A Deep Learning-Based Framework for Racket Sports Court Registration

Ahmed Jouini[1], Melek Elloumi[1], and Faten Chaieb[1]

Paris Panthéon-Assas University, EFREI Research Lab, F-94800, Villejuif, France
{ahmed.jouini,melek.elloumi,faten.chakchouk}@efrei.fr

Abstract. In this paper, we present a new framework that combines deep semantic segmentation with homography estimation to address challenges in racket sports court registration from broadcast videos. In particular, we deal with courts presenting the following problems: (a) brushed and occluded lines, (b) illumination variations, and (c) unknown camera parameters. Given an input frame from a broadcast video, our approach employs an encoder-decoder deep neural network to predict a precise pixel-level segmentation mask, which is then used to estimate the homography matrix between the input frame and its reference court model. For a comprehensive evaluation, we have developed two datasets for badminton and tennis that meet our specific needs. Since datasets and state-of-the-art methods with code are not publicly available, we compared our framework with a commonly handcrafted approach largely used as a baseline method in racket sports analysis. We show that our method outperforms the baseline in terms of registration accuracy and inference latency per frame.

Keywords: Court registration · Racket sports · semantic segmentation

1 Introduction

Sports-field registration and detection tasks have gained significant interest due to their crucial role in video sports analysis [26]. Early attempts [2–6] relied heavily on handcrafted features like field lines and keypoints, primarily focusing on court-net sports. As deep learning techniques proved their capabilities in sports video analysis [19], they have been widely applied across a diverse array of sports, far beyond just racket-based games [8, 23, 9, 24, 20, 15, 18].

First, a sports-field registration method employ homography estimation and non-linear Levenberg-Marquardt optimization [7] to refine camera calibration, enabling court tracking in video frames [2]. The authors propose a detailed process starting with white pixel detection, followed by line extraction using the Hough transform [22]. They then fit these lines to a court model using a homography matrix. For tracking across frames, the method adapts the predicted camera parameters to each new input frame, ensuring consistent court registration. The method's assumption of linearity might fail in the presence of lens distortion. Taking a more geometric approach, J. Han et al [3] calculate a 3x4

camera calibration matrix, paired with a line-pixel detection using RANSAC algorithm [21]. The method struggles in complex scenarios, such as when the net line closely aligns with a court line. Building upon these techniques, B. Dang et al [4] use luminance thresholding for white pixel extraction and RANSAC [21] for line parameter estimation. The tracking phase extends detected court lines from the previous frame to establish a search region. This method faces challenges with occlusions, variable lighting, and dynamic environments.

Most deep learning approaches use first Convolutional Neural Networks (CNN) to extract fields features, followed by an homography estimation step that matches the input frame to the court model. We can distinguish between handcrafted homography estimation-based methods [23, 9, 15, 20] and deep learning-based regression models [24, 8, 18].

In [23], authors propose a fully-convolutional U-Net architecture for keypoint identification in Basketball, Volleyball, and Soccer field images using multiple camera setups. They perform pixel-wise classification to identify semantic keypoints, which are intersections and corners of field lines, as well as player keypoints. Initial homography estimation is carried out based on these semantic keypoints by minimizing reprojection errors between the observed 2D image points and the transformed 3D ground points. However, this approach faces challenges in conditions with cluttered scenes, and its reliance on multi-camera setups or additional sensors may limit its applicability. In [9], a framework that comprises a multi-task deep network to detect both keypoints and dense frame-features was proposed. Built on a ResNet-18 encoder-decoder architecture, they use dilated convolutions and non-local blocks for better spatial context. Initial homography is estimated using the Direct Linear Transformation (DLT) [11] with RANSAC [21] and further refined via a weighted loss optimization scheme. Although the method performs well on various sports fields, it fails in the case of clay court. Furthermore, SFLNet [24], a multi-task network, was proposed for sports filed registration through single-shot regression. SFLNet was validated only on basket sport and faces limitations in handling uncommon court appearances not represented in the training dataset. In [8], a two-stage pipeline for sports field registration in broadcast videos to prevent overfitting was introduced. The first registration network, a modified ResNet-18 architecture, serves as initial homography regression. The second stage, the registration error network, estimates the error of the field model template wraping. This method was evaluated only in the case of soccer and hockey sports and the second stage seems to be time consuming when minimizing the estimated error. Similarly in the case of soccer sport, [15] propose an encoder-decoder architecture based on a grid of uniformly distributed keypoints as field-specific features to ensure feature representation under different camera poses.

