This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Pose-Guided R-CNN for Jersey Number Recognition in Sports

Hengyue Liu Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, Riverside, CA 92521
hliu087@ucr.edu, bhanu@cris.ucr.edu

Abstract

Recognizing player jersey number in sports match video streams is a challenging computer vision task. The human pose and view-point variations displayed in frames lead to many difficulties in recognizing the digits on jerseys. These challenges are addressed here using an approach that exploits human body part cues with a Region-based Convolutional Neural Network (R-CNN) variant for digit level localization and classification. The paper first adopts the Region Proposal Network (RPN) to perform anchor classification and bounding-box regression over three classes: background, person and digit. The person and digit proposals are geometrically related and fed to a network classifier. Subsequently, it introduces a human body keypoint prediction branch and a pose-guided regressor to get better bounding-box offsets for generating digit proposals. A novel dataset of soccer-match video frames with corresponding multi-digit class labels, player and jersey number bounding boxes, and single digit segmentation masks is collected. Our framework outperforms all existing models on jersey number recognition task. This work will be essential to the automation of player identification across multiple sports, and releasing the dataset will ease future research on sports video analysis.

1. Introduction

Broadcast sports are one of the most watched and studied videos in the world. Game analysis is performed in real time by professional commentators and videos are often recorded for coaching purposes. Analysis requires the review of thousands of hours of footage over the course of a season, and requires tasks that are impractical to be performed by human observer. Therefore, the automation of analysis is especially important. Tasks such as player detection, tracking, identification, as well as generation of game synopses, can be automated using computer vision algorithms to gather comprehensive sports match information without ever having to watch a minute of game video. Automated sports video analysis enhances the broadcasting experience for both the narrator and audience by providing auxiliary information of players location and identity at each time point. Match statistics from video analysis can be provided directly to coaches and players to improve strategy planning, opponent scouting, and player performance.

Figure 1. Illustration of two type of distractions (best viewed in color). Numbers bounded by green box are the jersey numbers of our interest while red-boxed numbers are noise. Our motivation partially comes from how to deal with various kinds of false positives.

Identifying players in sports matches is a key research challenge to make all the merits of automatic sports analysis come true. However, there are numerous problems in recognizing players in unconstrained sports video. The video resolution, viewpoint and motions of cameras, players pose, lighting conditions, variations of sports fields and jerseys, all these factors can introduce significant challenges for automatic video analysis. Traditional approaches for player recognition in sports can be organized into two categories: identifying players via face recognition or jersey number recognition. Both approaches have their own strength and flaws. Face recognition is robust given high resolution closeup shot, while infeasible for wide shots where faces are indistinguishable or low-resolution images. Jersey num-

ber recognition can be achieved under most cases as long as the numbers can be detected or segmented, but suffers from human pose deformation, shooting angles, motion blur, illumination conditions, *etc*. Moreover, the detection result is influenced by not only these factors but also distractions within or around the playground, such as yard markers, house numbers (illustrated in Figure 1), clocks, commercial logos and banners, *etc*.

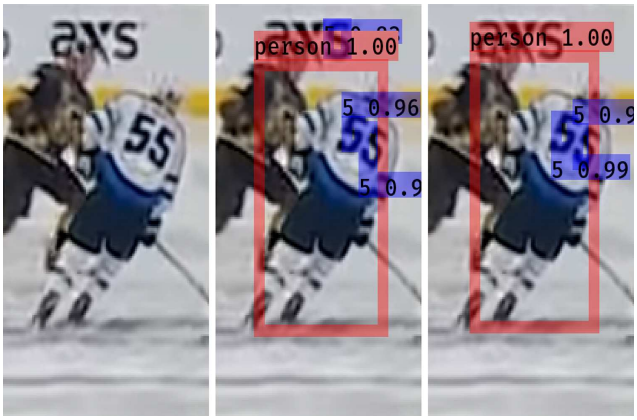

Figure 2. Demonstration of robustness to false positives. Left: original image; middle: Faster R-CNN results; right: proposed pose-guided R-CNN results. Our framework prevents the wrong detection of the character '*s*' in the background.

