Recent Advances in Sports Computer Vision

Andrew Aikawa
asai@berkeley.org

Danqing Wang
danqingwang@berkeley.edu

Xiaoer Hu
huxiaoer@berkeley.edu

Abstract

In this work, we explore recent advances in computer vision within the context of analyzing sports broadcast video. We examine two problems within the context of different sports: player position estimation in soccer and play classification in baseball. For the first part of this work, we demonstrate 3D reconstruction of soccer scenes from soccer videos by combining human pose estimation with sports field localization and camera calibration. In the second part, we evaluate the performance of different video analysis architectures within the context of multi-classification of broadcast footage of baseball pitching.

1. Introduction

We explore the application of video analysis within the context of broadcast sports video for end to end 3d reconstruction of soccer scenes and play classification of baseball video.

1.1. 3D Soccer Recovery

To perform 3d reconstruction of soccer videos, we leverage existing tools for sports field estimation, instance segmentation, pose estimation and human mesh recovery. Our method uses a frame by frame analysis to generate bounding boxes for each player per frame using MaskRCNN. Cropped frames of players are used to first estimate 2D pose keypoints followed by a 3D human mesh recovery that also estimates the the weak perspective camera parameters. This is combined with another frame by frame homography estimation between a field template and the broadcast footage. We leverage this in conjunction with player pose/weak camera estimates to find an estimate of the players position on the field. We use these to create a synthetic scene which encodes in low fidelity the player positions in 3d, allowing us to create synthetic multiviews of the same game.

1.2. Baseball Video Analysis

We investigate the supervised learning task of multi-classification of Major League Baseball (MLB) clips of pitching. This particular dataset is challenging due to the discriminating portion of the video only being a very small fraction of the entire video and being very visually similar. Recent advances in video analysis include SlowFast networks, which are the focus of this work. We compare the performance of SlowFast against Inceptionv3 and I3D baselines and show that SlowFast remains competitive even with sparse view sampling.

2. Related Work

Player detection: Numerous deep learning models for object detection have been explored in the field. The popular architectures include region proposal based detectors, such as R-CNN [13], Fast R-CNN, [12] Faster R-CNN [38] and detectors that directly predict boxes for an image in one step such as YOLO [37] and SSD [29]. By applying a pre-trained model on a general dataset, one can extract out the human-class detections. Instead of training on general dataset, training on ground-truth player positions would further improve the accuracy for reasons such as motion blur [20]. While the manual annotation is costly, people have developed self-supervised techniques to improve the accuracy without the labeled ground truth data. These approaches involve synthetically generating training images or using distillation framework for transfer learning [20, 19, 21, 6, 39].

Pose estimation: The main part of human pose estimation is to model the human body. There are several common models to achieve pose estimation, such as skeleton-based model, volume-based model, and contour-based model [30, 14, 5]. Skeleton-based model [4, 34, 1] uses a set of joins like shoulders, knees, ankles, elbows, and limb orientations comprising the skeletal structures of human body. Volume-based model [42, 2, 31, 23] uses 3D human body shapes represented by volume based models with geometric meshes and shapes that were captured by 3D scans. Contour-based model [24, 7] consists of the contour and tough width of body part which are presented with boundaries and rectangles of people's silhouette.

Homography Estimation: Homography estimation in sports field have been extensively studies. One of the most common techniques is to track on manually annotated interest points [35, 41]. Improvement on this method has been

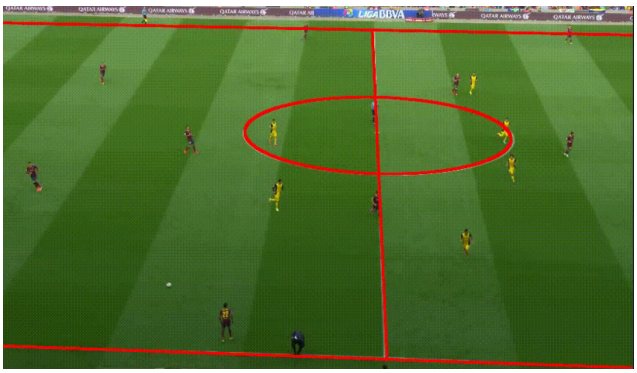

Figure 1. Field line layout after homography estimation. Homgraphy estimation takes in both a video frame and field line template as inputs and outputs the homography matrix that rectifies the video fram input to the template.

Figure 2. Player bounding boxes. Player bounding boxes are estimated using MaskRCNN

Figure 3. 2D pose annotations and recovered 3D mesh estimate. Player crops are obtained using bounding boxes from MaskR-CNN. 2D keypoints are generated from OpenPose. Left mesh corresponds to the original view perspective whereas the right mesh is a new synthetic view of the same mesh.

shown by using SIFT features augmented with line and ellipse information [15, 28, 33]. To eliminate manual initialization of correspondences, automated method based on SIFT correspondences has been proposed [17]. Although the latter improves the matching procedure, it might not apply to real games due to lack of visual features. Further improvements using deep network pre-trained on real life soccer videos have also been made [40].

