Player Identification in Different Sports

Ahmed Nady[1] and Elsayed E. Hemayed[2,3]
[1]*Department of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt*
[2]*Department of Computer Engineering, Faculty of Engineering, Cairo University, Giza,12613, Egypt*
[3]*Zewail City of Science and Technology, University of Science and Technology, Giza 12578, Egypt*
ahmednady@fci.helwan.edu.eg, hemayed@ieee.org

Keywords: Jersey Number Recognition, Player Identification, Sports Video Analysis.

Abstract: Identifying players through jersey numbers in sports videos is a challenging task. Jersey number can be distorted and deformed due to variation of the player's posture and the camera's view. Moreover, it varies in font and size due to the different sports fields. In this paper, we present a deep learning-based framework to address these challenges of jersey number recognition. Our framework has three main parts. Firstly, it detects players on the court using state of the art object detector YOLO V4. Secondly, each jersey number per detected player bounding boxes is localized. Then a four-stage scene text recognition is employed for recognizing detected number regions. A benchmark dataset consists of three subsets is collected. Two subsets include player images from different fields in basketball sport and the third includes player images from ice hockey sport. Experiments show that the proposed approach is effective compared to state-of-the-art jersey number recognition methods. This research makes the automation of player identification applicable across several sports.

1 INTRODUCTION

In recent years, automated sports video analysis has attracted a lot of attention especially in team sports such as ice hockey, basketball, soccer and volleyball due to the increasing demand by sports professionals and fans for extracting semantic information. Sports analysis results can be used in several applications such as storytelling on TV, adapting the training plan, game statistics generation, and evaluation of strengths or weaknesses of a team or a player. Sports video analysis includes ball and players' detection in each frame then their tracking over time and analysis of their interactions. Tracking multiple players is challenging due to the players' similar appearance within the team, occlusion, and players' complicated motion patterns. In the tracking phase, tracks may be lost and new tracks may be created throughout a game and tracking identities can be switched. Thus, player identification represents a major research challenge to realize the advantages of automatic sports analysis. Player identification includes linking the actual player to each track and associating it with his actions and statistics.

Player identification in broadcast sports video is challenging due to low video resolution, viewpoint and camera movements, players pose, illumination conditions, variations of sports fields and jerseys. The features that are employed for player identification on the court are face and jersey numbers. The approaches that rely on face recognition for identification operates in close-up shots where the player face appears clearly and became infeasible for overview shots. The other visual cue which being generic across sports is jersey number. Since jersey numbers occupy a large part of player back uniform and the rising of HD sports videos, the approaches that depend on jersey numbers are promising. The challenges of jersey number recognition are not limited to player tilting, motion blur and viewing angle but also include the distractions inside or surrounding the playground such as clocks, commercial logos and banners (Liu and Bhanu, 2019).

The past studies for jersey number recognition are grouped into two classes: Optical Character Recognition (OCR) based methods (Messelodi and Modena, 2013; Lu et al., 2013a; Šari et al., 2008) and Convolution Neural Networks(CNN) based methods (Gerke et al., 2015; Li et al., 2018). The former class employs hand-crafted features to localize text/number regions on the player uniform then the segmented regions are passed to OCR module to recognize the text/number.

The flaw of this class of methods is that the performance was not good enough. The latter class has no explicit localization of the jersey number. Moreover, the scope of these methods is limited to a specific sport such as soccer sport or basketball sport and are not tested on different sports such as ice hockey where the jersey number is bulky.

In this paper, we propose a compound deep neural network for player identification through jersey numbers across both games and sports. The proposed framework comprises three phases. In the first phase, players are detected using YOLO V4 (Bochkovskiy et al., 2020). In the second phase, the jersey number are detected using a fine-tuned Character Region Awareness for Text Detection(CRAFT) (Baek et al., 2019b) which is a character-level text detector that ensures a high level of flexibility in detecting involved scene text images such as arbitrary-oriented and distorted text. The third phase is responsible for the recognition of the jersey number regions using the scene text recognition model (Baek et al., 2019a). Similar works to the proposed framework were proposed by Nag et al. (Nag et al., 2019) and Wang et al. (Wang and Yang, 2020) in which they utilized the scene text detection and recognition in their work for runners bib number recognition. The bib number is easier to be detected because of its horizontal orientation, less variation in font stroke size, and the distinguishing appearance that results from number existence on pure color background. Therefore, the performance of these methods cannot be satisfactory for jersey number recognition.

