I.J. Image, Graphics and Signal Processing, **2024, 1, 61-80**
Published Online on February 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2024.01.05

When Handcrafted Features Meet Deep Features: An Empirical Study on Component-Level Image Classification

Tauseef Khan*
School of Computer Science & Engineering, VIT-AP University, Amaravati, 522237, Andhra Pradesh, India
E-mail: tauseef.khan@vitap.ac.in
ORCID ID: https://orcid.org/0000-0002-3359-9967
*Corresponding Author

Ayatullah Faruk Mollah
Department of Computer Science and Engineering, Aliah University, Kolkata, 700160, West Bengal, India
Center of New Technologies, University of Warsaw, Poland
ORCID ID: https://orcid.org/0000-0002-3445-7469

Received: 09 July, 2023; Revised: 12 August, 2023; Accepted: 08 September, 2023; Published: 08 February, 2024

Abstract: Scene text detection from natural images has been a prime focus from last few decades. Classification of foreground object components is an essential task in many scene text detection approaches under uncontrollable environment. As it heavily relies upon robust and discriminating features, several features have been engineered for component-level text non-text classification. Competency of such feature descriptors particularly in respect of deep features needs to be examined. In this paper, we present prospective feature descriptors applicable to component-level text non-text classification and examine their performance along with convolutional neural network based deep features. Series of experiments have been carried out on publicly available benchmark dataset(s) of multi-script document-type, scene-type, and combined text vs. non-text components. Interestingly, feature combination is found to put well-demonstrated deep features into tough competition on most datasets under consideration. For instance, on the combined text non-text classification problem, CNN based deep features yield 97.6%, whereas aggregated features produce an accuracy of 98.4%. Similar findings are obtained on other experiments as well. Along with the quantitative figures, results have been analyzed and insightful discussion is made to ascertain the conjectures drawn herein. This study may cater the need of leveraging potentially strong handcrafted feature descriptors.

Index Terms: Object recognition, competent handcrafted features, deep features, text non-text classification, text detection.

1. Introduction

Object recognition from images or video frames is an important research problem in computer vision and image processing. It is largely driven by appropriate extraction of intrinsic properties present inside the object region. Foreground text non-text object classification is a challenging task in scene text detection which is a prerequisite to many text-based computer vision applications like content-based image retrieval, license plate recognition, traffic sign detection etc. [1-2]. It is well known that detection of texts from images under complex scenario is challenging due to improper illumination, arbitrary text appearance, multi-lingual texts, noise, and other clutters. Till date researchers have developed several text detection methods to deal with these accompanied challenges. Though, efforts are being put forth to develop end-to-end methods for text detection, many text detection methods [1-4, 10, 12-14, 41] apply a two-stage approach which is driven by foreground component extraction from input images followed by classification as text or non-text based on some intrinsic features. Thus, text non-text component classification stands as an essential and non-trivial task for text detection from natural images. Traditional approaches to text non-text classification heavily rely upon development of discriminating low level handcrafted features. Though, design of such features is challenging under uncontrollable environment and they are often sensitive to noise and other factors, some handcrafted features have evolved as prospective and high performing in many application domains. Some works are recently reported [20, 69] that have employed such features and obtained pretty high performance.

With the emergence of deep learning, state-of-the-art methods have achieved incredible success in scene text detection [55-60]. Unlike handcrafted features, deep learning approaches extract high-level intrinsic features that make them highly effective. Several Deep Neural Networks (DNN) have been designed so far, which comprise of a series of hidden layers which automatically compute high-level complex features from component images and classify them. Convolutional Neural Networks (CNN) [23-26] are among the most successful and popular deep networks in the field of text detection. LeNet [48], the first CNN architecture was introduced for handwritten digit classification. Later, several different CNN architectures have evolved [49-54].

Despite high performance of deep networks in general, highly intensive computing and very high resource requirement of many deep models may render them unsuitable in many practical scenarios. Many researchers, therefore, adhere to classical approaches and report performance comparable to that of deep learning models [69, 70]. A few studies have been made that report text non-text classification using intrinsic feature extraction [17, 27, 29]. Besides, some comprehensive surveys on traditional handcrafted features as well as deep features on different image processing and analysis tasks [83, 84] have been presented recently. Ma et al. [83] have reviewed some classical and automated deep learning based techniques for image matching. Bekhouche et al. [84] have reported a performance comparison between handcrafted features and deep features on age estimation from facial images.

In this context, robust effective as well as light-weight traditional feature driven frameworks, that may compete with the state-of-the-art, may be justifiably considered as valuable facets for visual object detection and classification tasks. It is worth mentioning that existing methods have applied handcrafted features in random manner rather than trying to build relationship between the given problem and fundamental characteristics of such features, which leads to substandard performance. Hence, selection of appropriate and discriminant feature descriptors is the key aspect behind the success of such classical frameworks. However, to the best of our knowledge, no studies have been made so far on visual object classification in the context of complex scene images to provide a categorical evaluation and insightful analysis on component-level feature descriptors. Identifying prospective feature descriptors and assessing their performance in comparison to deep networks is indeed necessary to design robust image classification model.

In this paper, we present a comparative study, assessment, and insightful analysis on different component-level handcrafted feature descriptors under multi-script scene imagery environment. We also report the performance of deep features obtained with CNN at various depths on the same datasets to realize where the handcrafted feature descriptors stand in comparison to the ongoing buzz of deep networks. In short, the contributions of this paper are indicated below.

- A categorical study on state-of-the-art handcrafted feature descriptors is presented along with pros and cons in the context of component-level text non-text classification. Inspired by the study, some prospective handcrafted feature descriptors are computed and their performance is reported with multiple classifiers on component-level public dataset(s).
- Category-wise comparative analysis on performance of reported feature descriptors is presented in an effort to reveal their strength and robustness. In order to demonstrate the effectiveness and acceptability of above feature descriptors in comparison with deep features, CNN models with increasing depth have been employed and their performance on the same dataset(s) is reported.
- Such studies have been carried out in parallel on complex document-type images, scene-type images, and their combination i.e., combined texts to conform the findings on multiple experiments with different sets of images. Finally, performance of handcrafted feature descriptors from each category and their aggregation is compared with CNN generated deep features at various depths on combined texts and based on the outcome, an insightful discussion is made.

The rest of the paper is organized as follows. Section 2 reports a careful review on different relevant works. In Section 3, categorical illustration and computation of prospective feature descriptors are presented. A series of rigorous experiments are steered out on publicly available dataset(s) and insightful comparative analysis on obtained results is reported in Section 4. Finally, conclusion is drawn in Section 5.

2. Related Works

As evident in literature, traditional approaches to scene text detection are often based on either sliding-window (SW) [2, 62-64] or connected component (CC) [1, 3-4, 13-14, 65]. In SW-based methods, multiple text-candidate blocks are extracted by sliding multi-scale windows over entire image. Then, a pre-trained classifier is applied for classification which is driven by handcrafted features extracted from candidate blocks and potential text-contained blocks are identified. Pan et al. [2] have developed a text detection framework where foreground components are segmented using a text confidence map generated using SW technique. Then, extracted components are labelled as text or non-text using conditional random field model comprised of some unary and binary handcrafted features. Lee et al. [62] have proposed a text detection framework where multi-scale candidate text blocks are classified using pre-trained AdaBoost classifier driven by some intrinsic spatial features. This kind of methods mainly suffers from high computational cost due to large number of text-blocks and difficulties in designing discriminative feature descriptors from coarsely detected image blocks.

In CC-based approach, foreground components are identified using appropriate segmentation algorithms (e.g. intensity homogeneity, colour clustering, etc.) and classified as text or non-text using manually designed features (geometric properties, texture, text-stroke, edge gradient, etc.). Maximally Stable Extremal Region (MSER) [28] and Stroke-width Transform (SWT) [14] are the most representative and widely used algorithms for potential text-candidate extraction. MSER generates extreme stable regions within an image for a wide-range of thresholds which is suitable for appropriate text-candidate extraction [19]. In SWT, feature map is generated by analyzing stroke-pixels with similar width. It is found that CC based approaches typically extract bulk amount of non-text regions accompanied with true texts due to inappropriate feature selection. Thus, elimination of false-positives using laborious post-processing plays a decisive role for accurate text localization.

Design of robust component-level classifier driven by discriminative features is imperative behind the success of both the SW and CC based approaches. Gigantic exertions have been given by researchers to design effectual handcrafted feature-descriptors that may capture discriminatory information from image objects. Geometric feature descriptors extract the characteristics related to the geometry of the foreground object components [1-4, 22]. Yao et al. [66] have designed two-levels of feature descriptors viz. (i) component-level geometric features (occupancy ratio, contour shape, density etc.) that distinguish character-level texts, (ii) chain-level features to identify true-text lines using line-orientation, colour similarity, size variation etc. Le et al. [67] have segmented text and non-text components using component-shape information, stroke-width, and other properties. Texture is considered to be a vital information for component-level text non-text classification. Most of the existing works reported several feature descriptors to analyze textural pattern of text instances such as Co-Occurrence Matrix (COM) [5], HOG [6, 30], Local Binary Pattern (LBP) [9], wavelet transform [39], Discrete Cosine Transform (DCT) [46], etc. Minetto et al. [7] have proposed an advanced texture-based HOG (T-HOG) descriptor for text line detection from complex images. Later, HOG has been applied using Co-Occurrence Matrix (CO-HOG) for multi-script character recognition [8]. Recently, fuzzy based LBP is introduced in the context of feature extraction [37]. In [11], text localization method is proposed based on Gray-level Co-occurrence matrix (GLCM) for component-level classification. Shivakumara et al. [12] have used 2D-Haar wavelets that decompose the image in different sub-bands and adopted different statistical features like energy, entropy, intensity-homogeneity etc. to extract the textural properties of texts.

Generally, it is observed that text-characters often have uniform stroke width which may meaningfully separate them from other foreground object components. Huang et al. [10] have developed an enhanced version of SWT called Stroke Feature Transform (SFT) incorporating colour information for candidate component extraction. Then, Text Covariance Descriptor (TCD) is applied which is driven by some statistical stroke features (stroke-pixel colour, stroke-width, distance etc.) for text component and line-level classification. Stroke value distribution profile is also generated for text component extraction from complex video frames [15, 16]. Khan et al. [32] have designed a set of stroke-feature descriptors from distance transform map for text non-text classification. Recently, Khan et al. [61] have designed an area occupancy distribution using equidistant pixels computed from text strokes that significantly distinguish foreground components. Usually, texts have strong and sharp edges with directional gradient magnitudes that can potentially distinguish text instances from background [33]. Yu et al. [18] have segmented candidate components using edge analysis and later adopted edge-based features for non-text filtering. Huang et al. [20] have generated edge-feature map which is insensitive to illumination effect and it effectively detects scene texts. Lu et al. [21] have designed three text-specific edge-features viz. gradient variance, directional edge-cut and even number of edge-cuts for text non-text separation in order to localize texts. Exploiting the directional flow of gradient arbitrary-oriented text detection is reported in [22], where text-candidate components are grouped using geometrical features (text-line orientation, size, direction etc.).

