD OBAIDULLAH (Member, IEEE) received the Ph.D. degree in engineering from Jadavpur University, Kolkata, India, in 2017. He worked as a Postdoctoral Fellow with the University of Evora, Portugal, under the Erasmus Fellowship program funded by European Commission, from November 2017 to September 2018. He is currently an Associate Professor and the Head of the Department of Computer Science and Engineering, Aliah University. He is also working as a TARE Fellow at Indian Statistical Institute, Kolkata, funded by DST SERB, Government of India, for three years from 2019 to 2022. He has published more than 100 research papers in reputed international/national conferences and renowned journals and one edited book with CRC Press. His research interests include machine learning, deep learning, medical image analysis, biometrics, document image analysis, and audio signal processing. He is a fellow of IETE, a Life Member of IUPRAI, and a certified Chartered Engineer (CE) of The Institute of Engineers (India). He is also serving as an Associate Editor for the *IET Image Processing Journal*.

XIAO-ZHI GAO received the B.Sc. and M.Sc. degrees from Harbin Institute of Technology, China, in 1993 and 1996, respectively, and the D.Sc. (Tech.) degree from Helsinki University of Technology (now Aalto University), Finland, in 1999. He has been working as a Professor with the University of Eastern Finland, Finland, since 2018. He has published more than 400 technical papers in refereed journals and international conferences. His current Google Scholar H-index is 34. His research interests include nature-inspired computing methods with their applications in optimization, data mining, machine learning, control, signal processing, and industrial electronics.

KAUSHIK ROY received the B.E. degree in computer science and engineering from NIT Silchar, in 1998, and the M.E. and Ph.D. degrees in computer science and engineering from Jadavpur University, in 2002 and 2008, respectively. He is currently working as a Professor and the Head of the Department of Computer Science, West Bengal State University, Kolkata, India. He has published more than 200 research papers/book chapters in reputed conferences and journals. His research interests include pattern recognition, document image processing, medical image analysis, online handwriting recognition, speech recognition, and audio signal processing. He is Life Member of IETE, IUPRAI (a unit of IAPR), and Computer Society of India. He received the Young IT Professional Award from Computer Society of India, in 2004.