Detection Accuracy of Soccer Players in Aerial Images Captured from Several View Points

Takuro Oki[1], Ryusuke Miyamoto[2], Hiroyuki Yomo[3] and Shinsuke Hara[4]
[1]*Department of Computer Science, Graduate School of Science and Technology, Japan*
[2]*Department of Computer Science, School of Science and Technology, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Japan*
[3]*Department of Electrical and Electronic Engineering, Faculty of Engineering Science, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Japan*
[4]*Graduate School of Engineering, Osaka City University, 3-3-138 Sugimoto Sumiyoshi-ku, Osaka-shi, Japan*

Keywords: Player Detection, Aerial Images, Informed-Filters.

Abstract: To realize real-time vital sensing during exercise using wearable sensors attached to players, a novel multi-hop routing scheme is required. To solve this problem, image assisted routing that estimates the locations of sensor nodes based on images captured from cameras on UAVs is proposed. However, it is not clear where is the best view points for player detection in aerial images. In this paper, the authors have investigated the detection accuracy according to several view points using aerial images with annotations generated from the CG-based dataset. Experimental results show that the detection accuracy became best when the view points were slightly distant from just above the center of the field. In the best case, the detection accuracy became very good: 0.005524 miss rate at 0.01 FPPI.

1 INTRODUCTION

A real-time vital sensing system using wearable sensors with the wireless communication function attached to players is under development to enhance the efficiency of training or to manage player's condition during exercise(Hara et al., 2017). In several kinds of sports scenes, both the moving speed and the density of vital sensors attached to players may become high; therefore, widely used RSSI-based or GPS-based schemes for multi-hop routing cannot be adopted to this application. To solve this problem, Image-Assisted Routing (shortly IAR) that estimates the locations of sensors from the locations of players wearing them based in image processing is proposed(Miyamoto and Oki, 2016). To obtain image information necessary for the IAR, several cameras mounted on UAVs or fixed tripods are used.

To achieve such localization, player's location must be accurately estimated in sports movies by visual object detection that is one of the most challenging tasks in the field of computer vision. To solve this task, several kinds of schemes have been proposed(Zhang et al., 2017; Dollár et al., 2012; Liu et al., 2016; Zhang et al., 2015; Zhang et al., 2014). Recently, deep convolutional networks that showed good accuracy for the image classification task has begun to show good accuracy for also visual the object detection task. The most popular one showing good accuracy for visual object detection is R-CNN(Girshick et al., 2014), and some derivatives from it such as Fast R-CNN and Faster R-CNN are proposed to improve the computational speed and detection accuracy. In 2017, YOLOv2 (Redmon and Farhadi, 2017) won the best paper award at the CVPR conference, which showed good accuracy in several datasets such as COCO(Everingham et al., 2010), PASCAL(Everingham et al., 2010), etc. with good processing speed.

These object detection schemes showed good accuracy for human detection in urban scenes. Especially, detection accuracy by R-CNN can be improved if appropriate training and tuning are applied(Zhang et al., 2017). However, it is shown that detection accuracy of an object detection constructed with informed-filters using only color features is better than YOLOv2 and Faster R-CNN for player detection in soccer scenes(Nakamura et al., 2017). The

Oki T., Miyamoto R., Yomo H. and Hara S.
Detection Accuracy of Soccer Players in Aerial Images Captured from Several View Points.
DOI: 10.5220/0006960500710076
In *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2018)*, pages 71-76
ISBN: 978-989-758-325-4

computation speed of the scheme based on informed-filters can be improved by parallel processing on a GPU(Oki and Miyamoto, 2017) without degradation of detection accuracy; therefore, this scheme is more suitable for embedded systems than schemes based on deep learning that require huge computational resources for real-time processing.

The accuracy of the detection scheme based on informed-filters using only color features was evaluated using top-down view images generated from a CG-based dataset where the locations of players determined from an actual soccer scene. However, in the previous researches(Manafifard et al., 2017; Gerke et al., 2013), the relation between camera locations and detection accuracy was not investigated even though it seems important for player detection using aerial images. To find the optimal camera locations to detect players in aerial images obtained from UAVs this paper evaluates the detection accuracy of players when view points are changed.

The rest of this paper is organized as follows. Section 2 explains the dataset used in the experiment. Section 3 shows how to construct a detector with informed-filters and datasets corresponding to several view points. The heat map representing detection accuracy is shown in section 4 and section 5 concludes this paper.

2 DATASETS USED IN THE EVALUATION

This section details a dataset used in our experiment.

