

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**MULTIPLE-CAMERA MULTIPLE-OBJECT 3D
LOCALIZATION IN SPORTS VIDEOS**

by

Yukun Yang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Authorship/Originality

Yukun Yang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: *ProductionNote:
Signature removed prior to publication.*

Date: 15 / 03 / 2021

Acknowledgements

Throughout the writing of this thesis, I have received a great deal of support and assistance.

I would first like to thank my supervisor, Associate Professor Min Xu, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to acknowledge my colleagues from my research projects at UTS for their excellent collaboration. I would particularly like to single out Dr. Yu Peng, Ruiheng Zhang, Wanneng Wu to thank you for your patient support and for all of the opportunities I was given to further my research. Especially for the object identification contributed by Ruiheng Zhang and object tracking by Wanneng Wu.

Finally, I would like to thank my parents for their support and encouragement throughout my study.

Yukun Yang
Sydney, Australia, 2020.

List of Publications

Journal Papers

- J-1. R. Zhang, L. Wu, **Y. Yang**, W. Wu, Y. Chen and M. Xu, "Multi-camera multi-player tracking with deep player identification in sports video," *Pattern Recognition*, vol. 102, p. 107260, 2020. (Published, see Chapter 5)
- J-2. **Y. Yang**, R. Zhang, W. Wu, M. Xu, "3D localization for multiple players in multiview sport videos with deep identification reasoning," *Pattern Recognition*. (Submitted for publication, see Chapter 6)

Conference Papers

- C-1. **Y. Yang**, M. Xu, W. Wu, R. Zhang, and Y. Peng, "3D multiview basketball players detection and localization based on probabilistic occupancy," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1-8. (Published, see Chapter 4)
- C-2. **Y. Yang**, R. Zhang, W. Wu, Y. Peng, M. Xu, "Multi-camera sports players 3D localization with identification reasoning," in *Proceedings of the IEEE International Conference on Pattern Recognition*. (Published, see Chapter 6)

Contents

Certificate	ii
Acknowledgments	iii
List of Publications	iv
List of Figures	viii
Abbreviation	xiii
Notation	xiv
Abstract	xvi
1 Introduction	1
1.1 Background	1
1.2 Significance and Challenges	4
1.3 Research Objectives and Contributions	6
1.4 Thesis Organization	9
2 Literature Review	11
2.1 Multi-Camera Multi-Object Localization	11
2.1.1 Back-projection-based localization	12
2.1.2 Statistical modeling-based localization	14
2.1.3 Deep learning-based localization	16
2.1.4 RGB-D and point clouds-based localization	18
2.2 Person Re-Identification for Sports Players	22

3	Data Collection and Problem Formulation	25
3.1	Sports Video Recording and Data Preparation	25
3.1.1	Temporal synchronization for sports videos and image sequences	26
3.1.2	Multi-camera arrangement and calibration	27
3.1.3	Sports video datasets collection	29
3.2	Problem Formulation for 3D Localization	30
3.2.1	Modeling the sports space	32
3.2.2	Statistical modeling for locations of sports objects	35
4	POM+CNN+IniSet Localization Method	36
4.1	2D Monocular Segmentation for Multiple Sports Players	38
4.2	IniSet for The Bayesian Iteration	40
4.3	Experimental Evaluation	42
4.3.1	Datasets and metrics	43
4.3.2	Results	45
4.4	Conclusion	50
5	PomID Localization Method	52
5.1	DeepPlayer Model	53
5.1.1	Cascade Mask-RCNN	55
5.1.2	Pose-guided partial feature embedding	57
5.1.3	Obtain players' identification	58
5.2	PomID Localization Scheme	59
5.3	Experimental Evaluation	61
5.3.1	Datasets and experimental configuration	62

5.3.2	Results	62
5.4	Conclusion	65
6	PIOM Localization Method	67
6.1	Multi-Dimensional Bayesian Model	70
6.2	Image&ID Model	72
6.3	The Posterior Probabilities	75
6.4	An Efficient Iterative Process	78
6.5	Experimental Evaluation	79
6.5.1	Datasets	79
6.5.2	Metrics and baselines	80
6.5.3	Algorithm implementation and configuration	82
6.5.4	Results	83
6.6	Conclusion	92
7	Conclusion and Discussion	94
7.1	Conclusion	94
7.2	Discussion	96