The Contributions of This Work Are Listed as Follows:

1. Proposing a new framework for player identification that achieve high accuracy rate even across different sports.
2. Performing a transfer learning and fine-tuning character region awareness for text detection (CRAFT) (Baek et al., 2019b) for sports jersey numbers to account for player tilting, shirt deformation, sports fields and font of jersey numbers variations.
3. Adapting the scene text recognizer to address the challenge of not having a dataset of all possible jersey numbers.
4. Developing a benchmark dataset composed of three subsets in which the first subset contains 1872 basketball player images, the second subset includes 851 basketball player images but in a different arena and the third subset for ice hockey sport with 1317 player images. All images in the first subset are annotated with the jersey number bounding boxes and its class whereas the other subsets images are annotated with solely its class. We call this dataset Sports Jersey Number dataset (S^2JN).

The rest of the paper is organized as follows. Section 2 reviews the related work of player identification. Section 3 presents the proposed framework. Section 4 presents the sports jersey number dataset. The experimental results are presented and discussed in Section 5, followed by conclusions in Section 6.

2 RELATED WORK

Player recognition is one of the key components in automatic sports video analysis. The approaches of player identification can be placed into three categories: face recognition, jersey number recognition and person Re-Identification. Jersey number recognition can be further classified into two main groups: OCR-based and CNN-based approaches. Others have formulate the player identification as a person re-identification problem.

For OCR-based approaches, Messelodi et. al. (Messelodi and Modena, 2013) detect name or number on athlete's bib using prior knowledge about text background color and recognize candidate regions through OCR system. Lu et. al. (Lu et al., 2013a) locate jersey number regions in detected player bounding box in basketball videos by means of gradient difference and then adapt OCR scheme for recognition. Šari et. al. (Šari et al., 2008) precede the OCR module by localizing the number regions in HSV color space based on internal contours. The preceding OCR-based works have applicability limitations in wide circumstances because of adapting manually designed features.

For CNN-based approaches, Gerke et al. (Gerke et al., 2015)classify the cropped upper part of the soccer player bounding boxes using convolutional neural network architecture that composes three convolutional layers and three fully connected layers. Their finding showed that notably improved performance of number recognition compared to previous researches (Messelodi and Modena, 2013; Lu et al., 2013a; Šari et al., 2008). Misclassifications happen usually in classes (jersey numbers) that share at least one digit. The holistic number approach in which each number modelled as a separate class is better than a digit-wise approach where each digit is classified by a separate classifier. Li et al (Li et al., 2018) fuse the CNN model with spatial transformer network (STN) that brought attention and transformation to the number's region in the soccer player bounding boxes. They do not crop

Input image | detected players | cropped players | number detection | number recognition

Figure 1: Structure of proposed framework.

the upper half of the bounding box as input to CNN but they utilized STN for this purpose.

The digit wise approach (Gerke et al., 2015; Li et al., 2018) has a difficulty in separation of jersey number digits and the variability of camera perspective may make it more severe. Liu et al (Liu and Bhanu, 2019) proposed a joint framework that is based on faster R-CNN for player detection and jersey number recognition. They tackled the challenges of player pose and view-point variations associated to jersey numbers through a pose-guided regressor that utilizes prediction of player body key points. They designed Region Proposal Network (RPN) which produces candidate bounding boxes for background, player or digit and then associate person and digits proposals keeping solely digits' proposal that reside in person proposal. Their dataset is collected with pan and zoom which limits the camera field of view and makes the jersey numbers appear clearly and this is not the case in a broadcast video of soccer videos. Their attempt for model generalization showed good detections but the number classification performance is degraded due to font size variation.

