Jersey Number Recognition Project Proposal

Course: COSC 419B/519B
Instructor: Dr. Mohamed Shehata
Team Number: 8
Team Members: Varun Patel (88268834), Mark Liu (51438422), Pinjing Xu (25627175), Kaycee Goel (72079460), Dina Elkholy (50482579)

---

1. Introduction

A Jersey Number Recognition System is designed to automatically detect and recognize jersey numbers worn by athletes in sports footage. The primary goal of this system is to enhance sports analytics by providing accurate and efficient player identification, which can be used for performance tracking, game analysis, and automated statistics collection.

Accurate jersey number recognition is crucial in sports analytics, enabling real-time player tracking, assisting in tactical decision-making, and enhancing automated data collection for coaches, analysts, and broadcasters. However, this task presents several challenges, including motion blur, varying lighting conditions, occlusions from other players, and differences in jersey designs across teams and leagues. Additionally, factors such as low-resolution video feeds, complex backgrounds, and distortions caused by camera angles further complicate number detection. Another major challenge is the similarity between certain digits—such as 6 and 9 or 1 and 7—which can appear nearly identical under poor lighting, low resolution, or partial occlusion. These issues can lead to misclassification and errors in player identification, necessitating advanced computer vision techniques and effective data preprocessing methods to ensure high performance.

A valuable resource in addressing these challenges is the SoccerNet dataset (Giancola et al., 2018), which offers a collection of annotated soccer videos designed for player recognition systems. However, this dataset also suffers from the challenges mentioned earlier (e.g. motion blur, distortions, low resolution images). Many researchers have attempted to overcome these obstacles using various techniques, including sophisticated data preprocessing, robust data augmentation, and tailored deep learning architectures. Building on these prior efforts, we aim to address the shortcomings of existing approaches while trying new methods to tackle the problem.

2. Literature Review & Replicated Results

2.1. Related Work

Recent research in jersey number recognition has largely focused on overcoming the inherent challenges of dynamic sports environments—such as motion blur, low resolution, lighting variations, and distortions due to camera angles. By understanding prior approaches, we can integrate the most effective techniques while addressing unresolved challenges. Our review focuses on two main categories: image-level recognition methods—which include legibility classification, localization, and digit recognition—and tracklet-level recognition approaches that aggregate predictions across video frames to handle dynamic sports scenarios.

In Koshkina et al. (2024), a state-of-the-art jersey number recognition framework, the authors combine an image-level legibility classifier based on a ResNet34 model, torso localization via ViTPose (Xu et al., 2022), and digit recognition with the PARSeq STR model (Bautista et al., 2022), alongside tracklet-level techniques that aggregate predictions through Centroid-ReID-based (Wieczorek, 2021) filtering and iterative outlier removal. This integrated approach effectively addresses challenges like occlusion, motion blur, and low resolution, achieving up to 87.45% accuracy on the SoccerNet dataset. The details of this method will be discussed in subsequent sections.

An earlier approach in (Li et al., 2018) combined a CNN with a Spatial Transformer Network (STN) to correct distortions from camera angles and player motion. It used a Single Shot Detector (SSD) to detect players, followed by a CNN for jersey number classification. Two classification strategies were employed: separate digit classification and a sequence length predictor to minimize false positives. To improve STN performance, the authors used semi-supervised learning, annotating a small subset of images with quadrangles around jersey numbers for additional supervision. The final model was trained using multi-task learning, jointly optimizing the CNN and STN.

STNs enhance geometric invariance by dynamically learning spatial transformations (e.g., rotation, scaling, cropping) during training, improving robustness to variations in viewpoint and scale. Unlike manual preprocessing, STNs adapt in real-time, focusing on task-relevant regions. They are also more computationally efficient than pose estimation models like ViTPose (Xu et al., 2022), as they apply transformations directly to input data without detecting multiple key points. However, STNs may struggle with extreme occlusions or complex spatial relationships, where pose estimation models perform better.

Comandur et al. (2022) addresses the challenge of player re-identification in sports broadcast videos, a task complicated by factors such as similar team

uniforms, limited training samples per player, and issues like occlusion and low resolution. To overcome these obstacles, the authors proposed a hierarchical data sampling strategy that leverages metadata (including match, action, teams, and year) to group images into training batches that more closely mirror the conditions encountered during testing. Complementing this strategy, they introduced a straightforward centroid loss function that computes the Euclidean distance between the centroid of a specific player's embeddings and the centroid of all other players in the batch. This method encourages the formation of well-separated clusters and proves more effective than traditional triplet-based losses in data-scarce scenarios. Together, these changes yield significant improvements in mean average precision and rank-1 accuracy, achieving state-of-the-art results on the SoccerNet Re-ID Challenge 2022.

