Tracking and Identification of Ice Hockey Players

Qiao Chen[1] and Charalambos Poullis[1]

Concordia University, Montreal, Canada

Abstract. Due to the rapid movement of players, ice hockey is a high-speed sport that poses significant challenges for player tracking. In this paper, we present a comprehensive framework for player identification and tracking in ice hockey games, utilising deep neural networks trained on actual gameplay data. Player detection, identification, and tracking are the three main components of our architecture. The player detection component detects individuals in an image sequence using a region proposal technique. The player identification component makes use of a text detector model that performs character recognition on regions containing text detected by a scene text recognition model, enabling us to resolve ambiguities caused by players from the same squad having similar appearances. After identifying the players, a visual multi-object tracking model is used to track their movements throughout the game.

Experiments conducted with data collected from actual ice hockey games demonstrate the viability of our proposed framework for tracking and identifying players in real-world settings. Our framework achieves an average precision (AP) of 67.3 and a Multiple Object Tracking Accuracy (MOTA) of 80.2 for player detection and tracking, respectively. In addition, our team identification and player number identification accuracy is 82.39% and 87.19%, respectively. Overall, our framework is a significant advancement in the field of player tracking and identification in ice hockey, utilising cutting-edge deep learning techniques to achieve high accuracy and robustness in the face of complex and fast-paced gameplay. Our framework has the potential to be applied in a variety of applications, including sports analysis, player tracking, and team performance evaluation. Further enhancements can be made to address the challenges posed by complex and cluttered environments and enhance the system's precision.

Keywords: Object tracking · Text recognition · Player tagging.

1 Introduction

The computer vision community is becoming increasingly interested in automated sports video analysis because it can provide insights into the game plan decision-making process, aid coaching decisions, and make the game more exciting for spectators. Nevertheless, player tracking and identification in fast-paced sports, such as ice hockey, is difficult due to the rapid movement of players and the puck, as well as the occlusions and complex motion of the players. This paper proposes a player tracking and identification system for fast-paced sports based

on deep neural networks and demonstrates its applicability to ice hockey in order to address this challenge. The proposed system is comprised of three major elements: object detection, text detection and recognition, and player tracking.

The Faster R-CNN object detector is initially trained to recognise people in each video frame. To determine the jersey number of each player, we detect the region of the jersey number on the back of the player jerseys using a scene text recognition model [2] and fine-tune the CRAFT text detector [3]. The combination of object detection and text detection provides a more precise and reliable method for identifying players. Tracking multiple players is difficult due to the similar appearance of players on the same team, as well as the occlusions and complex motion of the players, which increase the difficulty. To address this challenge, we employ the Neural Solver Mot [6], a framework for visual multi-object tracking that can track multiple objects in video sequences using rudimentary data association and state estimation techniques.

We evaluated our proposed system using real-world data and obtained an 80.2 Multiple Object Tracking Accuracy (MOTA) score, which is superior to existing state-of-the-art methods. Moreover, we propose a practical method of transfer learning and fine-tuning a text detection model on player jersey numbers that achieves an accuracy of 87.19%. Our contributions consist of a comprehensive framework for player tracking and identification in ice hockey, which can provide valuable insights into the game plan decision-making process, and a practical method of transfer learning and refining a text detection model for player jersey numbers.

2 Related Work

2.1 Dataset

State-of-the-art techniques for player tracking, such as [8] and [35], have achieved impressive results using broadcast National Hockey League (NHL) videos. However, publicly available benchmark datasets that provide identification information for the teams and players are currently lacking. In this paper, we address this issue by using the McGill Hockey Player Tracking Dataset (MHPTD) [38], which is a publicly available dataset that consists of NHL broadcasting videos taken by the primary game camera. We use the MHPTD dataset and augment it with player identification jersey number labels to provide a publicly available benchmark dataset that enables accurate player tracking and identification in ice hockey videos. This will facilitate further research and development in this important area of sports video analysis.

