Part-based Player Identification using Deep Convolutional Representation and Multi-scale Pooling

Arda Senocak[1] Tae-Hyun Oh[2] Junsik Kim[1] In So Kweon[1]
Dept. EE, KAIST, South Korea[1]
MIT CSAIL, MA, USA[2]

Abstract

This paper addresses the problem of automatic player identification in broadcast sports videos filmed with a single side-view medium distance camera. Player identification in this setting is a challenging task because visual cues such as faces and jersey numbers are not clearly visible. Thus, this task requires sophisticated approaches to capture distinctive features from players to distinguish them. To this end, we use Convolutional Neural Networks (CNN) features extracted at multiple scales and encode them with an advanced pooling, called Fisher vector. We leverage it for exploring representations that have sufficient discriminatory power and ability to magnify subtle differences. We also analyze the distinguishing parts of the players and present a part based pooling approach to use these distinctive feature points. The resulting player representation is able to identify players even in difficult scenes. It achieves state-of-the-art results up to 96% *on NBA basketball clips.*

1. Introduction

Broadcast sports video analysis has many commercial applications. Among them, automatic player detection, tracking and identification are critical for enhancing broadcast sports videos; coaches and trainers can collect game data and statistics to make tactical analysis, and broadcasting companies can enhance the experience of audiences or provide interactive contents. However, most of the tasks are still done manually or semi-automatically, thus fully automating these tasks are very important.

While detection and tracking algorithms have been advanced, some of the commercial player tracking systems [24], such as STATS SportsVu [22] and Tracab, still largely rely on manual or semi-supported player labeling. Additionally, many of the tracking algorithms cannot follow the players consistently because of the interruptions, where it requires re-initialization or correction of the labels. Thus, fully-automated *player identification at a given moment in time* is a yet another important and open challenge. In the regard of few attempts on the player identification problem, it is crucial to focus on analyzing player identification task.

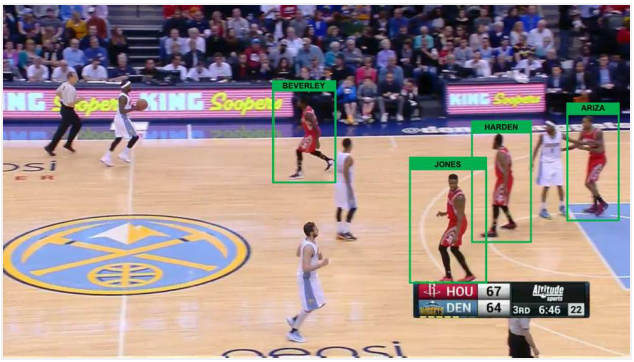

Figure 1. **Who's who on the court?** Automatic player identification results in broadcast basketball video.

Identifying players, especially in broadcast videos, is a challenging task. Previous player identification algorithms attempt to recognize players from close-up camera views. It is relatively easier problem because jersey numbers and faces of players are visible and players can be identified by using these visual cues. However in the case of videos captured by broadcast camera at any moment, identification process is getting harder by several challenges. The appearance of players look identical at a glance due to wearing the same uniform. Furthermore, distinctive features of players, such as jersey numbers, accessories and faces, typically have complex motion and deformation across time due to active actions and occupy a small fractional area in spatial visual domain. In addition, body poses and postures of players are continuously varying during the video. Therefore, it suggests to design a suitable player representation to improve visual identification so that system can capture high-level semantic discrepancy and distinguish such the subtle differences.