In this work, we propose a novel deep learning-based approach to register racket sports-fields from broadcast videos. We use deep semantic segmentation for court feature extraction and a handcrafted homography estimation method based on the extracted features.

Main contributions are as follows:

– Racket sports field registration datasets: To the best of our knowledge, there are no public datasets presenting homography annotations for racket sports. A tennis dataset presenting clay court challenges and a badminton dataset dealing with different field colors are created and detailed in section 3.1. Both datasets include semantic segmentation labels for specific field zones, lines, and ground truth homographies.
– A deep learning framework dedicated for racket sports-field registration. The proposed method was validated on our datasets.

The rest of the paper is organized as follows. In Section 2, we describe the proposed method. Experimental settings, results and discussions are reported in Section 3. Section 4 concludes the paper and summarizes a few directions for future work.

2 Method

As depicted in Figure 1, our proposed framework is structured into two main phases. : semantic segmentation and homography estimation. First, a semantic segmentation phase processes the input image and estimates the pixel regions corresponding to various labeled zones. These zones are defined according to the kind of racket sport. Then, an homography estimation phase is applied. It includes a corner extraction step that computes, for each zone in the output mask, a set of four points followed by an homography estimation using RANSAC algorithm.

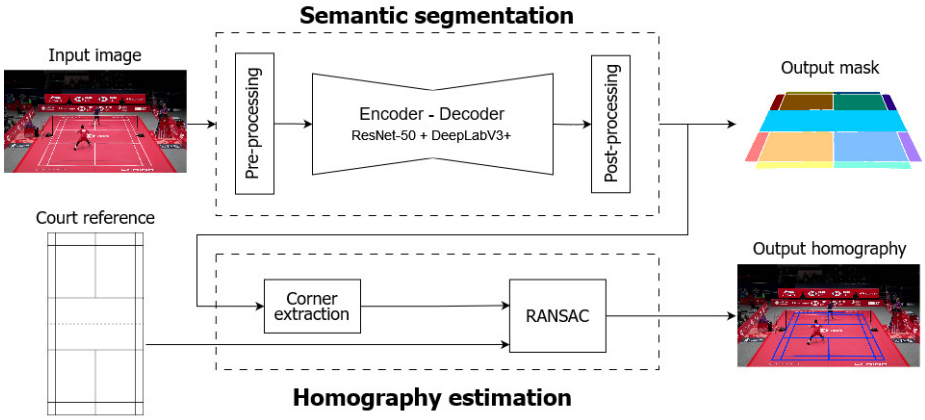

Fig. 1. Overview of the proposed framework.

2.1 Semantic Segmentation

The idea is to recognize the region of pixels for each predefined court zone by assigning the corresponding label in the output mask. As encoder, we rely on 50-layer ResNet [10] pre-traind on the extensive ImageNet dataset [25]. This backbone is the fundamental feature extractor of our semantic segmentation Network. Then a DeepLabV3Plus decoder [12] operates on the feature maps, employing a series of upsampling and concatenation operations to enhance spatial resolution, and to capture fine-grained details. For each pixel, the model generates a softmax score in a one-hot encoded mask, representing the probability of belonging to a class.

Training In this section, we will describe the training step for both badminton and tennis sports.

Badminton Case. In Figure 2, we show the 13 labeled zones for the badminton court.

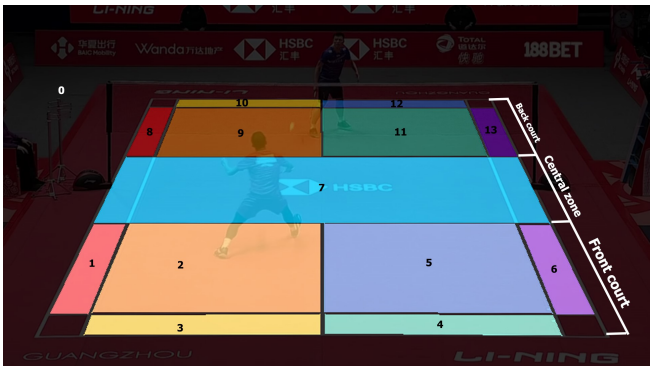

Fig. 2. Explanation of badminton mask configuration.