2.1 True Locations Obtained from Actual Motion of Players

Motions of players in a dataset used in our experiment is identical to them in the dataset created in the previous research(Miyamoto et al., 2017). The dataset was created using actual motions of players in a soccer game that were obtained from several image sequences captured by cameras located as Fig.1. To determine actual locations of players in a frame, rectangles that represents player locations in captured image sequences were marked manually. After mark-up of all players in frames included in the dataset, a two dimensional location in the two dimensional image plane is determined by the center of the bottom edge of a rectangle and then a location on the soccer field is obtained by coordinate transform. Fig. 3 shows the overview of the coordinate transform from an image plane to a soccer field.

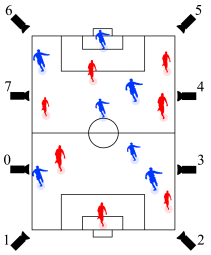

Figure 1: Camera locations.

Figure 2: An example of obtained image.

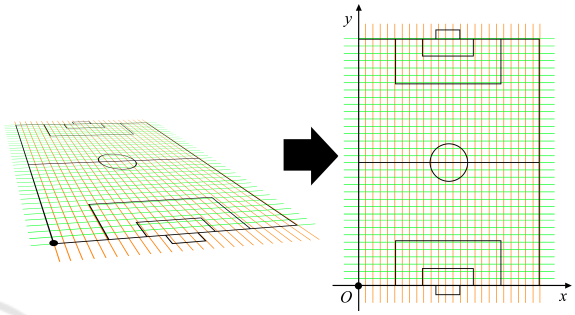

Figure 3: The overview of the coordinate transform of camera1.

After the creation of the ground truth about player locations, a three dimensional virtual space representing player's motions of the soccer game was created using Unity engine, where Unity-chan, a virtual character having a 3D model, is adopted to represent players on the soccer field. By using the dataset created on virtual dataset represented by the 3D CG technology, we can easily obtain images captured from arbitrary view points for a soccer scene.

2.2 Datasets Generated from the CG-based Dataset for Several View Points

The purpose of this paper is evaluation of detection accuracy according to several view points for aerial images obtained from a camera mounted on a UAV. Therefore, several kinds of datasets are created from the CG-based dataset (Miyamoto et al., 2017) changing locations and orientations of a camera in the virtual space. In addition to the images, locations of players in the generated images are also generated automatically.

Fig 4 shows the view points used to generate two dimensional images from the CG-based dataset. The height of the camera when it is located at just above the center of the soccer field was set to 50m in the virtual space. This location was indicated by "0". The camera location moved on the four circumferences: red, green, blue, and orange ones. Here, the angle be-

tween nearest radiuses was set to 15° and the camera orientation was set to the opposite of radial direction.

Fig. 5 shows examples of the generated images as the datasets for the evaluation in this paper.

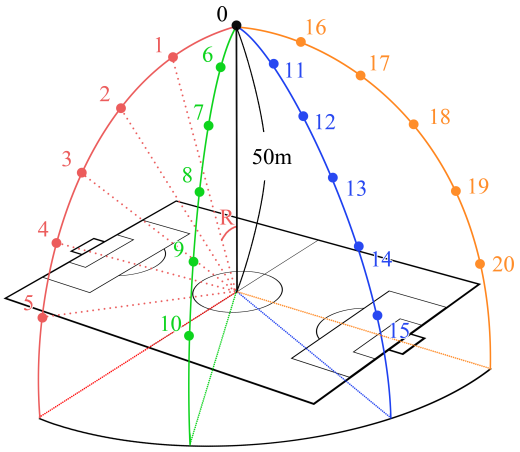

Figure 4: Camera locations.

3 HOW TO CONSTRUCT A DETECTOR BASED ON INFORMED FILTERS

This section details how to construct a soccer player detector using only color features with informed filters to evaluate the detection accuracy according to several kinds of view points. The rest of this section describes how to select training samples, how to design a template pool required for learning based on informed filters, and how to construct a final strong classifier.

3.1 Training Samples

Datasets used for the evaluation in this paper is composed of two dimensional image sequences according to several view points generated from the CG-based dataset as shown in the previous section. To train a classifier used as a detector, positive and negative samples must be extracted from the images generated from the CG-based dataset. Figs. 6 show examples of positive training samples where players are located in the center of cropped sub-images. On the other hand, negative samples do not include any players in cropped sub-images as shown in Figs. 7.

At the generation process of positive samples from image sequences using the actual location of targets obtained from the CG-based dataset, occlusions sometimes occur when view points moves drastically. Such occluded samples are not included in the training and evaluation samples because they are not suitable for appropriate evaluation.

3.2 Template Design for Informed Filters

The training process of a detector by a informed filters can be summarized as Figs. 8, and Figs. 9. Here, an average edgemap is generated from positive training samples. Then, the obtained edgemap is divided into rectangular regions called cells and labels are assigned to these cells considering their locations. Finally a template pool is designed using assigned labels.

It is obvious that positive training samples is necessary for the template design described as above. However, a good template pool may not be obtained if inappropriate samples are used for edge map generation. Especially, the selection of positive samples for template design becomes important because appearance of players become quite different when view points change drastically.