This work presents a scheme to automatically identify players from a broadcast video captured by a single broadcast camera, as shown in Figure 1. The proposed system

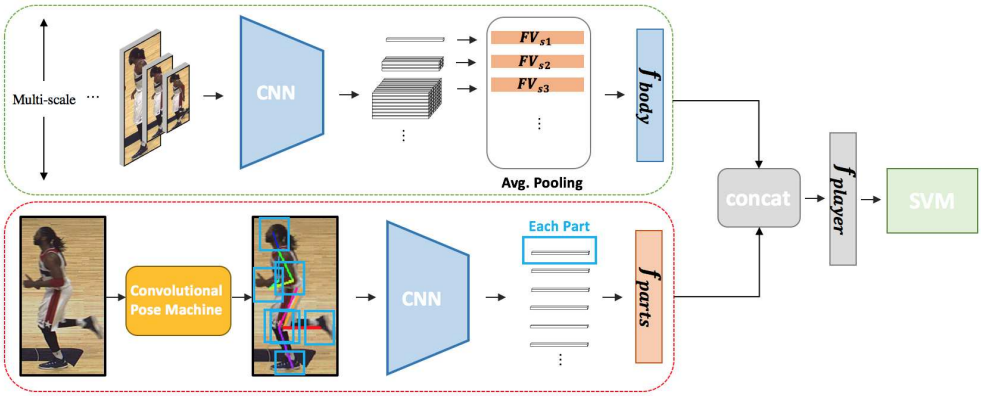

Figure 2. **Network Architecture.** This architecture is designed to tackle the player identification problem in the broadcast camera view setup. The green dotted box presents multi-scale deep convolutional representation pooling, where we model the body representation. The red dotted box shows the pipeline to model body part representation which is computed by locating body parts with CPM and pooling features from those parts. Final player representation is the concatenation of body and body parts representations.

takes an image of a player which is localized by detection algorithms in target video as an input, and produces identity information. Players are modeled by combining deep convolutional activations at multi-scale from the whole image of the player and also from the parts of the player, so that we can leverage both holistic representation and local salient representations. We effectively encode representations of the entire player body and the parts by Fisher vector separately, and these are fused into a single representation for a player as illustrated in Figure 2.

This work makes the following contributions:

- This work attempts to solve player identification problem by only using visual cues with deep convolutional representations.
- We analyze our approach by visualization where to look at to distinguish the players correctly. This suggests insights to model the player identification problem.

2. Related Work

The player identification can be viewed as a subset of person re-identification or fine-grained recognition problems. Since the player identification problem deals with players that wear same uniforms, the player identification problem only allows more limited information than the generic problems. In this work, we specify the scope of this work into the sport context, *i.e.*, player identification, and we focus on previous player identification techniques in this section.

Many of existing player identification approaches are concentrated on close-up camera views rather than broadcast camera views. In this close-up camera views, player identification can be done by detecting and recognizing faces and jersey number recognition. Face recognition is performed on the the detected face regions to label the players in [18, 2]. Jersey number recognition is also used widely for player identification as in [15, 28, 27, 23, 20, 6, 1]. They presented various effective ways to localize and segment the number part. A slightly different approach is presented by combining the textual cues with visual face information to have more robust and general player identification scheme [3, 4].

Somewhat against the trend, Lu *et al.* [17, 16] tracks and identifies basketball players by recognizing the entire player rather than face and jersey number recognition in a video captured by broadcast camera view. This is the first attempt to identify players by recognizing the entire body and its track. This work is the most relevant one to our proposed player identification system in flavor to recognizing players from the entire body in a broadcast camera view. They use low-level hand-crafted features such as MSER [19], SIFT [14] and color histograms at single scale to build a visual representation. In contrast, our method leverages deep convolutional feature at multi-scale to model the player representation. In addition, we exploit body parts to supplement coarse holistic body representation, which allows to to have a complete sports player representation.

While all the above mentioned approaches are prior to the recent advances of deep learning, Gerke *et al.* [9] presents jersey number recognition with a deep convolutional neural network. In this work, we use deep convolutional networks to automatically find the distinctive information from a entire body rather than using manually-designed targets such as jersey number or face information.

Some further works [8, 7] perform the player identification by using player's position. Gerke *et al.* [7] combines this location information as spatial constellation with jersey number recognition. Since trajectories are known, identification problem is formulated as an assignment problem. Lu *et al.* [16] also use detection and tracking results to build a CRF [12] model to label the players. In this work, we perform the identification without any location, tracking or

temporal information. In addition, player labeling is done in any frame at any moment in time.

