Automated top-view annotation of broadcast video from the American Ultimate Disc League (AUDL) CVPR Proceedings

Quang Hoàng
University of Rochester
qhoang@ur.rochester.edu

Nate August
University of Rochester
n.baker.august@gmail.com

Abstract

In this work we consider two approaches for automating top-view annotation of broadcast video from the American Ultimate Disc League (AUDL). The first approach (CONV) utilizes YOLOv5 for object tracking and a manually-constructed reference image for field registration and transformation of new frames. The second approach (E2E), formulates the task as a sequence-to-sequence problem where the coordinates on the top-view plane are predicted directly from video frames using a variety of deep encoding-decoding schemes. Both methods fail to attain an acceptable level of accuracy with CONV producing zero acceptable matches across dozens of frames and several field layouts. E2E similarly struggled. The VGG architecture achieved the best performance with a mean absolute error of 40.12 meters. We present several hypotheses and potential extensions to fuel further study of this problem.

github.com/BakerAugust/audl-cv

1. Introduction

The American Ultimate Disc Association (AUDL) is a developing semi-professional Ultimate Frisbee (Ultimate) league with teams across the United States and Canada [13]. Recent partnerships with DraftKings [15], a sports-betting platform and LSports [16], a provider of real-time sports analysis indicate that the AUDL is moving towards other highly-digitized and analyzed sports such as soccer, football and basketball. To help facilitate the analytical development, The AUDL began requiring each team to annotate the footage of every game with the top-view location of disc on the field throughout the point. The objective of this work is to automate the annotation of disc location by applying optical tracking methods to game footage. We propose two approaches: 1) a conventional approach using YOLO object detection and field registration through reference matching and 2) and end-to-end deep learning approach.

2. Related Work

Optical tracking in Soccer, Football and many other sports [1,11,12] has been widely studied. In "Ball Tracking in Sports, A Survey" [17], the others outline the task of ball tracking using traditional methods as seven sub-tasks: 1) camera placement, 2) field modeling and registration, 3) camera modeling and calibration, 4) ground truth, 5) foreground extraction, 6) ball detection, and 7) ball tracking. In terms of concrete methodology, a corresponding pipeline of commonly adopted techniques in tracking: Hough transform [1,2] and Canny edge detector [3], Tsai camera model [4,5], and particle filter [6,7] can be promising when pursuing this line of approach. On the other hand, utilizing deep learning models in computer vision [8,9] , we can solve this problem in an end-to-end fashion following [8] (bounding box approach) or [DETECRON2] and [9] (image segmentation).

The specific subset of work on optical tracking from broadcast video has focused on field registration as the core challenge when working with single-camera footage [19,20,21]. Hess and Fern [19] explored this problem in 2007 using conventional computer vision approaches. Their approach requires the construction of a top-view reference model by manually registering a small subset of frames from the broadcast video. The resulting reference serves as a surface model of the field against which new frames can be matched and registered. Invariant features are detected in both the reference image and new frames using the scale-invariant feature transformer (SIFT), resulting in 128-dimensional descriptors for each feature. The authors propose two methods for feature matching: 1) globally distinctive feature-matching and 2) locally-distinctive feature matching. Globally-distinctive matching is performed using the 2-nearest neighbor heuristic where a feature X is matched from reference feature set Π if the ratio of the euclidean distance between the descriptor $d(X)$ and it's nearest neighbor $d(\pi_1(X))$ and the euclidean distance between $d(X)$ and the second nearest neighbor $d(\pi_2(X))$ is less than a constant $\rho = 0.6$,

$$\frac{||d(X) - d(\pi_1(X))||}{||d(X) - d(\pi_2(X))||} < \rho \quad (1)$$

Selection of ρ allows some control over false positive matches.

Locally-distinctive matching extends the idea of globally-distinctive matching based on the observation that constraining the model feature space Π to a specific spatial region Π^R may improve the distinctiveness of features. Given the limited movement between frames in a broadcast video setting, a plausible spatial region R can be identified based on a strong registration the preceding or following frames. Thus, locally-distinctive matching can enhance both the tracking of features between frames and identifying new features that can be continually tracked as the video sequence progresses. Registration transforms are then fit using RANSAC [23]. Using this combination of local and global matching, Hess and Fern were able to achieve a minimum of 100 video-to-model correspondences across hundreds of frames of broadcast american football.