For person ReId approaches, the work in (Lu et al., 2013b; Senocak et al., 2018) formulate the player identification in a broadcast basketball video from a medium distance as a person re-identification problem where they recognize players from the entire body. Lu et al. (Lu et al., 2013b) make use of a mixture of maximally stable extremal regions (MSER) (Matas et al., 2004), SIFT features (Lowe, 2004), and color histogram features to form the player representation and then a logistic regression classifier is used for classification. Senocak1 et. al. (Senocak et al., 2018) model the player presentation by merging the deep convolutional representation from the entire player image at multi-scale and player parts. Player Re-Id approaches are not scalable across games and across sports where each player to identify must be included in the training dataset. Moreover, the jersey should be unified across all matches and this is difficult to achieve.

3 PROPOSED FRAMEWORK

In this section, we describe the proposed neural network model that detects and recognize jersey numbers across both games/matches and sports. Figure 1 shows the three main steps of our framework: sports player detection, text/number detection and text recognition. In the first step, object detector based on YOLO V4 (Bochkovskiy et al., 2020) is utilized to detect the players on the court then the text detector locates the jersey number region on each player in the second step. Finally, the detected candidate regions are recognized in the text recognition task. Details about the text detection and text recognition are provided in the following sections.

3.1 Scene Text Detection

Scene text detection has witnessed a huge development in the last years. The methods based on deep learning have shown promising results. Baek et al. (Baek et al., 2019b) introduced a scene text detection method through localizing character regions and linking these regions in a bottom-up manner. The method can detect text of various shapes such as horizontal, curved and arbitrary-oriented text. Motivated by the method's state-of-the-art performance and generalization ability, we adapted it to detect numbers on player T-shirt whether from back or front. The model architecture consists of a backbone network, which is VGG16-BN, and a decoding part in which the low-level features are aggregated. The model output is 2-channel score maps: region score that locates every character in image and affinity score that link successive characters into a single instance. The loss function L is defined as follows:

$$L = \sum_{p} \|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2) \quad (1)$$

where $S_r^*(p)$ and $S_a^*(p)$ indicate the ground truth region score and affinity map, respectively, and $S_r(p)$ and $S_a(p)$ indicate the predicted region score and affinity score, respectively. There could be other text instances than jersey number printed on player's shirt

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

Figure 2: Illustration of the used image augmentation techniques. (a) original player image (b-e) image after applying scaling, rotation, color manipulation and Gaussian blur respectively.

such as player name and its club. During inference, much of such text instances can be filtered and eliminated based on the aspect ratio where the aspect ratio of jersey number whether consisting of one-digit or two-digit is lower than 1.5 even for player pose tilt situations.

3.1.1 Implementation Details

In our implementation, the weights of CRAFT detector are initialized by the use of the general pre-trained model and then is trained with the first subset of S^2JN dataset to take into account the distortion of the number printed on the player's shirt. The first subset is splitted into training set containing 1274 player images, validation set having 317 player images and the remaining 281 player images are used for testing. Because of the lack of CRAFT (Baek et al., 2019b) training code, we supervised the training by providing the annotations for each digit in jersey numbers.

The model is trained for 35 epochs with a learning rate set to 3.2768e-5 and batch size set to 8 on image size 224 * 224. During training, the image augmentation technique is used by applying affine transformation, Gaussian blur and colour channels manipulation to both original player image and corresponding number b-box as shown in Figure 2.

The other two subsets are used for testing to validate our hypothesis that is the proposed method is generalized across games and sports. At testing phase, the value of the text confidence threshold, link confidence threshold and text low-bound score are set to 0.1. Different settings for input image size are utilized in experimentations.

3.2 Scene Text Recognition

The image sequence prediction techniques developed by Beak et al. (Baek et al., 2019a) has promising accuracy results and is able to recognize the number as a whole. Thus, it overcomes the difficulty of dividing a two-digit jersey number that is difficult to do due to non-up-frontal views and distortion. Beak et al. (Baek et al., 2019a) present a four-stage scene text recognition framework that most present STR models fit into. The first stage in the framework is transformation that employs the thin-plate spline (TPS) transformation to normalize the input text image. The second stage is feature extraction that extracts visual features from input or normalized image using CNN. The third stage is sequence modelling that uses Bidirectional LSTM (BiLSTM) to capture the contextual information within the sequence of features that were extracted in stage 2. The fourth stage is prediction that predicts the character sequence from the identified features of an image. Beak et al. (Baek et al., 2019a) implemented two methods of prediction module: Connectionist Temporal Classification (CTC) and Attention mechanism (Attn).