During training, we use the Generalized Dice Loss (GDL) [13] in three time phases t_k, $k \in \{1, 2, 3\}$ as shown in equation (1). Each time phase consists of a set of consecutive epochs.

$$\mathrm{GDL}^{(t_k)} = 1 - 2\frac{\sum_{z \in Z} w_z^{(t_k)} \sum_{i \in N_z} (p_i \cdot g_i) + \epsilon}{\sum_{z \in Z} w_z^{(t_k)} \sum_{i \in N_z} (p_i + g_i) + \epsilon} \tag{1}$$

where Z is the number of zone classes, $w_i^{(t_k)}$ is the loss weight assigned to zone z at time phase t_k, N_z is the total number of pixels for zone z, p_i is the predicted probability for pixel i, g_i is the ground truth label for pixel i, and ϵ is a small constant added to the denominator to avoid division by zero. In the first

time phase t_1, all loss weights of all zones $w_z^{t_1}$ are set to 1 so that the GDL is equivalent to the classic Dice Loss function (DL). The error is calculated using the prediction of all court zones classes. This time phase starts from the first epoch and ends when the model converges to a local minima. In time phase t_2, we enhance the detection of back zones, detailed alongside other zones in Figure 2, by setting their loss weights to 1 and canceling those corresponding to other zones. In fact, by introducing a back zone focused GDL in t_2, the network can effectively navigate away from local minima and guide the model weight updates towards more effective minima. At the end, we again perform a classic Dice loss function in order to correct the overall zones detection. The choice of GDL, rather than the commonly used cross-entropy loss, is motivated by their suitability for segmentation tasks. It specifically addresses the segmentation quality and overall agreement with the ground truth masks, leading to improved results for the zones as a whole rather than individual pixels alone.

Tennis Case. The brushed lines on clay courts cause poorly defined zone borders, making it difficult for a segmentation model to accurately distinguish between different court zones. Thus, we adopted a simplified labeling strategy for our segmentation model. Instead of segmenting each zone as a distinct label, we focused on two primary labels: the 'full court' and the 'background'. We implemented the Weighted Dice Loss (WDL) [13], detailed in equation (2), as a guided-attention on the full court zone.

$$\mathrm{WDL} = 1 - \sum_{c=1}^{C} w_c \times \frac{2 \times \sum_{i \in N_c} (p_i \cdot g_i) + \epsilon}{\sum_{i \in N_c} (p_i + g_i) + \epsilon} \tag{2}$$

Where C is the number of classes in the dataset. The weight assigned to each class c is represented by w_c, which is critical for balancing the model's focus across different classes. During training, w_c takes two values: 3.0 for the class 'full court' and 1.0 for the 'background'. N_c denotes the set of all pixels belonging to class c, essential for assessing class-wise segmentation accuracy. For each pixel i, p_i represents the model's predicted probability of that pixel being part of a specific class, while g_i stands for the actual ground truth label of pixel i. Lastly, ϵ is a small constant added to the denominator to avoid division by zero, ensuring numerical stability during the model's training phase. By assigning different weights to each class, we aim to guide the model's focus towards the more critical area, i.e., the court surface.

2.2 Post-Processing Operations

To extract the corner points for each zone, we start by applying four iterations of erosion followed by four iterations of dilation using a 3x3 kernel consisting of ones. We subsequently apply a linear regression model and a PCA-based model to accurately extract horizontal and vertical lines, respectively, from the set of contour points. Finally, we utilize the intersection points between the detected lines as inputs for estimating homography transformation using RANSAC [21] and DLT [11].

2.3 Homography Estimation

Corner Extraction. We propose to extract the corners of each zone in the predicted mask. To enhance the quality of corner detection, morphological operations are used to smooth mask zone borders. Subsequently, we convert the contours of the labeled zones into line segments, distinguishing between horizontal and vertical lines based on their orientation and alignment criteria. We employ linear regression [14] to fit lines through clusters of horizontal line segments. This regression provides two best-fit lines, each representing one of the horizontal border lines of the quadrilateral zone.

For vertical lines, we use orthogonal regression to find the line that better fits the vertical direction of cluster. Line intersections are the coordinates of labeled zones corners.