2.2 Player detection

The foundation of the majority of player detection algorithms has been the Viola-Jones Object Detection Framework [36] and Histograms of Oriented Gradients for Human Detection (HOG) [10]. These methods relied on hand-crafted features and segmentation and recognition of players. However, modern deep learning-based algorithms have eliminated the majority of the drawbacks of these early attempts at human recognition [25,31]. AlexNet [17], which won the Imagenet

Large Scale Visual Recognition Challenge (ILSVRC) [30], marked the beginning of deep neural networks. Since then, owing to the continual improvement and advancements in hardware and convolutional neural network methodology, numerous new robust solutions have been developed to address the challenges in sports videos. Today, object detection heavily relies on deep neural networks such as YOLO [26], or part-based approaches [32], which provide superior performance in terms of missed, false, duplicate, and unreliable detection boundaries. Chan et al. suggested a residual network (ResNet) [14] as the CNN base with recurrent long short-term memory (LSTM) [15] for player identification. Vats et al. [35] introduced a temporal 1D CNN with no other specialized networks for processing temporal information. Region-based Convolutional Neural Networks (R-CNN) [13], Fast R-CNN [12], and Faster R-CNN [27], which are considered to be state-of-the-art deep learning visual object detection algorithms, might be the most effective and widely used approach to object detection.

In our work, we utilize the Faster R-CNN [27] algorithm due to its rapid convergence when generating detection proposals. The Faster R-CNN algorithm consists of two parts: a region proposal network (RPN) and a region-based convolutional neural network (RCNN). The RPN generates a set of object proposals by sliding a small network over the convolutional feature map output by the backbone network. The RCNN then refines the proposals and performs classification. This two-stage approach allows the Faster R-CNN to achieve high accuracy while still being computationally efficient. In summary, while early player detection algorithms relied on hand-crafted features and segmentation and recognition of players, modern deep learning-based algorithms have significantly improved the accuracy of player detection in sports videos. Today, object detection heavily relies on deep neural networks such as Faster R-CNN [27], which allows for accurate and efficient player detection.

2.3 Player Tracking

Tracking-by-detection multi-object tracking frameworks such as SORT [5] and Deep SORT [37] have gained attention as they can track multiple objects in video sequences using rudimentary data association and state estimation approaches. The trackers compare detections using various measures such as features, Kalman filter, and person re-identification (ReID). SIFT[21], SURF [4], and ORB [29] are the most popular descriptors for feature extraction and matching in object tracking systems. ORB is frequently used for tracking, mapping, and relocalization due to its quick feature extraction and tolerance to picture rotation and noise. Kalman filter [16] is widely used to track moving objects by estimating their velocity and acceleration based on their locations. The primary function of the Kalman filter is to associate detections with trajectories [5,37,39]. In recent years, unscented Kalman filtering techniques have also been employed to track multiple moving objects with occlusion [9]. Another approach to tracking-by-detection is person re-identification (ReID), which is the process of identifying people across several images. For detection and ReID, ice hockey player tracking approaches such as [7] employ hand-crafted features. Ahmed et al. [1] present a method of learning features and accompanying similarity metrics

for person re-identification. If the images contain the same person, the network returns either a similarity score between the images or a classification of the images as identical. Neural Solver Mot [6] is a recent work that jointly learns features over the global graph of the entire set of detections and predicts final solutions. Our tracking system leverages the Neural Solver Mot [6] and employs ReID measures instead of face recognition and number detection since faces and jersey numbers are not always visible to the primary camera during sporting events.

2.4 Number Recognition

Jersey number recognition algorithms can be classified into two categories: OCR-based methods and CNN-based methods. OCR-based methods, such as those presented in [22] and [24], use hand-crafted features to localize the text or number regions on the player uniform and then pass the segmented regions to an OCR module for recognition of the text or number. These methods have been used for a long time but have been outperformed by CNN-based models, which have shown superior results in terms of number recognition, as reported in [11], [23] and [33]. However, CNN-based models have the disadvantage of being limited to the training set. Typically, jersey number detection follows a localization and recognition step. After character cells are detected, the recognition proceeds by first recognizing each word and then resolving ambiguous cases. Classes that share at least one digit are susceptible to erroneous recognition. To address this problem, digit-wise techniques presented by Li et al. [18] and Gerke et al. [11] have been proposed. These techniques fuse with the spatial transformer network (STN) to improve recognition. Jersey number recognition is difficult due to variations in player poses and viewpoints. To overcome this challenge, CRAFT [3], a scene text detection method, has shown promising results in challenging scenes with arbitrarily-oriented, curved, or deformed texts. In our work, we fine-tune the CRAFT text detector for jersey number recognition, which achieved an accuracy of 87.19%.

