Sports Field Localization using Memory Networks

Pascale B. Walters
David Clausi
Alexander Wong
University of Waterloo, ON, Canada
Email: {pbwalters}@uwaterloo.ca

Abstract

Sports analytics derived automatically from broadcast footage is a growing interest because it provides advantageous data to teams without the need for specialized equipment or trained staff. A fundamental step in automating sports video analytics extraction is registering the playing surface and transforming the broadcast footage to a top-down view. In this paper, a novel method is presented that performs automatic top-down registration of sports fields using temporal information. Using richer input data will increase the performance of the network and will not require an additional correction network.

1 Introduction

Video analytics of sports games can be used to provide teams with an advantage over their competitors, whereby they can gather more data about game events. These data can be used to influence coaching strategies and management decisions. In addition, the data can increase fan engagement as sports consumption becomes more digital.

Using readily available broadcast footage to derive analytics is advantageous as it does not require the deployment of cameras in calibrated locations in all arenas where games are played. The analytics can be generated by manual annotation, but this requires many hours of work with specially trained staff. Advances in the field of computer vision means that algorithms can be developed that rapidly and accurately collect the data. However, with the use of broadcast video for generating analytics automatically, the variable camera locations and parameters between different playing fields mean that registration of the broadcast video to the playing field is required to gather an understanding of the game from the moving camera.

There are several methods that have been proposed to calculate the 2D-to-3D projection problem for sports fields, but they are performed naively on individual frames from the broadcast video [1–4]. Fig. 1 shows the transform of a hockey broadcast video frame to a top-down view, based on a model of the ice rink. This paper proposes a method for sports field localization that includes cues from previous video frames to increase accuracy and approach real-time performance.

2 Related work

2.1 Sports field localization

Several methods have been proposed for registering broadcast video to a model of the sports field. Due to the difficulty of obtaining accurate annotated data, these methods attempt to maintain high performance with minimal training data. In addition, all of the following methods are performed on individual frames, with no inclusion of temporal information.

Homayounfar *et al.* segment the sports field and lines from each frame. This is used to calculate the registration transform based on the vanishing points of the field edges [1].

Sharma *et al.* formulate the problem as a nearest neighbours search from the segmented lines of the field. They generate a dictionary from a small amount of manually annotated frames by modeling the pan, tilt and zoom of a broadcast camera. Furthermore, smoothing of the camera motion is performed with Markov random field optimization and a camera stabilization algorithm [2].

Chen and Little have a similar approach to Sharma *et al.*, whereby they generate a dictionary of synthetic images. However, they propose a camera pose engine to build the dictionary based on prior knowledge of typical camera motion. Line segmentation is performed by two GANs and dictionary indexing is done with features from a Siamese network. Smoothing is performed with the Lucas-Kanade algorithm [3].

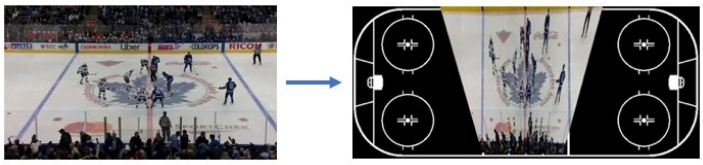

Fig. 1: 2D-to-3D projection from a broadcast hockey video frame to a top-down view of the ice rink.

Jiang *et al.* propose a network to regress the error of an initial homography estimation. They estimate the transform based on the correspondence of four points on the broadcast video frame and a model of the playing surface [4].

2.2 Memory networks for video analytics

The use of memory networks originated in natural language processing, especially for question answering. They have also found a use in visual tasks in videos, such as segmentation and tracking [5]. These methods are better able to handle occlusions, appearance changes and accumulated errors [6].

Yang and Chan use an LSTM with an attention mechanism as the memory encoder for object tracking [5]. Oh *et al.* use prior segmentation predictions in a space-time memory network to increase performance at later frames. They use an attention model for each pixel and achieve best performance by saving the intermediate prediction every 5 frames [6].

3 Proposed method

Previously proposed methods naively calculate a homography estimate based on a single frame from sports broadcast footage. Some methods perform an additional smoothing step afterwards that is based on adjacent frames [2], the retrieved transform from a dictionary [3], or a trained error network [4]. This correction is based on the assumption that the broadcast camera performs panning and zooming in a smooth fashion.

This paper proposes a homography estimation network that includes temporal information as input data. This increased richness of the input could lead to increased accuracy without the need for an additional smoothing step.

In the proposed method, homography would be inferred from four points on the broadcast footage frame, as in [4] or with other homography estimation methods [7, 8]. Additional priors from previous frames are used as input with a memory controller to select those that are most important, as in [5] and [6]. Adding in memory to the homography estimation network would increase accuracy without the need of an error correction method, thereby reducing computational cost.

Acknowledgments

This work was supported by Stathletes, Inc. through the Mitacs Accelerate Program.

References

[1] N. Homayounfar, S. Fidler, and R. Urtasun, “Sports field localization via deep structured models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5212–5220, 2017.

[2] R. A. Sharma, B. Bhat, V. Gandhi, and C. Jawahar, “Automated top view registration of broadcast football videos,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 305–313, IEEE, 2018.

[3] J. Chen and J. J. Little, “Sports camera calibration via synthetic data,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[4] W. Jiang, J. C. G. Higuera, B. Angles, W. Sun, M. Javan, and K. M. Yi, “Optimizing through learned errors for accurate sports field registration,” *arXiv preprint arXiv:1909.08034*, 2019.

[5] T. Yang and A. B. Chan, “Learning dynamic memory networks for object tracking,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 152–167, 2018.

[6] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, “Video object segmentation using space-time memory networks,” in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Deep image homography estimation,” *arXiv preprint arXiv:1606.03798*, 2016.

[8] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, “Unsupervised deep homography: A fast and robust homography estimation model,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2346–2353, 2018.