Who Is Your Favourite Player?
Specific Player Tracking in Soccer Broadcast

Tatsuya Nakamura and Katsuto Nakajima
Department of Information Systems and Multimedia Design, Tokyo Denki University, Tokyo, Japan

Keywords: Player Tracking, Player Identification, Digit Recognition, One-Shot Learning.

Abstract: In this paper, we propose a method to identify and track only a specific player out of a number of players wearing the same jersey in the video sequence of a soccer broadcast to make a summary video focusing on the plays of the specific player. In a soccer broadcast, it is not easy to track a specific player because many players on both teams come and go and move across the field. Therefore, we devised a method to overcome this difficulty by combining multiple machine-learning techniques, such as deep neural networks. Our evaluation was conducted for nine players in three different positions and wearing three different color jerseys, and it is shown that although there is room for improvement on the recall, our proposed method can successfully track specific players with a precision of over 90%.

1 INTRODUCTION

In order to make a summary video focusing on the plays of a specific player from a soccer broadcast, it is necessary to identify and track only a specific player from a number of players wearing the same jersey in the video. However, due to frequent scene changes (shot changes) and some telops on the screen, specific player tracking is very challenging.

There are three more reasons why it is difficult to track a specific player in a soccer game. First, because the playing field is very wide, the size of the players appearing in the video is relatively small, making it difficult to resolve the features effectively for the identification and tracking. Second, the players are often occluded by other players involved in the game. Third, the color appearance of the players may change dramatically due to direct sunlight and the shade of the stadium. Additionally, the possible unsteady motion of camera (“shaky camera”) in the broadcast makes it even more difficult to track a specific player because of excessive motion blur and the fact that the target player is not always in the frame.

In this paper, we propose a method of combining multiple machine-learning techniques such as deep neural networks to achieve reliable identification and tracking of the target player. Our evaluation shows that, though there is a scope for improvement on the recall, our proposed method is successful in tracking a specific soccer player with a precision of over 90% for nine players in three different positions and wearing three different color jerseys.

2 RELATED WORKS

Although various methods of player tracking have been proposed so far (Yang et al., 2017; Liu et al., 2009), few methods identify and track only a specific player. As a specific player in a soccer match must be identified among the ten players who wear the same jersey (except for the goalkeeper), only limited information can be utilized to distinguish that player.

As in a video broadcast, it is common to use the player’s face and jersey number for identification, as these do not change during a match. However, sufficiently large number of pixels are necessary to recognize them easily. For this reason, many existing methods did not work well in the field overview scenes (Figure 1(a)) but was successful in the close-up scenes (Figure 1(b)). In these scenes, the players can be captured with enough pixels to identify the player’s face and jersey number relatively easily. In Ballan et al. (2007), face identification was executed in the detected face region, and the label of the corresponding player was assigned. However, in close-up scenes, tracking is difficult due to the narrow angle. On the other hand, face identification is almost impossible in the field overview scenes because of insufficient resolution. Therefore, there has been no

Nakamura, T. and Nakajima, K.
Who Is Your Favourite Player? Specific Player Tracking in Soccer Broadcast.
DOI: 10.5220/0008910807330740
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 733-740
ISBN: 978-989-758-402-2; ISSN: 2184-4321

established method yet to both identify and track a specific player simultaneously.

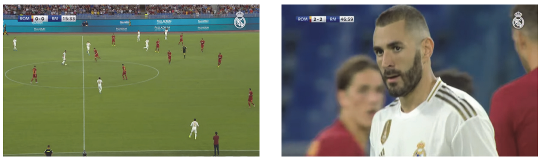

(a) Overview scene (b) Close-up scene

Figure 1: Differences in scene.

As a player's jersey number is uniquely assigned, it is a stronger identifier than faces. Therefore, it is widely used for player identification (e.g., Ye et al., 2005; Saric et al., 2008). Recently, it has become possible to recognize the jersey number in an overview scene by applying one of the several methods based on deep learning (Gerke et al., 2015; Li et al., 2018).

For these above-mentioned reasons, we propose a method to track a specific player by using (a) the jersey number identification before starting to track, (b) person detection for tracking, and (c) face re-identification when missing to continue to track the player.