3 System Overview

The proposed solution for player identification consists of four steps: player detection, player tracking, team identification, and jersey number recognition. The first step involves using the region proposals generated by Faster R-CNN to detect players in each frame of the video sequence. These detections are then used to track the players throughout the image sequence using similarity metrics for person re-identification. This produces a set of tracklets that describe the motion path of each player in the video. The next step involves identifying the team for each player in the tracklets. This is done by extracting the dominant color from the region containing the player and using it to classify the team to which the player belongs. Finally, for each player tracklet, the system identifies the jersey number of the player. This is accomplished using the CRAFT scene text detection method to recognize the digits on the back of the player's jersey. Figure 1 provides an overview of the proposed system's pipeline, depicting how the different components work together to achieve player tracking and identification. In the following sections, each component of the system is described in more detail.

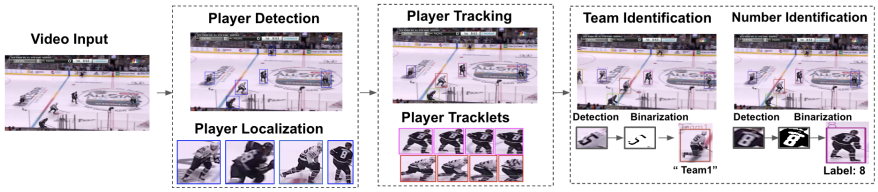


Fig. 1: Overview of player tracking and identification system.

3.1 Player detection

The first stage of the proposed pipeline is player detection, which is performed on a sequence of images obtained by converting a video of an ice hockey game. The detection process begins with preliminary person detection since players are instances of the class person. Subsequently, all further processing is performed solely on the regions where a human is detected. Several deep learning-based frameworks such as YOLO [26] and Faster R-CNN [27] have been used for accurate object identification. For this purpose, we used the Faster R-CNN Inception-V2-COCO model due to its higher performance. This model has been trained on 91 categories of objects from the Common Objects in Context (COCO) dataset [20]. Faster R-CNN produces plausible bounding boxes from an image using convolutional feature maps, which are region-based detectors. The Region Proposal Network (RPN) classifier is then applied, which simultaneously regresses region boundaries and objectness scores at each point on proposed regions/bounding boxes. RPNs can accurately predict region proposals with varying scales and aspect ratios. Finally, post-processing techniques such as non-maximum suppression are used to refine the bounding boxes, eliminate duplicate detections, and re-score the bounding boxes depending on other objects in the scene, after the region has been predicted. The Faster-RCNN loss function can be expressed as follows:

$$L(\mathcal{P}i, b_i) = \frac{1}{S_c} \sum iL_c(\mathcal{P}i, Gp_i) + w \times \frac{1}{S_r} \sum i\mathcal{P}_i L_r(B_i, Gb_i) (1)$$

where i is the index of an anchor in a batch and $\mathcal{P}_i$ is the probability of anchor i being an object. b_i is a vector representing the coordinates of the predicted bounding box. S_c and S_r are the normalization mini-batch size of classification and regression, respectively. L_c and L_r are the classification and regression loss, respectively. The ground-truth label Gp_i is 1 if the anchor is positive and is 0 if the anchor is negative. B_i is a vector representing the coordinates bounding box, and Gb_i is that of the ground-truth box. The two terms are weighted by a balancing parameter w.

3.2 Player Tracking

After detecting players in the first step, the second step of the proposed pipeline is individual player tracking. For this, the Neural Solver Mot [6] is used on the ice hockey dataset as the tracker to generate player tracklets. However, since perfect detection is challenging, errors in detection need to be considered using person re-identification (ReID) metrics. In multiple object tracking (MOT) techniques, external ReID databases are commonly incorporated. The network used in this step is pre-trained on three publicly available datasets for the ReIdentification (ReID) task: Market1501 [40], CUHK03 [19], and DukeMTMC [28].