This paper introduces a pose-guided R-CNN framework to address the challenges associated with player identification through jersey numbers. Faster R-CNN [27] is a two-stage object detector which can perform classification and bounding-box (b-box) regression, and Mask R-CNN [14] is an extension of it with predictions of segmentation masks. This work adapts and expands these concepts with re-designed region proposal and pose-guided b-box regression. The framework consists of two stages. The first stage addresses the digit-person proposal matching problem using a RPN which outputs candidate object b-boxes across three classes, background, player or digit (as opposed to vanilla RPN, which only proposes two, foreground, background). Person proposals and digit proposals are collected separately from a single RPN without adding many parameters. The second stage uses a modification of Faster R-CNN that replaces ROIPool with RoIAlign, and includes a human body key-point branch for predicting key-point masks. The classification and b-box regression are performed on pooled digit features concatenated with key-point masks. This framework improves localization performance of digits by associating person and digit Regions of Interest (RoI), as well as adding human pose supervision signal. Consequently, the model only targets digits inside person proposals with the help from keypoint locations. An example of efficacy of our framework is illustrated in Figure 2.

The main contributions for this paper are as follows:

- The RPN has been re-designed to better fit the jersey number recognition problem. The RPN outputs three classes, i.e., "background", "person" and "digit". By dividing into person and digit proposals, it is possible to match between them to jointly generate better proposals.
- A pose-guided supervision for digit bounding-box is proposed. It learns the offsets of proposals given the prediction of human body keypoints. This module is considered as the refinement of RPN proposals.
- State-of-the-art performance for the jersey number recognition task in comparison to previously established frameworks. Significantly different from previous works, ours is capable of locating and predicting multiple numbers from input images.
- A noval dataset of 3567 images that offers person and digit bounding-boxes, human body keypoints and digit masks. One or more players and digits are annotated per image. More images are being labeled, and the dataset will be made publicly available.

The rest of the this paper is organized as follows. Section 2 introduces the background of jersey number recognition and related research. Section 3 discusses the framework in details. Section 4 evaluates several models with the introduced dataset as well as other wild images from web, demonstrating the applicability across other sports. Several key conclusions are drawn in Section 5.

2. Related Work

Jersey number recognition problem: The problem of interest can be considered as the combination of person identification and digit recognition problem in the context of sports broadcast videos. Traditional approaches before the dominance of deep learning usually first build an Optical Character Recognition (OCR) system then classify numbers based on segmentation results. Šari *et al*. [28] introduce a complete OCR system to segment images in HSV color space with heavy pre-processing and post-processing. Ye *et al*. [34] combine tracking information of frames and a OCR system to predict jersey number based on voting. Lu *et al*. [21] take the person localizations of deformable part model (DPM) detector then performs OCR and classification with matching templates. These OCR-based methods have limited flexibility and robustness dealing with larger datasets. Switching to deep learning approaches, Gerke [10] designs a neural network for jersey number recgonition on small number-centered jersey images. A recent

work from Li *et al*. [18] embed Spatial Transformer Network (STN) modules [17] into a CNN architecture to localize jersey number more precisely and trains the network with additional manually-labeled transformation quadrangles in a semi-supervised fashion.

Some works take sports field into considerations. Delannay *et al*. [7] formulate ground plane occupancy maps from multi-views detection to perform localization, followed by a OCR system and multi-class Support Vector Machine (SVM). Gerke *et al*. [9] consider the player recognition problem as a classifier fusion of players' positional features and jersey number convolutional neural network (CNN) ones [10]. These works put strong assumptions on the hidden pattern of player's movement and mapping of real-world and image coordinates of players. These assumptions are neither well-constructed nor universal applicable.

The jersey number recognition problem can be formulated as person re-identification (ReID) as well. Some approaches favor performing player identification directly. Lu *et al*. [23] use handcrafted combination of features to create a player representation model, then builds a L1-regularized logistic regression classifier [25] for classification, and a Conditional Random Field (CRF) graphical model to predict unlabeled videos. Lu *et al*. [22] continue the work by introducing homography estimation and a weakly-supervised learning to reduce the labor of manual annotation via auxiliary text log information of game matches. Senocak *et al*. [29] tackle player identification problem by constructing a fused feature of multi-scale features extracted from whole body image and pooled features from body parts. We also consider player identification important since the jersey number features are highly correlated to human body ones.

Scene Text recognition: Regarding this similar research, Poignant *et al*. [26] propose a video OCR system for text recognition combing audio information to perform person identification. Goodfellow *et al*. [13] tackle number sequences recognition in constrained natural images with deep neural networks. Jaderberg *et al*. [16] proposed a complete text recognition system for natural scene images with heavily-engineered framework. [3, 30] use STNs for natural scene text detection and recognition. Bušta *et al*. [4] modify Region Proposal Network (RPN) [27] with rotation capability for better text localization. The above-mentioned literature addresses the issue of scene text being in irregular shapes which is also common but more complicated in jersey recognition problem. Jersey numbers are often distorted by player pose deformations and fast motion blur. Li *et al*. [18] adopt STN modules [17] in hope of improving localization and rectifying number distortion. However, the success of STN is built upon the fact of there being only one jersey number per image in their dataset. it is not applicable for complex scene with more involved people.