At the advent of deep neural networks, performance of image classification significantly improved due to their automatic high-level feature extraction ability from image objects [23, 24, 25, 26, 44, 45, 73, 74]. It may be stated that to some extent deep networks overcome some shortcomings of feature-driven approaches such as feature design in complex environments, tedious pre-processing, post-processing, etc. From the last few years, a number of works have been reported on text non-text classification from natural images using deep features. Bai et al. [34] have designed a multi-scale spatial partition network (MSP-Net) inspired by VGG-16 model [49] for image-level text non-text classification. MSP-Net comprises of five convolutional layers wherein the last three layers apply deconvolution to upsample the feature maps generated through convolution. Finally, concatenation of upscaled feature maps carry rich and semantic feature representation. Zhao et al. [35] have developed two CNNs for text non-text image classification, where a shallow CNN reduces the size of all features generated from each layer quite fast and the other deep CNN instructs the learning progress of shallow network to improve performance. However, these works [34, 35] detect whether text is present within the image or not, rather than identifying and classifying foreground components as text or non-text. Recently, Khan et al. [47] have reported a deep-CNN based method comprised of three convolution layers for multi-script component-level text/non-text classification.

Though, it is believed that deep networks outperform feature-driven methods in text non-text classification, adequate investigation is not yet made to check where the competent handcrafted feature descriptors stand in comparison to deep features. In related literature, it is noticed that some handcrafted feature descriptors have achieved significantly high performance for component-level image classification even in unconstrained environments. Moreover, computational overhead of deep networks discourages researchers with limited computing facilities, and in turn, encourages them to plunge into prospective feature driven pipelines. It is also noted that majority of the reported works

have tried to adopt feature descriptors by random assumption without sufficient analysis of the characteristics of target texts, which leads to compromised performance. On the contrary, careful adoption of handcrafted features conforming to text intrinsic properties may indeed improve the performance further. Hence, a curious mind may be keen to identify some potential handcrafted features for component-level classification, which may be competent to deep features as well. This paves the motivation of this work wherein we present an empirical study of some prospective feature descriptors and CNN generated deep features at various depths, evaluate them in the context of multi-script component-level text non-text classification in unconstrained scenarios, and finally based on the performance a comparative and insightful analysis has been presented.

3. Computation of Handcrafted Features

Image classification is generally driven by intrinsic features of object components. A discriminant feature set may lead to high accuracy for component-level classification or object classification. Features may be broadly divided into two categories viz. unary features and binary features. Unary features are extracted from single image component, whereas binary features are computed from a pair of objects based on their intrinsic information [1]. Let, $I \in \mathbb{R}^+$ be a digital image and $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$ be foreground components extracted from I using some segmentation methods. Then, component-level text non-text classification can be mathematically defined as $\mathcal{F}: \mathcal{C} \rightarrow \delta$, where $\delta \in \{\mathcal{C}_T, \mathcal{C}_{NT}\}$. Here, $\mathcal{C}_T$ refers to the foreground text component class and $\mathcal{C}_{NT}$ refers to the foreground non-text component class. So, the problem of component-level classification can be stated as: given a foreground component $\mathcal{C}_i$ to the function $\mathcal{F}$, the output element of $\mathcal{F}$ must belong to set δ. It may be noted that the said problem applies on pre-segmented foreground image components and the present work is primarily focused on component-level classification. Therefore, in order to have segmented components and compute feature descriptors from them, we convert colour images to grey scale using weighted average of RGB channels [31] and binarize with adaptive Otsu method followed by removal of tiny components considered as noise. Foreground components extracted from natural scene images are further labelled as text or non-text. Component-level feature descriptors, as evident in related literature, may be broadly categorized on the basis of their characteristics into five groups viz. (a) geometric features, (b) texture-based features, (c) stroke-based features, (d) gradient-based features, and (e) deep features.

3.1 Geometric Features

Geometry of text components largely differ from that of non-text components. Text components are generally aligned horizontally or vertically within an image, causing nearly uniform geometric shape, whereas alignment of non-text components may be arbitrary. Few popular geometrical feature descriptors are discussed below.

Normalized Aspect Ratio. This descriptor may contribute in filtering out too large or too small foreground components. It is normally presumed that text components have well-structured and nearly uniform shape. So, components having arbitrary shapes may be treated as non-text. It is defined as the ratio of height to width if height is less than width or else the vice-versa as in Eq. 1.

$$Normalized\ Aspect\ Ratio\ (r) = \frac{min(h,w)}{max(h,w)} \tag{1}$$

Occupancy Ratio. As expressed in Eq. 2, this ratio signifies the density of foreground components. It is defined as the ratio of number of pixels in the component to the area of its bounding box. Components with very small or large number of pixels within its bounding box may characterize non-text components. Fig. 1 shows the steps of computation of occupancy ratio and compactness.

$$Occupancy\ Ratio\ = \frac{nop(\mathcal{C})}{area(BB(\mathcal{C}))} \tag{2}$$

where nop() returns the number of pixels, area() gives the area of the object passed and BB() denotes the bounding box of component $\mathcal{C}$.

Compactness. Compactness is defined as the ratio of a component's perimeter to its bounding box area as shown in Eq. 3. The perimeter defines the number of pixels in the boundary of the component C. As text components are usually transient, compactness is often different in case of text from that of non-text. So, it may be used to characterize component objects. In order to compute compactness of components, canny edge-detection algorithm is applied that generates one pixel wide fine-scale edge-map.

$$Compactness\ = \frac{perimeter(\mathcal{C})}{area(BB(\mathcal{C}))} \tag{3}$$

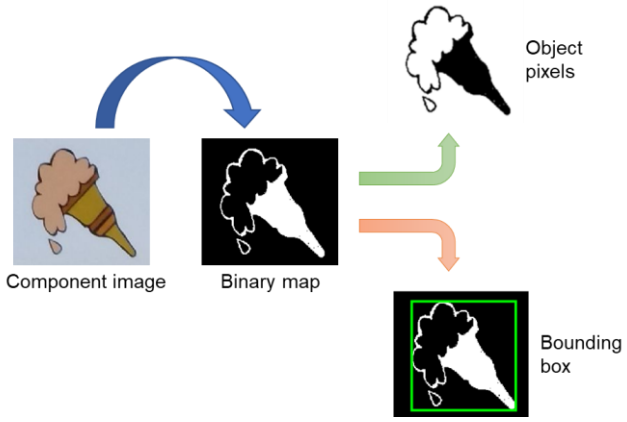

Fig. 1. Computation of occupancy ratio and compactness of object component.

Convexity. It denotes the perimeter of the convex hull (CH) to the original contour of the component image as shown in Eq. 4. Convex hull is the minimum convex polygon which encloses the entire component region. Fig. 2 depicts the comparison between convex hull of a text component and minimum bounding rectangle of the same.

$$Convexity = \frac{perimeter(CH(\mathcal{C}))}{perimeter(\mathcal{C})} \tag{4}$$

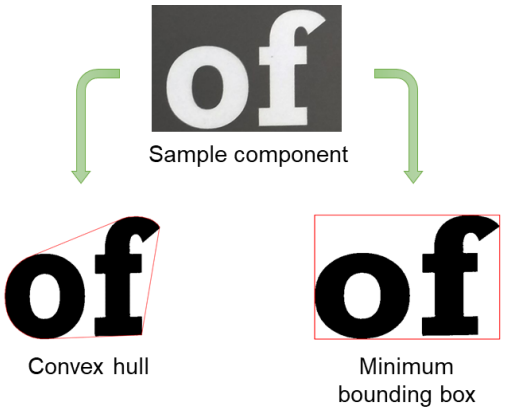

Fig. 2. Convex hull and minimum bounding rectangle of a sample component.

Solidity. This feature descriptor may help pruning out non-text components having too small or too large size. It is measured as the ratio of component's convex hull area to its bounding box area. Eq. 5 shows the mathematical expression of solidity. Solidity is discriminative in nature and may, in turn, produce higher classification accuracy.

$$Solidity = \frac{area(CH(\mathcal{C}))}{area(BB(\mathcal{C}))} \tag{5}$$

Elongation. It is the ratio of component minor-axis to its major-axis. Major-axis is the longest axis of the foreground component which passes through both the foci and minor-axis is the shortest axis that is perpendicular to the major-axis at a point equally distant from foci. Eq. 6 gives the mathematical expression of elongation of components. Text components normally have quadrangular shape whereas non-text components have arbitrary shapes. Elongation bears the shape information and hence contributes in component classification.

$$Elongation = \frac{minoraxis(\mathcal{C})}{majoraxis(\mathcal{C})} \tag{6}$$

Euler Number. Euler number is an important topological property to characterize image objects. If, in an object, the number of connected-components is n_{CC} and the number of holes present therein is n_{Hole}, Euler number is given as in Eq. 7. In this work, 8-neighborhood is followed for obtaining connected components.

$$Euler\ number = n_{CC} - n_{Hole} \tag{7}$$

Eccentricity. It measures the circularity of foreground components. It denotes the ratio of the distance between foci (c) and the length of the major axis (a) of the component. It may be realized from Eq. 8 that the value of eccentricity lies between [0.0,1.0]. A component having 0 eccentricity is a circle, whereas eccentricity tends to 1 for linearly elongated components.

$$\text{Eccentricity} = \frac{c}{a} \tag{8}$$

Pros and Cons. Geometrical features generally assume that texts may appear with certain geometry. This kind of feature descriptors may be effective and robust for texts with uniform-shape with near-homogenous background. However, designing discriminative geometric features may be challenging for (i) arbitrary-oriented texts, (ii) texts with non-uniform font-size, (iii) text appeared in complex background, and (iv) non-texts that resemble with texts.

3.2 Texture based Features

Texture is one of the crucial properties of objects and it is often applied in classifying text or non-texts. Texture represents the pattern of pixels orientation. In general, it is found that texts carry merely homogenous textural pattern compared to non-texts. The fundamental unit of repeating pattern of pixels is called texel. In this study, some important texture-based feature descriptors are presented below.

3.2.1 Histogram of Oriented Gradients

This feature descriptor calculates the distribution of gradient direction (orientation) of each pixel in the component. Yan et al. [40] have used HOG features to train a classifier for multi-script text detection from complex background. It is observed that orientation of gradient across strokes of each text-character is either horizontal or vertical, whereas for non-text components orientation of gradients is not specific. Fig. 3 represents the possibility of characterization of component images with extracted HOG descriptors through their distributions. It may be observed that text component generates a distribution where high gradient magnitudes are detected across 0 °, 90 ° and 180 ° approximately as change of intensity takes place either vertically or horizontally across text strokes. But non-text components have arbitrary gradient orientations due to non-uniform shape. In order to generate fixed length feature vector, input images are normalized using zero padding bits. In the current work, length of HOG feature vector and orientation bins are empirically chosen. Increasing the number of bins generates more detailed textural information. However, length of feature vector also increases with bin numbers that may incur additional computational cost.