Bhargavi et al.(2022) demonstrate the value of high quality synthetic training data, which other approaches have not used, to improve test scores compared to models trained only on real data. They also found significant drawbacks in using unfit training data such as the MNIST dataset. With a combination of synthetic data, ResNet50 backbone pose detector, and a non descript CNN for number identification, a test accuracy of 89.13% was achieved for identifying American football jersey numbers from broadcast footage. Due to the general similarities and issues in both problems, as well as the demonstrated test accuracy increase over models trained only on real data. Due to this demonstration of the benefits of a more complete training set, we aim to incorporate basic data augmentation to create a more comprehensive set of training data.

Aside from multi-step and CNN-based approaches, some research tackled transformer-based approaches. Vats et al. (2022) introduced a hybrid architecture, CNN model and a transformer, to identify NHL players in broadcast videos. The model architecture begins with a ResNet18 backbone that extracts 512-dimensional feature vectors from each video frame, capturing essential spatial details like jersey numbers. These features are then fed into a transformer encoder, which models the temporal dynamics across player tracklets to robustly recognize jersey numbers over time. Additionally, the system leverages NHL shift data by using OCR to extract game time, which is cross-referenced with play-by-play data to filter out off-ice players—yielding a 6% accuracy improvement. The entire framework is trained with a multi-task loss that combines whole-number and digit-wise classification, and further enhanced by data augmentation techniques. This integrated approach was validated on 3,510 annotated tracklets from 84 NHL games.

2.2. Replication of State-of-the-Art Results

We use the pipeline proposed by Koshkina et al. (2024) as our baseline, and thus our work on replicating the SOTA result mainly focuses on this pipeline.

For the task of image-level jersey number recognition the baseline pipeline consists of 3 major components: Legibility Classifier, Pose Detector and Scene Text Recognition. The legibility classifier is a binary CNN, ResNet34, used to determine if a player's jersey number is visible and legible in a given frame. The Pose detector is used to estimate body key points using a ViTPose model, to localize the torso region where the jersey number is typically located. And the torso region is cropped and padded to focus on the jersey number. The scene text recognition (STR) process Koshkina et al. (2024) use is the state-of-the-art STR model PARSeq to recognize the jersey number from the cropped torso region. They fine-tune the PARSeq model on a small dataset of hockey jersey number images for better performance.

For the tracklet-level recognition, the pipeline proposed in the baseline method extends the image-level approach with additional steps. Main subject filtering is achieved using a re-identification network (Centroid-ReID) to filter out frames where the main player is occluded by other players or objects. Prediction consolidation is performed to aggregate predictions from multiple frames in a tracklet, two methods were explored: Heuristic Consolidation uses confidence-weighted majority voting, with a bias toward two-digit numbers; and Probabilistic consolidation uses a Bayesian approach to combine predictions based on STR output probabilities.

The source code provided by Koshikina et al.(2024) consists of multiple submodules. Each submodule points to its original repository of work. To run the project, we need to download all submodules as instructed in their project README, and create anaconda environments for each of the modules, so that all modules can run with their own dependencies. This is the most tricky part when replicating the result of the baseline method. It consists of at least four different submodules from different authors, and setup for the codebase was on a challenging side especially with multiple environments and library versioning, such as Cuda, Python, Pytorch and other dependencies. This brought a lot of trouble to our team. In addition, running the project is a time consuming task, even for inference only.

Running inference for one round on a personal laptop with a **RTX 4060 GPU takes about 8h36m**, and reaches an accuracy of 86.79%, which is 1% lower than the value reported by Koshkina et al. (2024). We consider the data as a match. During this process, we have found some aspects that can be improved.

The Legibility Classifier proved to be the most computationally expensive stage, taking approximately 3.5 hours to process the dataset on a personal laptop equipped with an RTX 4060 GPU. Given its substantial computational overhead, removing or optimizing this classifier could significantly improve efficiency without compromising recognition performance.

The PARSeq model used during the STR step is based on Transformer architecture, which is well-known for its high computational cost, and requires careful tuning of hyperparameters. Koshkina et al.(2024) worked on fine-tuning the model to fit their needs.