R-CNN based approaches: With the successes of R-CNNs [12, 11, 27, 14], object detection and classification are unified with high practicality. Mask R-CNN [14] and Faster R-CNN [27] are built upon RPNs with pre-defined anchors to generate region proposals, then the features are pooled from these proposals and fed into regression and classification heads. Vanilla RPN has 3 scales and 3 ratios for each anchor, Ma *et al*. [24] extended the anchor design with rotation parameter for better text proposal alignment. Cai *et al*. [5] introduced a multi-stage Cascade R-CNN to address the issue of degraded detection performance when increasing Intersection-over-Union (IoU) threshold.

The main concern of recognition problem in nature scenes is: how to get robust region proposals. This work exploits the fact that locations of numbers and players are highly related, and achieves the state-of-the-art results. Our framework represents a strong advancement in automated analysis of multi-sport videos.

3. Approach

In this section, the jersey number recognition task is defined in details. A vanilla Faster R-CNN is replaced with a 3-class RPN and extended with additional key-point branch and human pose supervision, yielding the "Pose-guided R-CNN" framework shown in Figure 3. For real-time applicability, a corresponding light-weight model without sacrificing much performance that runs at 12 fps on a single NVIDIA GeForce GTX 1080 Ti GPU.

3.1. Task Definition

A jersey number is defined as the number worn on a player's uniform in order to identify and distinguish players. In our work, only numbers on the back are considered where player's jersey number is typically printed for most sports. Exact one number is associated to one player, and there can be multiple digits in a jersey number. Consider the input image to the model is a image in which at least one player presents with visible and recognizable jersey number. The task is to predict any human-recognizable digit instance $[0, ..., 9]$ displayed in the image. While this task has been modeled as an exact number classification problem [10] as well as a number length prediction problem [18], this work models the task as a 10-digit classification problem.

3.2. RoIAlign Faster R-CNN

From previous task definition, region-based methods are extremely suitable for our problem. One of the successful architectures is Faster R-CNN. It consists of a backbone feature extractor, a Region Proposal Network followed by a feature pooling module, and network heads for b-box regression and classification for each RoI. For an image,

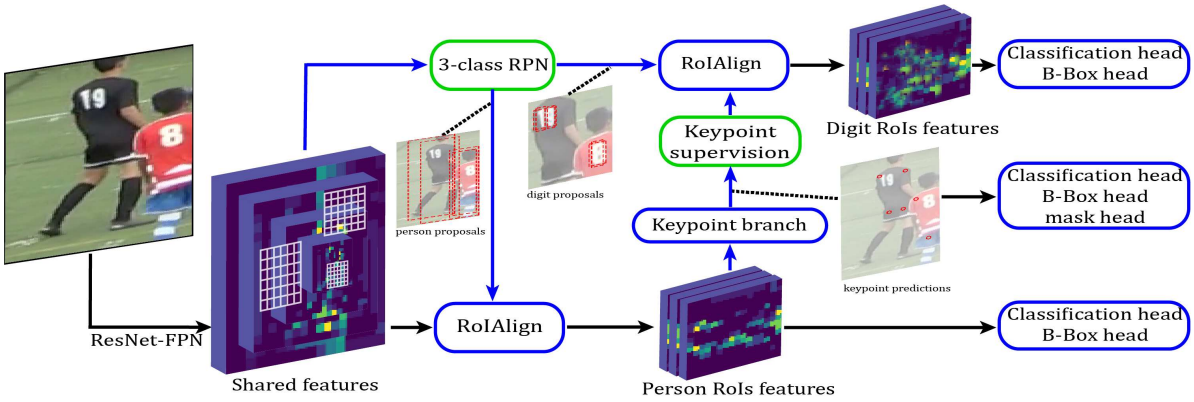

Figure 3. The architecture of proposed pose-guided R-CNN (feature maps are just for illustrations and not representing the actual results).

shareable convolutional (Conv) features are extracted first with choices of backbone architectures such as VGG-16 [31], ResNet [15] and ResNeXt [33] then the RPN generates a set of reference boxes (anchors) from an image of any size. For each pixel location, there can be arbitrary number of anchors given different scales and aspect ratios. A sliding network will traverse each pixel location and tries to predict if an object exists in the corresponding anchor and regress the b-box from shared features. After the proposals are generated, the pooled features for each RoI will be fed into the fully connected layers to perform detection and classification. Feature extraction from each RoI is done with RoI max pooling (RoIPool) such that a $h \times w$ Conv feature map is divided into numbers of $h/H \times w/W$ sub-windows then max-pooling is performed for each grid with quantization. For each detected b-box, non-maximum suppression (NMS) is used to filter out similar and close b-boxes.