3 PROPOSED METHOD

3.1 Process Flow

The overall process flow of the proposed method is shown in Figure 2. First, one frame is acquired from the input video, and the unnecessary region for player tracking is deleted. After that, the specific player tracking algorithm composed of the detection, identification, tracking, and search processes are followed. The successfully tracked frames are stored for the summary video. The following sections describe each process in detail.

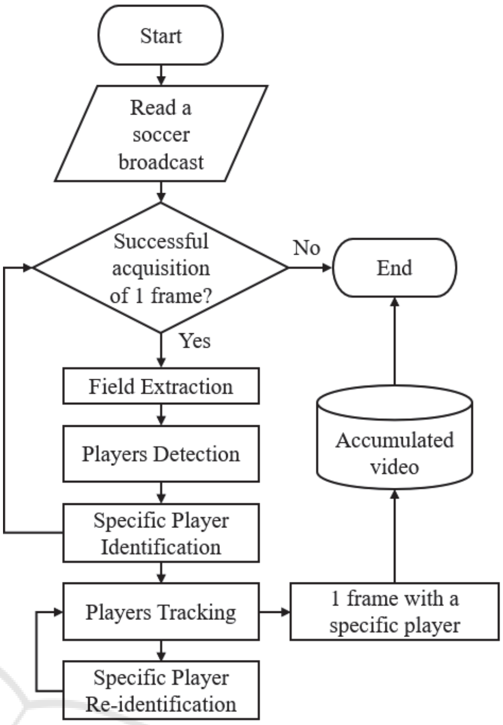

Start
Read a soccer broadcast
Successful acquisition of 1 frame?
No
End
Yes
Field Extraction
Players Detection
Accumulated video
Specific Player Identification
Players Tracking
1 frame with a specific player
Specific Player Re-identification

Figure 2: Flowchart of the proposed method.

3.2 Pre-processing (Field Extraction)

In each video frame of the broadcast, regions such as the spectator stands, which may become an obstacle for player tracking, are excluded. As the color of the soccer field must be green according to the IFAB law, only the field is left and other regions are masked out by a hue value from 80 to 160, which is regarded as the green of a field. To remove the noise, this mask is eroded and dilated before the masking procedure (Figure 3).

3.3 Detection Process

3.3.1 Players Detection

We must know the precise location of the target player in order to accurately identify and track them.

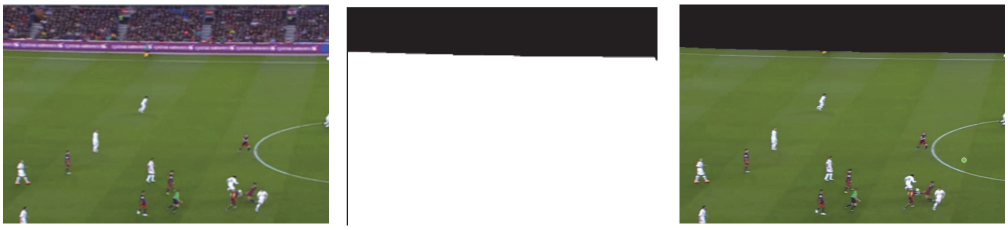
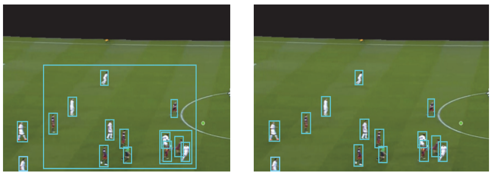

(a) Original image (b) Mask image (c) Field Extraction result

Figure 3: Procedure of field extraction.

Therefore, we utilize a deep-learning-based object detector, YOLOv3 (Redmon et al., 2018), to find the location and size of the player. YOLOv3 with a model trained by the COCO dataset has a good balance of detection accuracy and speed. It can also detect very small objects and output rectangles and show their locations and sizes (Figure 4(a)). However, it may occasionally output large rectangles as the result of a false detection. Therefore, we exclude such a large rectangle (having an area of 1/100 or more of the input frame) because the players would appear smaller than this on the field (Figure 4(b)).