Sharma et. al [20] propose an approach for improved automation of top-view registration by generating a synthetic dictionary of edge maps and homography pairs that new frames can be matched against. The synthetic dictionaries are generated by registering a small number of images to a top-view, but unlike Hess and Fern, the registrations are used to find an inverse homography matrix $\mathbf{H}^{-1}$ that transforms a generic top-view field model into video coordinates. The image registrations are perturbed to simulate pan, tilt and zoom to increase the breadth of the resulting dictionary. Since the top-view field model is generic, the features apply to all regulation-lined fields of a given sport and this process does not need to be repeated to apply to new fields as in Hess and Fern. The problem of registering a new frame then becomes reduced to finding the nearest entry in the dictionary. The authors find the greatest success by preprocessing the data using a stroke width transform (SWT) and then matching the nearest neighbor based on histogram of gradient (HOG) features.

Sha et al., [21] extend the dictionary-based approach further by utilizing a pipeline of neural network architectures to accomplish the tasks of 1) semantic field segmentation, 2) matching segmentation output against dictionary entries and 3) fine-tuned localization of the homographies. Our work embraces both conventional and deep learning approaches in our exploration of the broadcast video problem.

3. Methods

3.1. Data

Play-by-play data annotations for 134 games from the 2021 season are publicly-available through audlstats.com, a site maintained by the AUDL. A python package AUDL-Advanced-Stats has been developed to download and interact with the data [14]. The datasets contain the origin and destination for each pass as well as the thrower and receiver. Fig. 1 provides an example of the top-view disc location from a single possession. Accompanying broadcast-style film for at twelve games is publicly-available on YouTube. Fig. 2 shows some of the variety of camera height, lighting and field-linings present in the data set.

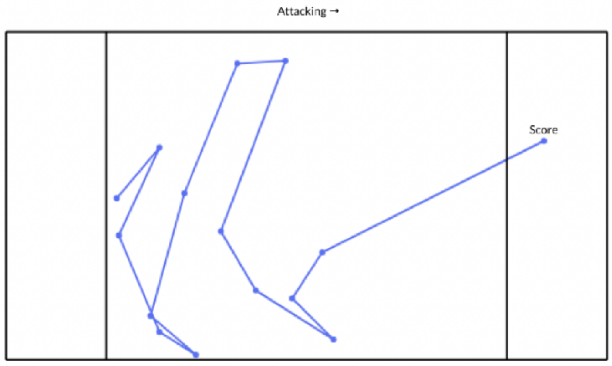


Figure 1. An example of a subfigure.

3.2. Approach 1 (Conventional)

3.2.1 Disc Detection

Disc detection was performed using a pre-trained instance of the You Only Look Once (YOLO) convolutional neural network architecture [8]. This architecture was chosen because of the inference speed and reduced background errors relative to other approaches like Fast R-CNN [23]. Specifically, an instance of YOLOv5s [24], a YOLO architecture with 7.2 M parameters pretrained on the Microsoft Common Objects in Context (COCO) dataset [25] was used.

3.2.2 Field Registration

The field template matching procedure presented by Hess Fern [18] was selected for field registration. While more manual and less accurate than more recent dictionary-based methods [20, 21], the variability of field appearances in the AUDL made these methods appear unfeasible; the field dimensions in the AUDL do not change, however, the games are rarely played in dedicated facilities and fields are often lined with markings for other sports like soccer and football. See Fig. 2 for examples.

To generate field reference images, a subset of clips were reviewed from each game and 10-20 selected frames were manually registered against the top-view model by identifying 4 point correspondences between the image frame and top-view field model and finding the resulting homogra-


(a) BOS-NY 8-20-21


(b) DAL-SD 8-28-21


(c) CHI-MAD 7-30-21


(d) RAL-DC 9-03-21

Figure 2. Examples of the diversity of camera angles and field environments of the AUDL broadcast dataset.

phies. The field reference images were then created as compilations of the registered frames transformed using their respective homographies.