To avoid this problem, a template pool for informed filters are generated as follows:

1. Divide all view points into five groups according to shooting orientations.
2. Generate edge maps for each group.
3. Create template pools according to generated edge maps.
4. Merge all template pools obtained in the previous procedure to generate a large template pool.

This procedure tries to maintain unique characteristics according to shooting orientations in the generation of a template pool.

After the generation of the large template pool, a strong classifier is constructed. In the training process, the Boosting algorithm selects effective weak classifiers from the large template pool.

4 DETECTION ACCURACY ACCORDING TO CAMERA LOCATIONS

This section evaluates the detection accuracy for several view points and find good one for player detection from a camera mounted on a UAV.

4.1 Test Images

In the evaluation, not to use the same target for both training and testing on the same frame, 1800 frames

Figure 5: Examples of dataset.

Figure 6: Positive samples.

Figure 7: Negative samples.

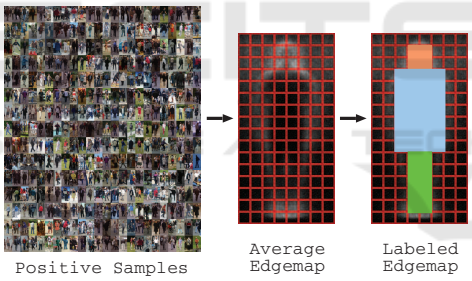

Figure 8: Generate edgemap.

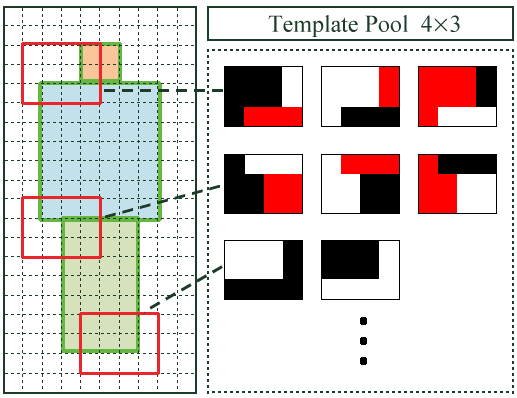

Figure 9: Template generation.

for each view point that were generated from the same 1800 frames in the CG-based dataset were kept for only testing. For the training of the detector, other 7200 frames for each view point were used.

4.2 Detection Procedure

Detection is performed by the exhaustive search with sliding windows that extracts huge numbers of sub-images from an input image to determine whether an extracted sub-image includes a detection target or not. In this process, scaled images are generated to find targets whose resolution is different from the size of detection window. Because the size of detection targets do not change drastically in aerial images, this evaluation uses only three scales: $\times 1.0$, $\times 1.05$, and $\times 1.1$. The size of stride to move sliding windows was one, which means all locations of an input image were evaluated.

For the evaluation, a detected sub-window was accepted as true positive if $Score_{overlap}$ computed by the following equation was greater than 0.65:

$$Score_{overlap} = \frac{BB_{GT} \cap BB_{DET}}{BB_{GT}}, \qquad (1)$$

where BB_{GT} and BB_{DET} mean a sub-window defined in ground truth and a detected sub-window, respectively.

4.3 Accuracy

Tables 1,2,3,4, and 5 show the evaluation results of detection accuracy. In these tables, miss rates at some false positive per images (FPPIs) are shown. The "nan" in the table means that there was no false positive in the evaluation process.

These results show that very accurate detection can be achieved at all view points. The best accuracy was obtained at cam17. The primary reason to improve detection accuracy must be the visual cues increases according to the increase of the projected area of detection targets that becomes the smalls at cam00. However, the detection accuracy become worse when

Table 1: Accuracy table, cam00.

	MR@cam00
FPPI=1.0	nan
FPPI=0.1	nan
FPPI=0.01	0.006490
FPPI=0.001	0.812667

Table 2: Accuracy table, cam01-05.

	MR@cam01	MR@cam02	MR@cam03	MR@cam04	MR@cam05
FPPI=1.0	nan	nan	nan	nan	nan
FPPI=0.1	nan	nan	nan	nan	nan
FPPI=0.01	0.006414	nan	0.008468	0.015025	0.031540
FPPI=0.001	0.558980	0.007738	0.014591	0.033021	0.037637

Table 3: Accuracy table, cam06-10.

	MR@cam06	MR@cam07	MR@cam08	MR@cam09	MR@cam10
FPPI=1.0	nan	0.005433	nan	nan	nan
FPPI=0.1	nan	0.005479	nan	nan	0.021780
FPPI=0.01	nan	0.005484	0.006148	0.009432	0.027652
FPPI=0.001	0.005678	0.005683	0.007467	0.012366	0.036157

Table 4: Accuracy table, cam11-15.