List of Figures

3.1	An example of the sport court and its multi-camera arrangement method, based on a basketball court.	27
3.2	An example of the marked distinguished points on the sport court, based on a football dataset. We take the points of the black and white grids as the distinguished points, and mark their 3D world coordinates and 2D image pixel coordinates for each camera.	29
3.3	An example of the original image frames of our collected STU basketball dataset. Here shows four out of eight camera views. The camera arrangement method is illustrated in Figure 3.1.	31
3.4	An example of the original image frames of our collected LH0716v2 youth football dataset. Here shows eight camera views. The camera arrangement method is illustrated in Figure 3.1.	31
3.5	Top view of the ground plane and the discretized grid cells, every square grid has the same width \mathcal{W} , the number of these grids is G . The number of cameras varies from different datasets.	33
3.6	The 3D world coordinate system with a cube. We use $(\mathcal{X}, \mathcal{Y}, \mathcal{W}, \mathcal{Z})$ to denote the 3D coordinate of the cube.	33
3.7	Back-projection of the cube on location k in camera view c . The ground plane is shown inside the blue lines. The occupancy of a target is described as a rectangle at the corresponding location. Note that some cubes' back-projected rectangles may not be visible if the camera views don't include the field where the location belongs.	34

4.1	This figure is an illustration of the implementation of 2D monocular segmentation for multiple sports players. We use CNN-based object segmentation to generate clear and correct foreground masks for the sports players that are captured by multi-camera. The outcomes of this method contain sports players' segmentation masks and bounding boxes. Those masks are removed when the players are standing outside of the sports courts.	39
4.2	Illustration of the IniSet model with indicative values, localization results are improved by eliminating large-scale miss-detection.	42
4.3	An example of the original image frames of the APIDIS basketball dataset. Here shows seven camera views.	43
4.4	Examples of experimental results of our proposed method. The green masks show where the players are presented. The black rectangles represent localization with qualified probabilities, while the blue areas denote likely occupied locations.	47
4.5	Examples of experimental results of our proposed method, the left column is the implementation of the POM method, while the right side is ours. Different rows present for different frames selected from various periods. The green lines denote qualified localization with probabilities higher than 0.8, while the red lines represent locations that are less likely to be occupied, with probabilities ranging from 0.2 to 0.8.	48
4.6	Precision and Recall curves in terms of the different thresholds. The first column is Precision versus different thresholds of three methods. The second column demonstrates Recall various from different thresholds. Additionally the third column is the Precision-Recall curves of these three methods.	51

5.1	The architecture of the DeepPlayer model. This model consists of two parts: (1) the Cascade Mask-RCNN for coarse-grained player detection(Cascade Mask-RCNN-P) and fine-grained jersey number recognition(Cascade Mask-RCNN-J); (2) the player mask embedding into the deep representation using PoseID. Finally, the player identity is decided by the jersey number class, the team class, and the deep representation.	54
5.2	PoseBox construction. Given a mask, the player pose is estimated by OpenPose. PoseBox1 = torso + arms + legs; PoseBox2 = head + torso + arms + legs; PoseBox3 = head + arms + legs.	59
5.3	Overview of the PomID model. The input of the PomID model includes objects' segmentation masks and ID labels. The 3D localization algorithm processes the players with identified ID and the players with ambiguous ID, respectively, followed by a post-process procedure with a set of experimentally defined thresholds. The output is the final 3D locations with distinguished identities.	60
5.4	Illustration of the PomID 3D localization results. The results are back-projected onto the original image sequences. Different colors indicate different team classes, while the labels of numbers indicate the different identities of multiple players across frames.	63
5.5	Illustration of the APIDIS dataset in Camera 3 and 6 indicates that the proposed method can avoid identity switches among Player 11, Player 14, and Player 15 in the dashed box.	64
5.6	Illustration of our localization results on the STU dataset in Camera 1, 4, and 6. The blue dotted boxes show that our method avoids identity switch between Player 6 of the white team and Player 11 of the black team.	65