3. Player Identification by Deep Representation

We design a system to address the player identification problem in the broadcast camera view setup. In order to deal with similar appearance of players, we use the entire body of the players and body parts as well with a shared network. The system composes of two main modules: body representation at multi-scale and body parts representation as illustrated in Figure 2. We describe the main components of each module in this sections.

3.1. Body Representation at Multi-scale

As we mentioned earlier, differences in basketball players are very small. Thus, it is important to capture subtle differences between players for identification task. Generally, using multi-scale features can contain more detailed information. This motivates us to extract multi-scale features for each input player image and combine these activations to have a better representation. Our coarsest scale is 227×227 and it is upscaled by a factor of 2. We use 7 scales to get activations.

We use an architecture, the Alex-Net [11] with modification. The fully connected layers of the network is changed with convolutional filters as in [13, 25, 29] to be able to extract multiple activations from the images that has a larger size than the original input size. The seventh layer of the network is used to extract activations from each scaled player image and every activation has dimension $D = 4096$. Each activation vector is ℓ_2-normalized and reduced in dimension from 4096-D to 128-D by the PCA which is learned from the activation vectors that are randomly sampled among training images. Also, a GMM with $K = 256$ components is learned from the same sampled activation vectors. One single Fisher vector is computed from the activations at each scale s. This architecture is similar to [5, 29] that combine pooling strategies with CNN activations at multi-scale. The output Fisher Vectors are followed by average pooling and final body representation is denoted as $\mathbf{f}_{body}$. Body representation is shown in (green-dotted box) in Figure 2.

3.2. What makes a player a player?

It would be interesting to analyze what makes a player different from the other players. For the goal of finding regions which distinguish a player from rest of the team, we weight each local activation and accumulate all the weights. As the nature of Fisher vectors, it gives fixed dimension vector regardless of the number of features.Thus, we compute a final Fisher Vector representation as in Figure 2 for every local activation vector (a local feature). Then, each representation is classified with pre-trained SVMs that are trained with using (green dotted box) in Figure 2. The weights of representation vectors that are identified as target players, are accumulated in spatial domain based on their corresponding locations. Figure 3 shows the characteristic parts of the same players. Distinctive details appear consistently on similar locations for each player. For example; what considered very distinctive for player `Harden` are his beard and sleeves on the knees. Similarly socks, shoes, sleeves, jersey numbers and head part look like distinctive parts for the players. It is interpretable that these distinctive parts are mostly the parts that are not covered by the jersey. Focusing on those areas can make the player representation more complete. Thus, we propose the part based representation as an addition to the body representation to have a final player representation.

3.3. Body Parts Representation

One can see from Figure 3 that the most prominent information lies usually in the highly specific semantic parts of the players that usually jersey doesn‘t cover such as head, leg, foot and *etc*. Thus, localizing these distinctive parts and extracting more information from these areas may give additional benefits that make the identification task more accurate. In this work, we use Convolutional Pose Machines [26] to detect player‘s body parts. It is a sequential architecture with convolutional networks. We note that there is no manually defined parts or part annotations for training our proposed approach. Adding this additional visual cue on the top of body representation makes a complete player representation.

These discriminative information are more important to distinguish players than general ones which are the common features shared across all players. To this end, we propose body parts representation as (red-dotted box) in Figure 2. Our architecture for part representation (1) takes an input image at single scale, (2) locates the body parts with Convolutional Pose Machines (CPM) [26], (3) extracts fixed-length activations from the every part by feeding them into CNN, (4) aggregates all the activations from every part into one single Fisher Vector as it is explained in Section 3.1. This final Fisher Vector is the body part representation and it is denoted as $\mathbf{f}_{parts}$.

Final player representation $\mathbf{f}_{player}$ is the fusion of body and part representations in a late fusion manner. The fusion is performed by concatenating the $\mathbf{f}_{body}$ and $\mathbf{f}_{parts}$.