Algorithm 1: Player Tracking

Input : Player Detections $P = \{p_1, ..., p_n\}$
MOT Graph $G = (V, E)$,
Fractional solution $\hat{F}$
Output: set of *tracklets* $T' = \{T_1, \ldots, T_m\}$ $e = 0$ for all e in $G = (V, E)$ #Initialization
for $p_i \in P$ **do**
  node v represents p_i, $v \in V$
  if p_i *has the same trajectory* T & *is temporally consecutive* **then**
    $e = 1$ in $G = (V, E)$
  else
    $e = 0$ in $G = (V, E)$
end
for $\{(e_1, e_2), ..., (e_{n_{i-1}}, e_{n_i})\} \in E$ **do**
  if $\hat{F}_(e_{n_{i-1}}, e_{n_i}) \geq \tau_\theta$ **then**
    T_i =1, $T_i \in \{T_1, ..., T_m\}$
  end
  else
    T_i =0
  end
end

The pseudocode for player tracking is given in Algorithm 1. The input consists of a collection of player detections $P = p_1, \ldots, p_n$, where n is the total number of objects across all frames. Each detection is represented by $p_i = (a_i, c_i, t_i)$, where a_i denotes the raw pixels of the bounding box, c_i includes its 2D image coordinates, and t_i its timestamp. A tracklet is defined as a collection of time-ordered object detections $T_i = p_{i_1}, \ldots, p_{i_{n_i}}$, where n_i is the number of detections that comprise the trajectory i. The objective of MOT is to identify the set of tracklets $T' = T_1, \ldots, T_m$, which provides the best explanation for the observations O. This problem can be modelled as an undirected graph $G = (V, E)$, where $V = 1, \ldots, n, E \subset V \times V$, and each node $i \in V$ represents a unique detection $p_i \in O$. The set of edges E is formed so that each pair of detections, i.e., nodes, in separate frames is connected, thus enabling the recovery of tracklets with missed detections.

In the player tracking step, the pseudocode checks whether the detection has the same trajectory and is temporally consecutive. If these conditions are met, a node v is added to the graph G. The set of edges E is then formed by connecting each pair of detections in separate frames. The evaluation of each tracklet is based on a fractional solution $\hat{F}$, which is determined by a similarity metric. If the similarity is greater than or equal to a threshold τ_θ, a tracklet is generated for that player. Finally, the output of this step is a set of tracklets T' that describes the motion path of each player in the image sequence.

3.3 Player Identification

Players are identified by their jersey numbers. Recognizing jersey numbers is challenging since jerseys are deformable objects that can appear somewhat distorted in the image. Moreover, there is great variation in the players' poses and camera view angles, which has a significant impact on the projected area and perceived font of jersey numbers.

Number/text detection Following the extraction of the tracklets for each player, the jersey number is identified. We apply a scene text detection algorithm [3] to each image in a player's tracklet, in order to localize the jersey number region. The model architecture has a VGG-16 backbone network [34] and is supervised to localize character regions and connect the regions from the bottom up. Using the pretrained CRAFT model [3], we identify texts of diverse horizontal, curved and arbitrary orientations. The model generates two-channel

score maps: a region score for each character's location and an affinity score for associating characters to instances. The loss function L is defined as follows:

$$L = \sum_i [S_r(i) - S'_r(i)]_2^2 + \sum_i [S_a(i) - S'_a(i)]_2^2 \tag{2}$$

where $S'_r(i)$ and $S'_a(i)$ indicate region score and affinity map of the ground truth respectively, and $S_r(i)$ and $S_a(i)$ indicate the predicted region score and affinity score, respectively. We further improve the robustness by extending it with a post-processing step that filters out text instances that are unlikely to be a jersey number based on the aspect ratio of the detected region.

Method	MOTA↑	IDF1↑	MT↑	FP↓	FN↓
SORT [5]	55.1	76.3	404	615	1296
Deep SORT [37]	56.3	77.1	435	487	968
MOT Neural Solver [6]	67.3	80.2	656	422	917

Table 1: Comparison of different approaches for multiple object tracking in a video clip.

Number identification All number-detected regions are further processed for (i) team identification and (ii) number identification to identify each player's tracklets.