Homography Estimation. The court reference model consists of the lines that are drawn onto the ground to define the play-field geometry. The lines are defined in the model coordinate system. Once the valid corner points are identified with their corresponding reference points (in the court reference), we use the RANSAC in conjunction with the Direct Linear Transform (DLT) method to compute the homography matrix defined by equation 3. It maps points from the model plane to corresponding points in the the image plane.

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \tag{3}$$

where each element h_{ij} corresponds to a parameter that influences the transformation. Specifically, h_{11}, h_{12}, h_{21}, and h_{22} contribute to scaling and rotational effects, h_{13} and h_{23} manage translations, while h_{31} and h_{32} control the perspective distortion. The transformation enabled by the homography matrix H allows us to convert the coordinates of a point in the court model, (x, y), to its new position, (x', y'), in the image captured by the camera, as shown in equation 4

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = H \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

where (x', y') are the coordinates after applying the homography and w' is the scaling factor that is used to bring the coordinates back to the conventional 2D form.

3 Experiments

3.1 Datasets

Tennis Case. Our tennis dataset, specifically focused on clay courts, comprises 486 images extracted randomly from YouTube broadcast match videos, with 100

designated for the test phase. These images were extracted from 50fps broadcast video frames. The choice of clay courts was intentional, as they present unique analytical challenges. One of the primary difficulties in clay court analysis is that the lines are often obscured due to player movement, making the edges of areas, corners, and sometimes even entire lines difficult to discern. To ensure a diverse and comprehensive dataset, we incorporated a variety of frames characterized by differing levels of brightness, different camera positions, with occluded court lines, and presence of shadows.

We annotated the tennis court using polygons, with each polygon representing a specific court zone and an additional one for the full court. Our annotation methodology provides the versatility to train our semantic segmentation model on various label definitions within the tennis court, accommodating multiple design choices for in-depth analysis. Additionally, it facilitated the creation of binary test images, distinctly representing court lines in white against non-line areas in black. These binary images serve as the ground truth for homography estimation.

Badminton Case. This new dataset is composed of 564 images of badminton courts. The annotation consists of a precise mask outlining the different court zones. Recognizing the crucial role of high-quality training data, the dataset cover diverse court environments and recurrent edge cases.

We manually annotated a test set of 92 images, producing corresponding binary ground truth images to accurately quantify overlap accuracy. For data collection, we selected 564 RGB images from public YouTube videos featuring broadcast badminton matches with 1920x1080 resolution. To mitigate potential biases towards specific court colors, the dataset includes images of both green and red badminton courts. Furthermore, we selected images featuring varying numbers of players to enable the model to learn and understand the impact of player occlusions on court detection. Addition- ally, as badminton match videos may contain segments where the court is not visible, we incorporated a subset of images with black masks to distinctly differentiate between what constitutes a court and non-court regions.

3.2 Evaluation Metrics

IoU. The Intersection over Union (IoU) is computed between ground truth binary mask and its projection using the predicted homography. The IoU reaches a value of one when there is an exact match between the two lines regions and drops to zero in cases of no overlap.

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{5}$$

For each class i, the IoU is calculated as the ratio of the number of true positive predictions TP_i to the sum of true positives TP_i, false positives FP_i, and false negatives FN_i.

Latency. To evaluate the speed of the framework, we measure the elapsed time from introducing the input image to receiving the overlay output. It is important to note that the recorded values are dependent on the performance of the used hardware. Nevertheless, these measurements enable us to compare different approaches executed on the same hardware, providing insights into their relative efficiency.

3.3 Baselines

To the best of our knowledge, there is no available source code for State-of-the-art sports-field registration methods. Consequently, to facilitate a comprehensive and meaningful comparison within our research, we benchmark our results against the Farin et al. method [2] frequently used in rackets sport analysis work. This baseline extracts line parameters through handcrafted filtering techniques and estimates a homography from line-to-line correspondences. We use the OpenCV implementation of DLT [11] algorithm to solve a set of linear equations to find the homography matrix. Additionally, we compare our method against the KaliCalib framework [20] for basketball court registration. The authors propose an encoder-decoder architecture to predict the locations of field keypoints and background heatmaps. These predictions are used along side a predefined field template as inputs to estimate the homography matrix with RANSAC.