(a) Before delete (b) After delete

Figure 4: Procedure of leaving only small detection rectangles.

3.3.2 Team Discrimination

Out of the person detection rectangles, only the ones that seem to correspond to the team to which the specific player (target player) belongs (the target team) are left as the objects to be tracked in the subsequent process. This improves the tracking accuracy. It is done based on the RGB value of the team jersey registered beforehand (Figure 5). The discrimination is made based on whether the number of pixels in a certain range around the registered RGB value of the jersey (± 6% this time) is equal to or greater than a threshold in the person detection rectangle. In order to cope with the fluctuation of the jersey color observed in the video, the range of the discrimination is dynamically shifted according to the average brightness of the region corresponding to the upper body, which is defined as the position slightly above the center of the rectangle. More specifically, the region from 67% to 50% of the height of the player rectangle with a half width of the rectangle is examined. This region is horizontally shifted by 10 % of the rectangle to the attacking direction (Figure 6). All of these parameters were determined according to our preliminary experiments. This enables the discrimination of the color of the player's jersey even if the player moves away from the area under the sunshine to the shadow area of the field or vice versa. We define the player rectangle in which the player is determined as the target team.

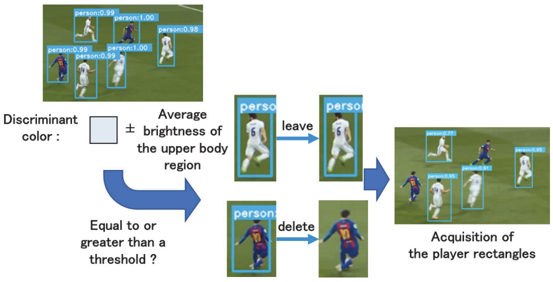

person:0.99
person:1.00
person:0.98
person:0.99
person:1.00
person:0.99
Discriminant color :
±
Average brightness of the upper body region
Equal to or greater than a threshold ?
perso
leave
perso
person
delete
Acquisition of the player rectangles

Figure 5: Procedure of team discrimination.

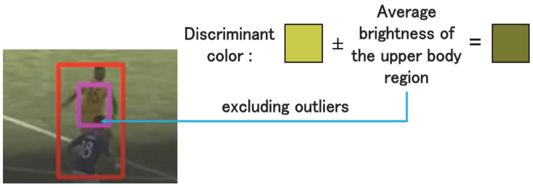


Figure 6: Dynamically fluctuation of discriminant color.

3.4 Identification Process (Jersey Number Recognition)

As we mentioned in Section 2, there are some deep-learning-based methods that can recognize a jersey number in an overview scene. However, at present, a sufficiently large dataset for deep-learning-based recognition of the jersey number is not publicly available, and the preparation of the dataset requires a large amount of labor. Therefore, in this work, we employ a digit recognition model for the house numbers, which was created by a public dataset of house numbers. The dataset contains various typefaces and appearances such as illumination, blur, and distortion, and many of those digits resemble the jersey numbers on the back of players.

3.4.1 Creation of Digit Recognition Model

The SVHM dataset (Netzer et al., 2011) is a collection of house number images from the Google Street View, and it was used to create their digit recognition model. The SVHM dataset consists of a group of RGB images of multiple digit fonts, including the background. It contains a variety of digits with different colors, shapes, and defects (scratches, chips, etc.), many of which look similar to blurred jersey numbers in a soccer broadcast. Figure 7 shows some examples of the jersey number images, and the recognition result of our tentative evaluation using the model trained by SVHM.

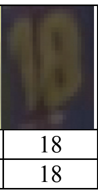
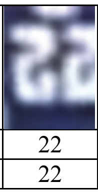
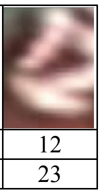
input					
GT	4	18	10	22	12
output	4	18	10	22	23

Figure 7: Jersey number image and recognition result.

These results are sufficient for our purpose. Various color digits with low illumination, low resolution, and deformation can be recognized. Though the jersey number with excessive inclination and blurring causes a mis-recognition, it is possible to reduce the risk of such a mis-recognition in the specific player tracking by accepting only the result with a high confidence for the registered jersey number of the target player.