	MR@cam11	MR@cam12	MR@cam13	MR@cam14	MR@cam15
FPPI=1.0	nan	nan	nan	nan	nan
FPPI=0.1	nan	nan	nan	0.006938	0.017012
FPPI=0.01	0.006399	0.005524	0.006243	0.012166	0.024646
FPPI=0.001	0.007465	0.006269	0.012894	0.059582	0.039167

Table 5: Accuracy table, cam16-20.

	MR@cam16	MR@cam17	MR@cam18	MR@cam19	MR@cam20
FPPI=1.0	nan	nan	nan	0.006879	nan
FPPI=0.1	nan	nan	nan	0.006982	0.015134
FPPI=0.01	nan	nan	0.005956	0.008233	0.020135
FPPI=0.001	0.006808	0.005627	0.009709	0.043803	0.039258

the height of view points be too lower because heavy occlusions may be caused at such view points.

4.4 Heat Map

Fig. 10 shows the heat map that represents the detection accuracy at all view points defined in Fig. 4. In this figure, the deep red color means lower miss rate (higher accuracy) and the miss rate becomes greater if the red components becomes lower. Regions colored with gray are not used in the evaluation because detection accuracy becomes worse if the height of view points becomes too low due to occlusion caused by other targets. Some regions near the top is white; the detection accuracy is quite bad in these regions.

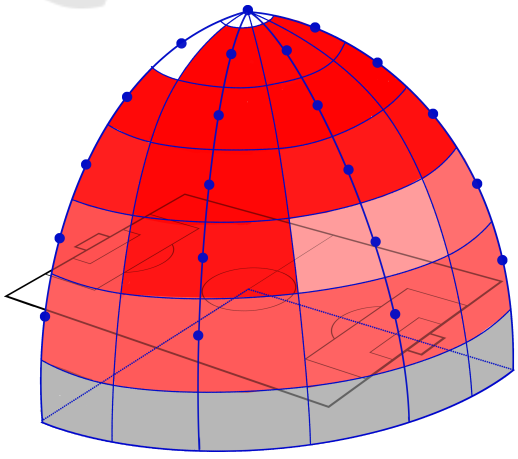

Figure 10: Heatmap.

5 CONCLUSION

This paper evaluated the detection accuracy of players for many view points to find the best UAV locations to capture aerial images. For the evaluation, several kinds of two dimensional images with annotations about player locations were generated from the CG-based dataset about a soccer game considering orientations and locations of a camera mounted on a UAV. To train a strong classifier, a large template pool was created from several template pools one of which was generated from some view points whose shooting orientation was the same. Experimental results using the generated two dimensional images show that the detection accuracy become best when a camera is located at a view point slightly distant from just above the center of the field.

ACKNOWLEDGEMENTS

The research results have been partly achieved by Research and development of Innovative Network Technologies to Create the Future, the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN.

REFERENCES

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Gerke, S., Singh, S., Linnemann, A., and Ndjiki-Nya, P. (2013). Unsupervised color classifier training for soccer player detection. In *Visual Communications and Image Processing*, pages 1–5.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 580–587.

Hara, S., Yomo, H., Miyamoto, R., Kawamoto, Y., Okuhata, H., Kawabata, T., and Nakamura, H. (2017). Challenges in Real-Time Vital Signs Monitoring for Persons during Exercises. *International Journal of Wireless Information Networks*, 24:91–108.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *Proc. European Conference on Computer Vision*, pages 21–37.

Redmon, J. and Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pages 65176525.

Manafifard, M., Ebadi, H., and Moghaddam, H. A. (2017). A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46.

Miyamoto, R. and Oki, T. (2016). Soccer Player Detection with Only Color Features Selected Using Informed Haar-like Features. In *Advanced Concepts for Intelligent Vision Systems*, volume 10016 of *Lecture Notes in Computer Science*, pages 238–249.

Miyamoto, R., Yokokawa, H., Oki, T., Yomo, H., and Hara, S. (2017). Human detection in top-view images using only color features. *The Journal of the Institute of Image Electronics Engineers of Japan(in Japanese)*, 46(4):559–567.

Nakamura, Y., Nakamura, T., Oki, T., and Miyamoto, R. (2017). Comparison of various approaches for object detection. In *Fall Meeting of Federation of Imaging Societies*, pages 94–98.

Oki, T. and Miyamoto, R. (2017). Efficient GPU implementation of informed-filters for fast computation. In *Image and Video Technology*, pages 302–313.

Zhang, S., Bauckhage, C., and Cremers, A. (2014). Informed haar-like features improve pedestrian detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 947–954.

Zhang, S., Benenson, R., and Schiele, B. (2015). Filtered channel features for pedestrian detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1751–1760.

Zhang, S., Benenson, R., and Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4457–4465.