Chapter 2

Related work

In this chapter, a brief description of the existing literature related to scene text analysis is presented. The text analysis in natural scene images is broadly categories into three main groups:

- Scene Text Detection.
- Scene Text Recognition.
- Scene Text Spotting.

The aforementioned groups are further classified into sub-groups according to the kind of scene text analysed, *i.e.,* horizontal, oriented, and arbitrary-shaped text.

2.1 Scene Text Detection

EAST [15] utilizes a fully convolutional network to perform word-level and text-line-level predictions for horizontal text. The distance of an oriented bounding box from a point to its every side is predicted. Deep Direct Regression [16] utilizes a fully convolutional network to directly project the coordinate offsets from a related bounding box. SegLink [17] decomposes a text instance into segments connected by links, which are detected densely by a fully convolutional neural network at multiple scales. An oriented box that covers a part of a word is a segment and such segments are combined

by links. DMPNet [18] recalls text instances with a higher overlapping area by sliding quadrilateral windows in intermediate convolutional layers. It incorporates a shared Monte-Carlo process and a sequential protocol that helps in relative regression of text instances with quadrangles. The authors of [19] proposed a bootstrapping technique with image augmentation for semantic aware border detection of arbitrary oriented long scene text instances at the word and text-line level. The authors of [20] proposed a fast arbitrary oriented text detector in which they perform pixel-level classification for identification of text and non-text instances and also predict word-level bounding boxes via a fully convolutional network. RefineText [21] processes features at multiple levels to produce dense text regions with higher semantic value. TextEdge [22] uses the text-region edge map for classification, edge prediction, and boundary regression, which are relevant to text instances. The authors in [23] obtain inclined proposals that have the information of orientation angle of text instances. They use oriented region proposal network and oriented region-of-interest pooling layer to map arbitrary-oriented region proposals to a feature tensor for text classification. Tian *et al.* project pixels onto an embedding space, where they consider pixels of same text instances appear closer to each other [24]. The authors in [25] incorporate normalization of scale and orientation of text instances to map to a desired canonical geometry range. Liu *et al.* make use of a tightness-aware intersect-over-union metric that quantifies completeness of ground truth, tightness of matching degree, and compactness of word and text-line detection [26]. The authors in [27] describe visual and geometrical correlations to exploit texts having higher pairwise similarity using cycle consistency constraint and permutation matrix.

In [28], a teacher-student learning based method and proposal-free multi-level feature mimicking approach are introduced to improve the accuracy by mimicking multi-level convolutional feature maps. The authors in [29] utilize instance segmentation network that combines prototype masks, per-instance mask coefficients, and self-distillation

for precise text detection. To address the issue of complex background in scene images, GISCA [30] uses adaptive soft attention to capture the context of salient areas in U-Net architecture and make the gradient back-propagation process stable. IncepText [31], introduces a deformable position sensitive region of interest pooling for multi-oriented text detection in scene images. TextSnake [32], describe text as a sequence of order with overlapping disks centered at symmetric axes. Each disk is associated with variable radius and orientation for detecting curved text using fully convolutional network. .In [33], a multilingual multi-oriented text detector is proposed exploiting instance segmentation and context information through channel and spatial attention. LATD [34] makes use of learnable anchors to refine scales and locations for regressing the offsets. The authors in [35] localize corner points of bounding boxes in test time and segmenting text regions in relative positions and generate candidate boxes using sampling and grouping corner points during the interference stage. Liao *et al.* utilize rotation-sensitive features for regression and rotation-invariant features for classification [36]. WordSup [37] is a weakly supervised approach exploiting word annotations to train a character detector and generate loose bounding boxes. CRAFT [38] estimates affinity between characters in the detection of arbitrarily-oriented, curved, or deformed text instances in scene images by providing character level annotations. The authors in [39] utilize two convolutional neural network in cascaded manner to perform word-level spotting of scene text without any post processing. In [40], authors focus on learning of strong features by using global and local information for detection of occluded and long text at word-level. The authors of [41] develop a semantic rich information feature map using feature pyramid text with novel loss function for end-to-end trainable system to detect small text instances. In [42], text components are identified by exploring most significant bit information of a bit plane slice. Also the authors fixed the window for character components of arbitrary oriented words which is based on angular relationship between sub-bands and a fused band for oriented scripts detection and recognition.