3.4.2 Application to Jersey Number Recognition

In order to incorporate the created digit recognition model into our identification process, it is necessary to extract the jersey number region in the player rectangle detected from every frame automatically. In this extraction, we assume that the player stands upright to some extent. In this assumption, we define the jersey number region to extend from 75% of the height of the player rectangle to 50% of the height of the player rectangle. The width is one-third the width of the player rectangle. This size in the rectangle was also based on our preliminary study. In this work, only one kind of jersey number is recognized, and the identification is considered to be a success when the confidence is over 0.85 (Figure 8).

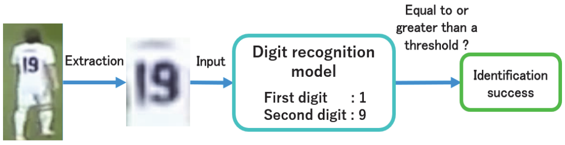


Figure 8: Procedure of jersey number recognition.

3.5 Tracking Process

3.5.1 Forward Centroid Tracking Algorithm

The first frame in which the jersey number identification succeeds is set as the base frame, and the tracking process starts with that frame. The procedure is described below.

Step 1: Find the centroid of the player rectangles

The centroid (center position) of the player rectangle is calculated by the upper left and lower right coordinates, and each centroid is assigned a unique ID. A flag is also assigned only to the player rectangle in which jersey number identification was successful so that the player can be distinguished from others.

Step 2: Associate the player rectangles between the previous and current frames

As the player detection process is executed for every frame, the centroid at each frame can be calculated. If the frame rate is about 30 frames per second, the distance players move between two adjacent frames is very small. Therefore, the ID of the closest centroid in the previous frame is handed over to the centroid in the current frame if the distance between those centroid locations is shorter than a threshold (200 pixels in this implementation), which is set at about two times the longest rectangle height.

During tracking, the number of players detected in a frame may increase or decrease. The player rectangle that cannot be associated with any rectangle in the previous frame and has no ID, will be registered as a new tracked player and assigned a new ID. On the other hand, if some pre-assigned ID cannot be associated with any others in the current frame, either the ID recognition process has failed or the player himself disappeared from the current frame. For this ID, the centroid position is kept unchanged and the number of absent frames is counted. If its counts exceed a certain threshold (30 frames = 1 second in this implementation), the ID is deleted from the list of tracked players. If the ID is successfully associated, its absent count is reset to zero. If the deleted ID is of the target player, the tracking process ends, and the search process starts.

3.5.2 Backward Centroid Tracking Algorithm

It is often the case that the target player is not found for a while after the camera shot is changed. On the other hand, the player can often be found when his jersey number is seen. As the purpose of this work is to make a summary video offline, we decided to trace back from the frame (the base frame) in which the target player is identified. We call it the backward tracking (Figure 9). As described above, tracking all the player rectangles from the beginning of a shot enables this backward tracking.

The backward tracking is carried out as follows. While a specific player cannot be identified, all the player rectangles acquired in the detection process are stored in the computer memory ("memorized") with their IDs. Once the jersey number identification is successful, the backward tracking is performed by the same algorithm as in 3.5.1 using the memorized player rectangles in the previous frames. When the player rectangles run out or the target player is deleted from the tracking player list during backward tracking, the forward tracking starts from the base frame in which jersey number identification succeeded.

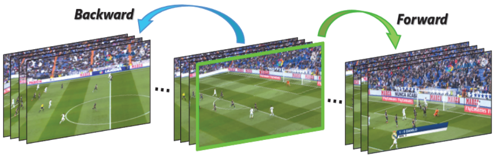


Figure 9: Bidirectional tracking image.

