Identifying and Extracting Football Features from Real-World Media Sources using Only Synthetic Training Data

Jose Cerqueira Fernandes and Benjamin Kenwright
Heriot-Watt University

Abstract

Real-world images used for training machine learning algorithms are often unstructured and inconsistent. The process of analysing and tagging these images can be costly and error prone (also availability, gaps and legal conundrums). However, as we demonstrate in this article, the potential to generate accurate graphical images that are indistinguishable from real-world sources has a multitude of benefits in machine learning paradigms. One such example of this is football data from broadcast services (television and other streaming media sources). The football games are usually recorded from multiple sources (cameras and phones) and resolutions, not to mention, occlusion of visual details and other artefacts (like blurring, weathering and lighting conditions) which make it difficult to accurately identify features. We demonstrate an approach which is able to overcome these limitations using generated tagged and structured images. The generated images are able to simulate a variety views and conditions (including noise and blurring) which may only occur sporadically in real-world data and make it difficult for machine learning algorithm to 'cope' with these unforeseen problems in real-data. This approach enables us to rapidly train and prepare a robust solution that accurately extracts features (e.g., spacial locations, markers on the pitch, player positions, ball location and camera FOV) from real-world football match sources for analytical purposes.

CCS Concepts

• **Social and professional topics**; • **Software and its engineering** → Software creation and management;

Keywords

graphics, machine learning, football, analytics, machine learning, generating, training, procedural

1 Introduction

Manual feature tagging and extraction in machine learning is tedious (if not impractical) and error prone. We demonstrate that it is possible to perform computer vision feature extraction on raw football game images using a machine learning solution that was trained using only synthetic data. The football community has long enjoyed the benefits of using machine learning techniques to extract and analyze details (from individual player performance to the overall game statistics [Herold et al. 2019]). Common to use real-world data sources (recordings of live games) as they captures details which are vital for analytics. Identifying and extracting the features for these analytical processes is often difficult and painstaking slow (especially for raw/live football matches). While researchers have tried to improve the methods for tagging and extracting information from football images, little research has been done on using synthetic models for training the system. We show that it is possible to synthesize data with minimal domain gap (difference between the synthetic and real-data), so that models trained on synthetic data generalize to real in-the-wild datasets.

We describe how to combine 3-dimensional graphical models (environments, objects and players) with a rendering framework to create image datasets with unprecedented realism and diversity. We train a machine learning system using the generated dataset to identify feature such as landmark localizations. We demonstrate that synthetic data can be used to accurately locate key features in real-world data as well as open up new approaches where manual labelling would be impossible.