Some modules are improved by Mask R-CNN. First it incorporates the Feature Pyramid Network (FPN) [19] with the backbone to generate multi-scale features. It then replaces RoIPool with RoIAlign which interpolates the sampled feature for better alignment between RoI and input feature maps. Beside, it adds an extra branch to generate object masks in parallel in addition to classification and b-box regression. The output mask is represented as a $m \times m\ px$ binary mask from each RoI without losing the spatial layout of convolutional features. For additional details, we refer interested readers to [27, 19, 14]. Faster R-CNN is referred to this improved implementation unless specified for the rest of this paper.

The loss for this baseline is defined as a multi-task loss both for final prediction and RPN proposals:

$$L = L_{cls} + \lambda L_{reg}, \tag{1}$$

where L_{cls} is classification loss, L_{reg} is the b-box regression loss, and λ is the multi-task balance weight. We consider each digit from 0 to 9 as a class, a 'person' class and a 'background' ('BG') class, in total of $K = 12$ independent classes. Ground-truth class is denoted by u where $u_{BG} = 0$ by convention. For each RoI, the output layer will produce a discrete probability distribution $p = (p_0, p_{K-1})$, then the class loss is define as log loss for true class

$$L_{cls}(p, u) = \log p_u. \tag{2}$$

The localization loss is defined as

$$L_{reg}(t^u, v) = \sum_{i \in \{x,y,w,h\}} smooth_{L_1}(t_i^u - v_i), \tag{3}$$

' where $u > 0$ ('BG' class does not contribute to the loss), and t_i^u is predicted bounding-box offsets four-tuple (x, y, w, h) for class u. (x, y) is the top-left corner coordinate, (w, h) is the predicted dimension of the b-box. $v = (v_x, v_y, v_w, v_h)$ as the ground-truth b-box. $smooth_{L_1}$ is a robust L_1 loss against outliers defined in [11].

3.3. Proposal Association

Up to this point, we have generated proposals of either one digit or a person and same for final detections. To collect the final results in terms of jersey numbers, we reduce our problem into a graph matching problem [32] with some relaxations. Nodes of the graph are the person and digit proposals, and the edges are all possible connections between pairs of person and digit proposals. The weight of each edge is computed by the Euclidean distance between the two centers of bounding boxes. And for each person node, there must exist k edges matched with digit nodes, where $1 \leq k \leq 2$. So each person node can be matched with up to two other digit nodes which is not necessarily bipartite matching. The problem is then solved by choosing the top-2 digit proposals for each person proposal.

3.4. Three-class Region Proposal Network

The original RPN only estimate the probability of each proposal being an object or not. It takes shared features to

Figure 4. Some examples from the dataset. (g), (k), and (l) are examples of multi-jersey annotations; (a), (f) and (g) are illustrations of jersey numbers under common conditions; (c) and (h) exhibit contrast lighting conditions ; (i) shows a close-view image where number aspect ratio is distorted; (b), (d) and (j) are examples of numbers influenced by pose deformation; (e) is highly distorted but still recognizable.

perform classification and bounding-box regression of anchors. Our motivation is simple: instead of just 2 classes, this work uses 3 classes to represent 'BG', 'person' and 'digit' by adding very few parameters. In this way, anchors are not treated independently. Anchors are divided into person and digit anchors that are then correlated by their spatial relationships.

No modifications are made to the pre-defined anchor settings in [27] that there are lots of overlaps among anchors. Each anchor is actually associated with many other anchors in terms of location. For example, if an anchor is of scale 512, some anchors of scale less than 512 will be inside it. The proposal scheme is modified to accommodate this anchor association. For training vanilla RPN, each positive anchor is assigned based on two criteria. The following conditions are provided along with three-class RPN:

- Anchor(s) that has/have the highest Intersection-over-Union overlap with certain ground-truth box.
- Person anchors with IoU higher than 0.7.
- Digit anchors with IoU higher than 0.7 and inside any person anchor.

After filtering and assignment of anchors, we associate each digit anchor to its closest person anchor based on Euclidean distance between centers of the two boxes.