3.6 Searching Process

When the tracking of a specific player is interrupted due to an occlusion with other players or telop on the broadcast screen, or due to a frame out, it is very difficult to search (re-identify) with manually designed features because the players in an overview scene have a very small number of pixels. Therefore, we employ the features learned by a deep neural network. For this purpose, it is necessary to collect the training data for the player during the tracking and to carry out the training in order to construct the model that can re-identify any player in a variety of appearances. There is a possibility of mis-training due to false tracking. Therefore, we introduced a mechanism for periodically initializing the training data and re-learning model using the Siamese Network (Bromley et al., 1993).

3.6.1 Siamese Network + One-Shot Learning

The Siamese Network is a deep metric learning algorithm, in which two networks share weights and each receive inputs such as vectors and images. It learns the difference (distance) between two inputs. For images, the Siamese Network learns whether the class of two image pairs is the same or different. It is not used here for the classification of input images, and so it does not need a large amount of training data. In an extreme case, one training data per class is enough. For that reason, it is used for One-shot learning and Few-shot learning (Koch et al., 2015).

In this work, we assume that the appearance of the head and face (what we call "face" hereafter) does not change much when the target player is framed in again soon after (within 30 frames) he is framed out. Therefore, the "face" image in the last frame just before the failure to track is used as the training data of One-shot learning for the Siamese Network.

3.6.2 Re-Identification by Siamese Network

The Siamese Network requires two networks that share weights. For them, we employ the VGGFace2-ResNet 50 (Cao et al., 2018) that was pre-trained by VGG-Face2. It works as a "face" feature extractor, and our Siamese Network calculates the similarity between two "face" images.

Our preliminary experiment showed that it was necessary to extract accurately the "face" region from the player rectangle in order to carry out the re-identification of the target player with a good accuracy using the similarity computed by the Siamese Network. As the "face" region in an overview field scene is too small to recognize even by a state-of-the-art object detector, we adopt a simple way to extract the "face" region. Assuming that the players are standing upright, we use an elliptical binary mask (Figure 10 (a)) that approximates the "face" region. By examining the player rectangles, we decided that the elliptical mask is from the top to 86% of the height of the rectangle and has a width of one-fifth the width of the rectangle. It is applied to the original image (Figure 10 (b)) to get the result (Figure 10 (c)), which is then used as an input image to the Siamese Network.

(a) Mask (b) Original image (c) Result image

Figure 10: Face region extraction using a prepared mask.

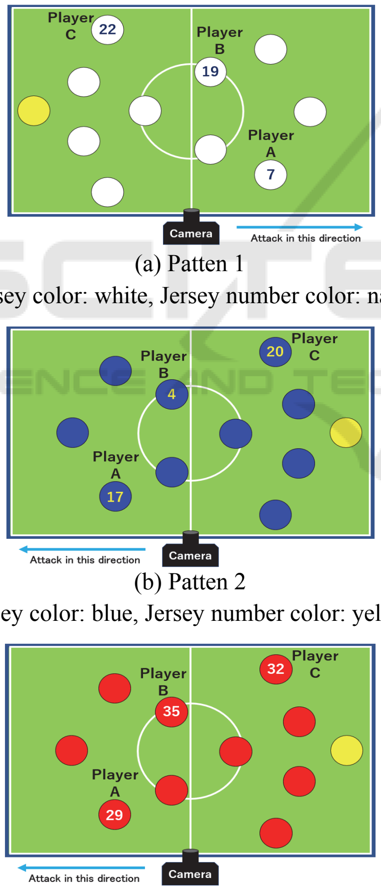
The training of the Siamese Network was conducted in a brute-force way. The "face" image of the specific player (true data) and other players' "face" images (false data) were used for training. We chose a player closest to the specific player as the false data. All possible pairs allowing the same data to be used for both inputs are learned. The training was carried out with a batch size of 1, 20 epochs, and Adam optimizer according to (Kingma et al., 2015). These hyperparameters were chosen for taking a balance between the speed and the accuracy of our on-line learning of the Siamese Network. We define the re-identification to be successful when the similarity for the pair of true data and false data is over 0.85, as a threshold for strict re-identification. When the re-identification is successful, the centroid of the player rectangle is assigned with the ID of the specific player and the tracking process resumes. After that, both the weights of the Siamese Network model and the training data are initialized.