Action recognition Action recognition can be used to identify the label of action and activity observed in a video clip. Video recognition presents new challenges relative to image classification from the additional temporal plane as well as the inclusion of audio information in some instances. There are many popular 3D convolution architectures for action recognition, including ResNeXt, Two-Stream Inflated 3D ConvNet (I3D), SlowFast, R(2+1)D, etc [25, 43]. In 3D CNN architecture, filters are designed in 3D, and the channels and temporal information are used as different dimensions. SlowFast architecture has 2 streams, called slow and fast paths. Slow stream operates at low frames and focuses on spatial information, while the fast stream operates at high frames and focuses on temporal information. There is information flow from the fast stream to the slow stream [10, 11]. However, 3D CNNs require huge computational costs and memory comparing with their 2D counterparts.

3. Soccer Game Reconstruction

In this section, we describe the components for performing 3D soccer game recovery.

3.1. Homography Estimation

We estimate the homography between a soccer field template and individual frames from soccer video as in Figure 1. Usual automated homography estimation of planar surfaces relies on keypoint detection and matching. However, this makes homography estimation heavily reliant on the visibility of keypoints (i.e. field line intersections/corners), which can result in failure from occlusion or being out of frame. We instead use a deep network pretrained on a World Cup 2014 Dataset for estimating homography of soccer video frames [18]. This is achieved by using two networks, one for making an initial estimate of the homography and another for learning the errors of the initial network, each trained independently of each other [22]. Adding the residuals from the second network further refines the homography estimate.

3.2. Camera Calibration

To estimate camera parameters, we choose the camera parameters that minimize the projection error of 3D points on the soccer field onto the 2D video frame. We have camera matrix using the usual intrinsic, extrinsic decomposi-

Figure 4. Input video frame and synthetic render from a new perspective of the same scene. Player meshes were shaded using texture samples from the original video frame

tion.

$$P = K[R|t]$$

By defining the 3D world space to be such that the $y = 0$ plane lies on the soccer field surface, we use the homography estimation from the previous step to create four 3D-to-2D keypoint pairs from which we can estimate the camera extrinsic parameters, R, t, given the camera intrinsics, K. To get the camera intrinsics, we perform a grid search over candidate focal lengths and pick the one that generates the smallest projection error of the 3D point, X_i, to the control point in the image, x_i.

$$\text{projection error} = \sum_i |x_i - PX_i|$$

3.3. Player Detection and Pose

To find all the players in a given video frame, we first create bounding boxes for every player instance. We accomplish this by using a pretrained MaskRCNN [16] for human detection as in Figure 2. With the bounding boxes, we produce crops of individual players which we use inputs into OpenPose to generate 2D keypoints [3].

2D keypoints can be used with recent work developed by Kanazawa *et al*. [26] to produce both 3D pose and shape of players [32] and weak camera parameters as in Figure 3. Human mesh recovery extracts the parameters for pose, θ, and shape, β as well as the camera scale, s, global rotation, R, and translation t. The projection of the 3D keypoints, $X(\theta, \beta)$ is given by

$$\hat{x} = s\Pi(RX(\theta, \beta)) + t$$

where Π is an orthographic projection. Note that the scale factor, s, is related to the average coordinate in Z by

$$s = \frac{f}{Z_{avg}}$$

Since the player crops were produced from the same video camera footage, weak camera scale combined with the fixed focal length estimate from camera calibration can be used to find the distance from the camera to the player, which we use to generate a 3D coordinate for the player's center of mass.

3.4. Scene Reconstruction

The previous step enables us to place the player meshes in 3D space by shifting the mesh such that the mean of the mesh points corresponds to the player center of mass. To generate synthetic multiple views, we can just shift our camera around the scene. To shade meshes, we texture sample by projecting the mesh face back onto the original video frame as in Figure 4.

4. Predicting Baseball Pitches

Here we describe how we performed video analysis of Baseball Videos

4.1. Dataset

The dataset we use is a labeled dataset of activity segmented videos of 20 baseball games from the 2017 MLB post-season available on YouTube [36]. Video segments are clipped around baseball pitches and somewhat shortly after. However, many clips, being part of broadcast footage contain no activity segments (e.g., panning over the spectators). As such, unlike other video datasets (i.e. Kinetics, Charades), it is difficult to correctly determine the activity from a single sampled video frame. Moreover, differences between videos even mutually exclusive labels are visually similar as in Figure 5.

The entire dataset contains 2828 videos for training and 962 for test. Clips are splits such that all clips from any particular game are only in either the test set or training set but never both. While the original paper describing this data set reported a larger number of total video segments,

Figure 5. Representative frames of a video clip labeled ball and strike respectively. Even mutually exclusive labels are visually very similar requiring either detection of umpire signals or good ball detection for effective prediction

at the time of writing this reports, a number of segments were not recoverable using Youtube-DL which was the suggested method from the original authors for downloading the dataset, explaining the slightly smaller dataset size.

Labels describe the outcomes of pitches (e.g. ball, strike, foul, hit, etc.). While certain labels are mutually exclusive, we frame our prediction task as an eight-way multi-label classification.

4.2. Implementation Details

Data was preprocessed offline to be downsampled to 256x720. The training set was 80/20 split for training and validation.