3.5. Pose-guided Supervision

Mask R-CNN can also perform human body keypoints estimation as stated in [14]. Similar to the binary mask representation of objects, each body keypoint is modeled as an object except that there is only one pixel labeled in the mask. For K types of keypoints, *e.g.* right shoulder, left hip, *etc.*, there are K individual one-hot masks. Human body modelling is not required in Mask R-CNN framework to achieve fair results. In the case of jersey number recognition, it is reasonable and achievable to perform jersey number localization better given body keypoints layouts. Though Faster R-CNN is capable of bounding-box regression for jersey numbers, there are limitations under more sophisticated scenarios. For example, complex jersey patterns, different number fonts, and numbers on the court introduce difficulties for RPN to generate satisfactory proposals. To tackle the problem, a pose-guided supervision branch is proposed for refining number localization.

A keypoint branch for predicting key-point mask is added similar to [14, 2]. The keypoint detection is only applied on person RoIs. At this point, each person RoI is associated with multiple digit RoIs as a result of three-class RPN. The keypoint mask is fed into a shallow network to obtain b-box offsets with which we can correct the RPN proposals. Features of refined proposals are then pooled via ROIAlign. It involves a transformation from keypoint locations to associated digit b-box regression in a hidden space. Finally, a digit branch is formulated that is responsible for digit recognition on refined RoIs. This cascade design provides digit RoIs with more information outside their regions.

The proposed pose-guided network takes predicted keypoints mask from each person RoI as inputs, and output the b-boxes offsets of corresponding jersey numbers. It is a small but effective network consisting of three fully connected layers.

The loss function 1 can be modified accordingly by

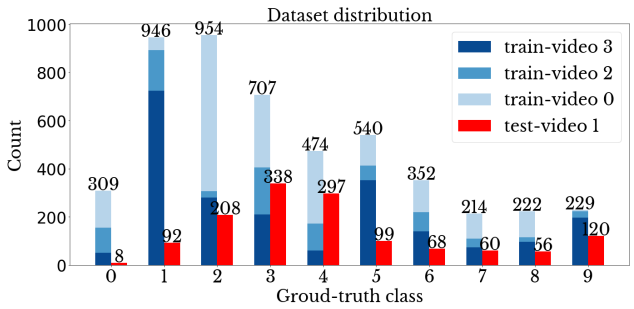

Figure 5. Distribution of digits in the data.

adding related keypoint classification and regression loss $L_{cls}^{keypoint}$, $L_{reg}^{keypoint}$. Then the regression loss for digit b-box is computed from the RoI refined by keypoint mask. The final loss function is

$$L = L_{cls} + \lambda L_{reg} + \eta\lambda L_{cls}^{keypoint} + \gamma\lambda L_{reg}^{keypoint}, \quad (4)$$

where η and γ are hyper-parameters similar to λ.

4. Experimental Results

The proposed pose-guided R-CNN, as well as related models are evaluated on the collected dataset, since there is no publicly available dataset on jersey numbers. The evaluation metrics used are standard Average Precision (AP) with IoU thresholds set to 0.5 and 0.75, and AP average (mAP) across IoU from 0.5 to 0.95. Number-level and digit-level accuracies are also reported.

4.1. Dataset

The dataset is gathered from four full soccer matches. The recording device used is a single Canon XA10 video camera which is installed 15 feet high, and 10 to 20 feet away from the horizontal baseline of the soccer field. For better video qualities in terms of recognizable jersey numbers, the camera operator is allowed to pan and zoom accordingly. Next, we convert the collected videos into frames by two different ways. One is to perform a human detector over frames scaled by 2 to get reliable images containing players. OpenPose [6] is used for person detection. In order to collect more difficult images, Random shifts and paddings are added to detected areas. The detection results are padded by 150px and a random shift of 20px. After data collection was complete, two professional annotators labeled any legible jersey numbers via VGG Image Annotator [8]. As a result, there are arbitrary number of ground-truths (GT) per person per image.

A total of 3567 images are annotated with ground-truth (GT) digit masks resulting in 6293 digit instances, see the distribution in Figure 5. All images are also labeled with person bounding-boxes and four human body key-points, namely left shoulder (LS), right shoulder (RS), left hip (LH) and right hip (RH). There are 114 images contain multiple numbers, and each digit is labeled with its associated person box. Figure 4 shows a few examples for our dataset. Dataset statistics are illustrated in Table 1.