4. Experimental Results

For training and evaluation, we first construct a new basketball player identification dataset. Based on this dataset, we evaluate the capability of the proposed method qualitatively and quantitatively against to various settings. As

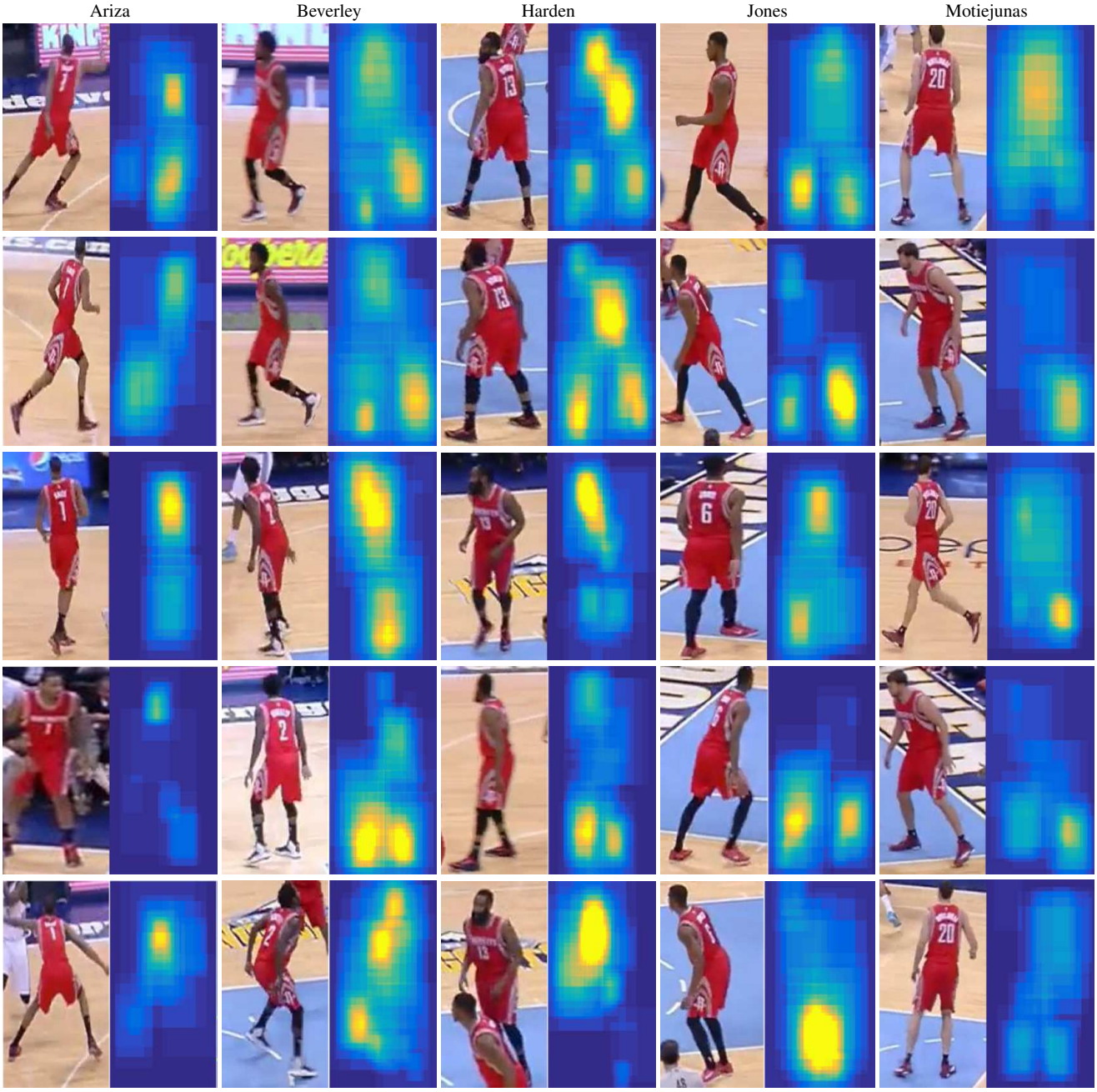

Figure 3. **Where to look at to distinguish the players from each other?** We visualize some of the distinctive parts of the players. We feed each activation at multi-scale as a Fisher Vector and classified with the learned category-specific linear SVMs. The locations of the activations that are identified as given players are uniformly weighted.