Article, August, 2022, Synthetic Football Training Data

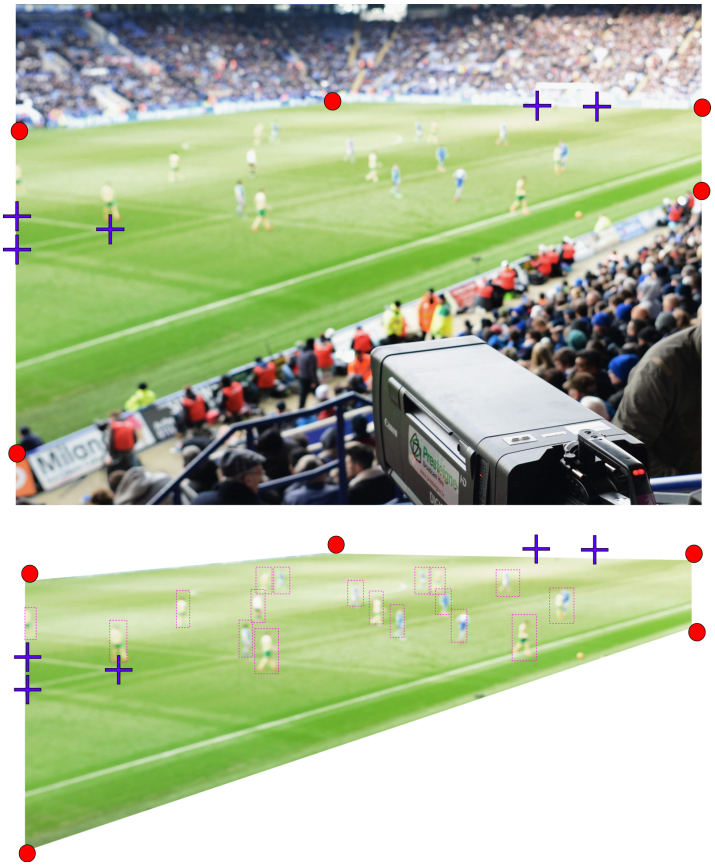

Figure 1: Live Image Data - Ability to identify and extrapolate key areas from real-world images (football pitch and object markers) in real-time without human intervention (this includes images with poor resolution, blurring, artifacts and occluded details).

Contributions The main contributions of this work are: (1) novel technique for synthesizing labeled football images for training machine learning algorithms; (2) demonstration of the capability of synthesizing features of arbitrary geometries and their corresponding labeled images; (3) use of the synthesized data for training machine-learning based feature parameter extractors; and (4) open source datasets for testing of synthetic training data.

2 Related Work

Feature matching machine-learning algorithms require a huge pool of labeled data for proper training, which is often unavailable [Konnik et al. 2021]. To resolve this shortage generated images can be used in place of real-world images. The concept of using generated graphical images for training machine learning systems is not a new one, and has been used other areas, such as face analysis [Wood et al. 2021] and feature extraction (x-rays images [Konnik et al. 2021]).

The most common approach for automatically identifying features in computer-vision (CV) techniques is to segment the images into distinct partitions [Blaschke and Lang 2006; Kaur and Kaur 2014; Zaitoun and Aqel 2015] which hopefully provide meaning for extracting geometric information.

Currently all of the publications on machine learning and football feature extraction have used real-world data as the source for the training models

[Herold et al. 2019]. Most of these data extraction projects have focused on understanding the game mechanics (predicting and understanding the game) [Herold et al. 2019; Rommers et al. 2020].

Our work: As far as we are aware no other work has attempted to fully generate a synthetic training set using computer graphics for use on real-world football game feature extraction.

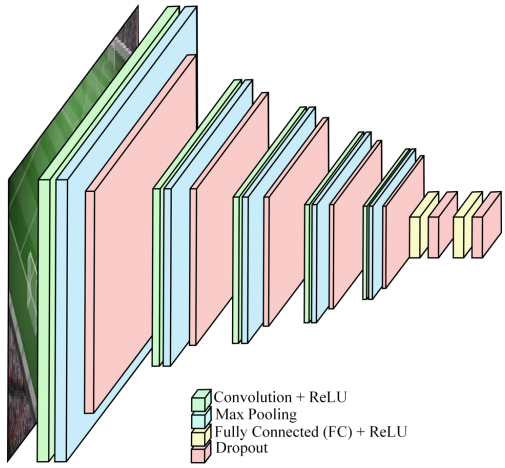

Figure 2: Network Architecture - Keypoint extraction using a CNN. The figure summarizes the blocks of the convolution layers of the approach presented in this paper. For example, given a input of 256x256 pixels we extracted 52 dimensional features (26 x-y points identifying key locations on the pitch - as shown in Figure 4).

3 Method

Training methodology We rendered multiple image datasets with a focus on specific features. During training, we perform data augmentation with overlay images. These augmentations were especially important for synthetic images which would otherwise be free of imperfections (noise and other artifacts). While some of these could be done at render time, we performed them at training time in order to randomly apply different augmentations to the same training images. We implemented neural networks with PyTorch and trained them with the Adam optimizer. Since the football pitch sits on a flat ground plan, a top down layout was used to define the pitch keypoint markers (shown in Figure 4).

