An Episodic Learning Network for Text Detection on Human Bodies in Sports Images

Pinaki Nath Chowdhury, Palaiahnakote Shivakumara, Ramachandra Raghavendra, *Senior Member, IEEE*, Sauradip Nag, Umapada Pal, *Senior Member, IEEE*, Tong Lu, and Daniel Lopresti, *Senior Member, IEEE*

***Abstract*— Due to the proliferation of sports-related multimedia content on the WWW, effective visual search and retrieval present interesting research challenges. These are caused by poor image quality, a wide range of possible camera points of view, pose variations on the part of athletes engaged in playing a sport, deformations of text appearing on sports person's clothing and uniforms in motion, occlusions caused by other objects, etc. To address these challenges, this paper presents a new method for detecting text on human bodies in sports images. Unlike most existing methods, which attempt to exploit locations of a player's torso, face, and skin, we propose an end-to-end episodic learning approach that employs inductive learning criteria for detecting clothing regions in an image, which are, in turn, then used for text detection. Our method integrates a Residual Network (ResNet) and Pyramidal Pooling Module (PPM) for generating a spatial attention map. The Progressive Scalable Expansion Algorithm (PSE) is adapted for text detection from these regions. Experimental results on our own dataset as well as several benchmarks (like RBNR and MMM which contain images of runners in marathons, and Re-ID which is a person re-identification dataset) demonstrate that the proposed method outperforms existing methods in terms of precision and F1-score. We also present results for sports images chosen from natural scene text detection datasets such as CTW1500 and MS-COCO to show the proposed method is effective and reliable across a range of inputs.**

***Index Terms*— Clothing detection, residual network, region proposal network, text detection, sports video retrieval, multimedia content management.**

Manuscript received December 30, 2020; accepted June 18, 2021. Date of publication June 28, 2021; date of current version April 5, 2022. This work was supported by the Natural Science Foundation of China under Grant 61672273. The work of Palaiahnakote Shivakumara was supported by Fundamental Research Grant Scheme (FRGS) Grant, Ministry of Higher Education, Malaysia, under Grant FP104-2020. This article was recommended by Associate Editor X. Li. *(Corresponding author: Tong Lu.)*

Pinaki Nath Chowdhury and Umapada Pal are with the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: pinakinathc@gmail.com; umapada@isical.ac.in).

Palaiahnakote Shivakumara is with the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia (e-mail: shiva@um.edu.my; hudempsk@yahoo.com).

Ramachandra Raghavendra is with the Faculty of Information Technology and Electrical Engineering, IIK, Norges Teknisk-Naturvitenskaplige Universitet (NTNU), 7491 Trondheim, Norway (e-mail: raghavendra.ramachandra@ntnu.no).

Sauraip Nag is with the Kalyani Government Engineering College, Kalyani 741235, India (e-mail: sauradipnag95@gmail.com).

Tong Lu is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: lutong@nju.edu.cn).

Daniel Lopresti is with the Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: lopresti@cse.lehigh.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3092713.

Digital Object Identifier 10.1109/TCSVT.2021.3092713

I. Introduction

WHEN there is a rapid growth in the field of communication and internet technologies, usage of multimedia content especially sharing sports information at anytime and anywhere over variety of devices increases exponentially [1]. In this context, users prefer to access relevant information rather bulk information from the large amount of data. For example, accessing a particular goal in case of soccer [1], hitting fours and sixes in case of cricket, etc. In addition, in cricket bowler and batsman require to access specific videos to analyze their mistakes and find weakness of opponents. In order to retrieve relevant information according to user interests, it is necessary to annotate data accurately at semantic level [2].

To identify the players or to trace actions of the players or tracing marathon runner in video for the purpose of indexing and retrieval, text on clothing play a vital role for retrieving desired information [1] because it provides information which is close to content of video images. Information of text and visual content can then be further combined for annotating videos at semantic level. Therefore, there is a need for accurate text detection in sports images. There are methods for detecting text in natural scene, sports and marathon images by exploring deep learning models [3]–[8]. However, these methods may not be effective for sports images because of intrinsic bias, such as selection bias, capture bias and negative set bias [9]. In addition, the multi-modal methods that use face, torso and skin information reduce complexity of the work well when it detects face, skin and torso accurately. Due to large variations of camera viewpoints, pose variations, and occlusions, there are high chances of missing and losing face, torso or skin information [9]. This observation motivates us to introduce cloth detection directly in contrast to face, skin and torso detection as a kind of context information for text detection in this work. This is because clothing detection is robust to the aforementioned challenges and it does not affect much by the above mentioned challenges. Furthermore, it is noted that usually the text is embedded on uniform or jersey for every player in the sports images, which provides unique information about the player and context of the situation.

(a) Text detection by TextSnake [3] method

(b) Text detection by PSENet method [4]

(c) Text detection by Shivakumara et al. [5] method

(d) Text detection by the Kamlesh et al. [6] method

(e) Text detection by the proposed method

Fig. 1. Example for text detection by the existing and the proposed methods.

In Fig. 1, natural scene text detection methods [3], [4], which use deep learning models, do not work well for sports images as shown in Fig. 1(a)-(b). Similarly, Fig. 1(c) and (d) show that the methods [5], [6] that work based on multimodal concepts do not detect all the texts in sports images. However, the proposed method detects all the texts as shown in Fig.1(e). It is also observed from Fig. 1(a)-(e) that all the existing methods fail to detect "7" in the first image, while the proposed method detects it correctly. This is the advantage of our method that considers detection of clothing as context information. Since the clothing is non-rigid material, it poses many challenges, such as distortions due to folding, different colors, partial occlusion and irregular sized characters. Therefore, text detection through clothing information in sports is complex and interesting.

The main contribution of this work is a new end-to-end episodic learning approach for text detection in sports images that unifies residual and region proposal networks, pyramidal pooling, and progressive scalable expansion networks under a single architecture. To the best of our knowledge, this is the first work exploring episodic learning for this application. In contrast to existing methods which use faces, torsos, and skin, we use deformable clothing regions to improve the accuracy of text detection. The way the proposed method integrates spatial features given by ResNet, regions of interest obtained by PPM, and character shapes given by PSENet is novel comparing to the state-of-the-art methods. Finally, we have developed a new dataset well suited for this purpose and it will be released to the research community.

II. Related Work

Since text detection in sports images is related to natural scene text detection, we review the methods of natural scene images, marathon images and sports images.

Wang *et al.* [4] proposed a Progressive Scale Expansion Network (PSENet) for text detection in natural scene images. The approach involves in segmentation-based detectors for predicting multiple text instances. However, the method may not perform well for texts of few characters. Ma *et al.* [10] proposed a rotation region proposal network for text detection in natural scene images. It considers rotation of region interest as pooling layers for the classifier. Angular or directional information is good for text of many character, else direction may yield incorrect results. Long *et al.* [3] proposed a flexible architecture for text detection in natural scene images. It considers text instances as a sequence of ordered, and finds symmetric axes with radius and orientation information. When a text is short and only contains a few characters, the method may perform well. The presence of single character or digit and short names are common in case of sports images.