	H	W	h	w	Digit mask area	Digit mask center
Mean	315.06	214.53	34.70	18.90	424.40	(0.50, 0.29)
Std	92.11	38.47	15.16	7.85	20.69	(0.12, 0.09)

Table 1. Dataset statistics. H, W, h and w are image height, image width, digit b-box height, and digit b-box width respectively. For heights and widths, the unit is pixel; mask area counts the number of pixels on the object; mask center is normalized within range [0, 1].

Framework	Backbone	Input	ACC_{number}	ACC_{digit}
Gerke[10]	-	40^2	65.04%	-
Li *et al.* [18]	-	200^2	74.41%	77.86%
Li *et al.* [18]	ResNet-50	512^2	77.55%	80.23%
Faster R-CNN	ResNet-FPN-50	256^2	86.13%	89.32%
Faster R-CNN	ResNet-FPN-50	512^2	88.74%	90.09%
Faster R-CNN	ResNet-FPN-101	512^2	89.02%	91.11%
Pose-guided (Ours)	ResNet-FPN-18	512^2	81.66%	83.97%
Pose-guided (Ours)	ResNet-FPN-50	256^2	90.84%	92.13%
Pose-guided (Ours)	ResNet-FPN-50	512^2	91.01%	93.29%
Pose-guided (Ours)	ResNet-FPN-101	512^2	**92.14%**	**94.09%**

Table 2. Comparison of results among approaches. Our method achieves the best accuracy (ACC) for both number-level and digit-level recognition. Input is cropped grayscale image for Gerke's [10], and original RGB image for all other approaches.

Bounding-box sizes are sorted into small (area $< 32^2$), medium ($32^2 <$ area $< 96^2$) and large (area $< 96^2$) objects like COCO dataset [20]. For person b-boxes, there are 4111 large, 213 medium and 1 small objects; for digit ones, there are 7 large, 1210 medium and 5076 small objects.

4.2. Implementation Details

The hyper-parameters in the loss function 4 are all set to one. All the experimented models make use of image augmentation technique by applying random affine transformation and hue/saturation manipulation to both original image and corresponding b-box. The backbone feature used in all experiments is ResNet-FPN. We use ResNet features at 4 different stages $[C2, C3, C4, C5]$ to build the feature pyramid. The constructed RPN features are $[P2, P3, P4, P5, P6]$. The light-weight model removes $C5$ and $P6$. For RPN anchors, 5 scales $[32, 64, 128, 256, 512]$ and 3 ratios $[0.3, 0.5, 1]$ are used. For the classification network $head$, $P6$ is not used as input. Partial implementation is adopted from [1].

Person and keypoint branches: The settings for person branch are same as described in [14]. The keypoint branch is based on mask prediction in Mask R-CNN, except that the keypoint mask is up-sampled to 32×32.

Digit branch: The pose-guided supervision module consists of two 512 Fully-Connected (FC) layers, and a $N \times 4$

FC layer with $linear$ activation as digit b-box regression $head$. N is the number of proposals, so it outputs the b-box offsets for each digit RoI. The rest of the branch resembles person branch except for the pooling size to be 16×16 in digit classification $head$. It gives better performance since digits are relative small in images.

Different settings including but not limited to changing the backbone features, input image size, image operations (re-sizing, padding, cropping, *etc*.), number of image channels are used in experimentation. ResNet-FPN-18, ResNet-FPN-50 and ResNet-FPN-101 with/without proposed pose-guided module are investigated. For collecting convincing results, the dataset is divided video-wisely, with video 0, 2, 3 for training and video 1 for testing.

Pre-train: To accommodate the lack of person-keypoint data in the collected dataset, the network is pre-trained on the COCO dataset [20] with a frozen digit branch. In this dataset, 17 human body keypoints are annotated, but four of them are used for less parameters and better convergence. Person and keypoint branches are then unfrozen, and the digit branch is trained with Street View House Number (SVHN) dataset [13]. This large-scale dataset consists of digit sequences with each digit labeled with bounding box. The model benefits from this dataset for training the backbone feature extractor.

Training: The model is trained for 100 epochs with starting learning rate (LR) 0.01. Learning rate is reduced by 10 every 20 epochs. The rest hyper-parameters are same with Mask R-CNN [14].

Testing: The settings are the same as training except that less (set to 100) detections are kept.