Training Data and Test Cases The experiments comprised of different datasets with 3000 images for training and 100 images for testing. The different datasets evaluated the value and impact of different synthetic features (for example, the value of adding or removing lighting from a synthetic scene also the importance of a realistic surroundings vs blank or randomly placed graphics).

- Flat pitch (blank background, no lighting or players) as shown in Figure 4.
- Pitch with random pictures in the background (add noise and feature distractions to the training)
- Pitch with random lighting conditions
- Pitch players randomly placed on pitch
- Pitch with stadium model and random lighting conditions
- Samples with visual artifacts added to simulate various conditions (e.g., weather or image corruption), using overlay textures (to noise, corruption, blurring).

Landmark (Pitch) localization Landmark localization finds the position for the pitch (2d locations on screen and their corresponding 3-dimensional locations relative to the camera). We evaluate our approach using synthetic data (both controlled test cases and generate ones to evaluate a diverse range of problems around extracting features from football images).

The trained network used a mean squared error loss to directly predict. We use the provided feature points to extract and identify regions of interest from the image (e.g., pixel regions-of-interest from each image). As shown in Figure 4, the ground markers provide essential details (e.g., location of players during the game). Importantly, multiple image sources could be used to extrapolate the complete game data, which means, different images sources may only show specific regions of the game at specific times (including different viewing angles, lighting, image artifacts and so on). Most of the time from live media footage of football games, it is rare to see all the players and all the information in high resolution at the same time in a single image.

2-dimensional data to 3-dimensional data The football pitch is flat and conforms to a international standard, so the 2d pitch key-point data can be used as a point of reference for reconstructing a relative 3-dimensional layout of the game. This data feeds through to player position data, using the key-point data lets us calculate the location of the player in real-world coordinates.

Our synthetic football images are realistic, diverse and scalable. Starting with a basic pitch template, we randomize features (orientation and viewing position), random placement of objects and backgrounds, and the 'visual' characteristics (lighting and weather conditions). We finally render the scene to a database with the associated keypoint data (i.e., the location features within the image).

Player Body/Pose Detection There is already a substantial body of work that has focused specifically on the detection and extraction of human poses (describing the bounding area and body keypoints). For the tests in this paper, we used the PyTorch ResNet101 pre-trained model which produced acceptable results (and could be swapped out for other versions at a later date). For example, the DeepCut model provides a robust solution for multiple close proximity interactions [Pishchulin et al. 2016].

To ensure the body/pose phase did not detect individuals in the crowd (only providing details for the players), after the football pitch outer marker regions were detected, these were used to 'cull' the image to the field area (see Figure 1.

4 Results

Only synthetic (generated) data was used for training. The generated image datasets where created using web-based tools (i.e., WebGPU API for hardware acceleration [Kenwright 2022]). We used a web-based solution due to the fact that new technologies offered by modern browsers have greatly increased in capabilities, not to mention, ubiquity across platforms, easy to make iterative updates (during early exploratory stages). The WebGPU API [Kenwright 2022] offered GPU-accelerated client-side rendering within the browser (generated 3000 images within a few minutes). The training of the network was done offline using PyTorch. While we experimented with different network configurations initially, the final network topology is shown in Figure 2. This configuration was used for all of the tests and final results. The web-based resources, generated training data and scripts are accessible online (Git Repository [Project Repository 2022]).

- Synthetic images for training (sample set for training and a corresponding set for validating/checking errors)
- Simple 'clean' image dataset (perfect images with no noise or inconsistencies) - first check to ensure the system is working
- Complex and diverse dataset (larger range of views) - including 'ambiguous' scenes (difficult to extract and identify the pitch data from what the image shows)
- Mix lighting conditions and noise
- Add objects and players (ball and players on pitch)
- Overlay 'identified' features on top of 'real-world' data images (compare visually the errors). Such as the location of field markers and the camera frustum.
- Use different colors to visualize the information (ball, players and the pitch)

Evaluating the landmark detectors from the image database, which exhibits environmental, lighting and camera information, we see that models trained with synthetic data alone can generalize well to real data of diverse sources (broadcast media). In order to generate diverse synthetic data, our generative models must be trained with diverse range of configurations.