Feng *et al.* [11] proposed a method for arbitrarily-oriented text spotting in natural scene images. The approach introduces RoISlide operator, which connects series of quadrangular texts. The method follows the idea of Long *et al.*'s method [3] for extracting instances of quadrangles. Since it depends on directions of texts, it may not work well for short texts, which are common in spots images. Baek *et al.* [12] proposed a character awareness-based method for text detection in natural scene images. It finds the relationship between characters for detecting texts in images. However, it is not sure how the method works for single characters. Raghunandan *et al.* [13] proposed a method for text detection in natural scene, video and Born digital images. The method uses bit plane and convex deficiency concepts for detecting text candidates. If a character misses a few pixels due to clutter background, the method may not work well. Xu *et al.* [14] proposed a method for irregular scene text detection in natural scene images. It finds the relationship between current text and its neighbor boundary to fix the bounding box of any orientation. However, the method is not tested on images with single characters or digits in sports images. Cai *et al.* [15] proposed an Inside-to-Outside Supervision Network (IOS-Net) for text detection in natural scene images. It designs a hierarchical supervision module to capture texts of different aspect ratios, and then multiple scale supervision for a stacked hierarchical supervision module.

The Mask-R-CNN has been studied for text detection in natural scene images as it is popular for improving instance segmentation by predicting accurate masks [16], [17].

Lyu *et al.* [18] proposed an end-to-end trainable neural network for text spotting in natural scene images. This method generates shape masks of objects and detects text by segmenting instance regions. This approach may not be effective for shorter test strings.

Roy *et al.* [19] proposed a method for text detection from multi-view of images of natural scenes. The method uses Delaunay triangulation for extracting features from estimating similarity and dissimilarity between components in different views. However, the method is limited to multiple views. Wang *et al.* [20] proposed a quadrilateral scene text detector for text detection in natural scene images. The approach uses two stages for achieving better results. However, for arbitrary oriented and irregular texts, the quadrilateral proposal network may not be effective. Liu *et al.* [21] proposed a method for text spotting in natural scene images based on an adaptive Bezier curve network. The method focuses on fixing tight bounding boxes for arbitrary oriented text lines to improve text detection performance. However, when points are extracted from irregular shaped characters due to non-rigid clothing in the sports images, the performance of the method may degrade. Wang *et al.* [22] proposed text detection from multi-view of natural scene images. It finds correspondence between multi-views for achieving results. To find correspondence, the method uses similarity and dissimilarity estimation between text components in multi-views. The method requires multiple views.

In the above discussions on the methods for text detection in natural scene images, it is noted that most methods use direction and aspect ratio of character information for designing deep learning architectures to address the problem of arbitrary orientation. It is observed that none of the methods considered clothing information in the sports images for text detection. In case of clothing contained text, one can expect words of short length compared to text in natural scene images, single digit numbers, partial occlusion due to clothing folding, deformation and body movements. As a result, characters may not preserve actual structures. Therefore, natural scene text detection methods may not be good enough to addresses the challenges of sports images.

To reduce the complexity of the problems in sports images, some methods are proposed to use multimodal concepts like face, skin, torso, and human body parts information for achieving better results for sports images. Ami *et al.* [23] proposed a method for text detection in marathon images using face information. The approach first detects face and then torso which contains bib numbers for detection. As long as face is visible in the images, the method works well. Otherwise, it does not. To overcome this limitation, Shivakumara *et al.* [5] proposed torso segmentation without face information for bib number detection in marathon images. However, the above methods only detect texts in torso regions but not from the other parts of human bodies. To improve bib number and text detection performances for sports images, Nag *et al.* [24] proposed a method to detect human body parts rather than relining only on torso regions. However, the performance of the method depends on the success of human body parts detection. Similarly, Kamlesh *et al.* [6] used text information in marathon images for person re-identification. The method recognizes text in the marathon images for person re-identification. It is observed from experimental results that the method works well for high quality images but not for poor-quality ones.

In summary, the methods that use multi-modal concept addressed a few issues of text detection in sports images and the performances of the methods depend on pre-processing steps like face, torso and human body parts detection. In addition, text is present usually on of uniform or jersey but not all parts of human in sports images. The methods ignore clothing information for detecting text in sports images. This observation motivates us to use clothing information as context for detecting text in sports images in this work. We note from this review that none of these methods has examined episodic learning for detecting text in sports images. Most use torso, face, and skin information for text detection. These methods may not be robust for text detection in the case of deformations (text appearing on clothing).The interesting text in sports images usually appears on clothing, then it is very natural to try to detect the clothing in such images.

The main advantage of episodic learning is that the proposed method can be trained with samples of dataset, which is other than testing dataset for achieving the results. In this work, for clothing detection and text detection, we use the samples of different clothing and text detection datasets but not the datasets used for evaluation in this work. Therefore, the proposed method is capable of addressing challenges of text on deformable clothing regions in sports images. To extract the above observations, we propose to integrate Residual Network (ResNet) [25], Pyramid Pooling Module (PPM) and Progressive Scalable Expansion Network (PSENet) [4] for text detection on human body in sports images irrespective of adverse effects deformable regions. The ResNet is used for finding significant information in images as it helps in extracting spatial relationship between clothing pixels. Due to clutter background and poor quality, ResNet alone is not enough to capture deformable clothing information. Therefore, PPM is proposed to improve the results of clothing detection because it is good for extracting regions of interest. In the same way, PSENet is good for studying shapes of characters regardless of font, text size and color. As a result, PSENet helps us to detect texts in the clothing regions. The way the proposed work integrates spatial features given by ResNet, regions of interest obtained by PPM, and character shapes given by PSENet is a new idea comparing to the state-of-the-art methods.

III. Proposed Method

As discussed in the previous section, for every input sports image, clothing of uniform or jersey is an essential component of persons in the image, which contains unique id, name, and advertisements. If we are able to detect clothing regions, it is as good as detecting regions of interest in the image. Therefore, for text detection in sports images, we consider clothing region as a context feature. Inspired by the great success of deep learning models, we employ Residual Network (ResNet) for clothing detection because it can generate strong semantic