- 6.1 An overview of the PIOM 3D localization framework. Firstly, we use the DeepPlayer model (see Section 5.1) that consists of a Cascade Mask-RCNN model and a PoseID model to extract the sports players’ segmentation masks and identification labels at pixel-level. At the same time, with the 3D world coordinate settings for the sports space mentioned in Section 3.2, we introduce an Image&ID model and an image distance norm to fuse the multiview pixel-wise segmentation and ID labels together with their 3D spatial relations. The synthetic images link the occupancy probabilities with the visible and computable image pixels, while the synthetic ID modules associate the identification inputs from all camera views with accurate spatial coordinates. With our proposed PIOM 3D localization algorithms, we then obtain sports players’ 3D locations and their unique ID labels. The localization results are finally given as the probabilities of locations that are occupied by the specifically labeled players. As shown above, different colors refer to different ID labels. 69
- 6.2 An example of the synthetic unit image, synthetic image and synthetic average images. (a) A synthetic unit image at location k from camera view c , the black area represents the back-projected ground plane grid on that location. (b) indicates the synthetic image where $X_{1,3} = 1$ and $X_2 = 0$. (c) and (d) are 2 examples of the synthetic average images $\bar{A}_{k,\zeta}^c$, when q_k has multiple values: $q_2 = 0.6, q_3 = 0.8, q_4 = 0.2$. But in (c) $q_1 = 1$, while in (d) $q_1 = 0$. . . 73
- 6.3 An example of the synthetic ID module. Three different colors refer to 3 different ID attributes. The first column shows the three synthetic unit ID \mathcal{R}_k s with different ID attributes from 3 different views at location k . The second column is the identification inputs from these views that including ID attributes and pixel information. The third column represents the calculated synthetic ID module R_k . . 74

- 6.4 Illustration of our localization results implemented on LH0716v2 dataset. The red cube refers to the generated location, while the yellow label indicates the detected identification outcome. In row No. 5 and 6, different colors refer to unique ID labels. Row No. 5 illustrates the localization results in 3D space, while No. 6 shows the results on the ground plane of bird-eye view. 84
- 6.5 Illustration of our localization results implemented on LH0928 dataset. Three rows indicate back-projection of the localization results from frame 3361, 3401, and 3441 on camera 2, 3, 6, and 7. The red cube refers to the generated location, while the yellow label indicates the detected identification outcome. 86
- 6.6 Illustration of our localization results implemented on LH0716v2 dataset that can overcome some extreme negative conditions, such as extremely crowded scenes, full/partial body occlusion, and inaccurate 2D detection from tiny objects. The first and third columns show the original back-projection of our localization results, while the second and fourth columns are obtained by zooming the crowded scenes inside the green borders. 87
- 6.7 Precision/Recall curves under multiple threshold settings. Results are obtained by implementing both PomID and our PIOM method on the LH0716v2 dataset. (a) and (d), P/R curves under BV distances range from $200 \sim 2000mm$; (b) and (e), P/R curves under IoU_{2d} ratios range from $0.05 \sim 0.95$; (c) and (f), P/R curves under IoU_{3d} ratios range from $0.05 \sim 0.95$. It illustrates that our proposed method PIOM outperforms the previous method PomID generally. . . 89

Abbreviation

2D: Two-dimensional

3D: Three-dimensional

AP: Average Precision

BV: Bird-eye View

CNN: Convolutional Neural Network

FPN: Feature Pyramid Network

IniSet: Initialization Settings

IoU: Intersection over Union

K-NN: K Nearest Neighbour

MCMOL: Multiple Camera Multiple Object 3D Localization

MODA: Multiple Object Detection Accuracy

MODP: Multiple Object Detection Precision

PIOM: Probabilistic and Identified Occupancy Map

POM: Probabilistic Occupancy Map

RCNN: Region Based Convolutional Neural Network

RPN: Region Proposal Network

ToF: Time-of-flight

Nomenclature and Notation

$c/C/\mathbf{C}$: the index/amount/set of cameras.

G/\mathbf{G} : the number/set of discretized grids.

k the index of grids.

$(\mathcal{X}, \mathcal{Y}, \mathcal{W}, \mathcal{Z})$: the center point coordinate, the width, and the average height of a 3D cube, respectively.