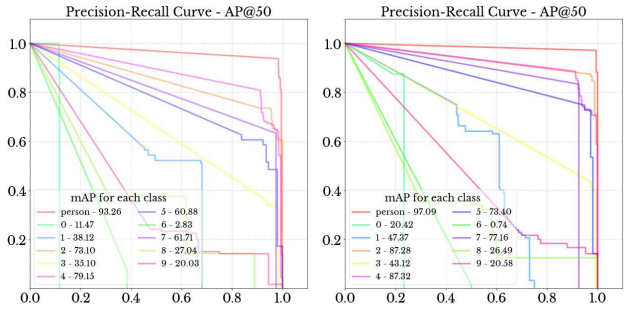

Figure 6. Precision-Recall Curves over each class. Left figure shows the Faster R-CNN results; right one shows ours results with improvement.

4.3. Main Results

The proposed model is compared to available methods in the field of jersey number recognition, see Table 2. All variants of our model outperform previous state-of-the-art models including Gerke [10] and Li *et al*. [18]. These two approaches can only perform image-level recognition. For fair comparison, multi-number images are removed during training and testing. Each image is grayscale, cropped and re-sized to 40×40 in accordance with [10]. Without access to the dataset of [18], this architecture is implemented without axis supervision. Its variant with ResNet-50 is also implemented. Faster-RCNN is also a strong baseline which already outperforms [10, 18]. The proposed model achieves even better performance that is highly robust to post variations. Figure 7 visualizes the recognition results against different poses. We evaluate both digit-level and number-level accuracies for our model and [10, 18]. The results are illustrated in Table 2.

Method	ACC_{number}	ACC_{digit}	mAP	AP_{50}	AP_{75}
Faster R-CNN	87.23%	89.04%	40.60	67.21	45.58
Pose-guided (Ours)	90.44%	93.12%	44.74	73.31	48.77

Table 3. Comparison of Faster R-CNN and our pose-guided R-CNN results. The backbone used is ResNet-FPN-50, and input image size is 512×512.

Evaluation metrics including number-level and digit-level accuracies, mean average precision (mAP), AP_{50}, and AP_{75} are used to compare variants of the R-CNN approaches. APs for different object scales are not used since most 'person' boxes are large and most 'digit' are small. The results are shown in Table 3. The proposed pose-guided R-CNN gives the best overall results.

4.4. Ablation Study

In this section, we only consider ResNet-FPN-50 as our backbone given several reasons: it has around $19M$ less parameters; we have a small dataset so ResNet-50 is more suitable; we did not fine-tune the models so better performance can be achieved through regularization. Therefore, we choose ResNet-FPN-50 over ResNet-FPN-101 without sacrificing much performance. Multi-number images are included for experiments in this section.

Input size: To build feature pyramid for ResNet-50, we need to resize the image so that its width and height can be divided by 2 at least 5 times. We need the image size to be large enough since the numbers in the dataset are mostly small objects. For simplicity, we re-size to square image with paddings while keeping the aspect ratio. We did experiments with several sizes: 128, 256, 512, 1024. When the input size is 512, it achieves the best performance of mAP 44.74, which outperforms 10.20, 3.12 and 0.56 points with respect to size 128, 256, and 1024.

Does 3-class RPN solely help: With the baseline of Faster R-CNN, we want to evaluate if replace the vanilla RPN with our 3-class RPN help improve the performance. We use image size of 512×512 as input, and ResNet-FPN backbone for this experiment's settings. Three-class RPN has -0.09, 0.12 and -0.14 gain respectively over vanilla

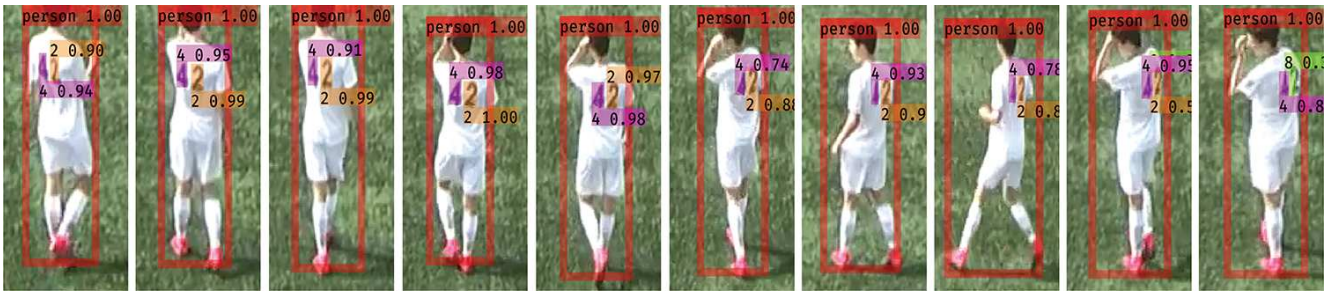

Figure 7. Recognition results across different poses. Most-left and most-right poses are extreme cases in our test set for this identity.

