Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.033062
Article

Tech Science Press

Automatic Team Assignment and Jersey Number Recognition in Football Videos

Ragd Alhejaily[1], Rahaf Alhejaily[1], Mai Almdahrsh[1], Shareefah Alessa[1] and Saleh Albelwi[1,2,*]

[1]Faculty of Computing and Information Technology, University of Tabuk, Tabuk, 47321, Saudi Arabia
[2]Industrial Innovation & Robotics Center (IIRC), University of Tabuk, Saudi Arabia
*Corresponding Author: Saleh Albelwi. Email: sbalawi@ut.edu.sa
Received: 06 June 2022; Accepted: 14 November 2022

Abstract: Football is one of the most-watched sports, but analyzing players' performance is currently difficult and labor intensive. Performance analysis is done manually, which means that someone must watch video recordings and then log each player's performance. This includes the number of passes and shots taken by each player, the location of the action, and whether or not the play had a successful outcome. Due to the time-consuming nature of manual analyses, interest in automatic analysis tools is high despite the many interdependent phases involved, such as pitch segmentation, player and ball detection, assigning players to their teams, identifying individual players, activity recognition, etc. This paper proposes a system for developing an automatic video analysis tool for sports. The proposed system is the first to integrate multiple phases, such as segmenting the field, detecting the players and the ball, assigning players to their teams, and identifying players' jersey numbers. In team assignment, this research employed unsupervised learning based on convolutional autoencoders (CAEs) to learn discriminative latent representations and minimize the latent embedding distance between the players on the same team while simultaneously maximizing the distance between those on opposing teams. This paper also created a highly accurate approach for the real-time detection of the ball. Furthermore, it also addressed the lack of jersey number datasets by creating a new dataset with more than 6,500 images for numbers ranging from 0 to 99. Since achieving a high performance in deep learning requires a large training set, and the collected dataset was not enough, this research utilized transfer learning (TL) to first pretrain the jersey number detection model on another large dataset and then fine-tune it on the target dataset to increase the accuracy. To test the proposed system, this paper presents a comprehensive evaluation of its individual stages as well as of the system as a whole. All codes, datasets, and experiments are available on GitHub (https://github.com/ragadalhejaily/project).

Keywords: Football video analysis; player detection; ball detection; team assignment; jersey number recognition

1 Introduction

Football is one of the most popular team sports. It generates global interest and is the most-watched sport in many countries [1], with more than 350 million players in 200 countries, worldwide [2]. Competitive tournaments, and the evaluation of teams and individual players, have evolved over the years, which has led to the emergence of various evaluation technologies. Sports analysis specialists, coaches, and assistants watch matches to evaluate players' performances, which requires a thorough knowledge of how to monitor players, as well as the target points, for analysis. These include the physical aspects of the player, how effective the player is during contact with members of his team, and how he deals with the opposing team. Each analyst has his own perspective in judging the player's performance, his strengths and weaknesses, how to best develop his skills in the future, and which playing strategies are best against a certain team. The current performance evaluation methods are either based on: (1) direct observation—that is, attendance at the match, or (2) watching recorded videos from several matches throughout the season [3].

Though current analysis is performed manually, there has been increased interest in automated systems. These aim to improve the performance of players, discover talent, determine the value of a player to a team, and target future deals based on a player's past performance, all without the need for direct observation or the watching of prerecorded videos. Automated tools can improve match preparations, player statistics, and even the way that TV personalities, such as sports anchors, tell stories about players [4]. To carry out such analyses, automatic systems must be able to detect key things—such as the soccer pitch, the players, and the ball—as well as identify players' teams, track both the players and the ball, compute the players' location within the pitch, and perform activity recognition. Each step of football analysis has several challenges that affect the subsequent steps and the final outcome of the analysis.

Accurate player detection is a significant challenge for automatic analysis systems [5]. Another important task is identifying the players, specifically doing so via their jersey numbers, which can be particularly challenging in video footage due to poor camera angles, changing player posture and movement, low-quality video feeds, and the lack of labelled datasets to automate the process [6]. Despite these difficulties, tracking both the ball and the players is critical for match analysis [7]. The automatic labelling of players to their teams is essential and will allow for easier player tracking, activity recognition, and player configuration analysis in the future [8].