- **Team identification.** In team identification, the aim is to separate the detected jersey numbers into two groups corresponding to the two teams playing in the video. This is necessary because in some cases, two players may have identical jersey numbers. To achieve this, the input patches are binarized and the dominant color of each patch is analyzed. Patches with a dark foreground color on a bright background are classified as white team jerseys, while patches with a bright foreground color on a dark background are classified as black team jerseys. If a team roster is available, the process also eliminates false detections by removing jersey numbers that do not exist in the team roster.
- **Number identification.** The second step, number identification, involves recognizing the jersey numbers for each player. This is achieved through a pre-trained model for TPS-ResNet-BiLSTM-Atten text recognition, which is a four-stage framework for scene text recognition. This model is capable of recognizing the jersey number as a whole, which is essential for identifying multiple-digit jersey numbers taken from non-frontal, distorted views.
  The approach of Baek et al. [2] describes two types of implementations for the text recognition model: Connectionist Temporal Classification (CTC) and Attention mechanism (Attn). CTC involves computing the conditional probability by summing the probabilities that are mapped onto the label sequence, as in equation 3:

$$\mathcal{P}(S_l|S_i) = \sum_{\pi:M(\pi)=S_l} \mathcal{P}(\pi|S_i) \tag{3}$$

  where S_l is the label sequence, S_i is input sequence and $\mathcal{P}(\pi|H)$ is the probability of observing either a character or a blank at a point in time, and M is the mapping of π onto S_l. The Attn approach uses an LSTM attention

decoder to predict the output at each time step, using trainable parameters, context vectors, and hidden states from the LSTM decoder as follows,

$$O_t = softmax(W_0 h_t + p_0) \tag{4}$$

$$h_t = LSTM(O_{t-1}, c_t, h_{t-1}) \tag{5}$$

where W_0, p_0 are the trainable parameters, c_t is a context vector, and h_t, h_{t-1} represent the decoder LSTM hidden states at time steps t and $t-1$, respectively.

4 Results

4.1 Dataset

The most relevant state-of-the-art methods to our tracking system are [35] and [8], however direct comparison is not possible because the authors do not provide their datasets. Thus, we report on the MHPTD dataset[38], a publicly available dataset which consists of 25 NHL gameplay video clips of resolution 1280 × 720 pixels. Each clip consists of a single shot of the gameplay from the overhead camera position comprised of a sequence of frames that run continuously without a cut scene or camera view change. The clips have mixed frame rates, including both 60 and 30 frames per second, which are standard NHL broadcast video frame rates available on the market. To facilitate the evaluation of the player tracking and identification, we augment the ground truth tracking information provided by MHPTD with manually labeled tracking IDs containing the jersey number and team label.

(a) (b) (c)

Fig. 2: (a) A visual comparison of output player tracklets using YOLO_v3 model (top-row) and Faster-RCNN model (bottom-row). (b) Comparison of Text Detection without(top row) or with(bottom row) fine-tuning. (c) Some failed detections, typically occurring when there are complex backgrounds such as banner advertisements which may contain text, player collisions and occlusions, low contrast, and contours resulting from stripes and other logos on the players' jerseys.

4.2 Player Tracking

A Faster-RCNN network [27] pre-trained on the COCO dataset [20] is utilised for player detection. We compared two different object detectors for player detection, the detector presented in Faster-RCNN [27] and the one in YOLO_v3 [26], respectively. There can be erroneous detections owing to misclassification,

occlusions, and the presence of the audience, but the majority of them can be filtered with the *tracklets* length and patch size. Figure 2a depicts an example of a *tracklet*, a sequence of images of a tracked player, with YOLO_v3 in the top row, and Faster-RCNN in the bottom row. The Faster-RCNN model recognizes a more comprehensive player zone. For the test videos, the object detector of Faster-RCNN achieves an average precision (AP) of 66.8, whereas YOLO_v3 achieves an average precision (AP) of 53.32. We tested three cutting-edge tracking algorithms on a dataset of hockey players, and used Multiple Object Tracking Accuracy (MOTA) and IDF1 Score (IDF1) as the main evaluation metrics. The authors also mention a third metric, Mostly Tracked (MT) trajectories, which refers to the trajectory coverage. The study reports the results of the evaluation in Table 1. According to the results, the MOT Neural Solver tracking model with person re-identification (reID) re-trained on the hockey dataset achieved the best tracking performance. The reported average for MOTA was 56.3, and the average for IDF1 was 60.67, according to the authors. However, based on the experiments conducted in the study, the MOT Neural Resolver algorithm achieved the highest average MOTA and IDF1 scores on the test videos, which were 67.3 and 80.2, respectively.