Additionally, PARSeq's use of an internal language model is not helpful in any way to the identification of digits only. PARSeq also intended to address the issue of text being organized in non left to right orientations (Bautista & Atienza, 2022). Most of the data is in the left to right orientation already, or could be rotated to be in the left to right direction, which lowers the importance of another one of PARSeq's advantages in this problem. PARSeq also is capable of context aware STR, which is not needed in jersey number identification. Given these factors not contributing to jersey number identification, we propose replacing PARSeq with a lightweight CNN-based model.

3. Proposed Pipeline

3.1. Overview of Proposed Approach

Similar to Koshkina et al. (2024), our pipeline consists of main player filtering for each tracklet, a legibility classifier, a pose detector, and a digits classifier to identify the jersey number. Our approach diverges from theirs in the following ways:

1. We will test Koshkina et al.'s (2024) pipeline without the legibility classifier, as it is one of the most computationally expensive steps.
2. We will apply various data preprocessing techniques to enhance contrast and resolution, improving the clarity of key features.
3. We will use data augmentation techniques such as random cropping, flipping, and rotation to increase dataset diversity.
4. We will replace the PARSeq model with a lightweight alternative for jersey number identification.
5. We will modify the player re-identification model to reduce computational costs.

Our methods converge when the final tracklet prediction is computed using probabilistic methods.

3.2. Model Architecture

1. Potential Replacements to the PARSeq Model

We will try different CNN models pre-trained on ImageNet dataset, and fine-tune the network on both SoccerNet and the Hokey dataset, prepared by Koshkina et al.(2024). We plan to try models such as ResNet34, ResNet50, VGG16, ConvNeXt-Tiny and EfficientNet. When fine-tuning the network, we will freeze the parameters in the network, and only train the last fully connected and output layers. For the output layer, since the output format of the pretrained network will not match our case, which classification of numbers from 0 to 99. We are planning to test and comparing the following methods:

- Replace the final fully connected layer with a new layer containing 100 neurons.
- Only output to 8 neurons, and using binary bits 0 and 1 to represent the numbers.
- We could use three parallel FC layers. The first layer only outputs 0/1 to indicate whether it is 1 or 2 digits, and the second and third layers each output a number between 0 and 9. This time we only need 21 neurons, but we need to test whether parallelism will consume more time when doing backpropagation

2. **Potential Changes to the Re-identification and Legibility Network**

The re-identification and legibility classification process consists of three steps. First, Centroid-ReID, pre-trained on Market1501, extracts discriminative feature vectors for each frame using ResNet34. Next, an isotropic Gaussian model iteratively prunes outlier frames (>3.5σ) to remove occlusions and background clutter. Finally, a hockey-trained classifier generates weak pseudo-labels for legibility classification, which is then fine-tuned on soccer tracklets to mitigate domain gaps like camera angles and jersey styles. Since this process is the most time-consuming, we plan to:

- Test removing the entire legibility process and evaluate its impact on classification accuracy.
- Adjust the number of Gaussian model iterations to 2 or 4 and assess its effect on accuracy.
- Experiment with a different architecture, such as ResNet18, instead of ResNet34.

3.3. Data Preprocessing & Augmentation

For data processing, we plan to explore various techniques to enhance image quality and diversity, as we hypothesize that higher-quality images will yield more representative features for the classification task.

- **Data Augmentation**

  We will apply transformations such as rotation, scaling, and contrast adjustments. To address the imbalance in jersey numbers, we may incorporate synthetic data, and small rotations will be applied to pose detector-cropped images to expand the training dataset. After pose detection, images will be converted to monochrome to reduce computation costs while preserving accuracy. Following Koshkina and Elder (2024), we will crop images to focus on the player's torso. Additionally, as Bhargavi et al. (2022) emphasized the importance of diverse and high-quality training data, we may supplement our dataset with NHL tracklet data from Chen et al. (2023).

- **Improving Image Quality**

  We plan to try simple image processing techniques and filters to enhance image quality and improve model performance. Specifically, we can apply high-pass kernels to sharpen the images before they enter the pipeline. We may explore more sophisticated approaches such as super-resolution models.