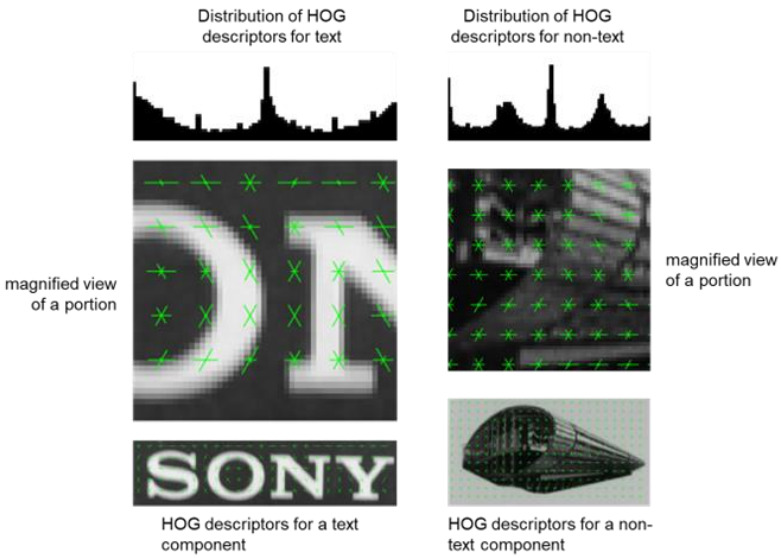

Fig. 3. Characterization of component images with HOG descriptors (Evident difference in the distributions may be noted).

3.2.2 Local Binary Pattern

Local binary pattern is another feature descriptor that analyses the local textural property of components [9, 36, 38, 77, 83]. LBP divides the whole component image into M × N blocks and computes binary threshold values across its center pixel. Fig. 4 shows the step-by-step process for computing LBP feature descriptor with fixed size window. Expression for LBP value of a given window is shown in Eq. 9 [83].

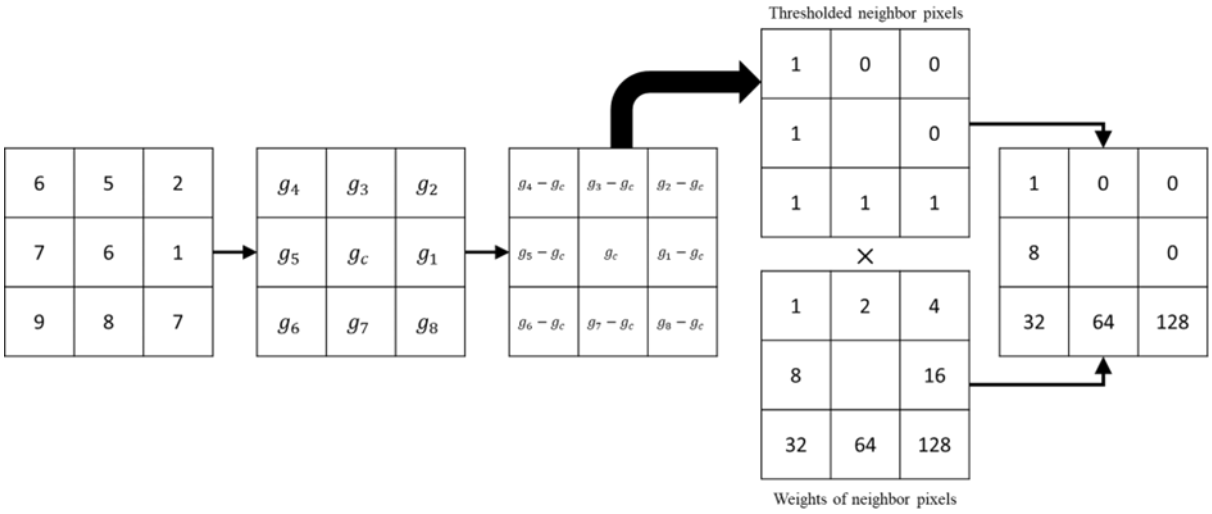

Fig. 4. Step-wise computation of LBP features. (a) 3×3 image window (b) circular 8 neighbor pixels of window, (c) subtraction method for achieving gray-scale invariance, (d) obtained thresholded neighbor pixels are multiplied with corresponding weights, (e) resultant values.

$$S(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad where\ S(x) = S(g_i - g_c)$$

The LBP value of the given window is measured by,

$$LBP_8 = \sum_{i=1}^{8} S(g_i - g_c) \times 2^{i-1} \tag{9}$$

where g_i is the i^{th} neighbor pixel and g_c is the centre pixel of the window. The obtained neighbor pixels after thresholding are multiplied by weights given to the corresponding pixels (see Fig. 4(d)). Finally, cumulative summation is performed on the values of all 8 neighbors (see Fig. 4 (e)) to obtain final LBP value of the window. In this study, the size of the LBP feature vector is set to 59. To compute LBP values of each center pixel an 8-neighbor kernel is chosen with radius 1. A number of variants of conventional LBP descriptor have been designed by different researchers for different problem domains. An outline of some variants of LBP is reported in Table 1.

Table 1. An outline of different LBP variants of texture based classification

Variants	Histogram size for 8-neighbor	Color information	Noise invariant	Illumination invariant	Rotation invariant
LBP [9, 82]	59	No	No	No	No
Center symmetric LBP (CS-LBP) [75]	16	No	No	No	No
Rotation invariant LBP (RILBP) [76]	256	No	Yes	Yes	Yes
Uniform RLBP (uRLBP) [77]	59	No	Yes	Yes	Yes
Modified LBP (MLBP) [78]	256	Yes	Yes	Yes	Yes
Median Robust Extended LBP (MRELBP) [76]	256	No	Yes	Yes	Yes
Intensity LBP (iLBP) [79]	256	Yes	No	Yes	No
Local Ternary Pattern (LTP) [80]	256	No	Yes	Yes	No
Local Directional Pattern (LDP) [81]	56	Yes	Yes	Yes	Yes

3.2.3 Discrete Cosine Transform

DCT reveals some levels of the textural property of objects. It decomposes the image into multiple sub-bands according to frequency level of pixels. DCT coefficients from each sub-bands carry frequency information of the images. Interestingly, few coefficients of this transform can reproduce the original image quite well, and may, therefore, be used as feature descriptors as well. In order to generate fixed number of DCT coefficients component images are normalized to a fixed dimension of 50×150 pixels using zero padding technique. We consider the first 25 coefficients as DCT descriptor of the component image.

3.2.4 Gray-level Co-occurrence Matrix

GLCM analyses the textural pattern of foreground components by calculating how often pairs of pixels with specific intensity appear in the image that demonstrates the spatial relationship between adjacent pixels. Few statistical

feature descriptors are designed from GLCM that may distinguish textural pattern between text and non-texts as outlined in Table 2 [71].

Table 2. Summary of different statistical feature descriptors derived from GLCM

Feature Descriptor	Mathematical Expression
Contrast. It measures the intensity difference between target pixel and its neighbors through entire image. It yields 0 value for image with true textual-homogeneity. Let, N denotes the number of gray-levels of image component and $P_{i,j}$ denotes the normalized gray-value with coordinates i, j of GLCM.	$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$
Correlation. It measures the correlation among target pixel and its neighbors over the entire image. Here, μ and σ are the mean and standard deviation of all the pixels contributed in GLCM.	$\sum_{i,j=0}^{N-1} P_{i,j}\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2\sigma_j^2}}$
Energy. It signifies the uniformity of textural pattern of component image. The value yields sum of all squared pixels of GLCM. Energy is obtained as the square root of an angular second moment.	$\sqrt{\sum_{i,j=0}^{N-1}(P_{ij})^2}$
Entropy. This descriptor signifies the randomness of texture patterns within image.	$\sum_{i,j=0}^{N-1} -ln(P_{ij})P_{ij}$
Homogeneity. This descriptor measures the homogeneity of pixels intensity within GLCM. Measurement of closeness among pixels distribution is defined by homogeneity. Homogeneity is more sensitive to the presence of near diagonal elements in the GLCM. The value of homogeneity is maximum when elements in the image are the same.	$\sum_{i,j=0}^{N-1}\frac{P_{ij}}{1+(i-j)^2}$

Pros and Cons. Texture information appears to be effective feature descriptor for component-level text non-text classification due to textural-homogeneity of text components unlike non-texts. Texture-based feature descriptors perform comparatively well under noisy, marginally complex-background and other clutters. However, texture classification incurs high-computational cost due to pixel-wise scanning. Moreover, it often struggles to perform well in case of multi-colored texts and rotated texts.

3.3 Stroke-based Features

Stroke is one of the important features to recognize characters in each text lines. Text components contain set of characters having distinct and near-uniform strokes whereas for non-text components stroke width may be non-uniform and arbitrary. In this section, few recent stroke-based feature descriptors have been computed for text non-text classification reported in [47, 61].

Stroke Distribution Profile. Khan et al. [47] have designed a stroke-based feature descriptor using distance values of medial skeleton maps of component images. Generation of medial skeleton of component images is an improved version of skeletonization technique. A novel burning rope algorithm is designed to generate medial skeleton map by removing spurious branches from skeleton maps. In this descriptor, a distribution profile of distance values of potential stroke pixels of text and non-texts has been generated and then analyzed for component classification. It is observed that for text components distance values of stroke-pixels are dispersed in normal distribution manner with one peak value due to their uniform nature of stroke-width. However, for non-text components, distance values are dispersed in a wide range due to non-uniformity of stroke-width, resulting arbitrary-nature of distribution profile. In this study, 16 feature descriptors are chosen for evaluation.

Area Occupancy Profile. Khan et al. [61] have generated another stroke-based feature descriptors for component level classification. The descriptors have been computed from area occupancy profile of pixels carrying equal distance values of distance transform map of component images. Initially, distance map is generated from input image using distance transform method [72]. Rectangular bounding box area is computed from four farthest pixels containing the highest distance value from distance map and then divided by component's bounding box to obtain the area occupancy ratio. In feature profile of texts, a steep increment of area occupancy ratio is often observed after few hops and then gradually it gets raised up to 1. Whereas for non-texts the nature of profile is arbitrary due to its inherent irregularity. In order to generate fixed length feature descriptors, area occupancy profile is partitioned by binning method.

Pros and Cons. Stroke being an important property, stroke-based feature descriptors significantly contribute in discriminanting object components. Generally, texts appear in near-uniform stroke width and maintain its regularity across the component. Stroke-based features are less sensitive to noise and arbitrarily oriented texts under complex enviroment. However, stroke measurement itself turns to be challenging in certain situations like (i) artistic texts having non-uniform stroke, and (ii) texts closely blend with background.