4.3 Player identification

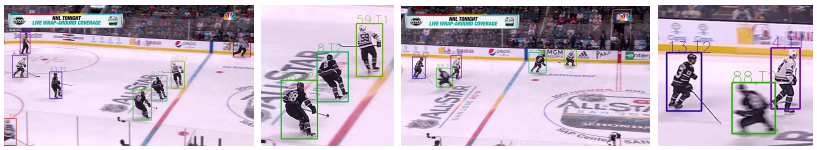
Text detection We adjust the pretrained weights of the CRAFT detector to the ice hockey dataset by performing fine-tuning. The fine-tuning process involved training the model for 30 epochs using a learning rate of $3.2e-5$, with 500 images from the dataset used for this purpose. The remaining subsets were used for testing and validation. During training, the authors performed image augmentation techniques, such as affine transformation, Gaussian blur, and modulation of the color channels, on both the original player images and the corresponding bounding boxes of the jersey number regions. This helped to improve the robustness of the model to variations in the input images.

Figure 2b provides an example of text detection using the fine-tuned CRAFT detector on the ice hockey dataset. The authors report that fine-tuning the pretrained model on the ice hockey dataset resulted in enhanced detection of jersey number regions, which is important for identifying players in the video.

However, the authors also note that there were some unsuccessful detections, which typically occurred in the presence of complicated backdrops such as banner advertisements containing text, player collisions and occlusions, low contrast, and contours arising from stripes and other logos on the players' jerseys. Figure 2c provides examples of such failed detections.

Team Identification Using the text region recognised for each player, we binarize and convert the image to black and white. This step simplifies the image and helps to identify the dominant colors in the image. Next, the dominant patch color is identified. If the text is white on a dark backdrop, then the dominant color in the patch will be dark, and the player is assigned to the home team. If the text is black on a bright background, then the dominant color in the patch will be bright, and the player is assigned to the visiting team. This process is based on the assumption that each team has a distinct color for their jerseys and

that the numbers on their jerseys are a contrasting colour. The proposed technique achieved an accuracy of 82.39% in classifying teams based on the color of their jerseys. However, the authors note that some inaccuracies were observed, primarily due to officials being misidentified as players, and colors with poor contrast leading to erroneous detections. Misidentifying officials as players is a common challenge in sports analysis, as officials often wear uniforms that are similar to those of the players. This can lead to inaccurate results if not properly accounted for in the algorithm. Poor contrast between the color of the jerseys and the background can also affect the accuracy of the algorithm by making it difficult to detect the colors of the jerseys accurately.



(a) Example of visualization (b) Close-up of (a) (c) Example of motion blur (d) Close-up of (c)

Fig. 3: Visualization of the output of the tracking system. If a player is tracked, a random coloured bounding box is drawn, and the jersey number label and the identified team are annotated above the box.

Jersey Number Identification For jersey number recognition, we use the pretrained text recognition model TPS-ResNet-BiLSTM-Atten [2]. This model uses an attention mechanism to handle problematic situations such as jersey numbers that share at least one digit, variations in player position, and shifts in camera perspective. The tracking result includes 459 player tracklets extracted from 15,083 player photos, along with the jersey number bounding box annotation and a per-player class. For each player's tracklet, the jersey number label with the most votes is assigned. The recognition accuracy of the jersey numbers is reported to be 87.19%. The authors provide an example of the output of their system in Figure 3, where team and player jersey number identification are overlaid on the input video. If a player is tracked, a bounding box of a random color is generated, and the jersey number label and team are marked above the box. However, the authors note that some unrecognised poses within the same tracklet may be assigned the jersey number label from other frames within the same tracklet, as player tracking does not use jersey number information.