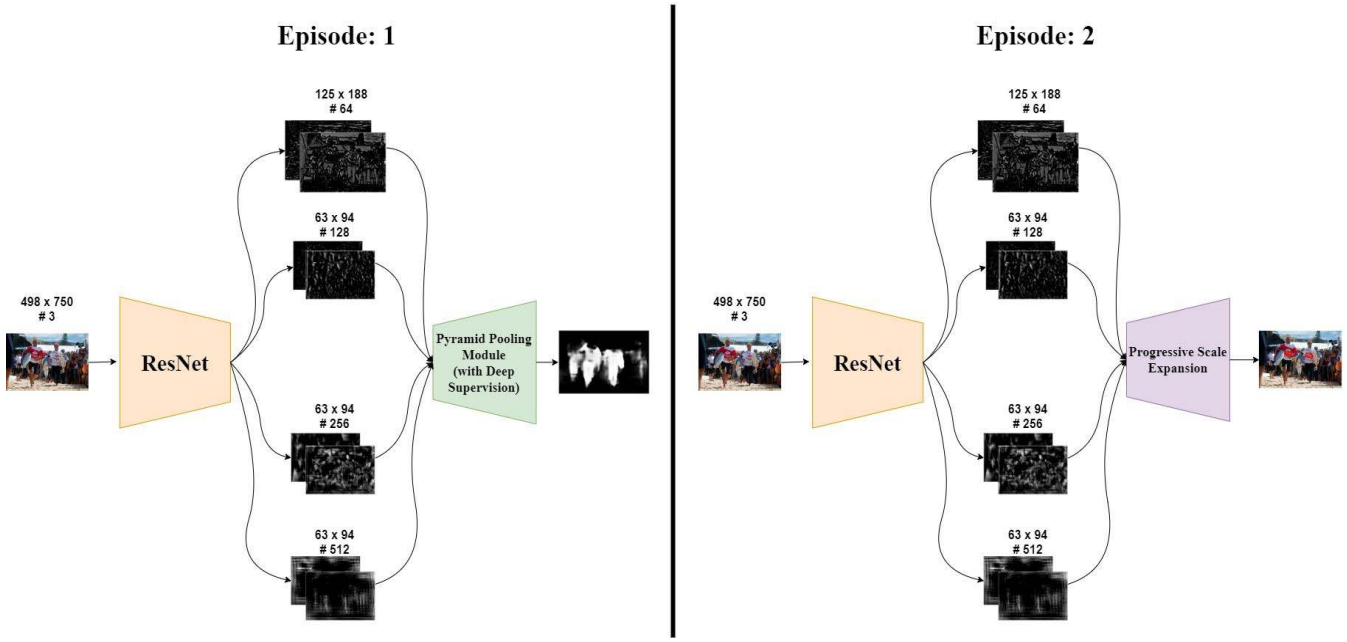

Fig. 2. Network Architecture during training episodes.

features by defining the spatial relationship between clothing pixels instead of conventional networks like VGG-16 [20]. It is noted that detecting texts on clothing of the sports player is challenging because of pose variations, deformable, multiple camera viewpoints, and the presence of single digits or characters. ResNet alone is not sufficient to cope with the above challenges. Therefore, motivated by the special property of PPM which extracts regions of interest, we explore PPM for strengthening the features of ResNet for accurate clothing detection in this work. Similarly, Progressive Scale Expansion Network (PSENet) [4] is used for studying shapes of objects and characters irrespective of fonts, text size and color. This property motivated us to explore PSENet for text detection from clothing regions detected by the combination of ResNet and PPM. The proposed work uses episodic training mechanism, which integrates the features of ResNet, PPM and PSENet for text detection in sports images. The architecture of the model for learning and evaluation are, respectively, shown in Fig. 2 and Fig. 3, where we can see ResNet, PPM and PSENet for text detection in sports images.

A. Attention ResNet for Clothing Detection From Human Body

For a given input image $I \in \mathbb{R}^{3\times H\times W}$, we employ a pertained CNN backbone, i.e., Residual Network (ResNet) to extract a 3D convolutional feature map as $\{\Psi \in \mathbb{R}^{C\times H'\times W'}, \psi_1 \in \mathbb{R}^{c_1\times h_1\times w_1}, \psi_2 \in \mathbb{R}^{c_2\times h_2\times w_2}, \psi_3 \in \mathbb{R}^{c_3\times h_3\times w_3}\}$. Here Ψ is the output of the last residual block of CNN backbone, ψ_1 the second last, ψ_2 is the third last, and ψ_3 the fourth last. $\{C, c_1, c_2, c_3\}$ represents numbers of channels, $\{H', h_1, h_2, h_3\}$ represents height, while $\{W', w_1, w_2, w_3\}$ represents width of output feature maps. The objective of Spatial Attention Network F_{atten} is to generate a heat map $\pi \in \mathbb{R}^{1\times H'\times W'}$, where each element of the attention map that represents a clothing region has values closer to 1, while the remaining elements of the attention map have values close to 0.

Inspired by the method in [26], where pyramid pooling has been used for handling multi-font size texts, we propose a similar 4-level pyramid scaling operation by passing Ψ to four adaptive pooling layers, which generates a feature map with the dimensions of $\{\mathbb{R}^{C\times 1\times 1}, \mathbb{R}^{C\times 2\times 2}, \mathbb{R}^{C\times 3\times 3}, \mathbb{R}^{C\times 6\times 6}\}$. Since average pooling outperforms max pooling, we prefer average pooling in this work. For the output of each adaptive pooling layer, the proposed method performs 1×1 convolutional layer to reduce the dimension to 512 output channels, which include a batch normalization layer, ReLU activation, and a bilinear interpolation layer. The extracted features are map to $\mathbb{R}^{512\times H'\times W'}$. The proposed method concatenates the output of 4-level pyramid pooling along with Ψ to derive the global prior $\theta \in \mathbb{R}^{(C+2048)\times H'\times W'}$ that captures information of different sub-regions. This process also follows the same as above to obtain the attention map π.

Motivated by the method [26], the proposed method performs deep supervision during training phase, where ψ_1 is fed to two convolutional layers to generate $\pi^* \in \mathbb{R}^{1\times H'\times W'}$. Therefore, the loss function is to train F_{CNN} along with the F_{atten} as defined in [26]:

$$\mathcal{L}_{atten} = \mathcal{L}_{NLLLoss}\left(\pi, \hat{\pi}\right) + \lambda_{dsup} \cdot \mathcal{L}_{NLLLoss}\left(\pi^*, \hat{\pi}\right) \quad (1)$$

where $\hat{\pi}$ is the ground truth of containing clothing regions, and $\mathcal{L}_{NLLLoss}$ is the negative log likelihood loss function. λ_{dsup} is a hyper parameter whose value is chosen similar to [4].

B. Region Proposal Network for Text Detection From Clothing Region

The features obtained by the network presented in the previous section $\{\Psi, \psi_1, \psi_2, \psi_3\}$ are fed to the Region