Network Architecture The final network architecture configuration was reduced to a reasonable size using a combinations of Conv2d and MaxPool2d filter layers before using a series of Dense layers. It required 5 Conv2d+MaxPool2d layers to bring the image dimension down to an acceptable size. Dropoff was added between each layer in order to reduce overfitting (increasing the dropoff probability gradually in later layers).

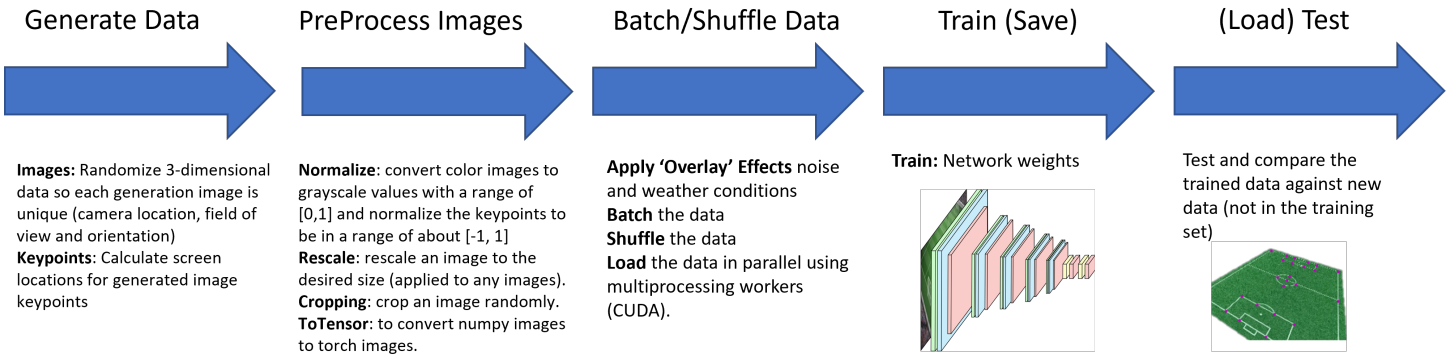

Figure 3: Stages - Process of creating the dataset through to training and evaluation.

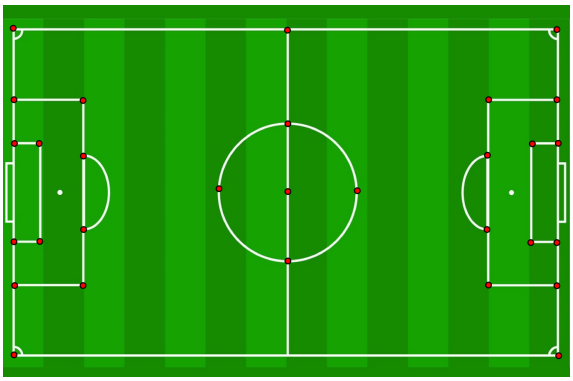

Figure 4: Pitch Markers (KeyPoints) - The field markers are essential as they provide vital information (e.g., players location during the game). As shown in Figure 9, the view of the game might only show a limited region (field markers help the machine learning algorithm identify the environmental positions). Consists of 7 horizontal lines, 10 vertical lines and 3 arcs.

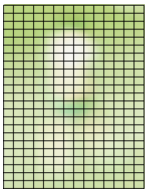

Figure 5: Resolution - Often features are low-resolution (e.g., players); usually only taking up a few pixels in the bigger picture. For instance, the image shows a zoomed in view of one of the players from Figure 1. Note the player is more identifiable when the image is considered in context of the surrounding's (zoomed out versus in isolation).

Only a single architecture is presented here but other models could be explored for further work (see Figure 2). Training times where generally in the range of half a day with CUDA acceleration (500 epochs).