In CTC, The conditional probability is computed by summing the probabilities of all π that are mapped by M onto Y, as in equation 2

$$P(Y \mid H) = \sum_{\pi : M(\pi) = Y} P(\pi \mid H) \tag{2}$$

where Y is the label sequence, H is input sequence and $P(\pi \mid H)$ is the probability of π defined as

$$P(\pi \mid H) = \prod_{t=1}^{T} y^t_{\pi_t} \tag{3}$$

where $y^t_{\pi_t}$ is the probability of observing π_t which is either a character or a blank (-) at timestamp t. During inference, the greedy decoding scheme is adopted by taking character π_t with highest probability at each time step t, and map the π_t onto Y

$$Y^* \approx M(arg\,max_\pi P(\pi \mid H) \tag{4}$$

In attention mechanism, the output y_t at time step t is predicted using LSTM attention decoder as follows:

$$y_t = softmax(W_0 s_t + b_0) \tag{5}$$

where W_0, b_0 are trainable parameters and s_t represents the decoder LSTM hidden state at time step t and is defined as

$$s_t = LSTM(y_t - 1, c_t, s_t - 1) \tag{6}$$

and c_t is a context vector and defined as

$$c_t = \sum_{i=1}^{I} \alpha_{ti} h_i \tag{7}$$

where $\alpha_t i$ is attention weight.

In our implementation, we used the pre-trained model for TPS-ResNet-BiLSTM-Atten text recognition framework (Baek et al., 2019a).

Table 1: First subset (training set) statistics. W, H, w and h are player image width, player image height, number b-box width and number b-box height respectively. Std is the standard deviation. The used unit is pixel.

	W	H	w	h
Mean	89.73	210.37	23.51	26.92
Std	24.06	44.95	6.70	5.94

4 SPORTS JERSEY NUMBER DATASET

To appraise the efficiency of the proposed compound neural network model, we performed experiments on the introduced **S**ports **J**ersey **N**umber dataset (S^2JN), since there is no publicly available dataset for jersey numbers. S^2JN dataset has three different subsets. Yolov4 object detector is used to detect players in each of subset videos. The first subset contains 1872 basketball player images that are extracted from set of video clips (cameras number 2, 5 and 8) from SPIROUDOME dataset. The video resolution is 1600 * 1200 and the framerate is 25 fps. The jersey number bounding boxes (b-box) annotations and its class per player b-box are provided. The first subset statistics are illustrated in Table 1. In the second subset, 851 basketball player images from the one-minute video clip of Camera 7 APIDIS dataset sampled at 5 fps with 1600 * 1200 resolution are annotated with their identities. The third subset for ice hockey sport with 1317 player image with their class from the video clip of CANADA vs FINLAND match in Lausanne 2020 Youth Olympic Games sampled at 5 fps with 1920 * 1080 resolution. The S2JN dataset presents detected players in various cases and thus jersey number can be influenced by pose tilting, blurring and severe camera-views as shown in Figure 3

5 EXPERIMENTAL RESULTS

In this section, we presented and discussed the results obtained when using the proposed framework and comparing them to the existing state-of-the-art jersey number recognition developed by Gerke et al (Gerke et al., 2015) and Li et al (Li et al., 2018). These two methods consider only numbers on the back of the player shirt. Therefore, we removed the small-number player images during training and testing for a fair comparison.

We implemented methods (Gerke et al., 2015; Li et al., 2018) based on the details provided in their papers. In Gerke et al (Gerke et al., 2015) method, the upper part of each player b-box is converted to grayscale, then cropped and resized to 40 * 40. The used image augmentation techniques are scaling and image inverse. Without access to the dataset of (Li et al., 2018), we carried out their base network architecture. The baseline framework is composed of pre-trained general CRAFT detector followed by TPS-ResNet-BiLSTM-Atten text recognition model (Baek et al., 2019a). From Table 2, we can notice that the baseline framework outperforms both related-methods (Gerke et al., 2015; Li et al., 2018). The introduced framework accomplishes even better performance due to its robustness to player pose and camera-view variations. Figure 4. shows jersey number detection results across different sports using pre-trained CRAFT and the fine-tuned one. To evaluate the number detection quantitatively, we did an experiment using the testing set of the first subset where bounding boxes of jersey number are provided and results are shown in Table 3.