aforementioned, the proposed architecture targets the player identification task at *a time point t in time* without using any temporal consistency or assignment over reference information. The source code, dataset and all of the results will be made available to the public.

4.1. Dataset

There are no publicly available datasets of basketball videos with corresponding player detections, trajectories and identities. For training and testing the performance of our proposed player identification system, we collect a new dataset from the 2014 - 2015 NBA regular season Denver Nuggets versus Houston Rockets game. It is a commercial broadcast video. Thus, it consists of various different views. However, we only use the sequences that are captured from the broadcast camera view. Three human annotators are asked to draw bounding boxes around the players

and tag the players inside of the boxes in regularly sampled consecutive frames. For training, we select some part of the first half of the game. Training set consists of detected bounding boxes across 44649 frames. From the set of training, we randomly sample 2500 for each player to train our architecture. Test set includes clips varied in length from 500 to 4500 frames that are selected from the second half of the game as in [17, 16]. Each clip includes different level of difficulties for identification. Annotating both training and test sets took us significant amount of time.

4.2. Results and Analysis

Quantitative results. We use the manually prepared test set with the ground-truth bounding boxes to evaluate our proposed system. We ignore the temporal information between detections and the identification is done on a per-frame to see the players identifiability at a single time point. We also do not use any assignment problem or mutual exclusion. Although each team has 12 players, we use only 5 players from Houston Rockets to evaluate the player identification framework for simplicity. Our evaluation is conducted in terms of the average classification accuracy across all players but also we present average accuracy for team. It is preferable to get balanced individual accuracies for each player rather than varying accuracies among players. This will make the player identification system more robust so that players can be identified at random time point with the same success rate.

We conduct through experiments to compare various methods and settings for to measure the capability of the proposed identification architecture. Following methods are used for evaluation and comparison.

AlexNet Activation Vector at Single Scale. This framework computes the one single 4096-dimensional CNN activation vector over the entire image. This vector is obtained from the seventh layer, $FC7$, of the AlexNet [11] architecture. One-vs-all linear SVMs are learned on the training images using these activation vectors.

ResNet-50 Activation Vector at Single Scale. This method results one single 2048-dimensional activation vector over the entire image. This is based on the feature extracted from the last fully-connected layer before the softmax layer of the ResNet-50 [10].

VGG-16 Activation Vector at Single Scale. This setting takes the activation vector from the $FC7$ layer of the VGG-16 [21] architecture.

Bag of Words Model with DSIFT. As we earlier mentioned, Lu *et al.* [17, 16] is the most related work to our proposed approach. However, their dataset and source code is not publicly available. Thus, we implement their approach in most similar way as we can. Also note that, this approach is not using any deep learning features rather using low-level hand-crafted features. SIFT features are extracted over the entire image and their image location (x_i, y_i) is appended at the end of the feature vector as $[f_i, x_i/W - 0.5, y_i/H - 0.5]^T$ where W x H are the size of the image. We learn a k-means codebook with $k = 2048$ centers and aggregate all features with BoW encoding method.