3.4 Gradient based Features

Gradient is a powerful feature descriptor to discriminate text and non-text components. Normally text regions have high gradient magnitude across edges whereas for non-text components gradient magnitudes across edges are arbitrary due to their complex structures. Some of the statistical measures based on edge information may be utilized to evaluate the goodness of text components [2, 42].

3.4.1 Statistical Gradient Descriptors

Horizontal Edge Score. It measures the edge-score of the entire image component in horizontal direction. In order to compute edge-score, Sobel filter [43] is convolved over entire image in horizontal direction and computes the cumulative edge-score. It is found that, edge-score is high in horizontal direction for text components as text generally have linear orientation.

Vertical Edge Score. It measures the edge-score of entire image component in vertical direction. Here, Sobel vertical filter is convolved over image pixels and computes and computes cumulative edge score. It has been found that vertical edge score is also high for text components due to linear orientation.

3.4.2 Corner-point based Gradient Descriptors

Corner point is an intersection point of two dominant edges from different directions. In general, due to linear-orientation of text instances, direction of corners points is either vertical or horizontal. Moreover, density of corner-points is high in word-level text instances due to linear-alignment of character-components. Some popular corner-point detectors are discussed below.

FAST. Features from Accelerated Segment Test (FAST) is an effective corner-point detector from gray-scale images. FAST identifies a target pixel as corner-point if intensity values of all its neighbor pixels are either higher or lower considering a threshold value. Fig. 5 illustrates the corner-point detection of text and non-text components using FAST algorithm. Some statistical measures are obtained from corner-points of foreground components.

(a) (b)

Fig.5. Corner-point detection from component images. (a) Text and non-text components respectively (top-down), (b) corresponding corner-point detection using FAST algorithm (corner-points marked in green colour).

Linearity. It measures the orientation of corner-points of foreground components. In general, most of the text-level corner-points are aligned in horizontal direction. However, corner-points of non-texts are arbitrarily-oriented. It is measured by variance of distance values of all corner-points across horizontal-direction of a component image. Let $p(x,y)$ denote a detected corner-point in horizontal direction and d_p is the corresponding distance value of pixel p. Then, linearity may be expressed as shown in Eq. 10.

$$Linearity = \frac{\sum_{i=1}^{N}(d_{p_i} - \overline{d_p})^2)}{N} \tag{10}$$

where, N is the number of corner points across horizontal direction of a component image and $\overline{d_p}$ is the mean of distance values of all corner points across horizontal direction.

Strength. It measures the strength of detected corner-points. Normally, texts have high gradients across sharp edges, which yields relatively stronger corner-points unlike non-text components. This feature computes the mean strength of detected corner-points using determinant of hessian approximation.

Pros and Cons. Combination of edge and gradient features may perform well for text and non-text separation from high-quality images like well-scanned document, born-digital and other computer-generated images. Moreover, edge-features are less sensitive to illumination effect. However, this feature suffers from detecting appropriate edges for complex scene images. Detection of corner points is also difficult for texts closely mingled with non-texts and other backgrounds.

3.5 Deep Features

Deep neural networks extract high level deep features from input images using convolutional approach. Several deep neural networks have been designed till now like Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) [53], Deep Belief Network (DBN) [54] etc. CNN models contain a series of convolution layers followed by activation and sub-sampling/pooling layers that generate fixed-sized feature maps. Let, $I_{M\times N}$ be a 2-D digital image and $K_{P\times Q}$ denote a kernel. Then, the produced feature map can be articulated as $F_{R\times S} = I_{M\times N} * K_{P\times Q}$. Small kernels are usually convolved across the image, responses are passed through activation, and pooling is applied to obtain the corresponding feature maps. In convolution layer, different combinations of filters are adopted to extract deep and high-level features like horizontal and vertical edges, corner points, Gaussian blur, noise, etc. Pooling layers further reduce the dimensions of feature maps generated from convolution layers while retaining its utmost vital spatial information. In the flattening layer, two dimensional feature maps are converted to one dimensional feature vector that are subsequently fed to fully connected layer where outputs of previous layer neurons are treated as inputs to the neurons of the next layer. Finally, softmax activation function provides a vector comprised of probabilities of each output class. Overall architecture of the employed CNN is depicted in Fig. 6.

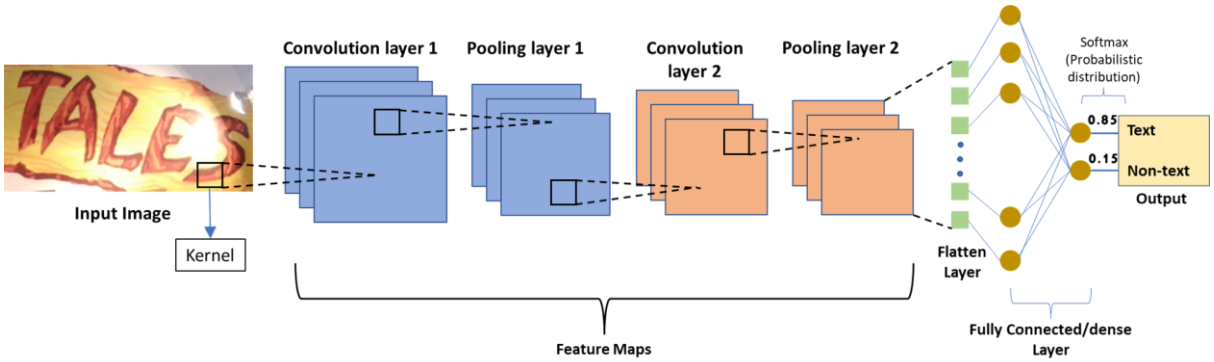

Fig. 6. Overall architecture of the employed Convolutional Neural Network.

At first, component-level training images are normalized to a fixed dimension of 1:2 ratio (height: width) using zero padding bits in order to train the employed CNN model. To validate the training model, 10% training samples are used. In the employed CNN, Adam optimizer is used as an optimization algorithm. Adam optimizer is an updated version of stochastic gradient decent algorithm that updates learnable-weights in more adaptive way based on training samples. Number of iterations for training is decided experimentally until the learning curve converges at a point.

Table 3. List of different parameters and associated values of the employed CNN model

Hyper-parameter	Values
Optimizer	Adam
No. of iteration	Experimentally decided (70)
Learning rate	0.001
Batch size	250
Decay rate 1st and 2nd momentum	0.9 & 0.99 respectively
Loss function	Categorical crossentropy
Padding mechanism	Same

Table 3 shows the list of different parameters and its associated values used in the CNN. Experimental environment is equipped with Intel i5 processor, 1 TB HDD, 16 GB RAM and NVIDIA GeForce GTX-1060 (6 GB) graphics memory. In the model with three conv-pool layers, the first two applied 32 and the third layer applied 16 learnable filters with dimension of 5×5 pixels. Table 4 illustrates layer-wise model configuration with trainable parameters using 1, 2 and 3 number of convolution-layers.

Pros and Cons. Deep neural networks have the ability to extract deep semantic features that are invariant to scale, text-size, illumination effect, complex backgrounds, text-orientation and other obstacles. Automatic extraction of deep features without a priori knowledge of image nature is convenient for scene text detection. However, few downsides are also present for deep features such as: (i) high-computational environment is a pre-requisite, (ii) it requires large dataset for training, and (iii) image rescaling is required for optimum learning.

Table 4. Layer-wise model configuration with trainable parameters. Model configuration is illustrated with increasing number of convolution-layers (from left to right).

Layer	Output shape	#para	Layer	Output shape	#para	Layer	Output shape	#para
Conv-1	(50, 100, 32)	832	Conv-1	(50, 100, 32)	832	Conv-1	(50, 100, 32)	832
Activation-1	(50, 100, 32)	0	Activation-1	(50, 100, 32)	0	Activation-1	(50, 100, 32)	0
Max-pool-1	(25,50, 32)	0	Max-pool-1	(25,50, 32)	0	Max-pool-1	(25,50, 32)	0
Flatten	40,000	0	Conv-2	(21,46,32)	25632	Conv-2	(21,46,32)	25632
Dropout	40,000	0	Activation-2	(21,46,32)	0	Activation-2	(21,46,32)	0
Dense-1	128	5120128	Max-pool-2	(10,23,32)	0	Max-pool-2	(10,23,32)	0
Activation	128	0	Flatten	7360	0	Conv-3	(6,19, 16)	12816
Dense-2	50	6450	Dropout	7360	0	Activation-3	(6,19,16)	0
Activation	50	0	Dense-1	128	942208	Max-pool-3	(3,9,16)	0
Output-layer	2	102	Activation	128	0	Flatten	432	0
Softmax	2	0	Dense-2	50	6450	Dropout	432	0
-	-	-	Activation	50	0	Dense-1	128	55424
-	-	-	Output-layer	2	102	Activation	128	0
-	-	-	Softmax	2	0	Dense-2	50	6450
-	-	-	-	-	-	Activation	50	0
-	-	-	-	-	-	Output-layer	2	102
-	-	-	-	-	-	Softmax	2	0

4. Experimental Results and Analysis

To assess the performance, a series of rigorous experiments have been conducted on category-wise handcrafted features, CNN-generated deep features, and finally aggregated features (combination of all categories of handcrafted feature descriptors) for text non-text classification on three subsets of AUTNT dataset [68] viz. AUTNT-document, AUTNT-scene, AUTNT-combined (document & scene), and obtained results are reported accordingly in this section. Computed handcrafted features are categorized into four major groups viz. geometrical, texture, stroke, and gradient based features as outlined in Table 5. Four leading pattern classifiers are adopted to evaluate these feature descriptors. Moreover, all categories of handcrafted features are combined and evaluated by different classifiers. On the other hand, to demonstrate the strength of deep features, increasing number of convolution-layers is appended in the employed CNN model. It is worth mentioning that pattern classifiers and CNN are trained using pre-partitioned training images and evaluated on test images as per the dataset.

Obtained performance is measured using standard evaluation metrics viz. Accuracy (Acc), Precision (P), Recall (R), F-Measure (F-M) and classification error. Separate P and R are measured for each class and at the end weighted average is reported as the result, whereas F-M is simply the harmonic mean of obtained P and R. During training of pattern classifiers, text and non-text components of the training sets are labelled as 0 and 1 respectively. Out of the training samples, 20% of data is used for validation to generalize the trained classifiers. Trained classifiers are evaluated on images of test set for performance assessment.