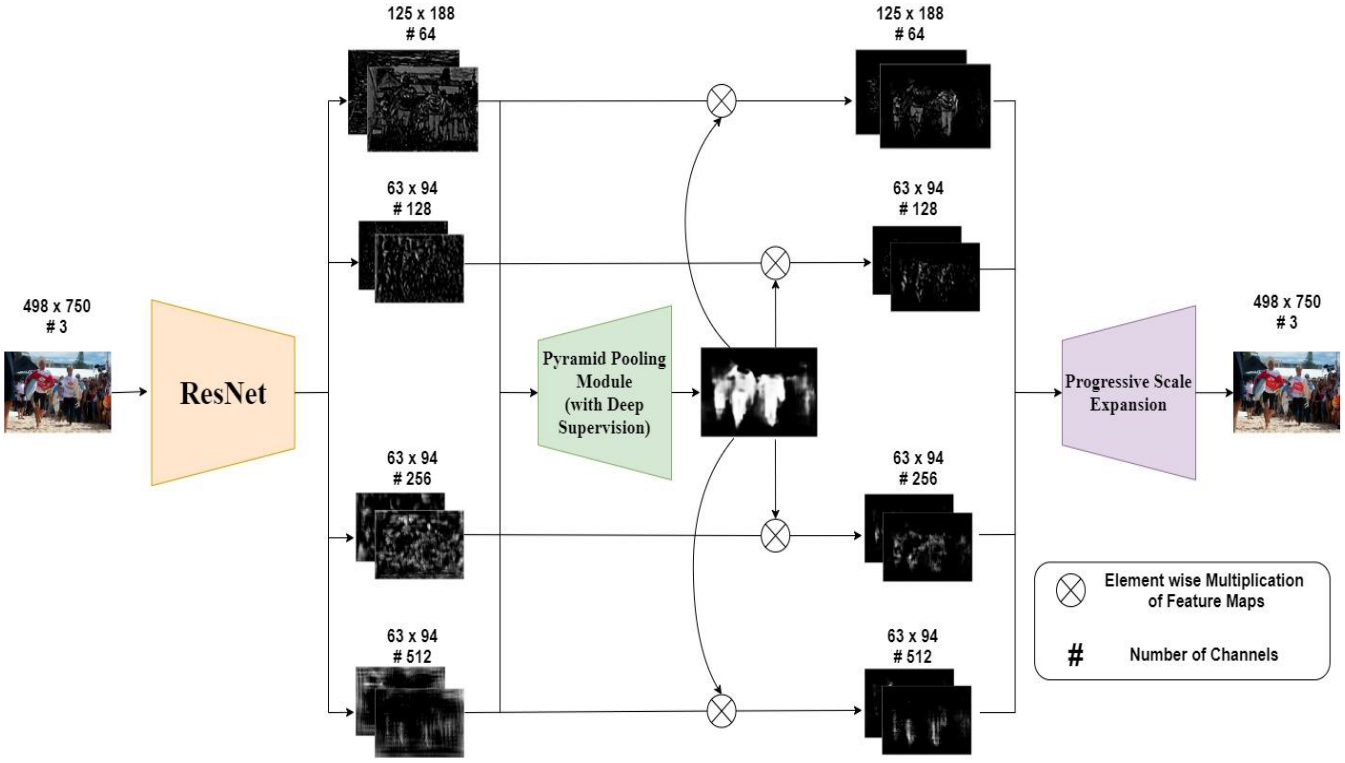

Fig. 3. Network architecture during evaluation (testing).

Proposal Network (RPN) F_{RPN} for text detection. Motivated by the method in [4], where a segmentation-based approach has been used to improve text detection performance, we propose the same for separating text instances by using a kernel-based framework called Progressive Scale Expansion Network (PSENet). For training PSENet, the loss function is derived as defined in Equation (2).

$$\begin{aligned} \mathcal{L}_{RPN} &= \lambda_{pse} L_c + \left(1 - \lambda_{pse}\right) \mathcal{L}_s \\ \mathcal{L}_c &= 1 - F_{dice}(S_n \cdot M, G_n \cdot M) \\ \mathcal{L}_s &= 1 - \frac{\sum_{i=1}^{n-1} F_{dice}(S_i \cdot W, G_i \cdot W)}{n-1} \\ W_{x,y} &= \begin{cases} 1, & if\ S_{n,x,y} \geq 0.5; \\ 0, & otherwise. \end{cases} \\ F_{dice}(S_i, G_i) &= \frac{2 \sum_{x,y} (S_{i,x,y} \times G_{i,x,y})}{\sum_{x,y} S_{i,x,y}^2 + \sum_{x,y} G_{i,x,y}^2} \end{aligned} \tag{2}$$

where $\mathcal{L}_c$ represents the loss for complete text instance, L_s denotes the loss for shrunk ones, F_{dice} represents the dice coefficient, $S_{i,x,y}$ represents the value of pixel (x, y) in n segmentation result S_i of PSENet, $G_{i,x,y}$ represents the value of pixel (x, y) in ground-truth G_i, M is the training mask, and W is a mask which ignores non-text pixels in S_n.

C. Episodic Training and Evaluation of Spatial Attention and Region Proposal Networks

As mentioned earlier, we propose Episodic training [27], which is an end-to-end text detection network for sports images, to train both the features extracted for clothing detection and the features extracted from clothing regions for text detection. The reason to propose to Episodic training is that it has the ability to extract robust features.

The proposed method follows two appropriately constructed episodes of training $\{F_{CNN}, F_{atten}\}$ using $\mathcal{L}_{atten}$ and $\{F_{CNN}, F_{RPN}\}$ using $\mathcal{L}_{RPN}$ that exposes F_{CNN} to a different statistics (i.e., either the statistics of F_{atten} or F_{RPN}) during each training episodes. While F_{atten} tries to capture clothing regions of a human, F_{RPN} endeavors to estimate the text regions in the input image. Since F_{CNN} performs feature extraction for both $\{F_{atten}, F_{RPN}\}$, the output feature maps $\{\Psi, \psi_1, \psi_2, \psi_3\}$ become robust to capture both clothing and text information.

To train $\{F_{CNN}, F_{atten}\}$ using $\mathcal{L}_{atten}$, we use DeepFashion2 Dataset [28] which provides a segmentation mask for different kinds of clothing regions. The proposed method unifies all the different kinds of clothing into a single class. The ground-truth used in $\mathcal{L}_{atten}$ is a binary mask, which indicates whether the region is a clothing region or not. The proposed method trains $\{F_{CNN}, F_{RPN}\}$ using $\mathcal{L}_{RPN}$ on ICDAR 2015 [29] dataset as it provides bounding boxes for texts, which are the ground-truth in $\mathcal{L}_{RPN}$. The algorithmic steps of Episodic Training are presented in Algorithm-1. Description of the variables used in the algrotihm-1 can be found in Table I.

The proposed method trains $\{F_{CNN}, F_{atten}, F_{RPN}\}$ using Algorithm-1, and then fuses the three components in a novel mechanism without any further training. As shown in Fig. 2, the output $\{\Psi, \psi_1, \psi_2, \psi_3\}$ from F_{CNN} is fed to F_{atten} which warps a spatial attention map π, resulting in a 2 dimensional

Algorithm 1 Episodic Training for F_{CNN}, F_{atten}, F_{RPN}

Input Clothing: $\mathrm{D_{cloth}} = [\mathrm{D}_1^{cloth}, \mathrm{D}_2^{cloth}, \ldots\ldots, \mathrm{D}_n^{cloth}]$
Input Text: $\mathrm{D_{text}} = [\mathrm{D}_1^{text}, \mathrm{D}_2^{text}, \ldots\ldots\ldots\ldots, \mathrm{D}_n^{text}]$
Initialize training episodes: $[N_{cloth}, N_{text}]$
Initialize model parameters for $[F_{CNN}, F_{atten}, F_{RPN}]$
Initialize learning rate α
flag $\leftarrow 0$ ▷ used for alternative training episodes
while model still converging **do**
if *flag* $== 0$ **then**
  for $i \in \{1, 2, \ldots\ldots\ldots, N_{cloth}\}$ **do**
    $\hat{\pi} \leftarrow$ *sample* D_j^{cloth} *from* D_{cloth}
  calculate $\mathcal{L}_{atten}$ *from* $[\hat{\pi}, \pi, \pi^*]$
  update $F_{CNN} := F_{CNN} - \alpha \cdot \nabla_{\text{atten}}(\mathcal{L}_{atten})$
  update $F_{atten} := F_{atten} - \alpha \cdot \nabla_{\text{atten}}(\mathcal{L}_{atten})$
 end for
 flag $\leftarrow 1$
end if
else if *flag* $== 1$ **then**
 for $i \in \{1, 2, \ldots\ldots\ldots, N_{text}\}$ **do**
  Sample D_j^{text} *from* D_{text}
  calculate $\mathcal{L}_{RPN}$ *from* $[\mathcal{L}_c, \mathcal{L}_s]$ *from* D_j^{text}
  update $F_{CNN} := F_{CNN} - \alpha \cdot \nabla_{\text{RPN}}(\mathcal{L}_{RPN})$
  update $F_{RPN} := F_{RPN} - \alpha \cdot \nabla_{\text{RPN}}(\mathcal{L}_{RPN})$
 end for
 flag $\leftarrow 0$
end if
end while
Output: Trained parameters $[F_{CNN}, F_{atten}, F_{RPN}]$