The authors in [43], proposed a ring radius transform (RRT) technique to perform detection of oriented and multi-script scene text. Histogram Oriented Moments (HOM) was introduced in [44] for text detection in video. It is invariant to rotation, scaling, font, and font size variations. The authors in [45] proposed a multi-scale text detection technique that uses feature pyramid network for small text detection. In [46], a character graph grouping algorithm is utilized based on local context information to distinguish background noise from scene texts. Tang and Wu include a combination of superpixel-based stroke feature transform, hand-crafted features, and deep learning based region classification for scene text detection [47]. In [48], two convolution neural networks are used for detection and classification for coarse segmentation of scene texts.

TextField [49] detects irregular scene texts by learning a direction field that points each text pixel away from the nearest boundary of the text instance. A morphological operation is then used as post-processing for final detection.TextContourNet [50] extract instance-level text contour to increase the accuracy of curve text detection. In [51], to enhance the feature representation ability, a pyramid attention network is used text detection tasks. Mask-Most Net [52] use instance-level mask approximation method through a combination of high-level semantic and low-level features. It applies the auxiliary regression task on center and corner points followed by a contextual information to increase the accuracy of detection. In [53], arbitrary shape text is detected by extracting text proposals, which are refined using a recurrent neural network (RNN) and an adaptive number of boundary points. Mask TTD [54] use text frontier learning and a tightness prior that refine pixel-wise mask prediction and assign polygonal boundary to each text region for arbitrary shaped text detection. In [55], feature enhancement and a region proposal network are explored to utilize the prior knowledge about the shape of the text instance and representations of enhanced features to generate the bounding boxes. A pyramid region-of-interest pooling attention is further introduced that extracts the features of fixed-size text segmentation. A bounding box refinement

network is also used to extract a curve text. OPMP [56] applies many arbitrary-shape fitting mechanisms to enrich the backbone layers followed by a re-classification of text instance using multi-grain classification. Yao *et al.* extract individual characters of a text region and their relationship as a part of a semantic segmentation problem using a single fully convolutional network for multi-oriented and curved text detection [57]. MSR [58] performs feature merging at different scales for text detection. PSENet [59] grafted kernels of different scale for each text instance, and in step by step procedure it find minimal scale kernel for complete shape text instance localization. PAN [60] is the low computational-cost detector for detecting horizontal, oriented text, and arbitrary shape text, respectively. ITN [61] encodes the unique geometric configurations of text instances whereas CRN [62] uses region based detection method. In [63], the authors present SPCNET, which is based on feature pyramid network and segmentation method for suppressing false positives in scene text detection procedure and precisely locate text regions in scene images. The authors in R-Net [64] and DSRN [65] aggregate features for scale-invariant text instances detection. HAM [66], and AAM [67] uses anchors mechanism to regress offsets to text regions.

LOMO [68] detects the extremely long and curved text using direct regression module, iterative refinement module, and shape expression module. In [69], the authors proposed a differentiable binarization model for arbitrary shape scene text detection. It is a segmentation based method that uses ResNet-50 as a backbone. It fails to detect text instance within a text instance. ContourNet [70] minimize the false positive in text detection by using scale-insensitive adaptive region proposal network and detect curve text with the help of local orthogonal texture-aware module that set contour points.