4 EVALUATION

4.1 Experiments

The proposed method was applied to the actual broadcast videos and its performance was evaluated. Figure 11 shows the positions and formations of the teams of the selected players in the experiment. In this evaluation, we chose nine players as the specific player, three players each from three different teams wearing different color jerseys. The three players in a team are selected from different positions: forward (player A), midfielder (player B), and defender (player C).

All videos in our evaluation contain 10 minutes from the start of the match and the resolution is 1280 × 720 pixels.



(a) Patten 1
Jersey color: white, Jersey number color: navy

(b) Patten 2
Jersey color: blue, Jersey number color: yellow

(c) Patten 3
Jersey color: red, Jersey number color: white

Figure 11: Players' position and jersey number used for evaluation. Yellow circles represent goalkeepers.

As there is no modification to the video, it includes both close-up scenes and field overview scenes. Therefore, we detected the shot switching using the optical flow technique. When a shot is switched, a specific player is left in an unidentified state, and all the states for track and search are initialized. The scenes except for the overview ones are excluded from our tracking.

The proposed method is evaluated by the precision, recall, and F-measure of the specific player identification in the overview scene. The definition of the categories for the evaluation index is shown in Table 1.

Table 1: The definition of the categories for the evaluation index.

True Positive (TP)	False Positive (FP)
Identify a specific player correctly	Identify a specific player not correctly
False Negative (FN)	True Negative (TN)
Not identified when a specific player is in the frame	Not identified when there is no specific player in the frame

4.2 Results and Discussion

4.2.1 Performance Comparison of Each Player Position

Table 2 shows the results of three videos. For each video, the precision, recall and F-measure are compared across all three player positions (forward, midfielder, and defender).

Table 2: Performance evaluation (%).

(a) Video 1

Player	Precision	Recall	F-measure
A	93.6	37.7	53.8
B	94.3	44.1	60.1
C	96.2	19.4	32.3

(b) Video 2

Player	Precision	Recall	F-measure
A	95.8	41.0	57.4
B	92.6	46.1	61.6
C	97.4	18.3	30.8

(c) Video 3

Player	Precision	Recall	F-measure
A	97.6	54.8	70.2
B	97.8	60.2	74.5
C	97.9	13.7	24.0

As there were many shots in which no jersey number identification was achieved at all in any frame during the shot, the recall can not be high for all players. Once a number identification was done, the precision exceeds 90% for all players. Judging from the F-measure, the mid-field player B had the best tracking performance in all videos.

The reason for this is probably because player B is often in a position where the jersey number can often be identified, and the frame-out frequency is moderate.

Forward player A is in the position where offensive players almost always turn their back to the camera in an overview field shot. Therefore, the identification of the jersey number recognition was highly successful. However, frame-outs of player A occurred very frequently, which made it very difficult to increase the recall performance. In order to improve the re-identification process, we must study the number and selection method of the training data used for the Siamese Network. For example, the hair style of the player is not included in the masked “face” region. The elliptical mask is too simple to extract this feature of the player. Effective feature extraction for re-identification is an item left for future work.

For the defense player C, our result shows that even if players look small in the video, it is possible to recognize their jersey number correctly by our method. However, because they are almost always facing the camera in an overview scene, their jersey numbers are rarely seen and difficult to identify, and the recall becomes very low.

For all positions, in order to improve the recall (correct tracking of the target player), we must employ other information to find and track a specific player. This can include the color of soccer cleats, hair, etc. of the individual player. In addition, the playing trajectory of players may also be helpful to determine the “position” of the players, which would be a strong cue to find the specific player. We leave how to incorporate them to future work as well.

Even though the recall is not satisfactory as mentioned above, our method is effective for the tracking of a specific player because the precision of all players exceeds 90%.

4.2.2 Performance Comparison of Tracking Methods

For the player B, whose performance was the best in terms of the F-measure, we verified the effectiveness of backward tracking. The results are shown in Table 3.

Table 3: Performance comparison with tracking methods using Player B only (%).

