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

List of Figures

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.

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

$\mathbf{B} = \{B^1, B^2, ..., B^C\}$: 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 ith 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 DeepPlayer 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.