TABLE I
THE LIST OF VARIABLES USED IN THE ALGORITHM-1

Variables	Meanings
Ψ	Output of last residual block from ResNet
ψ_1	Output from the Second last residual block from ResNet
ψ_2	Output from Third last residual block from ResNet
ψ_3	Output from the Fourth last residual block from ResNet
C, H', W'	Number of channel, Height and Width of Ψ
c_1, h_1, w_1	Number of channel, Height and Width of ψ_1
c_2, h_2, w_2	Number of channel, Height and Width of ψ_2
c_3, h_3, w_3	Number of channel, Height and Width of ψ_3
π	Attention Maps from Spatial Attention Network
$\mathcal{L}_{atten}$	Loss function to train the Spatial Attention Network
$\mathcal{L}_{NLLLoss}$	Negative Log Likelihood to detect ground-truth clothing region
$\lambda_{dsup}, \lambda_{pse}$	Hyper-parameters evaluated empirically
α	Learning rate
∇	Partial Differentiation for backpropagation

matrix. We use bilinear interpolation to resize π to have the same shape with that of each output feature map from F_{CNN}, and performs element-wise multiplication with all the channels of $\{\Psi, \psi_1, \psi_2, \psi_3\}$ to derive a new rectified feature map $\{\Psi^*, \psi_1^*, \psi_2^*\}$, ψ_3^* as defined in Equation (3).The proposed method feeds the rectified $\{\Psi^*, \psi_1^*, \psi_2^*, \psi_3^*\}$ to

Fig. 4. Samples results for detecting text in deformable cloth region.

F_{RPN} for detecting text instances.

$$\left.\begin{aligned} \Psi^* &= \pi \otimes \Psi \\ \psi_1^* &= \pi \otimes \psi_1 \\ \psi_2^* &= \pi \otimes \psi_2 \\ \psi_3^* &= \pi \otimes \psi_3 \end{aligned}\right\} \tag{3}$$

where $\otimes$ represents element-wise multiplication in each channel.

For training, the proposed method uses samples of the DeepFashion2 dataset [28] for clothing detection, and for text detection, it uses samples of the ICDAR 2015 dataset [29]. In other words, we follow across dataset validation procedure for training and testing in this work. As a result, we do not use samples from the training dataset when testing the method. We believe this leads to an approach that is more robust. It is noted from the results shown in Fig 4, where one can notice the effect of deformable clothing region on text. Those examples illustrate the way the method works.

Overall, the key idea is to avoid conventional pipeline approaches for text detection by removing unnecessary background information. This leads to a heavy dependence on a prior module to RPN network. Instead, the proposed method uses an attention mechanism that overcomes this limitation. In other words, we can conclude that F_{atten} acts like a guiding signal for F_{RPN} such that it focuses on clothing regions as compared to the background, as shown in Fig. 5, where the results of intermediate steps of the backbone ($\mathrm{F_{CNN}}$) are presented. As a result, where there is clothing, the pixels are sharpened by suppressing other pixels as shown in Fig. 5(a). The effect of attention network (ResNet-16 + PPM) can be seen in Fig. 5(b), where these regions are detected well. This helps PSENet to detect text on deformable clothing region. This is illustrated Fig.5(c) of respective intermediate results in Fig. 5(a).

IV. EXPERIMENTAL RESULTS

Experimental study includes dataset creation because there is no standard dataset for text detection in sports images. To show the effectiveness of the proposed method, we also test it on the benchmark datasets of marathon, namely, RBNR [23], MMM [5] and R-ID [6] as these images are part of sports

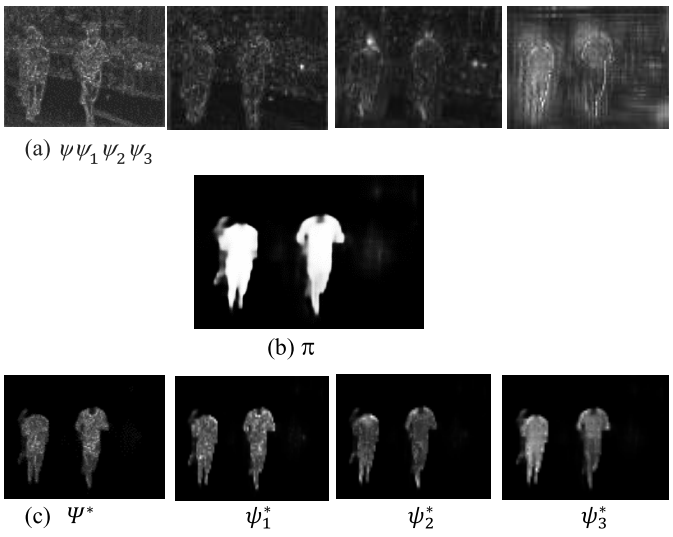

Fig. 5. Spatial attention network guiding F_{RPN} for clothing region detection. (a) Intermediate results of backbone, F_{CNN}, (b) Generated by spatial attention map and (c) Rectified feature maps of respective intermediate results in (a) are fed to FRPN to detect text instances.

images where we can see bib number and name of the runner on the jersey. In addition, since the scope of the proposed work to detect text in sports images, we choose sports images which are available in benchmark natural scene datasets of CTW1500 and MS-COCO [19] for experimentation.

A. Dataset Creation and Evaluation

We create a dataset by collecting images from different sources, namely, soccer, tennis, cricket, marathon, internet, YouTube and our own collections. This dataset comprises images affected by pose variations, camera viewpoints and variations on distances between target and camera. Due to the above challenges, one can expect images suffering from perspective distortion due to text on non-rigid material like uniform or jersey. In addition, sports images contain both scene texts and bib numbers especially for athletes. Such variations increase the complexity of the dataset.

Similarly, in order to show the robustness of the proposed method, we consider three standard datasets, namely, RBNR [23], MMM [5] and R-ID [6], which provide marathon images, in which we can see bib numbers prominently. In the same way, to test the utility of our method, we also collect images that contain human bodies with text information from the benchmark datasets of natural scene images, namely, CTW1500 [19] and MS-COCO Text [19]. Since these two datasets are constructed for text detection in natural scene images but not sports images, the sports images chosen from these two datasets are challenging for achieving high results compared to the others. More details of our and the benchmark datasets are listed in Table II, where one can see images with different ranges of resolution in different datasets. In total, 22392 images are considered for experimentation in this work. Sample images for each dataset are shown in Fig.6, where it can be noted that images of each dataset have their own complexities according to the nature of the datasets. It can be seen

TABLE II
DETAILS OF DIFFERENT DATASETS CONSIDERED FOR EXPERIMENTATION

Datasets	Number of images	Resolution	
		Min	Max
Our Data-Sports and Marathon	13200	150 ×150	1280 × 720
RBNR Data-Marathon [23]	217	342 × 479	1260 × 850
Shivakumara et al-Marathon [5]	212	318 × 479	2939 × 1959
Re-ID [6]	8706	300 × 300	1200 × 800
CTW1500 [19] (subset)	33	620 × 437	1728 × 2592
MS-COCO Text [19] (subset)	24	401 × 375	640 ×x 640
Total	22392		

from Table II that our dataset includes images of low resolution starting from 150×150. Therefore, our dataset provides images with large variations in resolution and contrast.