3.4 Implementation Details

For all our experiments, the semantic segmentation architecture was implemented using the PyTorch framework. We employed the Adam optimizer with a batch size of 2 images and an initial learning rate of 0.0001. The encoder layers of the ResNet50 [10] backbone were initialized with pre-trained weights from ImageNet [25]. We did not freeze them so they subsequently get fine-tuned to enhance the detection of badminton courts. The decoder layers of the DeepLabV3Plus [12] architecture were initialized with default PyTorch uniform distributions. Data augmentation was implemented using the Albumentations library. We applied a spatial transformation with a random probability of 0.6, utilizing a scaling parameter of 0.05 and a rotation parameter of 3 degrees. Additionally, a color jitter transformation with a probability of 0.5 was applied, modifying saturation, brightness, and contrast. Gaussian blur augmentation was applied with a probability of 0.1 to enhance robustness.

For badminton, The first Generalized Dice Loss function was used for t_1 from epoch 1 to epoch 14 and for t_3 from epoch 16 to epoch 50. The second GDL function was used for t_2 in epoch 15. We chose epoch 15 with experimentation as that's when the model converges to a local minima using only the first GDL. All experiments were conducted on a system equipped with an Intel CPU i9-13900KF (3.00GHz) and a Nvidia RTX A6000 GPU.

3.5 Results

In Table 1, we report mean, median, min, and max values of IoU scores for all the samples within the test set as obtained by the methods of Farin et al. [2], KaliCalib [20], and ours. For tennis and badminton datasets, our approach achieves enhanced IoU scores, leading in mean, median, and minimum values. Furthermore, the latency of our framework is significantly lower than that of Farin et al.'s method and comparable to that of KaliCalib. This significant latency of the Farin et al. method on the tennis dataset can be logically attributed to the challenges posed by the brushed lines and the poor conditions of clay courts. These factors introduce considerable noise and an excess of white pixels that necessitate extensive filtering. In Figure 3, we visually demonstrate the qualita-

Table 1. Quantitative results of homography estimation on our private datasets.

	Method	IoU↑				Latency(s)
		Median	Mean	Min	Max	
Tennis	Ours	**0.56**	**0.548**	**0.147**	**1.0**	**0.195**
	KaliCalib	0.477	0.473	0.0	**1.0**	0.21
	Farin et al.	0.497	0.407	0.0	0.652	20892.3
Badminton	Ours	**0.787**	**0.781**	**0.704**	0.822	0.404
	KaliCalib	0.717	0.727	0.635	**0.836**	**0.371**
	Farin et al.	0.648	0.674	0.567	0.811	3.5

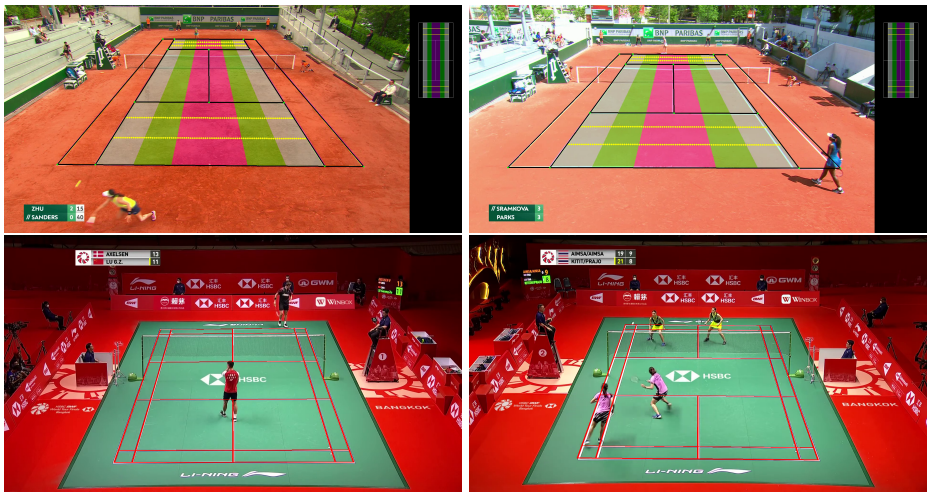

Fig. 3. Qualitative results of our approach for different examples from tennis and badminton datasets. The second column presents an instance of imprecise projection for each sport, primarily caused by the random interference of players.

tive performance of our deep learning approach applied to tennis and badminton datasets. The figure showcases the overlay of field lines (in black) and keypoints (in green) onto the tennis frames. The accuracy of our method is evidenced by the precise alignment of these overlays with the court's geometry. For badminton, the court template is segmented into colored zones and lines, serving as a foundational element for subsequent analysis. A side-by-side comparison within the figure provides a stark contrast between the precision of our model's projection in the first column and instances of less accurate projection in the second column.