Table 5. Brief outline of category-wise feature descriptors distribution of aggregated features

Category of feature descriptors		No. of feature descriptors
Geometrical feature descriptors		5
Texture-based feature descriptors	HOG	16
	LBP	59
	DCT coefficients	25
	GLCM	5
Stroke-based feature descriptors	Stroke distribution profile of medial skeleton [47]	8
	Area occupancy ratio of distance transform map [61]	8
Gradient-based feature descriptors		5
Total feature descriptors		130

4.1 Dataset Outline and Classifiers

AUTNT [68] is a multi-purpose dataset comprised of component-level text and non-text images of multiple scripts and image types. As the focus of the present work is component classification, samples of this dataset have been organized into three folds – (i) document-type texts, (ii) scene-type texts, and (iii) non-texts. Text images consist of three different scripts viz. Latin, Devanagari, and Bangla for both document and scene type images. Samples of all folds come with pre-partitioned training and test set. All the images are acquired with high-resolution built-in cell-phone camera under diverse sources and conditions to ensure unconstrained working environment and pertinence in real-world states. A detailed outline of AUTNT dataset along with image particulars is reported in Table 6. Some text images of three different scripts and non-text images of AUTNT dataset are shown in Fig. 7.

Table 6. A brief outline of AUTNT dataset and its image particulars

Component image type	Category	Script	Training	Test	Image source	Text orientation
Text	Document	Latin	1258	314	Newspaper, book cover, files, book pages	Horizontal, angular
		Bengali	1002	251	Newspaper, book page	Horizontal, angular
		Devanagari	1004	250	Newspaper, book page	Horizontal
	Scene	Latin	1759	439	Street-view image, house-hold commodity image, sign-board	Horizontal, angular, curved
		Bengali	1011	251	Street view image, house-hold commodity image, wall poster, banner, sign-board	Horizontal, angular curved
		Devanagari	280	71	Street view image, traffic sign-board	Horizontal, angular curved
Non-text	Document & scene	-	2305	576	Newspaper, book cover, street-view image, commodity image, etc.	Arbitrarily-oriented

4.1.1 Relevance of AUTNT

To ensure the relevance of AUTNT dataset in representation of real-world scenarios, miscellaneous cases such as image acquisition, hardwares and environments, nature of image objects, illumination effects, etc. are discussed. Moreover, inclusion of multi-script texts certainly augments more complexity in the dataset. Some of the salient features of the dataset are mentioned here.

- *Uncontrollable Environments* - Images of AUTNT have been captured in uncontrollable indoor-outdoor scene environments such as improper lighting effects, complex backgrounds, perspective distortion, and other environments clutters, that are pertinent to real-world scenarios.
- *Multi-resolution Images* - Images of AUTNT are acquired with mobile cameras of different resolutions such as 5, 8, and 13 mega pixels. Besides, images are taken from different angles and distances. Camera devices closer to the image object produce high-resolution images and far from image object generates low-resolution images.
- *Multi-script Images* - AUTNT dataset comprises of images of three popular scripts specially in Indian context, which represents the real-world scenarios of Indian subcontinents.
- *Flexible Image Acquisition*- Images are captured with handheld digital camera built-in cellphone, which is more convenient in daily-life due to high availability and easy usage.
- *Multi-level Image Annotation* –Images of AUTNT are annotated in three different levels to make them applicable in different research problems i.e., text non-text classification, script identification, and character recognition.

Fig. 7. Sample text and non-text component-level images of different scripts from AUTNT dataset [68]. (a) Document-level text components (viz. row-wise Latin, Bangla, and Devanagari scripted text components), (b) scene-level texts (viz. row-wise Latin, Bangla, and Devanagari scripted text components) and (c) non-text components.

4.1.2 Classification Models

Four different pattern classifiers are employed in this study viz. Multi-layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Adaptive Boosting (AdaBoost). A concise implementation details of different classifiers and their suitability are given below.

MLP. It is one of the widely recognized classification algorithms for image classification. Model learns through backpropagation mechanism and number of neurons in hidden-layer is (length of feature vector + output class size)/2. Number of epochs and learning-rate are empirically set to 500 and 0.5 respectively to achieve the optimal training accuracy. MLP is suitable for complex non-linear problems, handling large number of data samples, faster prediction, etc. However, algorithm is equally effective for small number of data samples.

SVM. This algorithm handles high-dimensional data, which is suitable for image classification. SVM adopts kernel-trick method to transform input data samples into higher dimension. The regularization and tolerance parameters are fixed to 1 and 0.001 respectively. SVM is less prone to overfitting problem and memory efficient as well.

RF. This algorithm fits multiple decision-tree classifiers on training samples and computes the average readings to improve accuracy and overcome the over-fitting problem. RF is capable to deal with noisy images and provides more stable results through majority-based voting technique. RF uses bagging mechanism for training. Here, 100% data is filled in each bag and number of decision tree is set to 100 for training purpose.

AdaBoost. It adopts ensemble machine learning method which primarily converts a bunch of low-accuracy classifiers into a high-accuracy classifier by sign-magnitude flipping mechanism. In AdaBoost decision stamp tree is applied as base classifier and epoch number and weight-threshold is set to 30 and 100 respectively during training process. This algorithm improves classification accuracy and reduces overfitting problem by using weighted combination of multiple weak classifiers. AdaBoost is comparatively fast and efficient than other ensemble approaches.

4.2 Document-type Text Non-text Classification

All four categories of feature descriptors, deep features, and aggregated features are evaluated on document-type texts containing all three scripts and non-texts. In order to assess the performance of deep features, employed CNN model is evaluated with increasing number of convolution-layers starting from one upto three. Performance of different experiments are reported in Table 7. Commenting on individual performance of each category of feature descriptors, it is observed that highest classification accuracies of geometrical, texture and stroke-based feature descriptors are 92.1%, 94.4% and 94.6% respectively for RF classifier. Gradient-based feature descriptors have obtained 92.5% accuracy using AdaBoost classifier. Other pattern classifiers have also performed well and achieved more than 90% accuracy in most of the cases. On the other side, using CNN-generated deep features classification accuracy is reached to 98.56% for three convolution-layers. It is observed that performance of deep features gradually increases with increasing number of convolution-layers which is obvious as model extracts higher number of deep and semantic features from input images.

Table 7. Performance of different category of handcrafted feature descriptors and CNN-generated deep features on document-type texts and non-texts using standard evaluation metrics.

Category	Feature descriptors	Classifiers	Accuracy	P	R	F-M	Error
Geometrical	Aspect ratio, Occupancy ratio, compactness, solidity, eccentricity, Elongation	MLP	0.907	0.897	0.906	0.901	0.093
		SVM	0.901	0.906	0.901	0.904	0.099
		RF	**0.921**	0.924	0.920	0.922	0.079
		AdaBoost	0.919	0.926	0.918	0.922	0.081
Texture-based	HOG, LBP, DCT-coefficients, GLCM (statistical measurements)	MLP	0.935	0.937	0.920	0.928	0.065
		SVM	0.902	0.905	0.903	0.903	0.098
		RF	**0.944**	0.949	0.937	0.943	0.056
		AdaBoost	0.933	0.934	0.931	0.932	0.067
Stroke-based	Stroke Distribution Profile, Area Occupancy Profile	MLP	0.899	0.899	0.898	0.898	0.101
		SVM	0.883	0.887	0.884	0.885	0.116
		RF	**0.946**	0.947	0.947	0.947	0.054
		AdaBoost	0.923	0.924	0.922	0.922	0.076
Gradient-based	Vertical-edge score, horizontal-edge score, Linearity of corner-points, strength of corner-points	MLP	0.923	0.931	0.927	0.929	0.077
		SVM	0.917	0.924	0.919	0.921	0.083
		RF	0.910	0.913	0.911	0.912	0.090
		AdaBoost	**0.925**	0.932	0.926	0.929	0.075
Aggregated features	Geometrical + Texture + Stroke + Gradient features	MLP	0.986	0.984	0.987	0.985	0.013
		SVM	0.985	0.986	0.984	0.985	0.014
		RF	**0.990**	0.988	0.989	0.989	0.010
		AdaBoost	0.989	0.987	0.989	0.988	0.011
Deep features	One convolution layer	CNN	0.956	0.952	0.961	0.955	0.043
	Two convolution layers		0.984	0.983	0.985	0.984	0.015
	Three convolution layers		**0.985**	0.984	0.986	0.985	0.014

On the contrary to the above results, classification accuracies of aggregated features (combination of four categories of features) are significantly high compared to each category of handcrafted features (by 5-7%). Classification accuracy has reached upto 99.0% for aggregated features with the RF classifier. Surprisingly, it is observed that performance of aggregated handcrafted features has surpassed deep features (98.56%), which indeed demonstrates the exceptional classification ability of reported features descriptors for different types of text-images.

4.3 Scene-type Text Non-text Classification

Similar experiments have been carried out for scene-type text component images comprised of all three scripts and non-text component images. Obtained results of different experiments are shown in Table 8. Performance on scene images demonstrates the prowess and effectiveness of handcrafted feature descriptors in practical scenario. Highest classification accuracy obtained with the geometrical and texture-based feature descriptors is 86.6% and 93.7% respectively. Besides, stroke and gradient feature descriptor have produced 93.33% and 90.6% classification accuracy using AdaBoost and MLP classifier respectively. CNN-generated deep feature has obtained highest classification accuracy of 95.96% with three convolution-layers. Like document-type texts, performance of deep features for scene-type images has also progressively improved with increasing number of convolution-layers.

On the other side, classification accuracy of aggregated features is remarkably improved compared to individual performance of different categories of features by around 3-10%. Classification accuracy of aggregated features is more than 96% for all classifiers with highest reading of 96.5% using AdaBoost classifier. Besides, obtained results have exceeded deep features in terms of classification accuracy, which certainly establishes the robustness and effectiveness of reported feature descriptors in unconstrained scenario. However, it may be observed that overall performance of all feature descriptors is slightly low for scene-type texts compared to document-type texts which is obvious due to complex nature of scene-type images.

Table 8. Performance of different categories of handcrafted features and deep features on scene-type text and non-text images.