4. Proposed Enhancements & Performance Improvements

4.1. Technical details

Our proposed approach aims to improve accuracy, speed, and resource efficiency. The original framework's Scene Text Recognition component relies on PARSeq, a Transformer-based autoregressive sequence model (Bautista & Atienza, 2022). Due to the redundancies PARSeq introduces for jersey number identification (see Section 2.2), we will replace it with a CNN-based model, which offers faster, more resource-efficient feature extraction and digit recognition in sports analytics (Alhejaily et al., 2023; Lin et al., 2024; Liu & Bhanu, 2022; Vats et al., 2021). We will implement and test various CNN backbones, such as ResNet34, VGG16, or EfficientNet, to introduce a lightweight number recognition model that enhances inference speed while maintaining accuracy.

Additionally, we will optimize speed and resource efficiency by modifying the legibility classification process, as outlined in Section 3.2. To further improve accuracy, we will adopt data processing and augmentation techniques such as random cropping, rotation, flipping, and contrast adjustments. Since image resolution directly affects accuracy, we also plan to enhance image quality using simple image processing techniques.

4.2 Justification

Transformer-based models like PARSeq offer strong accuracy but are computationally expensive. PARSeq was designed to address challenges in general Scene Text Recognition (STR) tasks that are less relevant to jersey number identification (Bautista et al., 2022). These factors make CNN-based methods a preferable alternative. ResNet is widely used for feature extraction in jersey number recognition, with studies demonstrating its effectiveness. Liu & Bhanu (2022) introduced JEDE, a ResNet-50-based framework integrating player detection, pose estimation, and jersey number recognition. Alhejaily et al. (2023) applied transfer learning with convolutional autoencoders to enhance recognition performance, while Lin et al. (2024) leveraged ResNet-50 with multi-task learning, incorporating body posture and digit structure for improved accuracy. Aligning with these approaches, we will either train a ResNet model from scratch or fine-tune a pre-trained model. Additionally, EfficientNet and ConvNeXt are strong alternatives, offering high performance with fewer parameters than transformers or larger networks.

Multi-task learning has proven effective in jersey number recognition by simultaneously optimizing multiple tasks. Vats et al. (2021) introduced a multi-task approach combining holistic jersey number classification with digit-wise recognition, improving generalization. Bhargavi et al. (2022) achieved their highest accuracy using this method, while Lin et al. (2024) refined it further by integrating orientation-guided weight refinement for greater robustness against rotation and occlusion. Inspired by these studies, our model will incorporate parallel output layers to merge multiple predictions, enhancing accuracy.

5. Experimental Plan & Evaluation Metrics

5.1 Experiment Design

We will continue to use the work by Koshinkina et al.(2024) as the baseline. As we are removing/alternating the computational heavy steps in the baseline, we are expecting to see much shorter time of use for both the training and inference phase. We will use the similar software setup as the baseline, and run the project on a GPU in the graduate student lab (A6000 GPU).

We will use a CNNs pretrained on ImageNet (e.g. ResNet34, VGG16, ..etc) and freeze the convolutional layers to fine-tune on the SoccerNet dataset and the NHL tracklet dataset prepared by Chen et al.(2023). We will use the same setup for splitting the training and testing dataset.

The different setups being measured will be:

1) Applying the data processing and augmentation techniques

2) Evaluating with and without the Legibility Classifier and based on the result, we will then have an idea on whether we should keep it or not
3) Based on 2, we will then try to modify the re-identification network as detailed in Section 3.2.
4) Based on the results from 2 and 3, we will then continue experimenting on the different CNNs to replace the PARSeq model:
    a) 100 node output layer where 0-99 is a jersey number, and 100 is a case for no number.
    b) The multitask parallel output layer as proposed by Vats et al. (2021). Where 2 layers identify digits separately and compare to a third layer that identifies whether we have one or two digits.

5) A ResNet34 backbone CNN in place of the PARSeq model
    a) Evaluated with and without the legibility classifier step in the pipeline
    b) With and without the legibility classifier in place, test
        i) 100 node output layer where 0-99 is a jersey number, and 100 is a case for no number
        ii) The multitask parallel output layer as proposed by Vats et al. (2021). Where 2 layers identify digits separately and compare to a third layer that identifies both digits together
6) A VGG16 backbone CNN in place of the PARSeq model
   With the legibility classifier in place, test:
    a) 100 node output layer where 0-99 is a jersey number, and 100 is a case for no number
    b) The multitask parallel output layer as proposed by Vats et al. (2021). Where 2 layers identify digits separately and compare to a third layer that identifies both digits together

5.2 Evaluation Metrics

Our main indicators of performance will be the overall testing & SoccerNet challenge accuracies, as well as the F1-score. Secondarily, we will measure the inference time on the same hardware used to reproduce the Koshkina et al. (2024) results; measure how performance and inference time are related, as one of our performance improvement goals is in computational costs.