$(X_{min}, Y_{min}, X_{max}, Y_{max})$: 4-element 2D coordinate of the rectangle.

t : timestamp.

$\mathbf{I}_t = \{I_t^1, I_t^2, \dots, I_t^C\}$: image sequence from camera \mathbf{C} with timestamp t .

$\mathbf{B} = \{B^1, B^2, \dots, B^C\}$: the information that is processed from the synchronized image frames from cameras \mathbf{C} .

$\mathbf{X} = \{X_k | k \in \mathbf{G}\}$: the set of Boolean random variables where X_k represents the presence and absence of a location k .

$\mathbf{Y} = \{Y_k | k \in \mathbf{G}\}$: the set of discrete random variable where Y_k represents the index of the identity of a location k .

\mathcal{A}_k^c : the synthetic unit image at location k in camera c .

A^c : the synthetic image in camera c .

$\overline{A}_{k,\zeta}$: the synthetic average image.

\mathcal{R}_k : the synthetic unit ID at location k .

R_k : the synthetic ID module.

q_k : the marginal probability at location k , also know as the posterior probability.

ε_k : the prior probability at location k .

i : index of the proposals.

g_i^c : groundtruth of the proposal region.

p_i^c : predicted classification of the proposal region.

p_i^l : predicted vector representing the offset between the i th proposal and its corresponding groundtruth bounding box.

g_i^l : the true offset value.

g_i^m : the groundtruth mask of the proposal region.

p_i^m : the predicted mask of the proposal region.

\mathcal{L}_{cls} : the loss of team classification.

\mathcal{L}_{loc} : the loss of player bounding box regression.

\mathcal{L}_{mask} : the loss of player mask.

ABSTRACT

MULTIPLE-CAMERA MULTIPLE-OBJECT 3D LOCALIZATION IN SPORTS VIDEOS

by

Yukun Yang

Sports video analysis and object 3D detection are extensively studied problems in computer vision. As one of the most important scenarios of object detection in 3D, multiple-camera multiple-object 3D localization (MCMOL) in sports videos has recently drawn much attention in the research community due to the growing trend of object detection from monocular to multiview, i.e., from 2D to 3D.

Due to heavy occlusion in crowded sports scenes and high-speed moving targets in sports games, MCMOL for sports objects tends to be extremely challenging. Existing solutions generally apply foreground extraction as input, design statistical or Convolutional Neural Network (CNN) models commonly to all visible targets to obtain objects' coordinates and/or location encoding. However, ambiguous foreground masks and heavy occlusion limit their performance by a large margin. Moreover, the obtained coordinates cannot be associated or retrieved back to the particular objects. There is no one-to-one relationship between the outcomes and the objects to be detected. Thus, the false-positive and false-negative rates increase.

To deal with the above-mentioned issues, in this thesis, we conduct comprehensive studies about the MCMOL problems in sports videos. Due to the challenges mentioned above, we develop three multi-camera multi-object 3D localization approaches that provide accurate, reliable, and distinguishable results. Firstly, we apply Convolutional Neural Network with Initialization Settings over the Probabilistic Occupancy Map (i.e., POM+CNN+IniSet). This approach applies CNN-based monocular segmentation jointly on multiple cameras and develops an indicative

parameter initialization scheme for the Bayesian iteration model. Afterward, we propose the POM with Identification (PomID) method and introduce the Deep-Player model including a Cascade Mask-RCNN model and a pose-guided partial feature embedding to conduct segmentation and identification simultaneously for multiple players. This method separately estimates locations for individuals with identified labels and the rest of the objects without specific identities. Finally, we propose the Probabilistic and Identified Occupancy Map (PIOM) method and develop an Image&ID model to mathematically describe the segmentation pixels and identification estimation as the likelihood probabilities. This method then creates a multi-dimensional Bayesian model to estimate the localization results as posterior occupancy probabilities with unique ID labels. Given the pre-defined prior probabilities, the Bayesian model is optimized by an efficient iterative convergence. Our work is the first attempt to take advantage of CNN-based object identification for object 3D localization applications.

Experimental results demonstrate that our proposed framework improves the localization performance by a large margin and outperforms the state-of-the-art in MCMOL sports video scenarios.