Category	Feature descriptors	Classifiers	Accuracy	P	R	F-M	Error
Geometrical	Aspect ratio, Occupancy ratio, compactness, solidity, eccentricity, Elongation	MLP	0.783	0.849	0.784	0.815	0.217
		SVM	0.805	0.855	0.806	0.829	0.195
		RF	**0.866**	0.891	0.866	0.878	0.134
		AdaBoost	0.827	0.865	0.827	0.845	0.173
Texture-based	HOG, LBP, DCT-coefficients, GLCM (statistical measurements)	MLP	0.932	0.938	0.933	0.935	0.067
		SVM	0.929	0.935	0.933	0.934	0.071
		RF	**0.937**	0.941	0.938	0.939	0.062
		AdaBoost	0.892	0.898	0.892	0.894	0.107
Stroke-based	Stroke Distribution Profile, Area Occupancy Profile	MLP	0.895	0.894	0.892	0.893	0.104
		SVM	0.885	0.886	0.884	0.885	0.114
		RF	0.931	0.930	0.926	0.927	0.068
		AdaBoost	**0.933**	0.933	0.931	0.932	0.067
Gradient-based	Vertical-edge score, horizontal-edge score, Linearity of corner-points, strength of corner-points	MLP	**0.906**	0.908	0.907	0.907	0.093
		SVM	0.872	0.876	0.872	0.874	0.127
		RF	0.862	0.870	0.863	0.866	0.137
		AdaBoost	0.874	0.880	0.874	0.877	0.126
Aggregated features	Geometrical + Texture + Stroke + Gradient features	MLP	0.962	0.964	0.963	0.963	0.037
		SVM	0.961	0.963	0.960	0.961	0.039
		RF	0.962	0.963	0.962	0.962	0.038
		AdaBoost	**0.965**	0.967	0.966	0.966	0.034
Deep features	One convolution layer	CNN	0.957	0.954	0.959	0.956	0.042
	Two convolution layers		0.958	0.956	0.961	0.958	0.041
	Three convolution layers		**0.959**	0.957	0.960	0.958	0.040

4.4 Combined Text Non-Text Classification

At the end, same set of experiments have been performed on combined text components irrespective of image-source that may reveal the actual robustness and range of applicability of different handcrafted feature descriptors specially in deep learning era. Table 9 illustrates the performance of different categories of handcrafted features, aggregated features, and deep features. Highest classification accuracy for geometrical, texture and stroke-based features are 92.4%, 93.9% and 95.2% respectively, that are obtained using RF classifier. Gradient-based features have performed well on overall texts with 92.5% classification accuracy obtained by MLP. Other side, CNN-generated deep features have achieved the highest classification accuracy of 97.63% for three convolution layers.

Talking about aggregated features, classification accuracy is again improved by 2-6% compared to individual category of feature descriptors. Without any surprise, performance of aggregated features has again surpassed deep features with the highest classification accuracy of 98.4% obtained using AdaBoost classifier.

Table 9. Category-wise results of different handcrafted feature and deep features on combined (document and scene) text images and non-texts.

Category	Feature descriptors	Classifiers	Accuracy	P	R	F-M	Error
Geometrical	Aspect ratio, Occupancy ratio, compactness, solidity, eccentricity, Elongation	MLP	0.906	0.897	0.905	0.901	0.094
		SVM	0.901	0.915	0.901	0.908	0.098
		RF	**0.924**	0.932	0.923	0.927	0.076
		AdaBoost	0.901	0.920	0.901	0.910	0.099
Texture-based	HOG, LBP, DCT-coefficients, GLCM (statistical measurements)	MLP	0.937	0.939	0.934	0.936	0.063
		SVM	0.918	0.921	0.917	0.919	0.082
		RF	**0.939**	0.944	0.937	0.942	0.061
		AdaBoost	0.922	0.926	0.922	0.924	0.078
Stroke-based	Stroke Distribution Profile, Area Occupancy Profile	MLP	0.913	0.912	0.913	0.912	0.086
		SVM	0.902	0.905	0.901	0.903	0.097
		RF	**0.952**	0.953	0.953	0.953	0.047
		AdaBoost	0.945	0.946	0.943	0.944	0.055
Gradient based	Vertical-edge score, horizontal-edge score, Linearity of corner-points, strength of corner-points	MLP	**0.925**	0.928	0.925	0.926	0.074
		SVM	0.908	0.914	0.907	0.911	0.091
		RF	0.812	0.810	0.813	0.812	0.188
		AdaBoost	0.877	0.875	0.878	0.876	0.123
Aggregated features	Geometrical + Texture + Stroke + Gradient features	MLP	0.976	0.978	0.976	0.977	0.023
		SVM	0.978	0.979	0.979	0.979	0.021
		RF	0.983	0.983	0.981	0.982	0.017
		AdaBoost	**0.984**	0.985	0.983	0.984	0.015
Deep features	One convolution layer	CNN	0.970	0.955	0.970	0.962	0.029
	Two convolution layers		0.971	0.960	0.966	0.963	0.028
	Three convolution layers		**0.976**	0.964	0.976	0.970	0.023

4.5 Comparative Analysis and Discussion

After a series of exhaustive experiments, it is found that handcrafted feature descriptors are reasonably competitive in terms of performance with deep features in practical scenario. Although, it is observed from the experimental results that in one-to-one comparison between specific category of handcrafted features and deep features, deep features with no surprise appear at the upper hand in terms of performance. However, most of the feature descriptors yield promising results on both document-type and scene-type text images, which demonstrates their robustness and effectiveness. On the contrary, aggregated features generated by combining all categories of descriptors not only compete with deep features but also surpass in all the three experimental dataset(s), which indeed is a tremendous achievement. Some true-classified and misclassified text and non-text components of AUTNT-combined dataset using different categories of handcrafted features and CNN-generated deep features are shown in Fig.8. Moreover, performance comparison between each category of handcrafted features vs. deep features for all the three subsets of AUTNT dataset (document-type texts, scene-type texts, and combined texts and nontext components) have been depicted in Fig. 9.

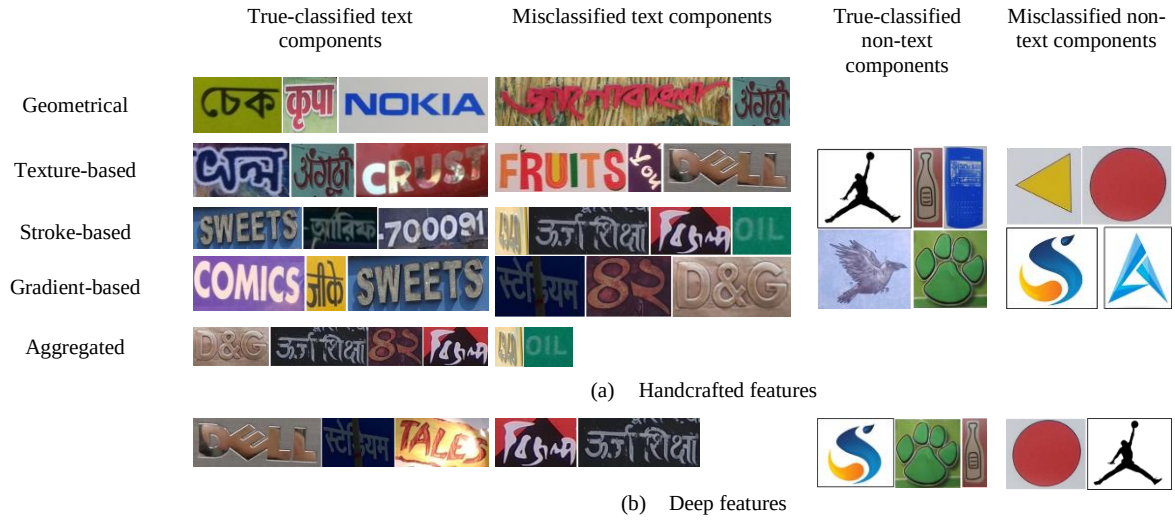

Fig. 8. Some true-classified and misclassified images of AUTNT-combined dataset using category-wise handcrafted features and CNN-generated deep features. (a) True-classified and misclassified text and non-text components using reported categories of handcrafted features, (b) true-classified and misclassified text and non-text components using deep features. Common true-classified and misclassified non-text components for all categories of handcrafted features have been considered.

It is observed from Table 7-9 that in most of the cases, stroke-based features have obtained the highest classification accuracy, which certainly reveals the discriminative strength of text-stroke uniformity irrespective of image sources. However, promising results have been obtained from other categories of feature descriptors as well. It

may be shown from Fig. 8 that geometrical features of object components perform well for well-structured document texts with near homogenous background. However, slight downfall is noted for scene images with arbitrary-orientation, cursive text-pattern, complex background, etc. Texture-based features are effective and robust for all kinds of texts, but fail in certain situations such as multi-color text-fonts, text-layout fully blend with backgrounds, etc. Stroke-based features are very effective for text instances of varying dimension and orientation in complex scenario. Still, accurate extraction of stroke features from degraded texts, perspectively distorted texts, artistic text-fonts, smudgy images, low-contrast images, etc. is challenging, that may lead to poor results (see in Fig. 8). Performance of gradient based features has found to be satisfactory as texts generally have high gradient magnitude across edges. However, gradient features yield poor performance for low-resolution images, degraded texts, etc. Additionally, linear orientation of corner-points from word-level texts has been exploited in this study for classification. On the other hand, for CNN-generated deep features, the design of weight balancing and pooling mechanism significantly reduces the number of parameters required to find local and global patterns from input object components, that certainly alleviates the overall classification performance. To investigate the performance of deep features, different number of convolution-layers are employed and the obtained results reflect that accuracy increases with increasing depth of features. However, intensive computational activities of deep learning approaches may be the possible hindrances for low resource environments. In both handcrafted and deep features, few non-text components are wrongly classified as texts due to their geometrical and symbol-like appearance.