Number of Epochs and Batch Size .
Experiment used a batch sizes of 50. We decided to go with batch size of 50 because a larger batch size gave much smoother loss trends over time, which made the loss curve a little easier to interpret, especially for initial experiments where we used a small number of epochs. Also, the computer faced no difficulties in performing the training steps when using batch size of 50. Initially we ran quick experiments for combinations of different architectures, batch sizes, optimizer and loss functions using 1-5 epochs. For the final setup, we settled on 500 as the training epoch (during testing we found that after 500 epochs the loss decreased very slowly - see Figure 8).

Figure 6: Stadium - Complexity and the size of the visual can be extended to encapsulate a wide range of visuals, including spotlights, crowds and shadows.

Limitations The training set is 'generated', so care needs to be taken that the 'generated' data does not 'miss' characteristics that may be visible in real-world data (but not in the generated versions). Also specific cases where the image does not contain enough information to identify pitch markers (e.g., only grass can be seen). If the resolution is low (e.g., Figure 5, it can be difficult to identify player/match details - especially when groups of players are in close proximity).

Future Work Future work would explore 'spatial' models, that is, not just about single frame snap-shots but taking into account multiple frames to extrapolate and extract extra details (e.g., motion analystics and behavioural attributes that maybe difficult or impossible to spot using individual images).
Continue to extend the dataset and provide an online resource (web-based solution) that can be used by anyone to upload and extract features from football images (also video). User could download and use these details for analysis.
Explore fusing multiple dataset resources together to obtain more accurate results, such as, images that record the same football game but from different cameras and viewing locations. Additional research on improving the keypoint approach, for instance using additional feature locations to improve accuracy [Widya et al. 2018], also the opportunity to explore the generation and fusion of other data sources (beyond just 2d images) as being explored in related fields [Fauzy Othman et al. 2022].

5 Conclusion/Discussion

The proposed technique has endless potentials and addresses the lack of automated robust, fast and accurate methods for extracting features from football match images.

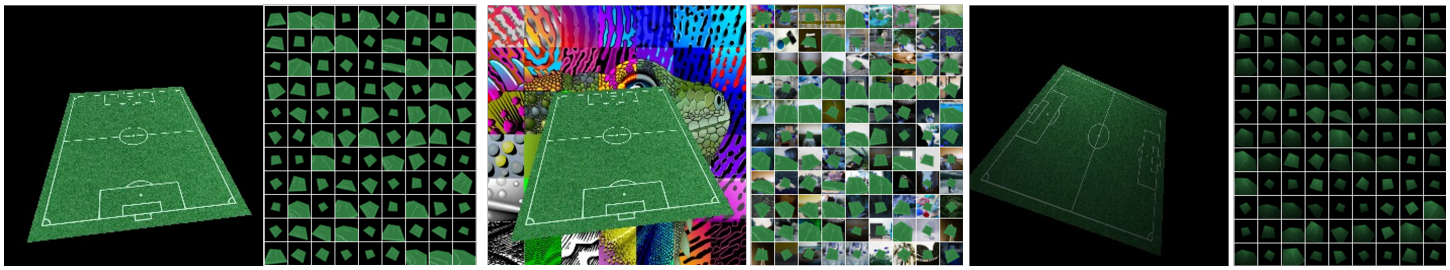

Figure 7: Multiple Datasets - Created multiple datasets, each dataset consisted of 3000 training images (256x256) with 52 keypoint markers. All images represent 3-dimensional views of a football pitch. (left) basic field with no lighting or background, (middle) random background images and (right) random lighting conditions.

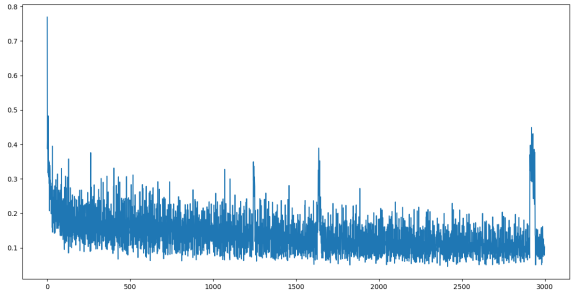

Figure 8: Training Plot - Fitness of the network as it is trained (500 epoch).