6. List of References

S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos," *King Abdullah University of Science and Technology (KAUST), Saudi Arabia*.

Alhejaily, R., Alhejaily, R., Almdahrsh, M., Alessa, S., Albelwi, S. (2023). Automatic team assignment and jersey number recognition in football videos. Intelligent Automation & Soft Computing, 36(3), 2669–2684. https://doi.org/10.32604/iasc.2023.033062

Lin, Y. H., Chang, Y. W., Shih, H. C., & Ogawa, T. (2024). Generalized Jersey Number Recognition Using Multi-task Learning With Orientation-guided Weight Refinement. arXiv preprint arXiv:2406.01033.

Liu, H., & Bhanu, B. (2022). Jede: Universal jersey number detector for sports. IEEE Transactions on Circuits and Systems for Video Technology, 32(11), 7894-7909.

Vats, K., Fani, M., Clausi, D. A., & Zelek, J. (2021, October). Multi-task learning for jersey number recognition in ice hockey. In Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports (pp. 11-15).

Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A. Clausi, and John Zelek. 2023. Jersey Number Recognition using Keyframe Identification from Low-Resolution Broadcast Videos. In Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports (MMSports '23). Association for Computing Machinery, New York, NY, USA, 123–130. https://doi.org/10.1145/3606038.3616162

Bhargavi D, Gholami S, Pelaez Coyotl E. Jersey number detection using synthetic data in a low-data regime. Front Artif Intell. 2022 Oct 6;5:988113. doi: 10.3389/frai.2022.988113. PMID: 36277169; PMCID: PMC9583843.

Bautista D, Atienza R. Scene text recognition with permuted autoregressive sequence models. InEuropean conference on computer vision 2022 Oct 20 (pp. 178-196). Cham: Springer Nature Switzerland. https://arxiv.org/abs/2207.06966

Chen, Q., Poullis, C. (2023). Tracking and Identification of Ice Hockey Players. In: Christensen, H.I., Corke, P., Detry, R., Weibel, JB., Vincze, M. (eds) Computer Vision Systems. ICVS 2023. Lecture Notes in Computer Science, vol 14253. Springer, Cham. https://doi.org/10.1007/978-3-031-44137-0_1

Koshkina, M., & Elder, J. H. (2024). A General Framework for Jersey Number Recognition in Sports Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3235-3244).

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). *ViTPose: Simple Vision Transformer baselines for human pose estimation*. arXiv preprint arXiv:2204.12484. https://arxiv.org/abs/2204.12484

Bautista, D., & Atienza, R. (2022). Scene text recognition with permuted autoregressive sequence models. arXiv preprint arXiv:2207.06966. https://arxiv.org/abs/2207.06966

Wieczorek, M., Rychalska, B., & Dabrowski, J. (2021). On the unreasonable effectiveness of centroids in image retrieval. arXiv preprint arXiv:2104.13643. https://arxiv.org/abs/2104.13643

Li, G., Xu, S., Liu, X., Li, L., & Wang, C. (2018). Jersey number recognition with semi-supervised spatial transformer network. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1864–1867). IEEE. https://doi.org/10.1109/CVPRW.2018.00231

Comandur, B. (2022). Sports Re-ID: Improving re-identification of players in broadcast videos of team sports. arXiv preprint arXiv:2206.02373. https://doi.org/10.48550/arXiv.2206.02373

Vats, K., McNally, W., Walters, P., Clausi, D. A., & Zelek, J. S. (2022). Ice hockey player identification via transformers and weakly supervised learning. arXiv preprint arXiv:2111.11535. https://doi.org/10.48550/arXiv.2111.11535

7. Timeline & Milestones

Feb 27, 2025: Submission of proposal

Mar 5, 2025: Test ImageNet Pretrained model, data processing and augmentation

Mar 12, 2025: Fine-tuning the pretrained model and testing on our dataset

Mar 19, 2025: Figure out the optimal solution to substitute the legibility classifier

Mar 26, 2025: Presentation and working demo

8. Team Contributions

We are splitting the work and making sure everyone in the group has a similar workload and contribution to the project. Our Github repository can be found at: https://github.com/Varun2711/COSC419B-Project

---