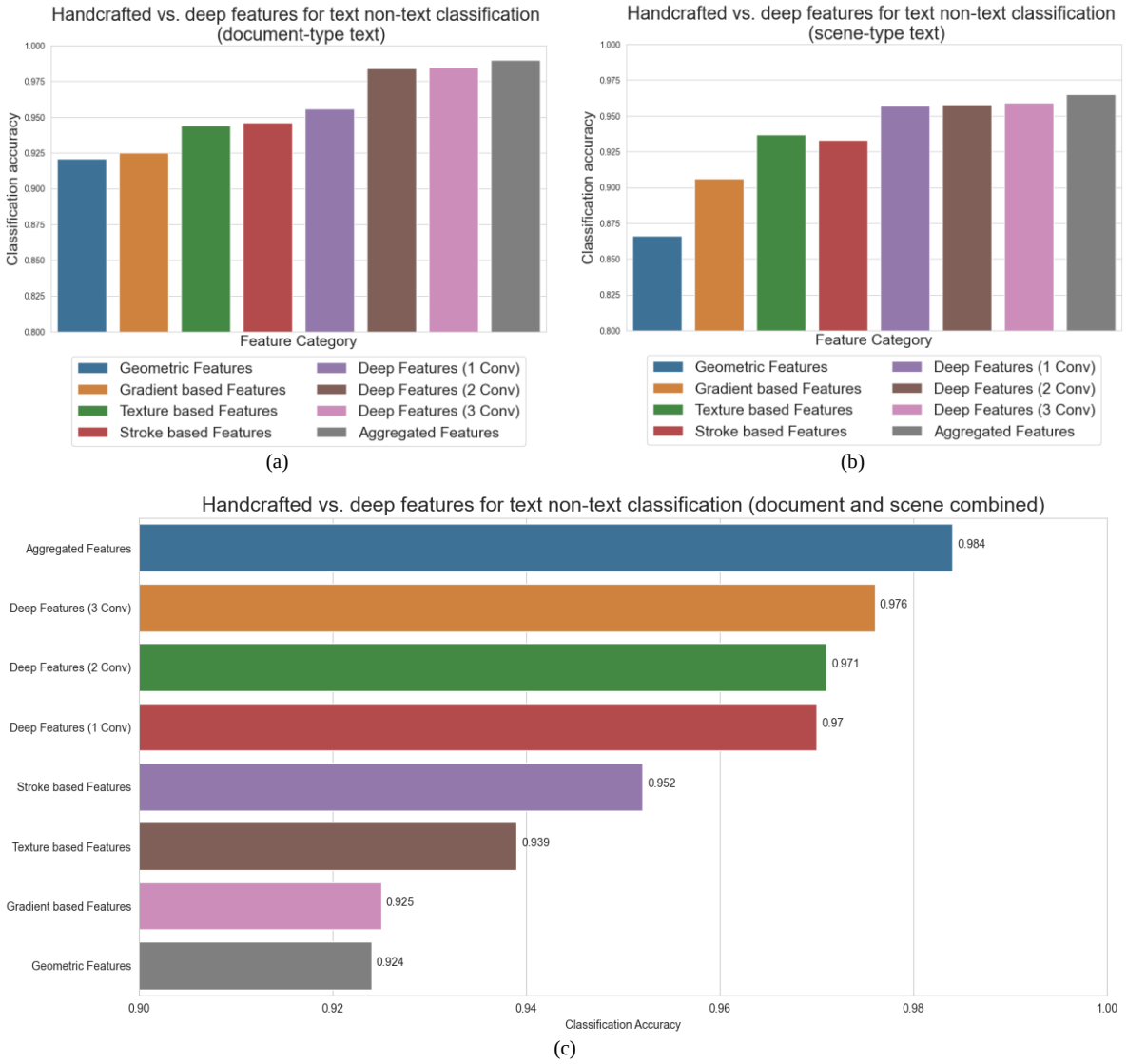

Fig. 9. Performance comparison between handcrafted features and CNN-generated deep features on three subsets of the experimental dataset (document, scene, and combined texts and nontext components). (a) Performance comparison between different categories of handcrafted features and deep features on document-type texts and nontexts. (b) Corresponding performance comparison on scene-type texts and nontexts. (c) Same performance comparison on combined texts and nontexts (Highest classification accuracy obtained by different classifiers for each category of features is considered for performance comparison. It may be noted that aggregated feature has surpassed other categories of handcrafted features as well as deep features in all the three cases.)

The key finding of the current study is the inspiring performance of aggregated features over deep features on multiple image types. It may be stated that individual category of feature descriptors may not justify their effectiveness for all types of text-images in diverse scenarios, but their combination may demonstrate high classification ability and justify the selection of reported feature descriptors. Performance of aggregated features has aggravated by at least 3-7% compared to individual category of feature descriptors for all three types of text images. It is worth mentioning that performance of different categories of handcrafted features are subjected to image types, whereas performance of aggregated feature is comparatively high and consistent for any kind image type irrespective of source that reveals its generality and robustness. Keeping view of limited resource constraint, such findings not only reveal the discriminant nature of such feature descriptors, but also encourage research community to plunge further into designing really discriminant handcrafted features.

5. Conclusion

In this paper, a comparative study between prospective handcrafted feature descriptors and deep features has been presented on text non-text component classification in document, scene, and unconstrained environments. A series of exhaustive experiments have been conducted on multiple image types to examine the competency of prospective handcrafted features as well as deep features. A significant outcome of this study is that aggregated feature descriptors have produced 99.0%, 96.5% and 98.4% classification accuracy for document-type, scene-type, and mixed-type (combined) text images respectively and in turn surpassed CNN-generated deep features i.e., 98.5%, 95.9% and 97.6% respectively. It is also observed that among individual categories of features, stroke-based descriptors outperform other feature sets on combined (document and scene) text images with more than 95% accuracy. Texture-based feature descriptors have produced high and consistent results (more than 93% accuracy) on both document and scene images that certainly establish their sturdiness irrespective of image source. Although, performance of deep features is in-agreement with the prevailing presumption, and in turn, reconfirms the outstanding performance of deep features, victory of aggregated features over automated deep features would not only benefit readers to choose appropriate feature descriptors but also encourage researchers to craft really discriminant feature descriptors. In future, amalgamation of more prospective handcrafted features and other deep features may be considered for different image classification tasks. Feature selection may also be incorporated to discard noisy features for improving the accuracy further.

Acknowledgment

The first author is thankful to School of Computer Science and Engineering (SCOPE), VIT-AP University for providing required support for writing the manuscript and Department of Computer Science and Engineering, Aliah University for providing necessary assistance to carry out problem analysis, design, coding, and experimental works. The second author is thankful to Department of Computer Science and Engineering, Aliah University for the same.

References

[1] H. I. Koo, D. H. Kim "Scene text detection via connected component clustering and nontext filtering." *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2296-2305, IEEE, 2013.

[2] Y. F. Pan, X. Hou, C. L. Liu "A hybrid approach to detect and localize texts in natural scene images." *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 800-813, IEEE, 2011.

[3] X. Chen, A. L. Yuille "Detecting and reading text in natural scenes" *in proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II-II, 2004.

[4] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao "Scene text detection using graph model built upon maximally stable extremal regions." *Pattern Recognition Letters*, vol. 34, no. 2, pp. 107-116, , Elsevier, 2013.

[5] R.M. Haralick, K. Shanmugam "Textural features for image classification." *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610-621, IEEE, 1973.

[6] N. Dalal, B. Triggs "Histograms of oriented gradients for human detection" *in proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886-893, IEEE, 2005.

[7] R. Minetto, N. Thome, M. Cord, N. J. Leite, J. Stolfi "T-HOG: An effective gradient-based descriptor for single line text regions." *Pattern Recognition*, vol. 46, no. 3, pp. 1078-1090, 2013.

[8] S. Tian et al. "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients." *Pattern Recognition*, vol. 51, pp. 125-134, Elsevier, 2016.

[9] T. Ojala, M. Pietikäinen, D. Harwood "A comparative study of texture measures with classification based on featured distributions." *Pattern Recognition*, vol. 29, no. 1, pp. 51-59, Elsevier, 1996.

[10] W. Huang, Z. Lin, J. Yang, J. Wang "Text localization in natural images using stroke feature transform and text covariance descriptors" *in proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1241-1248, IEEE.

[11] C.P. Sumathi, G.G. Devi "Automatic text extraction from complex colored images using gamma correction method." *Journal of Computer Science*, vol. 10, no. 4, pp.705-715, 2014.

[12] P. Shivakumara, T.Q. Phan, C.L. Tan "A robust wavelet transform based technique for video text detection" *in proceedings of the 10th International Conference on Document Analysis and Recognition*, pp. 1285-1289, IEEE, 2009.

[13] H. Li, D. Doermann, O. Kia "Automatic text detection and tracking in digital video." *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 147-156, Elsevier, 2000.
[14] B. Epshtein, E. Ofek, Y. Wexler "Detecting text in natural scenes with stroke width transform" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2963-2970, IEEE, 2010.
[15] Y. Zhao, T. Lu, W. Liao "A robust color-independent text detection method from complex videos" *in proceedings of International Conference on Document Analysis and Recognition*, 2011, pp. 374-378, IEEE.
[16] K. Subramanian, P. Natarajan, M. Decerbo, D. Castanon "Character-stroke detection for text-localization and extraction" *in proceedings of the 9th International Conference on Document Analysis and Recognition*, pp. 33-37, IEEE, 2007.
[17] S. Bhowmik, R. Sarkar, M. Nasipuri, D. Doermann "Text and non-text separation in offline document images: A Survey." *International Journal on Document Analysis and Recognition*, vol.21, no. 1-2, pp.1-20, IEEE, 2018.
[18] C. Yu, Y. Song, Y. Zhang "Scene text localization using edge analysis and feature pool." *Neurocomputing*, vol. 175, part. A, pp. 652-661, 2016.
[19] H. Cho, M. Sung, B. Jun "Canny text detector: Fast and robust scene text localization algorithm" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3566-3573, IEEE, 2016.
[20] X. Huang "Automatic video scene text detection based on saliency edge map." *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 34819-34838, Springer, 2019.
[21] S. Lu, T. Chen, S. Tian, J. H. Lim, C. L. Tan "Scene text extraction based on edges and support vector regression." *International Journal on Document Analysis and Recognition*, vol. 18, no. 2, pp. 125-135, 2015.
[22] P. Shivakumara, T. Q. Phan, S. Lu, C. L. Tan "Gradient vector flow and grouping-based method for arbitrarily oriented scene text detection in video images." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1729-1739, 2013.
[23] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang "EAST: an efficient and accurate scene text detector" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2642-2651, 2017.
[24] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue "Arbitrary-oriented scene text detection via rotation proposals." *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111-3122, 2018.
[25] Y. Dai, Z. Huang, Y. Gao, Y. Xu, K. Chen, J. Guo, W. Qiu "Fused text segmentation networks for multi-oriented scene text detection" *in proceedings of the 24th International Conference on Pattern Recognition*, pp. 3604-3609, IEEE, 2018.
[26] S. Prasad, A. W. K. Kong "Using object information for spotting text" *in proceedings of the European Conference on Computer Vision*, pp. 559-576, Springer, 2018.
[27] Q. Ye, D. Doermann "Text detection and recognition in imagery: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no.7, IEEE, pp. 1480-1500, 2015.
[28] L. Neumann, J. Matas "Real-time scene text localization and recognition" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3538-3545, IEEE, 2012.
[29] X. Liu, G. Meng, C. Pan "Scene text detection and recognition with advances in deep learning: a survey." *International Journal on Document Analysis and Recognition*, vol. 22, no. 2, pp. 143-162, 2019.
[30] J. Greenhalgh, M. Mirmehdi "Real-time detection and recognition of road traffic signs." *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, IEEE, pp. 1498-1506, 2012.
[31] A. F. Mollah, S. Basu, M. Nasipuri "Text detection from camera captured images using a novel fuzzy-based technique" *in proceedings of the 3rd International Conference on Emerging Applications of Information Technology*, pp. 291-294, IEEE, 2012.
[32] T. Khan, A.F. Mollah "Distance transform-based stroke feature descriptor for text non-text classification" *In Recent Developments in Machine Learning and Data Analytics*, pp. 189-200, Springer, 2018.
[33] J. Canny "A computational approach to edge detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, 1986.
[34] X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi "Text/non-text image classification in the wild with convolutional neural networks." *Pattern Recognition*, vol. 66, pp. 437-446, 2017.
[35] M. Zhao, R.Q. Wang, F. Yin, X.Y. Zhang, L.L. Huang, J.M. Ogier "Fast text/non-text image classification with knowledge distillation" *in proceedings of the International Conference on Document Analysis and Recognition*, pp. 1458-1463, IEEE, 2019.
[36] L. Liu, S. Lao, P.W. Fieguth, Y. Guo, X. Wang, M. Pietikäinen "Median robust extended local binary pattern for texture classification." *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp.1368-1381, 2016.
[37] M. D. Ansari, S. P. Ghrera "Intuitionistic fuzzy local binary pattern for features extraction." *International Journal of Information and Communication Technology*, vol. 13, no. 1, pp. 83-98, 2018.
[38] L. Liu, P. Fieguth, Y. Guo, X. Wang, M. Pietikäinen "Local binary features for texture classification: taxonomy and experimental study." *Pattern Recognition*, vol. 62, pp. 135-160, Elsevier, 2017.
[39] T.L. da Silveira, A. J. Kozakevicius, C. R. Rodrigues "Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain." *Medical & Biological Engineering & Computing*, vol. 55, no. 2, pp. 343-352, 2017.
[40] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai "Effective Uyghur language text detection in complex background images for traffic prompt identification." *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 220-229, 2018.
[41] T. Kasar, A.G. Ramakrishnan "Multi-script and multi-oriented text localization from scene images" *in proceedings of the International Workshop on Camera-Based Document Analysis and Recognition*, pp. 1-14, Springer, 2011.
[42] H. Goto, M. Tanaka "Text-tracking wearable camera system for the blind" *in proceedings of the 10th International Conference on Document Analysis and Recognition*, pp. 141-145, IEEE, 2009.
[43] I. Sobel, G. Feldman "A 3x3 isotropic gradient operator for image processing." *A talk at the Stanford Artificial Project*, pp. 271-272, 1968.
[44] Y. Wu, P. Natarajan "Self-organized text detection with minimal post-processing via border learning" *in proceedings of the IEEE International Conference on Computer Vision*, pp. 5010-5019, IEEE, 2017.
[45] D. He, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai "Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild" *in proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3519-3528, 2017.