To evaluate the performance of the proposed method, we use the standard measures, namely, Recall (R), Precision (P) and F-Measure (F) as defined in Equation (4), Equation (5) and Equation (6), respectively. In case of F-measure as defined in Equation (6), the value of α is fixed at 0.5 according to the instructions given in [5], [6], [18].

$$P = \frac{T_p}{T_p + F_p} \tag{4}$$

$$R = \frac{T_p}{T_p + F_n} \tag{5}$$

Here, T_p signifies the total number of texts detected correctly by the proposed method, F_p signifies the total number of the texts detected falsely, and F_n is the total number of texts that are missed.

$$F = \frac{P.R}{\beta.R + (1-\beta).P} \tag{6}$$

To show the effectiveness of our approach, we implement the three state-of-the-art methods that focus on bib number detection in marathon images for comparative studies in this work. For example, Ami *et al.*'s method [23], which uses the combination of face, and torso for bib number detection from Marathon images, Shivakumara *et al.*'s method [5], which detects torso without detecting face for bib number detection in marathon images, and Kamlesh *et al.*'s method [6], which uses text detection and recognition for person Re-Identification (Re-ID) in marathon images. We also implement the recent method [24], which is developed for jersey number detection in sports images. It detects human body parts as a preprocessing step for achieving results. To show that the methods developed for scene text detection in natural scene images may not work well for sports images, we use the recent deep learning based methods for comparative studies, that is, Long *et al.*'s method [3], which proposes TextSnake with flexible representation for text detection in natural scene images, and Wang *et al.*'s method [4], which proposes PESNet to overcome the problems of pixel wise segmentation based approaches for text detection in natural scene images. In the same way, to show the methods which use Mask-R-CNN for text detection are not effective in the case of deformed clothing regions, we compare our approach with the method of Lyu *et al.* [18], which was developed for natural scene images. The

reason to consider these three methods for comparative studies is that they address the challenges like arbitrary orientation, complex background, low contrast and low resolution, which are common for sports images.

In this work, for Episodic training, we determine the following values empirically with labeled samples. The same set up and values are used for all the experiments in this work for evaluation. The learning rate is of 0.02 and stochastic gradient descent optimizer for a total of 31122 iterations. F_{CNN}, F_{atten} is trained for $N_{cloth} = 73$ iterations followed by a second episode of training to optimize F_{CNN}, F_{RPN} for $N_{text} = 73$ iterations. The decay for learning rate is set to 0.0004. The value of λ_{pse} is set to 0.7, and λ_{dsup} is give a value of 0.4. For experiments on all six datasets including our own, the existing methods were trained with samples from the same dataset used for testing. So we could perform a comprehensive study across all of the datasets, and we programmed our own implementations of each of the methods compared here.

B. Ablation Study

In our method, as discussed in the Proposed Methodology Section, ResNet-18 is used for generating spatial attention on clothing and then PSENetis used for text detection. PSENet and ResNet-18 are the improved versions of baseline models, namely, VGG-16, Spatial Pyramid Pooling (SPP) Network [26]and Residual Network (ResNet-50) [20]. To study the effects of the proposed approach, we conduct experiments for both the basic models and the modified models used in the proposed work on our dataset as reported in Table III. For the first experiment, we use VGG-16 and PPM for clothing detection and PSE for text detection. This experiment shows poor results compared the proposed architecture. The reason for this is the inherent limitation of the VGG-16 architecture, which does not extract deep features and is more expensive to implement. Combination of VGG-16+PPM+PSE is not effective for handling the text detection challenges we are interested in. Note that ResNet outperforms VGG-16+PPM+PSE. For the second experiment, we use Pyramid Scene Parsing (PSPNet) [30] for clothing segmentation upon which text detection is performed using PSENet, which is used in [4] for text detection. It can be observed that the model does not achieve better Recall compared to the proposed model. However, the Precision is the same as the proposed model. This can be attributed that the model misses clothing regions in sports images and hence, PSENet fails to detect some text instances in sports images. Therefore, one can infer that the proposed combination is better than PSPNet + PSENet models especially in terms of Recall and F1-score.

Similarly, to show that ResNet-18 is better than ResNet-50 and contributes more for achieving better results, we conduct experiments for the combination of ResNet-50 + PPM + PSE and the proposed model, ResNet-18 + PPM + PSE as reported in Table III. Here, ResNet-50 and PPM are used for clothing segmentation, while PSE is used for text detection from detected clothing regions. The former combination achieves better Recall and F1-score compared to PSPNet + PSENet, while Precision drops compared to the proposed model and the model PSPNet + PSENet. Therefore, one can conclude that ResNet-18 has the ability to balance both detecting text instances and producing fewer false positives as it is evident from the F1-score, which is almost the same for both the combinations. Overall, we can confirm that PSPNet, PPM and RestNet have the ability to cope with the challenges of text detection in sports images.

TABLE III

ASSESSING THE CONTRIBUTION OF PSPNET AND RESNET-18 + PSENET FOR TEXT DETECTION ON OUR DATASET

Steps	Precision	Recall	F1-Score
VGG-16 + PPM + PSE	0.75	0.77	0.760
PSPNet[30] +PSENet	0.93	0.79	0.854
ResNet-50+PPM+PSE	0.91	0.85	0.878
Proposed ResNet-18+PPM+PSE	0.93	0.82	0.871

Since the clothing of uniform dress code or jersey detection is the key step of the proposed model, sample qualitative results of the proposed model on clothing detection for images of different datasets in Fig. 7(a) are shown in Fig. 7(b). It can be noted from the results in Fig. 7(b) that the proposed combination of ResNet-18+PPM+PSE works well for images of different deformable clothing reasons.

C. Evaluating the Proposed Text Detection

Qualitative results of the proposed model for text detection in sports and marathon images of different datasets are shown in Fig. 8. It is observed from Fig. 8 that our method detects texts quite well for images affected by of different complexities. This shows that the proposed approach is capable of handling challenges of sports and marathon images.

Quantitative results of the proposed and the existing methods for our and benchmark datasets of marathon and natural scene images are reported in Table IV. It is noted from Table IV that the proposed model is the best at Precision and F1-score for all the datasets including our and natural scene text datasets except MMM [5] compared to the existing methods. However, it reports poor Recall compared the existing methods because it misses text instances when an image contains too small fonts and occlusions. On the other hand, the existing methods, namely, Shivakumara *et al.*'s [5], Nag *et al.*'s [24], Long *et al.*'s [3], and Wang *et al.*'s [4] achieve higher Recall for different datasets compared to the proposed model. The methods in [5], [24] used face, torso and human body parts detection to reduce the complexity of the problem such that text detection step does not miss text instances. Similarly, since the methods [3], [4] are developed for text detection in natural scene images and explore deep learning models, the methods do not miss text instances. However, the above existing methods are the worst for Precision compared to the proposed model. This infers that the existing methods [3], [4], [5], [24] produce more false positives for sports images, and hence Precision is low compared to the proposed model. The main reason for the poor Precision of the existing methods [3], [4] is that the methods detect human body parts as text. However, the existing methods [5], [24] work well when an image provides complete faces and human body parts without missing due to occlusion. Otherwise,