Figure 8. Exemplary results from wild images collected from internet. Fails: '7' is classified wrongly as '2' in second last image; '6' in the last image is classified wrongly as '5'.

RPN on mAP, AP_{50}, and AP_{75}. Both give similar experimental results, so it suggest that by just switching to three-class RPN, the performance is not significantly influenced. RPN is a shallow 'neck' network for anchor classification and regression. Splitting 'object' class into 'person' and 'digit' does not introduce hardness for these two tasks, but we can not guarantee multi-class RPN will work for more classes. The key function of our three-class RPN is dividing then matching person and digit anchors. If the following structure remains the same with Faster R-CNN, the results are expected to be similar. However, as we already match the anchors in three-class RPN, the proposal association procedure for number-level prediction can be removed.

Pose-guided R-CNN: Table 3 suggests that there is 4.14 gain over Faster R-CNN. We also report AP_{50} for each class for these two models illustrated in Figure 6. It shows significant improvement achieved by adding pose supervision which has a keypoint mAP of 58.2. The reason of poor performance on '0' is that We have very few images contain '0' in test dataset, so it drops drastically even if only one of them is classified incorrectly. Figure 7 provides recognition results of our pose-guided R-CNN model against different poses. However, there are still some limitations under extreme poses as the last two examples shown in Figure 7. For testing our model's generalization, We also collected some images from internet videos for different sports: basketball, American football and hockey. The results are illustrated in Figure 8. Fair detection results are still obtained, but classification performance is reduced. Recognition is possibly simpler for soccer and basketball due to plain jerseys, while jerseys in American football and hockey are normally bulky with sharp contours. Better performance can be achieved by gathering more data across different sports.

5. Conclusion

In this work, a pose-guided R-CNN multi-task framework is proposed as an all-in-one solution for person detection, body keypoints prediction and jersey number recognition. It produces the best digit accuracy of 94.09% comparing with related literature. Three insights are used to achieve this performance: 1. re-designed three-class RPN for anchor association; 2. implementation of pose-guided localization network that can impose proposal refinement for jersey number location through human pose; 3. the generality of region-based CNN model. By combining the three components, the proposed approach is end-to-end trainable and can be easily extended to other sports.

Acknowledgements

This work was supported in parts by Bourns Endowment and funds and a gift from SEVAai to the University of California, Riverside. We thank our colleague Alex Shin for providing image annotations.

References

[1] W. Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[2] R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[3] C. Bartz, H. Yang, and C. Meinel. See: towards semi-supervised end-to-end scene text recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[4] M. Busta, L. Neumann, and J. Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2204–2212, 2017.

[5] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[7] D. Delannay, N. Danhier, and C. D. Vleeschouwer. Detection and recognition of sports(wo)men from multiple views. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–7, Aug 2009.

[8] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/~vgg/software/via/, 2016. Version: 2.0.0, Accessed: 7.1.2018.

[9] S. Gerke, A. Linnemann, and K. Müller. Soccer player recognition using spatial constellation features and jersey number recognition. *Computer Vision and Image Understanding*, 159:105–115, 2017.

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

[11] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016.

[13] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. In *International Conference on Learning Representations*, 2014.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[17] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*, pages 2017–2025, 2015.

[18] G. Li, S. Xu, X. Liu, L. Li, and C. Wang. 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, 2018.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

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

[22] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1704–1716, 2013.

[23] W.-L. Lu, J.-A. Ting, K. Murphy, and J. Little. Identifying players in broadcast sports videos using conditional random fields. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3249–3256. IEEE Computer Society, 2011.

[24] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.

[25] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 78. ACM, 2004.

[26] J. Poignant, L. Besacier, G. Quenot, and F. Thollard. From text detection in videos to person identification. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo*, pages 854–859. IEEE Computer Society, 2012.

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99, 2015.

[28] M. Šari, H. Dujmi, V. Papi, and N. Roži. 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)*, 2008.

[29] A. Senocak, T.-H. O. J. Kim, and I. S. Kweon. 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, 2018.

[30] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4168–4176, 2016.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[32] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996.

[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[34] Q. Ye, Q. Huang, S. Jiang, Y. Liu, and W. Gao. Jersey number detection in sports video for athlete identification. In *Visual Communications and Image Processing*, volume 5960, pages 1599–1606, 2005.