The failed cases were due to the distortion of the number, extreme pose variations, the distance between the camera and the player, in addition to the distraction in the playground such as clock and banner. In second basketball subset, there are 30 player images falsely recognized due to banner distractions as shown in Figure 5.a. These distractions can be filtered in post-processing step such as filtering detection based on aspect ratio, proving foreground mask of player besides its b-box or by providing the active player numbers. The number miss detected for 54 player images occurs in player tilting images because the number in those images is one-digit number printed on the player's shirt with a font that makes the digit appears discontinuity stroke (See Figure 5.c). By using closing morphological operation on the grayscale images, the accuracy became 87%, which enhance the method's performance with 1.72%. Adding samples for number with non-simple font strokes in various player poses especially for one-digit jersey number can achieve better performance. In ice hockey sports subset, the wide player pose variability and the bulky jersey number makes the number difficult to be detected and recognized. The recognition error results from recognizing a number either as a different number such as 1 and 7 or a sequence of characters such as 4 and y and 5 and s as in Figure 5.b.

5.1 Ablation Study

In this experiment, we investigate the following: input player image size and methods of prediction module of the four-stage text recognition. In this experiments, small-number player images are involved.

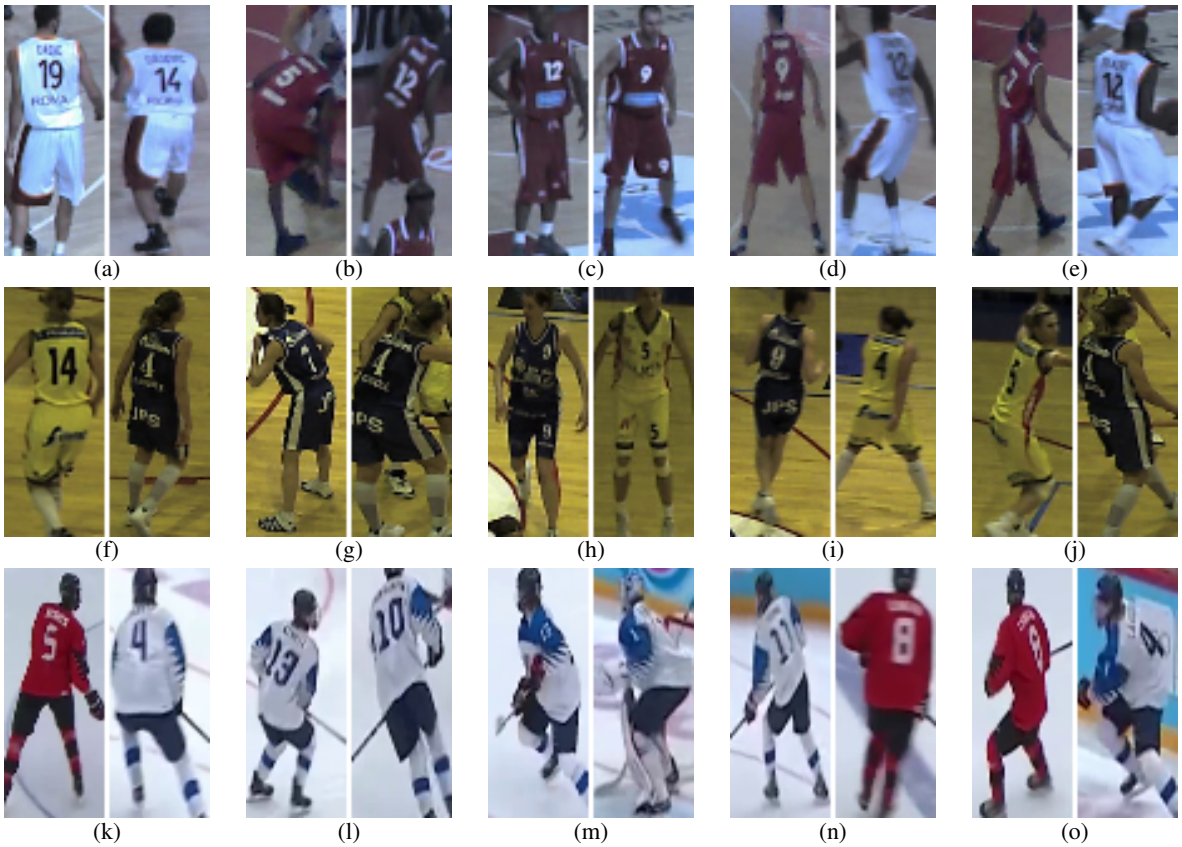


Figure 3: Illustration of S^2JN dataset. Sample Images in each row represent detected players from first subset, second subset and third subset respectively. The players can be detected in various situations: (a) (f) (k) indicate normal situations, (b) (g) (l) pose tilt, (c) (h) (m) Non back jersey numbers, (d) (i) (n) motion blur and (e) (j) (o) severe views.

Table 2: Comparison of number level accuracy among approaches.

Method	Test set of First Basketball Subset	Second Basketball Subset	Ice hockey Subset
(Gerke et al., 2015)	84.73%	40.11%	63.73%
(Li et al., 2018)	76.34%	66.76%	48.38%
baseline	65.64%	71.79%	78.49%
Our Framework	**95.41%**	**85.28%**	**85.86%**

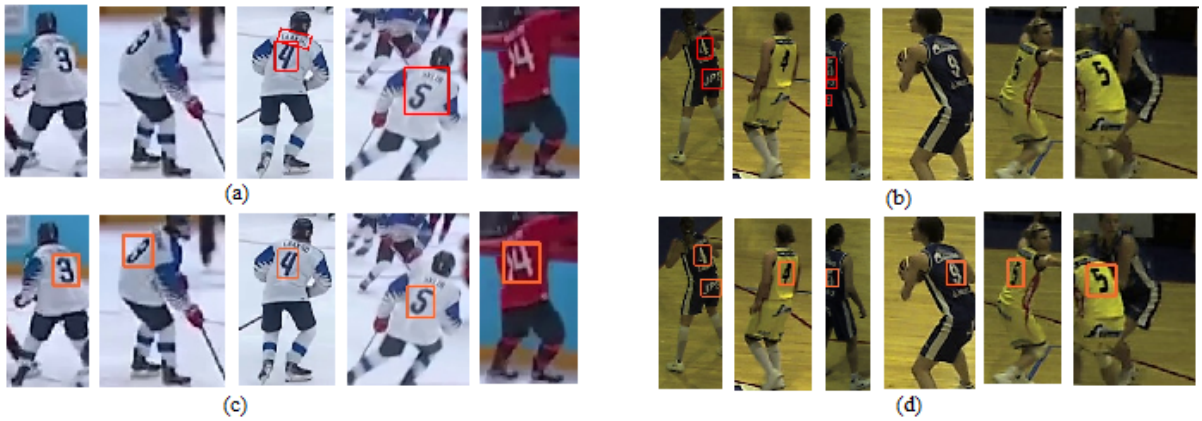


Figure 4: Number detection results on ice hockey and basketball subset using (a-b) CRAFT detector (c-d) fine-tuned CRAFT.

Table 4: Comparison of proposed method accuracy on S^2JN dataset based on longest input image side.

Longest side	Test set of First Basketball subset	Second Basketball Subset	Ice hockey Subset
160	90.74%	75.08%	**87.46%**
192	**95.01%**	82.02%	85.86%
224	93.95%	**82.49%**	83.43%
256	94.66%	81.90%	83.43%

Table 5: Comparison between attention-based and CTC-based recognizer on S^2JN subsets.

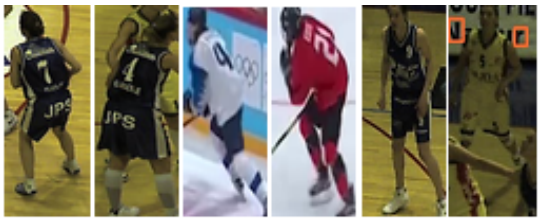
Method	Test set of First Basketball Subset	Second Basketball Subset	Ice hockey Subset
Attention-based	**95.01%**	**82.02%**	**85.86%**
CTC-based	93.59%	80.61%	84.72%

Table 3: Number detection results on testing set of first subset. R, P and H refer to recall, precision and H-mean.