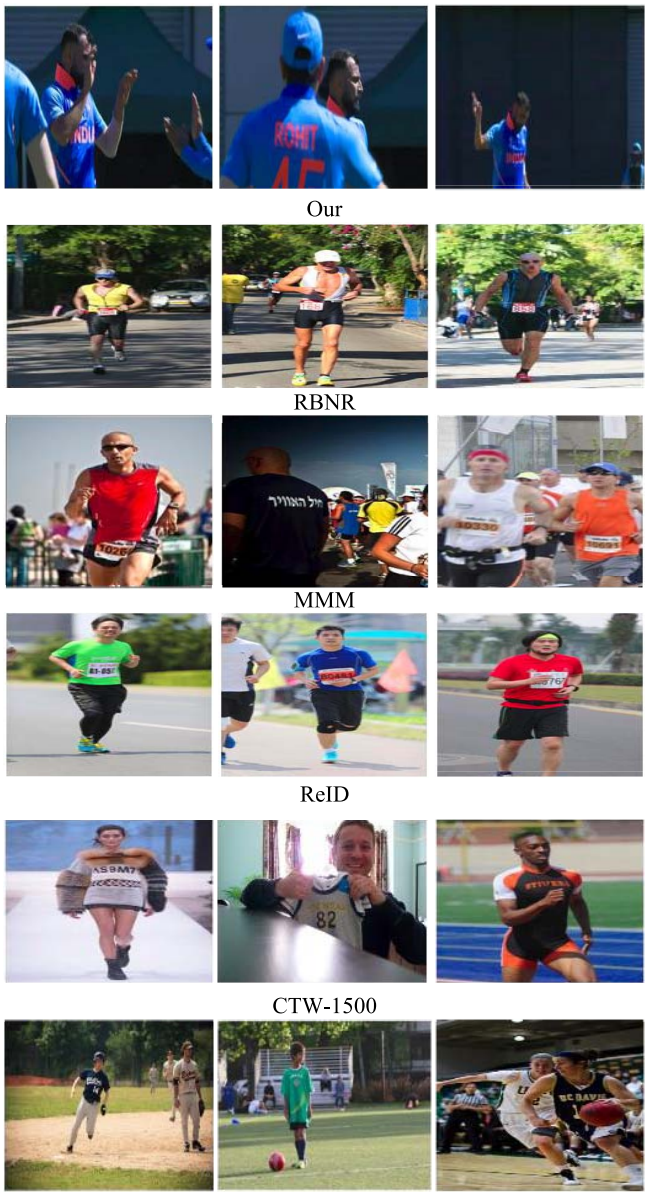

Fig. 6. Sample images of our, marathon and natural scene datasets.

(a) Input images of our, RBNR and MMM datasets.

(b) Uniform dress code and jersey detection of the proposed model

Fig. 7. Examples of cloth detection for images of different datasets by the proposed model.

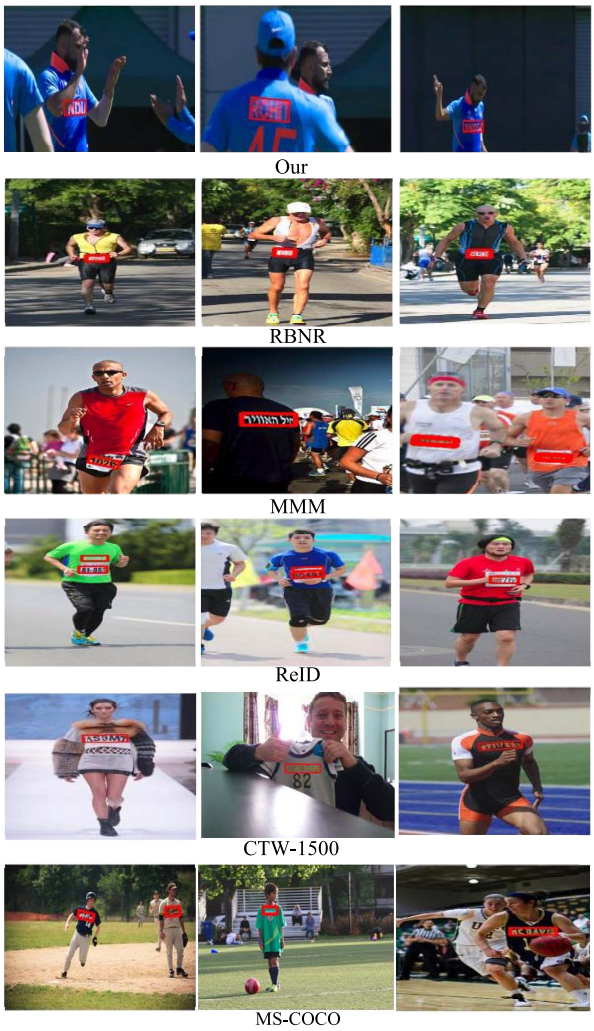

Fig. 8. Text detection of the proposed model on different datasets.

the methods detect falsely objects as faces and human body parts, which lead to more false positives for sports images.

Interestingly, it is observed from Table IV that the methods [5], [6], [23] developed for bib number detection in marathon images score poor results compared to the other and the proposed methods for almost all the datasets especially F1-score. Therefore, we can understand that the methods do not have the ability to handle the challenges of sports images, other than marathon and images chosen from natural scene datasets. The key idea of obtaining better results for the proposed model is clothing detection directly without detecting any other parts of human body including face, torso and skin.

Table IV shows that the proposed and the existing methods report poor results for CTW15000 and MS-COCO datasets compared to marathon and our sport datasets. This is because these two datasets provide small number of samples for learning the parameters of our model. However, Precision and F1-score of the proposed model are better than those of the existing methods for CTW1500 and MS-COCO natural scene datasets. When images are affected by longer distances between cameras and targets, the proposed model does not

TABLE IV
PERFORMANCE OF THE PROPOSED AND EXISTING METHODS FOR TEXT DETECTION ON DIFFERENT DATASETS

Methods	RBNR [23]			MMM [5]			Re-ID [6]			CTW-1500 (subset)			MS-COCO (subset)			Our Dataset		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Ami et al. [23]	0.53	0.40	0.45	0.37	0.38	0.38	0.67	0.63	0.65	0.30	0.33	0.31	0.20	0.21	0.24	0.46	0.50	0.48
Shivakumara et al. [5]	0.64	**0.88**	0.74	0.60	0.74	0.66	0.71	0.79	0.75	0.39	0.40	0.40	0.29	0.21	0.24	0.62	0.59	0.60
Kamlesh et al. [6]	0.70	0.72	0.71	0.74	0.73	0.73	0.90	0.85	0.87	0.72	0.64	0.68	0.34	0.29	0.31	0.81	0.76	0.78
Lyu et al. [18]	0.75	0.69	0.72	0.72	0.66	0.69	0.81	0.75	0.78	0.70	0.64	0.67	0.33	0.31	0.32	0.74	0.68	0.71
Long et al. [3]	0.65	0.78	0.71	0.64	0.76	0.70	0.78	0.82	0.80	0.64	0.73	0.68	0.32	**0.34**	**0.33**	0.73	0.72	0.73
Nag et al. [24]	0.80	0.77	0.78	0.83	**0.79**	**0.81**	0.91	**0.93**	**0.92**	0.72	0.64	0.68	0.34	0.29	0.31	0.84	**0.85**	0.85
Wang et al. [4]	0.89	0.67	0.76	0.86	0.59	0.70	0.79	0.85	0.82	0.66	**0.73**	0.69	0.35	0.29	0.32	0.83	0.82	0.81
Proposed Method	**0.93**	0.69	**0.79**	**0.91**	0.65	0.76	**0.97**	0.87	**0.92**	**0.83**	0.61	**0.70**	**0.36**	0.31	**0.33**	**0.93**	0.82	**0.87**