Table 1. **Per-player identification performances.**

Method	Ariza	Beverley	Harden	Jones	Motiejunas	mAP
AlexNet-Single	62.0	85.4	64.2	85.5	92.0	77.8
ResNet-50 Single	68.2	88.4	77.6	84.0	92.6	82.2
VGG-16 Single	52.3	82.0	58.6	83	88.1	72.8
DSIFT+BoW	47.1	64.5	29.0	68.8	81.8	58.3
Ours-Body	91.5	95.6	90.4	95.0	98.0	94.1
Ours-Body+Parts	94.1	97.7	93.3	96.4	98.1	96.0

Table 2. **Identification performance of our proposed system with different numbers of scales.**

Scale	Ariza	Beverley	Harden	Jones	Motiejunas	mAP
3	76.1	93	83.9	93.5	96.8	88.7
5	88.8	97.2	94.9	97.7	98.8	95.5
7	94.1	97.7	93.3	96.4	98.1	96
9	96.1	97.8	92.6	94.6	99.3	96.1

Comparison results for the players using different schemes and settings are summarized in Table 1. As expected, BoW model performs the worst for all the players. Our proposed methods, entire body representation at multi-scale and body parts combined representation, give the highest accuracy. Interestingly, accuracy for `Motiejunas` is always higher than the other players regardless of the architecture. We believe that color information helps to distinguish him among others because he is the only Caucasian player among Houston Rockets players.

We report results with the state-of-the-art network architectures such as ResNet and VGG-16. This justifies that even though these representations are very powerful in many domains, however simply using them to represent and distinguish the players is not enough in player identification task. The results show that the proposed methods benefit more from the multi-scale feature extraction, body parts information and Fisher Vector encoding though AlexNet is used for activations.

While our combined player representation performs the best in variety of settings, however improvement is around 2% for the team. A possible explanation is that Fisher Vector model is already localizing the parts and embed these discriminative information into vector since GMM centers may localize important parts. `Harden`, `Beverley` and `Ariza` are the players who mostly benefit from this combined method while `Jones` and `Motiejunas` performs almost similar. As Figure 3 shows, `Harden` and `Beverley` have very discriminative parts but `Motiejunas` does not contain much of this information.

We measure the effect of the number of the multi-scales in our proposed method in Table 2. The results show that