5 Conclusion

We presented a complete framework for player tracking and identification in ice hockey that exploits the high performance of deep learning neural networks. The framework consists of three main components, namely player detection, player tracking, and player identification. We extended the publicly available dataset called MHPTD with jersey number and team information and conducted experiments to evaluate the performance of the proposed framework. The results of the experiments show that the average precision (AP) for player detection using the method is 67.3, the Multiple Object Tracking Accuracy (MOTA) for

player tracking is 80.2, and the accuracies for team identification and player number identification are 82.39% and 87.19%, respectively. Our framework can track multiple players simultaneously in fast-paced games such as ice hockey and that its performance is equivalent to that of cutting-edge player tracking and identification systems. Overall, our results suggest that the proposed framework is effective in tracking and identifying players in ice hockey games using deep learning neural networks. This can be useful for various applications such as sports analysis, player tracking, and team performance evaluation. Further improvements can be made to address the challenges associated with complex and cluttered environments and improve the accuracy of the system.

6 Acknowledgements

This research is based upon work supported by the Natural Sciences and Engineering Research Council of Canada Grants No. RGPIN-2021-03479 (Discovery Grant) and ALLRP 571887-21 (Alliance). Special thanks to Livebarn Inc. for their support.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3908–3916 (2015) 3
2. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4715–4723 (2019) 2, 7, 10
3. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9365–9374 (2019) 2, 4, 6
4. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: European conference on computer vision. pp. 404–417. Springer (2006) 3
5. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016) 3, 7
6. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6247–6257 (2020) 2, 4, 5, 7
7. Cai, Y., Freitas, N.d., Little, J.J.: Robust visual tracking for multiple targets. In: European conference on computer vision. pp. 107–118. Springer (2006) 3
8. Chan, A., Levine, M.D., Javan, M.: Player identification in hockey broadcast videos. Expert Systems with Applications **165**, 113891 (2021) 2, 8
9. Chen, X., Wang, X., Xuan, J.: Tracking multiple moving objects using unscented kalman filtering techniques. arXiv preprint arXiv:1802.01235 (2018) 3
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 886–893. Ieee (2005) 2

11. Gerke, S., Muller, K., Schafer, R.: Soccer jersey number recognition using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 17–24 (2015) 4
12. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015) 3
13. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014) 3
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 3
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 3
16. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960) 3
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012) 2
18. Li, G., Xu, S., Liu, X., Li, L., Wang, C.: Jersey number recognition with semi-supervised spatial transformer network. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1783–1790 (2018) 4
19. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014) 5
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 5, 8
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004) 3
22. Lu, C.W., Lin, C.Y., Hsu, C.Y., Weng, M.F., Kang, L.W., Liao, H.Y.M.: Identification and tracking of players in sport videos. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service. pp. 113–116 (2013) 4
23. Lyu, P., Liao, M., Yao, C., Wu, W., Bai, X.: Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 67–83 (2018) 4
24. Messelodi, S., Modena, C.: Scene text recognition and tracking to identify athletes in sport videos. Multimedia Tools and Applications **63**, 1–25 (01 2012). https://doi.org/10.1007/s11042-011-0878-y 4
25. Okuma, K., Taleghani, A., Freitas, N.d., Little, J.J., Lowe, D.G.: A boosted particle filter: Multitarget detection and tracking. In: European conference on computer vision. pp. 28–39. Springer (2004) 2
26. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 3, 5, 8
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 3, 5, 8
28. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016) 5

29. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011) 3
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) (2012) 3
31. Šaric, M., Dujmic, H., Papic, V., Rožic, N.: Player number localization and recognition in soccer video using hsv color space and internal contours. International Journal of Electrical and Computer Engineering **2**(7), 1408–1412 (2008) 2
32. Senocak, A., Oh, T.H., Kim, J., So Kweon, I.: 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. pp. 1732–1739 (2018) 3
33. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016) 4
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 6
35. Vats, K., Walters, P., Fani, M., Clausi, D.A., Zelek, J.: Player tracking and identification in ice hockey. arXiv preprint arXiv:2110.03090 (2021) 2, 3, 8
36. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. vol. 1, pp. I–I. Ieee (2001) 2
37. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017) 3, 7
38. Yingnan Zhao, Zihui Li, K.C.: A method for tracking hockey players by exploiting multiple detections and omni-scale appearance features. Project Report (2020) 2, 8
39. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021) 3
40. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015) 5