Registering new frames against the reference images is then formulated as a task of matching SIFT features. Globally-distinctive features, as described in RELATED WORK was used in matching with the 2NN heuristic and a variety of values for ρ .

3.3. Approach 2 (End-to-end Deep Learning)

For the deep learning-based approach, we formulate this task as a sequence-to-sequence problem where we use a sequence of images to predict a sequence of coordinates from the overhead view. There are a few reasons why this formulation is desirable, first of all is not having multiple label sets. Sha et al. [20] design a multi-stage pipeline that, while can also be trained in an end-to-end fashion, would require a set of labeled data for each component and make the data requisition process expensive. Meanwhile, our formulation only requires labels in the form of the final 2D coordinates. Furthermore, assuming equal performance, a simpler model would limit error propagation and generalize to different domains better.

We split the dataset by possessions, with 5 possessions in the test set and 26 in the train set. Images are resized to 224 by 224, scaled to [0, 1], and normalized by the pretrained constants: mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225].

4. Experiments

4.1. Annotations

Three game videos were selected from the twelve available for their diversity in field appearance and camera angles and annotated to enable experimentation. First, the start and end time for each possession during the first quarter of play was noted and the film was clipped to produce continuous clips of game footage, excluding highlights from different camera angles, views of the crowd and close-ups on players. Since the AUDL-provided weak annotations were insufficiently linked to actual video frames, frame-by-frame annotations were added. This a 10x reduced frame rate on the clips and then linearly interpolated to apply labels to every frame in the clip.

4.2. Approach 1 (Conventional)

Initial experimentation with YOLOv5 showed promising results. Without any additional training the ”sports ball” would occasionally appear associated with the location of the disc. Given the success of additional training in other domains [8], the authors believe this problem to satisfactorily solvable given additional training and opted to focus more effort on the challenges of field registration.

New-frame registration was attempted using globally-distinctive features on dozens of frames and manually reviewed for three different values of ρ: 0.5, 0.6, 0.8. Across these parameterizations and all frames reviewed, not a single true match was identified. The failure of globally-distinctive features precluded the subsequent use of locally-distinctive features to improve consistency between frames because of the initialization requirement of globally-distinctive features.

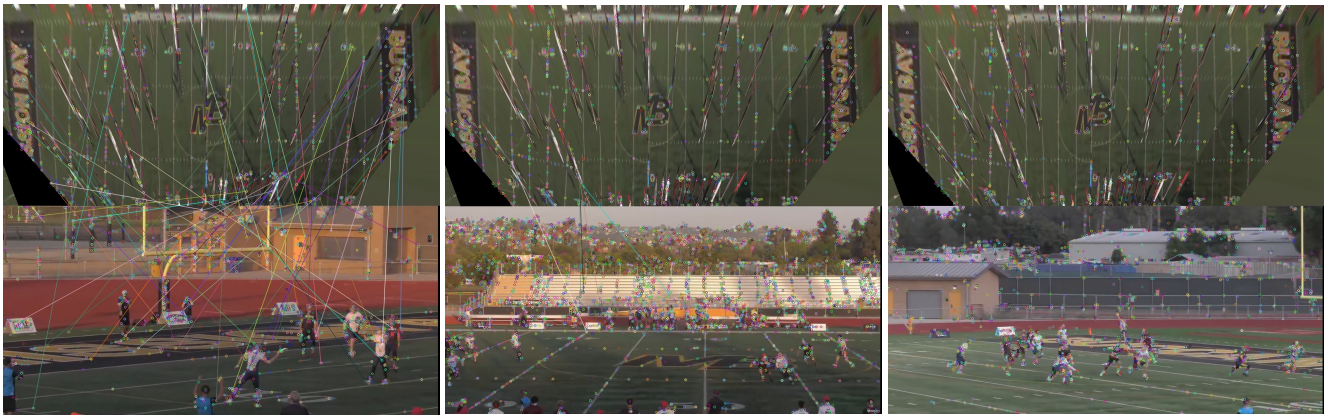
4.3. Approach 2 (End-to-end Deep Learning)