For our network architecture, we chose a SlowFast 16x8 [8] with a temporal ratio of $\alpha = 4$ and a spatial ratio of $\beta = 1/8$ and a ResNet50 backbone and global average pool before inference with a fully connected layer [9]. For training, data was augmented by random scaling such that the short side of the video could be 340 pixels long, which was followed by a 225x225 pixel crop. For regularization, the classification head was trained with a dropout rate of 50%. Due to hardware limitations, we trained with a batch size of 1 and on 1 GPU. We began with Kinetics-400 pretrained weights [27]. We train using SGD for 57 epochs with a base learning rate of 0.0375 and momentum 0.9 and 10x step-wise decay for validation loss plateaus.

Method	mAP
Random	16.3
Inceptionv3 + mean pool	35.6
I3D + mean pool	42.4
SlowFast 16x8, R50 + mean pool	**45.5**
Inceptionv3 + max pool	47.9
I3D + max pool	48.3

Table 1. Mean Average Precision for our SlowFast model against Inceptionv3 and I3D baselines

At test time, we uniformly sample 10 clips along the time axis of the video and instead use 3 256x256 spatial crops (instead of the 224x224 for training) resulting in 30 total views for inference.

4.3. Results

We evaluate our model using mean average precision (mAP) for each segment, following the usual practice for multi-label classification. Table 2 summarizes previously reported baselines for other one-stream temporally pooled architectures in comparison to our method [36]. We find that our method outperforms all other mean pooled baselines and is comparable to max pooled baselines. Our SlowFast implementation sparsely samples 64 frames total which was a fraction of most videos whereas the baselines described would use every video frame as part of the input. This suggests that SlowFast successfully captures temporal semantics even at much lower cost and with fewer views.

5. Conclusion

In summary, we have shown that recent methods have made it possible to perform end-to-end 3D scene recovery and better encode video semantics from broadcast sports footage. For 3D scene recovery, the methods presented here could in principle be extended to other sports, only requiring that the networks for homography estimation be trained for other sports fields. This necessarily requires a dataset with labeled homography estimates for other sports to perform supervised training as was the case for this implementation which at the time of the writing of this paper is not publicly available. Performance with different temporal pooling schemes prior to classification should be explored as older architectures have found success with using temporal pyramids.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference*

on Computer Vision and Pattern Recognition, pages 3686–3693, 2014. 1

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005. 1

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016. 1

[5] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, Mar 2020. 1

[6] Anthony Cioppa, Adrien Deliege, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. Arthus: Adaptive real-time human segmentation in sports through online distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1

[7] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 1

[8] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 4

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. 4

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573, 2016. 2

[12] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 1

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1

[14] Wenjuan Gong, Xuena Zhang, Jordi Gonzàlez, Andrews Sobral, Thierry Bouwmans, Changhe Tu, and El-hadi Zahzah. Human pose estimation from monocular images: A comprehensive survey. *Sensors*, 16(12), 2016. 1

[15] Ankur Gupta, James J. Little, and Robert J. Woodham. Using line and ellipse features for rectification of broadcast hockey video. In *2011 Canadian Conference on Computer and Robot Vision*, pages 32–39, 2011. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 3

[17] Robin Hess and Alan Fern. Improved video registration using non-distinctive local image features. pages 1–8, 07 2007. 2

[18] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4012–4020, 2017. 2

[19] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection, 2019. 1

[20] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. *CoRR*, abs/2011.10336, 2020. 1

[21] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation, 2018. 1

[22] Wei Jiang, Juan Camilo Gamboa Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan, and Kwang Moo Yi. Optimizing through learned errors for accurate sports field registration. *CoRR*, abs/1909.08034, 2019. 2

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies, 2018. 1

[24] S.X. Ju, M.J. Black, and Y. Yacoob. Cardboard people: a parameterized model of articulated image motion. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 38–44, 1996. 1

[25] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. Late temporal modeling in 3d CNN architectures with BERT for action recognition. *CoRR*, abs/2008.01232, 2020. 2

[26] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 3

[27] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4

[28] Ruonan Li and Rama Chellapa. Group motion segmentation using a spatio-temporal driving force model. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2038 – 2045, 2010/06// 2010. 2

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 1

[30] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: The body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015. 1

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), Oct. 2015. 1

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 3

[33] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 2

[34] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, 2017. 1

[35] Kenji Okuma, J.J. Little, and David Lowe. Automatic rectification of long image sequences. 01 2004. 1

[36] A. J. Piergiovanni and Michael S. Ryoo. Fine-grained activity recognition in baseball videos. *CoRR*, abs/1804.03247, 2018. 3, 4

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 1

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[39] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training, 2019. 1

[40] Rahul Anand Sharma, Bharath Bhat, Vineet Gandhi, and C. V. Jawahar. Automated top view registration of broadcast football videos. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 305–313, 2018. 2

[41] Jianbo Shi and Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994. 1

[42] H. Sidenbladh, F. De la Torre, and M.J. Black. A framework for modeling the appearance of 3d articulated figures. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 368–375, 2000. 1

[43] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition, 2020. 2