Text detection method	Test set of First Subset		
	R	P	H
Pre-trained CRAFT	0.46	0.54	0.5
Fine-tuned CRAFT	**0.99**	**0.95**	**0.97**

(a) Banner distraction

(b) False recognition



(c) Miss number detection

Figure 5: Samples images for failure cases: banner distractions,font stroke with extreme pose and recognition errors.

Input Size. How to select the suitable input size for number/text detection where it may be different for each sport? We performed experiments by resizing the longest side of the player input image to 160, 192, 224 and 256. Table 4 lists the accuracy of the proposed method on three subsets of S^2JN according to the longest input side for detection. As shown in Table 4, the longest image side 192 achieves better performance on basketball sport where in ice hockey sport, the better performance is gained by the longest image side 160. Our framework accuracy in Table 4 is slightly lower than what we reported in Table 2 as we include small-number player images (Fig. 3.c, 3.h). The added images are 19 image from the first basketball subset and 117 from the second basketball subset. The miss detection is not only due to the small size of the number but also due to image blurring results from the image motion as shown in Fig. 5.c.

Is Attention-based Text Recognition Better? We need to assess the performance of our framework by replacing attention-based with CTC-based text recognition. For the experiment's setting, the longest player image side that resized to 192 is used as an input for fine-tuned CRAFT model. The pre-trained model TPS-ResNet-BiLSTM-CTC is used CTC-based recognizer. The attention-based recognizer has 1.42%, 1.41% and 1.14% gain respectively over CTC-based recognizer on testing set of the first subset, second basketball subset and ice hockey subset (see Table 5).

6 CONCLUSION

Through this work, we present a compound deep neural network for sports jersey number detection and recognition. First, our method detects jersey numbers from the detected player using fine-tuned CRAFT model. Second, the detected regions are passed to the TPS-ResNet-BiLSTM-Atten text recognition model to get a readable number/text and then keep solely number with a high probability per player image. Thanks to a state-of-the-art character-based text detector, we can detect jersey number either from the frontal part or from the back of the player's uniform. The experiments demonstrate the efficacy of our method compared with competing ones on the introduced dataset that contains player images from different arena and sports.

REFERENCES

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S. J., and Lee, H. (2019a). What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4715–4723.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019b). Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Gerke, S., Muller, K., and Schafer, R. (2015). Soccer jersey number recognition using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 17–24.

Li, G., Xu, S., Liu, X., Li, L., and Wang, C. (2018). Jersey number recognition with semi-supervised spatial transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1783–1790.

Liu, H. and Bhanu, B. (2019). Pose-guided r-cnn for jersey number recognition in sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Lu, C.-W., Lin, C.-Y., Hsu, C.-Y., Weng, M.-F., Kang, L.-W., and Liao, H.-Y. M. (2013a). Identification and tracking of players in sport videos. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 113–116.

Lu, W.-L., Ting, J.-A., Little, J. J., and Murphy, K. P. (2013b). Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767.

Messelodi, S. and Modena, C. M. (2013). Scene text recognition and tracking to identify athletes in sport videos. *Multimedia tools and applications*, 63(2):521–545.

Nag, S., Ramachandra, R., Shivakumara, P., Pal, U., Lu, T., and Kankanhalli, M. (2019). Crnn based jersey-bib number/text recognition in sports and marathon images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1149–1156. IEEE.

Šari, M., Dujmi, H., Papi, V., and Roži, N. (2008). Player number localization and recognition in soccer video using hsv color space and internal contours. In *The International Conference on Signal and Image Processing (ICSIP 2008)*.

Senocak, A., Oh, T.-H., Kim, J., and So Kweon, I. (2018). Part-based player identification using deep convolutional representation and multi-scale pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1732–1739.

Wang, X. and Yang, J. (2020). Marathon athletes number recognition model with compound deep neural network. *Signal, Image and Video Processing*, 14(7):1379–1386.