Components Analysis of The Semantic Segmentation Model. In order to justify our network design choices, we conducted a model components analysis on the badminton dataset that focuses on the encoders, decoders and loss functions. Table 2 summarizes the results of different network architecture variants. We show the results of variants using different encoders which are respectively ResNet-34 [10], ResNet-101 [10], ResNet-152 [10], EfficientNet-B5 [27] and ResNet-50 [10]. We found that both ResNet-101 and ResNet-50 yielded good results. However, considering the doubled number of parameters and computational complexity of the former, the latter was deemed a better choice overall. In addition, we studied different decoders and found that DeepLabV3Plus [12] has good results against DeepLabV3 [28], UNet[16], UNet++[17].

Table 2. Comparison of the semantic segmentation model variants

	Variant	mIoU(%)↑		
		Train	Valid	Test
Encoder	With ResNet-18	99.66	99.41	98.67
	With ResNet-34	99.78	99.47	99.00
	With ResNet-101	99.64	99.14	98.40
	With ResNet-152	98.65	99.50	98.90
	With EfficientNet-B5	99.17	98.99	98.35
Decoder	With DeepLabV3	99.59	99.31	98.93
	With U-Net	99.73	99.05	98.81
	With U-Net++	99.79	99.26	98.40
	Ours	**99.79**	**99.50**	**99.08**

Dataset Composition Analysis. Table 3 shows the performances our model based on different datasets properties. The term non-court indicates that the dataset comprises images where court visibility is absent. The proposed method provides best results when the model is trained on a dataset composed with green and red courts, and non-court images. We also show that our model performs better on HSV color system.

Table 3. Impact of badminton court dataset characteristics on model performance.

Non-court	Green court	Red court	Color system	Augmentation	**mIoU(%)↑**		
					Train	**Valid**	**Test**
	X	X	HSV	X	98.94	94.24	96.64
X		X	HSV	X	99.52	97.75	97.15
X	X		HSV	X	98.91	85.03	78.14
X	X	X	RGB	X	99.69	99.41	99.01
X	X	X	HSV		99.63	99.42	98.57
X	X	X	HSV	X	**99.79**	**99.50**	**99.08**

4 Conclusion

In this work, we introduce a new deep learning-based framework that combines deep semantic segmentation with homography estimation step based on RANSAC for racket sports court registration from broadcast video frames. Our experimental evaluations, through two distinct datasets, show that our method not only surpasses the baselines approachs but also exhibits robustness in court registration tasks. In Future work, we focus on estimating homography parameters based on regression techniques.

References

1. Jones, C.D., Smith, A.B., Roberts, E.F.: Article Title. In: Proceedings Title, vol. II, pp. 803–806. IEEE (2003)
2. Farin, D., Krabbe, S., de With, Peter H. N., Effelsberg, W.: Robust camera calibration for sport videos using court models. In: Storage and Retrieval Methods and Applications for Multimedia 2004, LNCS, vol. 5307, pp. 80–91. International Society for Optics and Photonics (2003)
3. Han, J., Farin, D., de With, Peter H. N.: Generic 3-D Modeling for Content Analysis of Court-Net Sports Sequences. In: Advances in Multimedia Modeling, LNCS, Springer Berlin Heidelberg, pp. 279–288 (2006). https://doi.org/978-3-540-69429-8
4. Dang, B., Tran, A., Dinh, Ti., Dinh, Th.: A Real Time Player Tracking System for Broadcast Tennis Video. In: Intelligent Information and Database Systems, LNCS, Springer Berlin Heidelberg, pp. 105–113 (2010). https://doi.org/978-3-642-12101-2
5. Mihai, P., Andreea-Oana, P., Hassan, B. L., Bruno, F., Cédric, B.: Real Time Tennis Match Tracking with Low Cost Equipment. In: The Florida AI Research Society (2018). https://api.semanticscholar.org/CorpusID:44157580
6. Silvia, V. M.: Computer vision and machine learning for in-play tennis analysis: framework, algorithms and implementation (2018). https://api.semanticscholar.org/CorpusID:198358283
7. Jorge, J. M.: Levenberg–Marquardt algorithm: implementation and theory (1977). https://api.semanticscholar.org/CorpusID:203694768
8. Jiang, W., Gamboa Higuera, J. C., Angles, B., Sun, W., Javan, M., Yi, K. M.: Optimizing Through Learned Errors for Accurate Sports Field Registration. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 201–210 (2020). https://doi.org/10.1109/WACV45572.2020.9093581