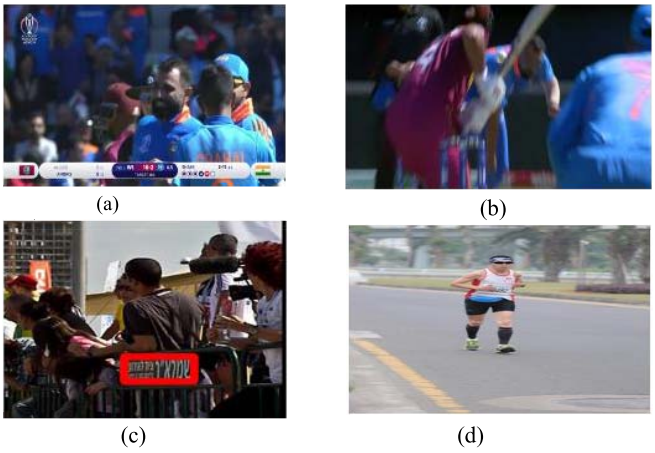

Fig. 9. Examples of some failure cases of the proposed method. (a) the player name is occluded by a scoreboard, (b) image suffering from server blur and occlusion, (c) text in a banner is confused as clothing text and (d) text occluded by the runner's.

work well. This is the limitation of PSENet used in the proposed work. To alleviate this limitation, the proposed method can combine an enhancement model through super resolution concept with the proposed text detection model, which is beyond the scope of the proposed work and plan to handle in future.

There are still some limitations with our proposed method because sports images present many challenges. When text is occluded by parts of the body or other objects, as shown in Fig. 9(a), the proposed method does not perform well. In this case, it loses key features. Similarly, when the image suffers from severe blur as shown in Fig. 9(b), the proposed method also does not work well. When the features extracted for clothing detection overlap with background objects (other than clothing) as shown in Fig. 9(c), the proposed method misclassifies them as text. Similarly, the presence of haze and occlusion of text as shown in Fig. 9(d), the performance of the proposed method degrades. These challenges are beyond the scope of the proposed work and thus there is a scope for the improvement of the proposed work in near future.

V. CONCLUSION AND FUTURE WORK

We have proposed a novel episodic learning-based architecture for text detection in sports images. It unifies Residual Network and Region Proposal Networks as a single architecture for clothing of uniform dress code and jersey detection, and then it uses PSENet for text detection. Unlike most existing methods that use face, skin, torso and parts of human body for reducing background complexity of text detection in sports images, the proposed model uses deformable clothing region for text detection. This is because face, skin, torso and other parts of the human body may disappear due to occlusions, while clothing may not disappear completely in case of sports. The proposed model employs Episodic training for unifying the features extracted from input images using ResNet for clothing region segmentation, and the feature extracted from clothing region using PSENetof RPN for text detection. Experimental results on our dataset and benchmark datasets containing marathon runners and sports images chosen from natural scene datasets show that the proposed approach outperforms existing methods in terms of precision and F-measure. To the best of our knowledge, this is the first work the uses an athlete's clothing as a means for text detection. While robust in many cases, our method performs badly for certain very challenging images; this is a subject for future research.

ACKNOWLEDGMENT

The authors thank anonymous reviewers and editor for their constructing comments and suggestions to improve the quality and clarity of the work.

REFERENCES

[1] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1342–1355, Nov. 2008.

[2] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 838–849, Jun. 2012.

[3] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. ECCV*, 2018, pp. 19–35.

[4] W. Wang *et al.*, "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9328–9337.

[5] P. Shivakumara, R. Raghavendra, L. Qin, K. B. Raja, T. Lu, and U. Pal, "A new multi-modal approach to bib number/text detection and recognition in Marathon images," *Pattern Recognit.*, vol. 61, pp. 479–491, Jan. 2017.

[6] P. Xu, Y. Yang, and Y. Xu, "Person re-identification with end-to-end scene text recognition," in *Proc. CCCV*, 2017, pp. 363–374.

[7] B. Bataineh, S. N. H. S. Abdullah, and K. Omar, "A novel statistical feature extraction method for textual images: Optical font recognition," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5470–5477, Apr. 2012.

[8] S. M. S. Ismail, S. N. H. S. Abdullah, and F. Fauzi, "Detecting and recognition via adaptive binarization and fuzzy clustering," *J. Sci. Technol.*, vol. 27, no. 4, pp. 1759–1781, 2019.

[9] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.

[10] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[11] W. Feng, W. He, F. Yin, X. Y. Zhang, and C. L. Liu, "TextDragon: An end-to-end framework for arbitrary shaped text spotting," in *Proc. ICCV*, 2019, pp. 9076–9084.

[12] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9365–9374.

[13] K. S. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal, and T. Lu, "Multi-Script-oriented text detection and recognition in video/scene/born digital images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1145–1162, Apr. 2019.

[14] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "TextField: Learning a deep direction field for irregular scene text detection," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, Nov. 2019.

[15] Y. Cai, W. Wang, Y. Chen, and Q. Ye, "IOS-net: An inside-to-outside supervision network for scale robust text detection in the wild," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107304.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2980–2988.

[17] Z. Huang, Z. Zhong, L. Sun, and Q. Huo, "Mask R-CNN with pyramid attention network for scene text detection," in *Proc. ICCV*, 2019, pp. 764–772.

[18] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. ECCV*, 2018, pp. 71–78.

[19] S. Roy, P. Shivakumara, U. Pal, T. Lu, and G. H. Kumar, "Delaunay triangulation based text detection from multi-view images of natural scene," *Pattern Recognit. Lett.*, vol. 129, pp. 92–100, Jan. 2020.

[20] S. Wang, Y. Liu, Z. He, Y. Wang, and Z. Tang, "A quadrilateral scene text detector with two-stage network architecture," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107230.

[21] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "ABCNet: Real-time scene text spotting with adaptive bezier-curve networ," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9809–9818.

[22] C. Wang, H. Fu, L. Yang, and X. Cao, "Text co-detection in multi-view scene," *IEEE Trans. Image Process.*, vol. 29, pp. 4627–4642, 2020.

[23] I. B. Ami, T. Basha, and S. Avidan, "Racing bib number recognition," in *Proc. BMCV*, 2012, pp. 1–12.

[24] S. Nag, R. Rmachandra, P. Shivakumara, U. Pal, T. Lu, and M. Kankanhalli, "CRNN based jersey number/text recognition in sports and Marathon images," in *Proc. ICDAR*, 2019, pp. 1149–1156.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[27] D. Li, J. Zhang, Y. Yang, C. Liu, Y. Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proc. ICCV*, 2019, pp. 1446–1455.

[28] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, "A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing ing images," in *Proc. CVPR*, 2019, pp. 5337–5345.

[29] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. ICDAR*, 2015, pp. 1156–1160.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. CVPR*, 2017, pp. 6230–6239.