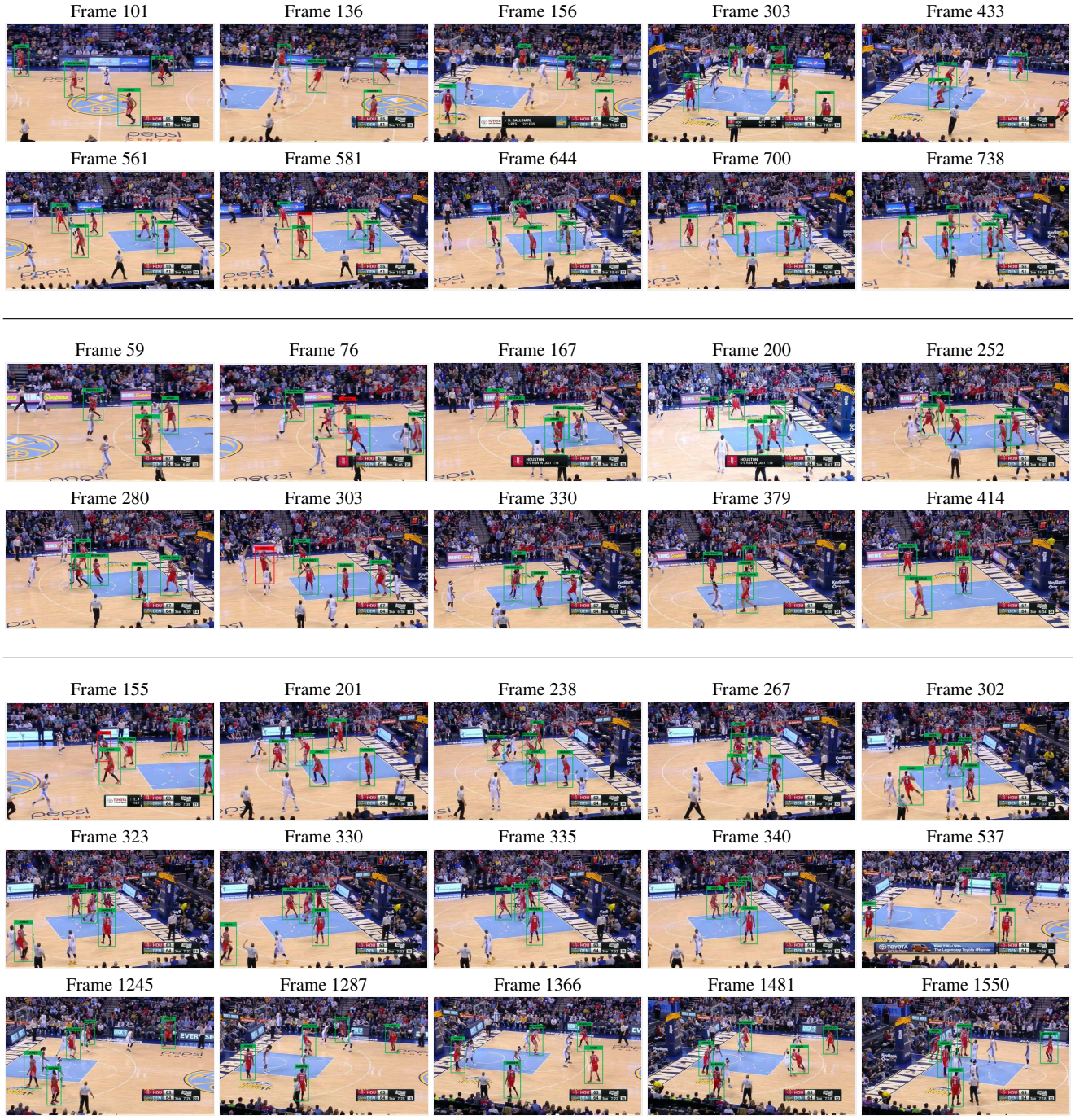

Figure 4. **Video sequence example.** We show results for identification of the players in three different sequences from our dataset. Text in boxes (players' name) are automatic identification results. Green bounding boxes show true identifications and red ones represent misidentification.

near 5 scales are sufficient for the identification performance since average accuracy of the proposed system is saturated and the computational time of the algorithm increases as well.

Qualitative results. We present some identification results from the test sequences of our dataset in Figure 4. Identification is done on the given detected boxes. Texts are written on the top of the boxes indicate the player name inferred by the proposed system. Green bounding boxes represent correct identifications and bounding boxes with the red color show misidentification. The results show that our proposed method is able to identify the players despite chal-

Figure 5. **Star-View camera example.** We show some example frames from our star-view camera. Our system allows to create virtual cameras for selected players. This virtual personal camera angles improves the video experience.

lenges such as similar appearances of players, occlusions and fast movement.

Additionally, our identification system allows to create star-view camera which is a virtual locked zoom camera for the interested player. Since we know who is who on the court, virtual camera angle can be created for that player. We show some results in Figure 5 from our star-view camera application. This application helps to improve the personalized and interactive sports video experience.

5. Discussion and Conclusion

We introduce a novel approach for identifying players in broadcast sports videos captured by single side-view camera. The system is designed to identify players in given detections. Player identification in sports videos is a challenging task because appearances of players are ambiguous and biggest part of the bodies is covered by jersey which is a generic feature shared among all the players.

To cope with these difficulties and have a complete player representation, we use Convolutional Neural Networks for feature extraction at the multi-scale levels and encode these local activations by Fisher Vector method to have body representation. We show that characteristic features of the players come from the body parts such as foot,legs,head and etc. Thus, we propose combined part based player representation to get more discriminative information from these parts.

Our proposed system has ability to identify players even in difficult scenes with high accuracy. Extensive experiments have shown that this high performance comes from using Convolutional Neural Networks with multi-scale local activations encoded by Fisher Vector and also part based extraction to capture characteristic parts of the players.

As aforementioned, we do not use any temporal cue or assignment problem. Applying assignment problem in this method would increase the overall accuracy. Additionally, using Recurrent Neural Networks (RNN) for consecutive frames of the players might improve the proposed method. With these improvements, we believe that proposed system would open potential directions for future research to combine player identification task reliably with different sports.