2.2 Scene Text Recognition

A comparative study of encoder-decoder approaches with attention module on large-scale text recognition tasks in natural scene images is performed in [71]. Shi *et al.*

[72,73] involve a spatial transformer network for rectifying and recognizing a text image using an attention-based sequence network. In [73, 74], the authors apply an attention mechanism that learns adaptively and selects the suitable features for recognizing text instances. Focusing Attention [75] learns character-wise annotations by supervising a learnable attention module. AON [76] applies an arbitrary orientation network to extract features of text instances in four different directions and the placement semantics of characters. Bai *et al.* [77] address expensive annotation problem by introducing an edit probability loss considering the occurrences of missing or unnecessary characters. Char-Net [78] utilizes an auxiliary dense detection task of characters to address the issue of having irregular text. The authors introduce an optimized joint network to perform detection and recognition of scene text at word and character level. They develop an iterative method for character detection, which transforms character-level detection capability learned from synthetic data to real-world scene images. ESIR [79] has involved line-fitting transformation recursively to eradicate text-line curvature and perspective distortion by estimating the pose of text-lines. In [80], text recognition is performed by extracting discriminative features and increasing the alignment between the target character region and attention region. The authors in [81] utilize text shape descriptors, such as center line, scale, and orientation to deal with highly curved or distorted text. NRTR [82] dispenses with recurrences and convolutions with a stacked self-attention module, where an encoder extracts features and a decoder perform the recognization of texts based on the output of the encoder. In [83], a realization of asynchronous training and inference behavior is performed to classify images irrespective of the presence of text instances, which leads to multimodal recognition tasks. CA-FCN [84], is an attention based text recognizer that uses Fully Convolutional Network (FCN) and perform arbitrary shape text detection by classifying text characters at each pixel location.

The authors in [85] presents a recognition framework for irregular scene text, where they uses ResNet CNN, LSTM encoder-decoder fallowed by an attention module. In

this work, the training is slow due to encoder-decoder mechanism also better visual feature can be obtained by performing attention mechanism on graph for complex structures. UnrealText [86], is region based detector that renders images from 3D world to access scene information more precisely by considering normal, depth, and object meshes in scene images for multilingual text detecting and recognition. It fails to recognise the small text and have human rules to selected parameters for the detection. The authors in [87] mitigates the issue of time-dependent decoding and one-way serial transmission of semantic context in RNN based text recognition methods by proposing an end-to-end trainable network that capture global semantic context through multi-way parallel transmission. In [88], the authors proposed a generic and efficient attack method for scene text recognition and develop connectionist temporal classification and attention based method with targeted and untargeted attack modes for text recognition. SEED [89] predicts an additional global semantic enhanced encoder-decoder framework that is supervised by the word embedding for robust recognition of low-quality scene texts. ABCNet [90] develop a BezierAlign layer for grafting accurate convolution features for arbitrary shape text spotting with lesser computational overhead. SCATTER [91] stands for selective context attentional text recognizer that uses stacked block architecture with intermediate supervision to train Bi-LSTM encoder. Also, for the decoding of contextual dependencies attention mechanism is used.

2.3 Scene Text Spotting

Jaderberg *et al.* enables feature sharing for detecting text instances. They further utilize character case-sensitive and insensitive classification and bigram classification using multi-mode learning. Downsampling is avoided for a per-pixel sliding window [92]. The authors in [93], localize and recognize text instances using region proposal networks and deep networks, respectively. Deep TextSpotter [94] obtains text proposals from a region proposal network, which are normalized by bilinear sampling to obtain

a variable-width feature map. Each region is then mapped with a sequence of characters. The authors in [95] apply region-of-interest (RoI) pooling to obtain feature maps only once and shared by both detection and recognition, where RNN encoder encodes feature maps of different lengths into the fixed-size. Curriculum learning is utilized to learn the character-level language and appearance models. He *et al.* uses a text-alignment layer to compute arbitrarily orientated text features [96]. An explicit supervision based on spatial information of characters and recurrent neural network for word recognition is also included. FOTS [1] introduces feature sharing with rotated text proposals to develop an end-to-end trainable system for detection and recognition of scene text instances. In a single network farword pass, TextBoxes [97] localize and recognize horizontal text and TextBoxes++ [98] deal with arbitrarily-oriented text instances. In [99], detection and recognition processes are serially connected yielding a straightforward relation between the text detection task and the followup text recognition task. A rectification of mask is used to adapt to incidental recognition of text instances with arbitrary orientation and shape. TextPlace [100] performs topological metric localization considering spatial-temporal dependence between text instances.