We ran various experiments to select the best pretrained model. We hypothesize that this is crucial for performance as the pretrained model would need to produce high-quality embeddings, i.e. vector representation that would need to contain information not just about the disc but also various features of the playing field, for the decoder component. The results of out trials is in the following table

Model	Epochs	MAE
SqueezeNet	5	42.63
VGG	4	40.12
DenseNet	4	41.28
ResNet	5	45.77
ResNeXt	6	41.79
MNASTNet	4	42.83

A model checkpoint was saved at lowest validation loss at the specified number of epochs of training. MAE can be directly converted to physical distance in terms of yards.

It is immediately clear that the model fails to reach anywhere near acceptable loss across different encoders. In addition, upon further close inspection of model outputs, we



(a) $\rho = 0.8$ (b) $\rho = 0.6$ (c) $\rho = 0.5$

Figure 3. Proposed matches using SIFT features and 2NN heuristic.

see that the model have converged to an average at of coordinates and constantly give that same prediction - one sign that the model cannot learn meaningful relationship from the data.

As we believe that the initial formulation makes sense, we theorize some possible reasons how the model failed:

1. Lack of data: With the task and domain drastically different from the pretraining scheme of the encoders we have used, we can expect to need a significant amount of data to perform transfer learning effectively. This also, to an extent, brings into question whether the interpolation performed for data annotation might have negatively impacted the training as the frisbee's flight paths are not always straight; this combined with the approximative nature of the original annotations might render the dataset extremely noisy for the model. While more fine-grained and precise annotation is one way to improve the data, we can also re-formulate the problem to be grid-based instead of coordinate-based, i.e. dividing the overhead playing field into a grid and predicting the square the frisbee is in. We propose this as a better approach as it alleviates labeling difficulty of having to pinpoint a coordinate while having minimal loss of practical usefulness.

2. Poor encoder performance: As mentioned above, we are highly dependent on the encoder to learn an effective representation of images to provide accurate prediction. However, through preliminary experiments with some object detection models (i.e. we use the pretrained model with no fine-tuning to qualitatively assess if it can pick up the frisbee consistently), we have found that even in ideal situations, e.g. the playing field takes up a majority of the frame, the frisbee is not occluded and not in motion nor in the air, pretrained models can only typically detect it half the time, with this rate dropping even more when in complex visual circumstances.

5. Conclusion

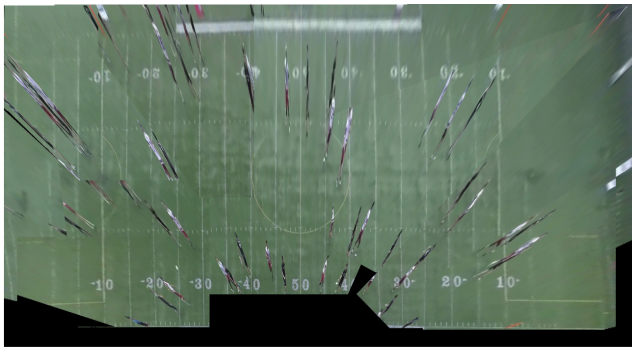
In this work we consider two approaches for automating top-view annotation of broadcast video from the AUDL.

The "conventional" approach utilizing YOLO showed some initial promise with object detection, but failed with field registration. We propose two hypotheses for why the feature matching under-performed relative to the results observed in Hess and Fern: 1) the extent of the pan and zoom in their American Football footage far exceeded that of AUDL, allowing the creation of a superior field reference image and 2) their field had improved lighting and more in-field logos leading to a richer field landscape. Both of these theories find support in Fig 4. where we can see greater stretching of AUDL players, the inability to register certain areas on the AUDL field and fewer logos. Furthermore, the locally-distinctive feature matching proposed by Hess and Fern shows much promise, but requires initialization with some global matches which could not be identified in the AUDL data.

The end-to-end deep learning similarly struggled. VGG achieved the best performance with a mean absolute error of 40.12 meters. We posit that improved performance could be achieved by including some additional pretraining on the disc-detection task alone and providing a larger dataset to learn useful sequence-to-sequence embeddings.