(a) Video 1

Method	Precision	Recall	F-measure
Forward only	97.4	32.8	49.0
Forward & backward	94.3	44.1	60.1

(b) Video 2

Method	Precision	Recall	F-measure
Forward only	95.1	30.9	46.9
Forward & backward	92.6	46.1	61.6

(c) Video 3

Method	Precision	Recall	F-measure
Forward only	98.5	49.7	66.1
Forward & backward	97.8	60.2	74.5

From Table 3, it is evident that adopting backward tracking slightly decreases the precision, but greatly improves the recall. The frames that show a specific player must not be rejected for making of a summary-video. Therefore, the adoption of our backward tracking must be very effective for this off-line purpose.

When the backward tracking is carried out, one of the reasons why the precision decreases is that in the backward time series of three videos, many players are often crowded together, and the identification is attempted even in such a situation. What is important in specific player tracking is not the tracking in the extreme crowding, but the tracking correctly after the crowd disperses. In many cases, most players participate in a set play, which means there is a high probability of the specific player being present in a frame. Therefore, it may be better to exclude the extremely crowded frames for the identification and tracking of individual players, and treat them by a different approach. For example, the crowd detection and the presumption of the existence of the target player in the crowd, could improve the actual recall performance in these cases, and implementing it is also left for future work.

5 CONCLUSIONS

In this paper, we proposed a method for identifying and tracking a specific player in a video sequence of a soccer broadcast for the purpose of making a summary video focusing on the plays of that specific player. The method is mainly based on three deep learning techniques: object detection for player detection, digit recognition for the jersey number identification, and similarity discrimination for re-identification of the target player.

Our evaluation results with nine players from three videos show that the precision of the target player identification is over 90% in all experiments. However, the recall performance could be improved.

In future work, we will extract more sophisticated features for the target player identification, such as hair style for the case of miss-tracking and frame-out. In addition, we will develop a mechanism to improve the actual recall performance by considering the players' position and by finding extremely crowded situations in order to stop tracking of the individual players but keep those frames in which the target player might be present in the summary video.

REFERENCES

Yang, Y., and Li, D. (2017). Robust player detection and tracking in broadcast soccer video based on enhanced particle filter. In *Journal of Visual Communication and Image Representation*, volume 46, pages 81-94.

Liu, Z. J., Tong, X., Li, W., Wang, T., Zhang, Y., and Wang, H. (2009). Automatic player detection, labeling and tracking in broadcast soccer video. In *Journal of Pattern Recognition Letters*, volume 30, no 2, pages 103-113.

Ballan, L., Bertini, M., Bimbo, A., and Nunziati, W. (2007). Soccer players identification based on visual local features. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR)*, pages 258-265.

Ye, Q., Huang, Q., Jiang, S., Liu, Y., and Gao, W. (2005). Jersey number detection in sports video for athlete identification. In *Proceedings of the SPIE Visual Communications and Image Processing (VCIP)*, volume 5960, pages 1599-1606.

Saric, M., Dujmic, H., Papic, V., and Rozic, N. (2008). Player number localization and recognition in soccer video using hsv color space and internal contours. In *Proceedings of the 10th WSEAS International Conference on Automation & Information (ICAI)*, pages 175-180.

Gerke, S., Müller, K., and Schäfer, R. (2015). Soccer jersey number recognition using convolutional neural networks. In *IEEE International Conference on Computer Vision Workshops (ICCV)*, pages 17-24.

Li, G., Xu, S., Liu, X., Li, L., and Wang, C. (2018). Jersey number recognition with semi-supervised spatial transformer network. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 1896-1903.

IFAB. International Football Association Board | IFAB. http://www.theifab.com/laws (2019-10-03 reference).

Redmon, J., and Farhadi, A. (2018). YOLOv3: an incremental improvement. In *arXiv:1804.02767*.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. (2011). Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems Workshops, (NIPS)*, pages 1-9.

Bromley, J., Guyon, I., LeCun, Y., Sickinger, E., and Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS)*, pages 737-744.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *Proceedings of the 32nd International Conference on Machine Learning Workshops (ICML)*, volume 37.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M, and Zisserman, A. (2018). VGGFace2: A dataset for recognising face across pose and age. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.

Kingma, D.P., and Ba, J. (2015). Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*.