Figure 9: Scene and Lighting Conditions - Easy to integrate parameters into the generation pipeline to control visual features (e.g., enabling/disabling crowds and lighting conditions). Left middle and right shows the same scene but with no crowd, crowd and amplified darkness).

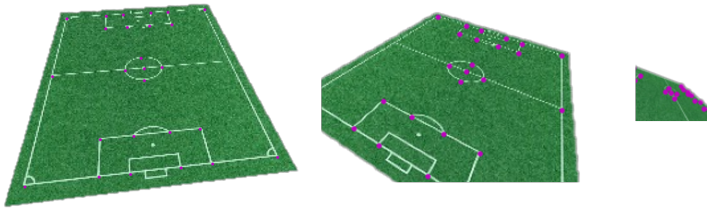

Figure 10: Example KeyPoint Data - When features are available, even limited views (image on right), the model is able to identify correctly the field points.

Synthetic training data will become more prominent over the years as generated content becomes indistinguishable from the real-world. Allows training data to be generated that would otherwise be impossible or rare using real-world data. While this article has focused specifically on football-games, the concepts are applicable to other areas such as medical operations, vehicle navigation data, flight simulations and even interactive games.

Figure 11: Players - Characters can be customized and placed in the scene, either to recreate existing games from other data or randomly to create a diverse and complex dataset (includes poses and textures).

References

Thomas Blaschke and Stefan Lang. 2006. Object based image analysis for automated information extraction–a synthesis. In *Measuring the Earth II ASPRS Fall Conference*. CD-ROM San Antonio, TX, 6–10. 1

Petronas Fauzy Othman, Phil Bartie, and Benjamin Kenwright. 2022. A Review of Point Cloud Deep Learning Fusion for Unmanned Aerial Data Systems. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (2022). 3

Mat Herold, Floris Goes, Stephan Nopp, Pascal Bauer, Chris Thompson, and Tim Meyer. 2019. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching* 14, 6 (2019), 798–817. 1, 2

Dilpreet Kaur and Yadwinder Kaur. 2014. Various image segmentation techniques: a review. *International Journal of Computer Science and Mobile Computing* 3, 5 (2014), 809–814. 1

Benjamin Kenwright. 2022. Introduction to the WebGPU API. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Courses (SIGGRAPH '22 Courses). Vancouver, BC, Canada.* https://doi.org/10.1145/3532720.3535625 2

Matthew Konnik, Bahar Ahmadi, Nicholas May, Joseph Favata, Zahra Shahbazi, Sina Shahbazmohamadi, and Pouya Tavousi. 2021. Training AI-based feature extraction algorithms, for micro CT images, using synthesized data. *Journal of Nondestructive Evaluation* 40, 1 (2021), 1–13. 1

Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4929–4937. 2

Project Repository. 2022. Identifying and Extracting Football Features from Real-World Media Sources using Only Synthetic Training Data. *Project URL: https://learningfootball.github.io Last Accessed Online (30/06/2022)* (2022). 2

Nikki Rommers, Roland Rössler, Evert Verhagen, Florian Vandecasteele, Steven Verstockt, Roel Vaeyens, Matthieu Lenoir, Eva D'Hondt, and Erik Witvrouw. 2020. A machine learning approach to assess injury risk in elite youth football players. *Medicine and science in sports and exercise* 52, 8 (2020), 1745–1751. 2

Aji Resindra Widya, Akihiko Torii, and Masatoshi Okutomi. 2018. Structure from motion using dense CNN features with keypoint relocalization. *IPSJ Transactions on Computer Vision and Applications* 10, 1 (2018), 1–7. 3

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. 2021. Fake It Till You Make It: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3681–3691. 1

Nida M Zaitoun and Musbah J Aqel. 2015. Survey on image segmentation techniques. *Procedia Computer Science* 65 (2015), 797–806. 1