References

[1] Chen H-T, Tsai W-J, Lee S-Y, Yu J-Y (2012) Ball tracking and 3D trajectory approximation with applications to tactics analysis from single-camera volleyball sequences. Multimed Tools Appl 60:641–667



(a) BOS-NY 2021/08/20

(b) From Hess Fern [2]

Figure 4. Comparing AUDL field reference and that of an American Football field from Hess Fern [19]

[2] Tong X, Liu J, Wang T, Zhang Y (2011) Automatic player labeling, tracking and field registration and trajectory mapping in broadcast soccer video. ACM Trans Intell Syst Technol 2:15

[3] Ren J, Orwell J, Jones GA, Xu M (2009) Tracking the soccer ball using multiple fixed cameras. Comput Vis Image Underst 113:633–642

[4] Kumar A, Chavan PS, Sharatchandra V, David S, Kelly P, O'Connor NE (2011) 3D estimation and visualization of motion in a multicamera network for sports. In: 2011 Irish machine vision and image processing conference (IMVIP). IEEE, pp 15–19

[5] Li Y, Dore A, Orwell J (2005) Evaluating the performance of systems for tracking football players and ball. In: IEEE conference on advanced video and signal based surveillance. IEEE, pp 632–637

[6] Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process 50:174–188

[7] Choi K, Seo Y (2005) Tracking soccer ball in TV broadcast video. Springer, pp 661–668

[8] Bochkovskiy, Alexey et al. "YOLOv4: Optimal Speed and Accuracy of Object Detection." ArXiv abs/2004.10934 (2020)

[9] Yu-Chuan Huang, "TrackNet: Tennis Ball Tracking from Broadcast Video by Deep Learning Networks,"

[10] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[11] B. Chakraborty and S. Meher, "A trajectory-based ball detection and tracking system with applications to shot-type identification in volleyball videos," 2012 International Conference on Signal Processing and Communications (SPCOM), 2012, pp. 1-5, doi: 10.1109/SPCOM.2012.6290005.

[12] W. Chen and Y. J. Zhang, "Tracking Ball and Players with Applications to Highlight Ranking of Broadcasting Table Tennis Video," The Proceedings of the Multiconference on "Computational Engineering in Systems Applications", 2006, pp. 1896-1903, doi: 10.1109/CESA.2006.4281948.

[13] American Ultimate Disc Association, "About the AUDL," Available: https://theaudl.com/. [Accessed: Oct, 26, 2021]

[14] J. Lithio., "AUDL-Advanced-Stats" Available: https://github.com/JohnLithio/AUDL-Advanced-Stats [Accessed: Oct, 26, 2021]

[15] C. Eisenhood, "DraftKings Now Taking AUDL Bets in Some States," Ultiworld.com, July 22, 2021. [Accessed: Oct, 26, 2021]

[16] American Ultimate Disc League, "American Ultimate Disc League and LSports Data Enter Into $3 Million Data Distribution and Co-Development Strategic Partnership," CISION PR Newswire, April 22, 2021. [Accessed: Oct, 26, 2021]

[17] Kamble, P.R., Keskar, A.G. Bhurchandi, K.M. Ball tracking in sports: a survey. Artif Intell Rev 52, 1655–1705 (2019). https://doi.org/10.1007/s10462-017-9582-2

[18] Li Y, Dore A, Orwell J (2005) Evaluating the performance of systems for tracking football players and ball

[19] R. Hess and A. Fern, Improved Video Registration using Non-Distinctive Local Image Features CVPR 2007

[20] R.A. Sharma, B. Bhat, V. Gandhi, C.V. Jawahar, Automated Top View Registration of Broadcast Football Videos, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)

[21] L. Sha, J. Hobbs, P. Felsen, X. Wei End-to-End Camera Calibration for Broadcast Videos CVPR 2020

[22] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381– 395, 1981.

[23] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. "You Only Look Once: Unified, Real-Time Object Detection" 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.

[24] Ultralytics, YOLOv5, available: https://github.com/ultralytics/yolov5

[25] T.Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollar, Microsoft COCO: Common Objects in Context, ECCV 2014