9. Nie, X., Chen, S., Hamid, R.: A Robust and Efficient Framework for Sports-Field Registration. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1935–1943 (2021). https://doi.org/10.1109/WACV48630.2021.00198
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
11. Harltey, A., Zisserman, A.: Multiple View Geometry in Computer Vision (2nd ed.). Publisher Unknown, Location Unknown (2006). https://api.semanticscholar.org/CorpusID:8641226
12. Chen, L-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Computer Vision – ECCV 2018, Springer International Publishing, Cham, pp. 833–851 (2018). https://doi.org/978-3-030-01234-2
13. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, pp. 240–248 (2017). https://doi.org/978-3-319-67558-9
14. Huang, M.: Theory and Implementation of Linear Regression. In: 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), pp. 210–217 (2020). https://doi.org/10.1109/CVIDL51233.2020.00-99
15. Chu, Yen-Jui, Su, Jheng-Wei, Hsiao, Kai-Wen, Lien, Chi-Yu, Fan, Shu-Ho, Hu, Min-Chun, Lee, Ruen-Rone, Yao, Chih-Yuan, Chu, Hung-Kuo: Sports Field Registration via Keypoints-aware Label Condition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3522–3529 (2022). https://doi.org/10.1109/CVPRW56347.2022.00396
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, pp. 234–241 (2015). https://doi.org/978-3-319-24574-4
17. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, pp. 3–11 (2018). https://doi.org/978-3-030-00889-5
18. Shi, F., Marchwica, P., Gamboa Higuera, J. C., Jamieson, M., Javan, M., Siva, P.: Self-Supervised Shape Alignment for Sports Field Registration. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3768–3777 (2022). https://doi.org/10.1109/WACV51458.2022.00382
19. Keerthana, R., Muhammad, A. A., Nur, A. R., Nurul, F. G., Saharudin, I.: Deep Learning in Sport Video Analysis: A Review. TELKOMNIKA Telecommunication Computing Electronics and Control **18**, 1926–1933 (2020). https://api.semanticscholar.org/CorpusID:216200176
20. Maglo, A., Orcesi, A., Pham, Quoc-Cuong: KaliCalib: A Framework for Basketball Court Registration. In: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, Association for Computing Machinery, pp. 111–116 (2022). https://doi.org/10.1145/3552437.3555701
21. Fischler, M. A., Bolles, R. C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM **24**(6), 381–395 (June 1981). https://doi.org/10.1145/358669.358692

22. Allam, S. H., Sherien, M., Mohamed, S., Mohammad, E. R.: A Survey on Hough Transform, Theory, Techniques and Applications. ArXiv, vol. abs/1502.02160 (2015). https://api.semanticscholar.org/CorpusID:11028590
23. Citraro, L., Márquez-Neila, P., Savare, S., Jayaram, V., Dubout, C., Renaut, F., Hasfura, A., Ben Shitrit, H., Fua, P.: Real-time Camera Pose Estimation for Sports Fields. Machine Vision and Applications **31** (2020). https://api.semanticscholar.org/CorpusID:214632673
24. Tarashima, S.: Sports Field Recognition Using Deep Multi-task Learning. Journal of Information Processing **29**, 328–335 (2021). https://doi.org/10.2197/ipsjjip.29.328
25. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**, 211–252 (2014). https://api.semanticscholar.org/CorpusID:2930547
26. Shih, Huang-Chia: A Survey of Content-Aware Video Analysis for Sports. IEEE Transactions on Circuits and Systems for Video Technology **28**(5), 1212–1231 (2018). https://doi.org/10.1109/TCSVT.2017.2655624
27. Tan, M., Quoc V. Le: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv, vol. abs/1905.11946 (2019). https://api.semanticscholar.org/CorpusID:167217261
28. Chen, Liang-Chieh, Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. ArXiv, vol. abs/1706.05587 (2017). https://api.semanticscholar.org/CorpusID:22655199