Acknowledgements: This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2017-0-01780). The part of this work was done when T.-H. Oh was in KAIST. We thank to Francis Zane and Kiswe Mobile for helpful discussions.

References

[1] E. L. Andrade, E. Khan, J. C. Woods, and M. Ghanbari. Player identification in interactive sport scenes using region space analysis prior information and number recognition. In *International Conference on Visual Information Engineering*, 2003. 2

[2] L. Ballan, M. Bertini, A. D. Bimbo, , and W. Nunziati. Soccer players identification based on visual local features. *International Conference on Image and Video Retrieval*, 2007. 2

[3] M. Bertini, A. D. Bimbo, and W. Nunziati. Matching faces with textual cues in soccer videos. *International Conference on Multimedia and Expo, ICME*, 2006. 2

[4] M. Bertini, A. Del Bimbo, and W. Nunziati. Player identification in soccer videos. In *International Workshop on Multimedia Information Retrieval*, MIR '05, pages 25–32, 2005. 2

[5] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[6] D. Delannay, N. Danhier, and C. D. Vleeschouwer. Detection and recognition of sports(wo)men from multiple views. In *ACM/IEEE International Conference on Distributed Smart Cameras*, 2009. 2

[7] S. Gerke, A. Linnemann, and K. Mller. Soccer player recognition using spatial constellation features and jersey number recognition. *Computer Vision and Image Understanding*, 2017. 2

[8] S. Gerke and K. Mller. Identifying soccer players using spatial constellation features. In *KDD Workshop on Large-Scale Sports Analytics*, 2015. 2

[9] S. Gerke, K. Mller, and R. Schfer. Soccer jersey number recognition using convolutional neural networks. In *IEEE International Conference on Computer Vision Workshops*, 2015. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012. 3, 5

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning*, 2001. 2

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 2

[15] C.-W. Lu, C.-Y. Lin, C.-Y. Hsu, M.-F. Weng, L.-W. Kang, and H.-Y. M. Liao. Identification and tracking of players in sport videos. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, 2013. 2

[16] W. Lu, J. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 2, 5

[17] W. Lu, J. Ting, K. P. Murphy, and J. J. Little. Identifying players in broadcast sports videos using conditional random fields. *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2, 5

[18] Z. Mahmood, T. Ali, S. Khattak, L. Hasan, and S. U. Khan. Automatic player detection and identification for sports entertainment applications. *Pattern Analysis and Applications*, 18(4):971–982, Nov 2015. 2

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *British Machine Vision Conference*, 2002. 2

[20] M. Saric, H. Dujmic, V. Papic, and N. Rozic. Player number localization and recognition in soccer video using hsv color space and internal contours. *ICSIP*, 2008. 2

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[22] STATS. Sportvu: Player tracking and predictive analytics, 2017. https://www.stats.com/publications/stats-sportvu-player-tracking-advanced-analytics/. 1

[23] S. W. Sun, W. H. Cheng, Y. L. Hung, C. L. I. Fan, J. Hung, C. K. Lin, and H. Liao. Who's who in a sports video? an individual level sports video indexing system. *International Conference on Multimedia and Expo, ICME*, 2012. 2

[24] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 2017. 1

[25] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Neural Information Processing Systems*, 2014. 3

[26] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[27] T. Yamamoto, H. Kataoka, M. Hayashi, Y. Aoki, K. Oshima, and M. Tanabiki. Multiple players tracking and identification using group detection and player number recognition in sports video. In *IEEE Industrial Electronics Conference*, 2013. 2

[28] Q. Ye, Q. Huang, S. Jiang, Y. Liu, and W. Gao. Jersey number detection in sports video for athlete identification. In *Proceedings of the SPIE Visual Communications and Image Processing*, 2005. 2

[29] D. Yoo, S. Park, J. Y. Lee, and I. S. Kweon. Multi-scale pyramid pooling for deep convolutional representation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015. 3