MLTS [101] is a multi-language text spotter that maps text proposals to a fixed height keeping the aspect ratio same, which is followed by detection and recognition. Each proposal uses a connectionist temporal classification to decode multi-language texts. E2E-MLT [102] is one of the popular multilingual optical character recognition for scene text detection and recognition. It uses a single shared fully convolutional network. OctShuffleMLT [103] is a hardware-efficient network for multi-language detection and recognition that also uses fully-convolutional neural network with a less number of parameters and layers. Mask TextSpotter [104] uses semantic segmentation to detect text of arbitrary shapes and spatial attention for handling text instances of irregular shapes by simultaneously considering local and global textual information. TextDragon [105] describes the shape of text with a sequence of quadrangles to handle

the text of arbitrary shapes and RoISlide that connect a deep network and connectionist temporal classification based text recognizer. The labeling of locations of characters is not needed. WACNET [106] applies a shared convolutional backbone between word-level segmentation and char-level detection and recognition. ASTS [107] customizes the mask R-CNN [108] to exploit the holistic-level semantics and pixel-level semantics for text spotting, simultaneously. It further delivers sequence-level semantics for text recognition using an attention-based sequence-to-sequence network. CSE [109] uses a region expansion process where a seed is initialized within a text region and based on extracted local features and contextual information; it combines neighborhood regions. Boundary [110] does not require character-level annotations for arbitrary-shaped text spotting. Scene text in the cluttered background is handled in Cluttered TextSpotter [111]. Text Perceptron [112] is an order-aware arbitrary-shaped text detector. The authors in [113] proposed an end-to-end trainable network using graph convolutional network. In this each text instance is divided into a series of small rectangular text components and these components are linked together by adopting graph based network.

2.4 Motivation

The motivations of this thesis are based on the challenges arises in capturing scene images in everyday life. The field of scene text analysis using deep network has been considered very useful function of early visual processing application where the primitive features such as text edges play a significant role in text detection. In the unconstrained environment the noise like occlusion, perspective distortion, blur (edge contrast), and fainted edges make explicit many behavioural transitions in surface properties of the captured scene text images. These behaviourally transition leads the loss of significant information about the presence of text in scene images. Unfortunately, most of the existing state-of-art only focus on the text detection based on text shape and its

orientation. Scene text analysis grows from horizontal text to oriented text followed by curve and arbitrary shaped text detection. However, the existing literature shows significant improvement in scene text detection process, but when these methods are tested for the images which are captured in unconstrained environments, these method performances degraded dramatically. Therefore, in this thesis we address three different problems that are frequently arises in real-world applications. We briefly describe the problems as follows:

PROBLEM 1: Scene images have many objects scattered in the background. Text instances are smaller in nature. The presence of background clutters, partial occlusion and truncation of texts are a typical issue in scene images. For accurate spotting of scene texts, it is important to overcome the background clutters and partial occlusion artifacts. The presence of cluttered background noise restricts in dense feature matching process and leads to misclassification problem.

PROBLEM 2: Scene images are captured very often using mobile phones, cameras, and webcams. The issue of camera shake and motion blur are commonly found while taking images unless images are captured by experts. The perspective distortion and blurry / shaky text edges make the process of spotting very complex.

PROBLEM 3: The effect of weather conditions in terms of the ambient lighting and noises impact a lot while capture a scene image. The presence of adverse weather conditions and ambient noises, such as fog, rain, poor contrast, and low illumination are the main reasons for the fainted text edges in the scene images. This enhances the problem of inter-class interference in classification.

PROBLEM 4: Multilingual arbitrary-shaped scene text spotting is as per with everyday applications, where scene images are captured randomly using mobile cameras and bizarre artistic styles are used to make the text instances more attracting. This variations in language, font, script, shape, and alignment make the problem of spotting text instances more challenging.