[46] F. Khan, M. A. Tahir, F. Khelifi, A. Bouridane, R. Almotaeryi "Robust off-line text independent writer identification using bagged discrete cosine transform features." *Expert Systems with Applications*, vol. 71, pp. 404-415, 2017.
[47] T. Khan, A.F. Mollah "AUTNT-A component level dataset for text non-text classification and benchmarking with novel script invariant feature descriptors and D-CNN." *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 32159-32186, 2019.
[48] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner "Gradient-based learning applied to document recognition" *in proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.
[49] K. Simonyan, A. Zisserman "Very deep convolutional networks for large-scale image recognition" *in proceedings of the 3rd International Conference on Learning Representations*, pp. 1-14, San Diego, USA, 2015.
[50] A. Krizhevsky, I. Sutskever, G.E. Hinton "ImageNet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.
[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich "Going deeper with convolutions" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.
[52] K. He, X. Zhang, S. Ren, J. Sun "Deep residual learning for image recognition" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.
[53] A. Sherstinsky "Fundamentals of Recurrent Neural Network (RNN) and Long Short-term Memory (LSTM) network" *Physica D Nonlinear Phenomena*, vol. 404, no. 8, pp. 1-43, 2020.
[54] G.E. Hinton "Deep belief networks." *Scholarpedia*, vol. 4, no. 5, pp. 5947, 2009.
[55] H. Liu, A. Guo, D. Jiang, Y. Hu, B. Ren "PuzzleNet: Scene Text Detection by Segment Context Graph Learning." *arXiv preprint arXiv:2002.11371*, 2020.
[56] C. Ma, L. Sun, Z. Zhong, Q. Huo "ReLaText: Exploiting Visual Relationships for Arbitrary-Shaped Scene Text Detection with Graph Convolutional Networks." *arXiv preprint arXiv:2003.06999*, 2020.
[57] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, J. Jia "Learning shape-aware embedding for scene text detection" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4234-4243, 2019.
[58] X. Wang, Y. Jiang, Z. Luo, C. L. Liu, H. Choi, S. Kim "Arbitrary shape scene text detection with adaptive text region representation" *in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6449- 6458, 2019.
[59] T. Kobchaisawat, T. H. Chalidabhongse, S. I. Satoh "Scene Text Detection with Polygon Offsetting and Border Augmentation." *Electronics*, vol. 9, no.1, pp. 117-132, 2020.
[60] P. Yang, G. Yang, X. Gong, P. Wu, X. Han, J. Wu, C. Chen "Instance Segmentation Network with Self-Distillation for Scene Text Detection." *IEEE Access*, vol. 8, pp. 45825-45836, 2020.
[61] T. Khan, A.F. Mollah "Text non-text classification based on area occupancy of equidistant pixels." *Procedia Computer Science*, vol. 167, pp. 1889-1900, Elsevier, 2020.
[62] J.J. Lee, P.H. Lee, S.W. Lee, A. Yuille, C. Koch "Adaboost for text detection in natural scene" *in proceedings of the International Conference on Document Analysis and Recognition*, pp. 429-434, 2011.
[63] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, A.Y. Ng "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning" *in proceedings of the IEEE International Conference on Document Analysis and Recognition*, pp. 440-445, 2011.
[64] K. Wang, B. Babenko, S. Belongie "End-to-end scene text recognition" in proceedings of the IEEE International Conference on Computer Vision, pp. 1457-1464, 2011.
[65] T. Khan, A.F. Mollah "A novel text localization scheme for camera captured document images" *in proceedings of 2nd International Conference on Computer Vision & Image Processing*, pp. 253-264, IIT Roorkee, Springer, 2018.
[66] C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu "Detecting texts of arbitrary orientations in natural images" *in proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1083-1090, IEEE, 2012.
[67] V.P. Le, N. Nayef, M. Visani, J. M. Ogier, C. De Tran "Text and non-text segmentation based on connected component features" *in proceedings of the 13th International Conference on Document Analysis and Recognition*, pp. 1096-1100, IEEE, 2015.
[68] Aliah University Text Non-text dataset, https://github.com/iilabau/AUTNTdataset
[69] S. Dey, P. Shivakumara, K. S. Raghunandan, U. Pal, T. Lu, G. H. Kumar, C.S. Chan "Script independent approach for multi-oriented text detection in scene image." *Neurocomputing*, vol. 242, pp. 96-112, 2017.
[70] A. Sain, A.K. Bhunia, P.P. Roy, U. Pal "Multi-oriented text detection and verification in video frames and scene images." *Neurocomputing*, vol. 275, pp. 1531-1549, 2018.
[71] GLCM Properties. https://in.mathworks.com/help/images/ref/graycoprops.html. Accessed December 21, 2022.
[72] H. Breu, J. Gil, D. Kirkpatrick, M. Werman "Linear time Euclidean distance transform algorithms." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 529-533, 1995.
[73] T. Khan, R. Sarkar, A.F. Mollah "Deep learning approaches to scene text detection: a comprehensive review." *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3239-3298, 2021.
[74] T. Khan, A.F. Mollah "A Novel Multi-scale Deep Neural Framework for Script Invariant Text Detection." *Neural Processing Letters*, vol. 54, no. 2, pp. 1371-1397, 2022.
[75] J. Xiao, G. Wu "A robust and compact descriptor based on Center-symmetric LBP" *in proceedings of the International Conference on Image and Graphics*, pp. 388-393, IEEE, 2011.
[76] T. Ojala, M. Pietikainen, T. Maenpaa "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002.
[77] M. Heikkila, M. Pietikainen "A texture-based method for modeling the background and detecting moving objects" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657-662, 2006.
[78] B. Zahran, J. Al-Azzeh, Z. Alqadi, M. A. Al-zoghoul, S. Khawatreh "A Modified LBP Method to Extract Features from Color Images. *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 10, 2018.
[79] B. Vishnyakov, V. Gorbatsevich, S. Sidyakin, Y. Vizilter, I. Malin, A. Egorov "Fast moving objects detection using iLBP background model" *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, no. 3, pp. 347-350, 2014.
[80] Tan, X. and Triggs, B "Enhanced local texture feature sets for face recognition under difficult lighting conditions" *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635-1650, 2010.

[81] A. Chahi, Y. Ruichek, R. Touahni "Local directional ternary pattern: A new texture descriptor for texture classification" *Computer vision and image understanding*, vol. 169, pp.14-27, 2018.
[82] T. Ojala, M. Pietikäinen, T. Mäenpää "Gray scale and rotation invariant texture classification with local binary patterns" *in Proceedings of the 6th European Conference on Computer Vision*, pp. 404-420, Ireland, Springer Berlin Heidelberg, 2000.
[83] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan "Image matching from handcrafted to deep features: A survey" *International Journal of Computer Vision*, vol. 129, pp. 23-79, 2021.
[84] S.E. Bekhouche, F. Dornaika, A. Benlamoudi, A. Ouafi, A. Taleb-Ahmed "A comparative study of human facial age estimation: handcrafted features vs. deep features" *Multimedia Tools and Applications*, vol. 79, pp. 26605-26622, 2020.

Authors' Profiles

Tauseef Khan has received the Ph.D. degree in Computer Science and Engineering from Aliah University, Kolkata in 2022. He completed Master of Technology (Gold) in Information Technology from Maulana Abul Kalam Azad University of Technology (MAKAUT, formerly WBUT) in 2013. He is currently working as an Assistant Professor in the School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh. His research interest includes computer vision, digital image processing, machine learning, pattern recognition, etc. He has published several articles in reputed journals and conferences on scene text detection, image classification, script identification, etc. He is a member of IEEE and life member of ISTE.

Ayatullah Faruk Mollah (Senior member, IEEE, USA) is an Assistant Professor and former Head (Officiating) of the Department of Computer Science and Engineering, Aliah University, India. He has completed his doctoral study from Jadavpur University, India. He is also a Life Member of Indian Unit of International Association for Pattern Recognition (IAPR). He was a Senior Software Engineer at Atrenta (I) Pvt. Ltd. Noida, India. He has also worked with ScanBiz Mobile Solutions, New York, USA. He was awarded prestigious European Union Erusmus Mundus cLink Fellowship, University Grants Commission (UGC) Research Fellowship for Meritorious Students in Science, and Postdoctoral Fellowship from University of Warsaw, Poland. He is actively engaged in research. His research interests include deep learning, data science, image and video analysis, machine learning, bioinformatics, etc. He is currently guiding a number of PhD students in frontier areas of image analysis. So far, he has authored nearly hundred articles in refereed journals and conference proceedings. Dr. Mollah is also a Co-Principal Investigator of the Multi-script project funded by Department of Science and Technology, Govt. of India.

How to cite this paper: Tauseef Khan, Ayatullah Faruk Mollah, "When Handcrafted Features Meet Deep Features: An Empirical Study on Component-Level Image Classification", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.16, No.1, pp. 61-80, 2024. DOI:10.5815/ijigsp.2024.01.05