Recent progress in deep learning and computer vision algorithms have proven efficient in multimedia, big-data analyses. These algorithms have led to developments in the field of sports analysis, particularly in the evaluation of video. Deep learning techniques have impacted how athletes plan for and play the game as well as how sports analysts monitor players, analyze performance on individual and collective levels, determine appropriate wages, and help the coach achieve better results [9]. This leads, in the end, to increased player performance and spectator enjoyment.

This paper proposes a system that analyzes players' performance using recorded videos. The system integrates all necessary stages, such as segmenting the football field, detecting the players and the ball, assigning players to their teams, and identifying players' jersey numbers. These steps are critical for accurate and effective video analysis, and the proposed system performs all required tasks. The main contributions of this paper can be summarized as follows:

This research proposes a new approach for team classification based on convolutional autoencoders (CAEs) that are trained, end to end, to learn feature representations via unlabeled data. The advantage of this approach is that it provides generalization for any football video; no pre-training (on team colors, etc.) is needed, so automatic analysis can begin immediately at the start of the game.

This paper also presents an approach to automatically detect and classify player numbers from their jerseys. It utilized transfer learning to pretrain the model on a large dataset and further fine-tuned it on the

target dataset. Furthermore, this paper introduced a new dataset for jersey numbers, which allowed this research to circumvent the lack of publicly available datasets for jersey number identification.

Finally, this paper proposed an effective method to detect the football based on a modified version of the You Only Look Once (Yolo) v5 object detection model, which improves football detection.

The rest of the paper is organized as follows: Section 2 presents the challenges of analyzing player performance from videos. Section 3 describes related works. Section 4 explains the proposed system in detail. The quantitative results and discussions are presented in Section 5. Lastly, the conclusions and directions for future work are presented in Section 6.

2 Challenges of Analyzing Player Performance from Video

Annotating a video stream to analyze player performance can be challenging. The objects (in this case, the players or the ball) are often small because the camera must record a large field of view in order to capture the full football match. This issue can be compounded by camera motion, which may further blur or distort the video feed. This movement, combined with changes in ball possession and direction as well as varying illumination on the pitch, can create challenges in performance analysis. For example, in a long-shot recording, the ball appears very small on the opposite side of the field, thus making detection and tracking difficult [10].

Camera motion isn't the only difficulty presented by recording technology. Other issues include differences in frame angles and the quality of shots, as rapid movements not only blur the players but can also cause differences in frame angles; in low-quality frames, then, detection is difficult or impossible [11]. Further complicating this is the size of the pitch. A sports video captures the entire playing area, including the players, goalkeepers, referees, and fans. Though this is important for capturing players' actions during the game, it can be challenging to distinguish players from other people in the video, such as referees or spectators, due to scaling issues, occlusion, and blur [12].

Other camera-based issues include the difficulty of identifying jersey numbers. High-quality videos may lessen this challenge, but a player's number may still be difficult to see in any given frame, based on their posture, for example. Identifying jersey numbers is important both when a player does and does not have possession of the ball, as every player should be detected, tracked, and identified in each frame for the best performance analysis. Though this data is useful, it is computationally expensive to identify every player in every frame when videos run at rates of more than 30 frames per second [13].

3 Related Works

3.1 Player and Ball Detection

Player detection is critical for analyzing athletes' performances. To obtain this data, an automatic analysis tool must detect both the players and the football. Current player detection methods based on deep learning are divided into two types: one-stage or two-stage detectors [14]. One-stage detectors, such as Yolo [15] and single-shot detector (SSD) [16], utilize a single convolutional neural network (CNN) that provides both the bounding boxes and the object classification at the same time. Yolo separates an image into grids, and then makes predications for each of those grids. Unfortunately, this type of detector may perform poorly when identifying objects in groups. SSD features three key improvements on Yolo, including its use of: convolutional filters and anchor offsets; pyramid features for predications; and defaults to account for objects' shapes.

Two-stage detectors, on the other hand, such as Faster R-CNN [17], break the classification task into two steps: first, they identify areas of interest; and second, they sort by these areas for regression and classification. These detectors are trained by region proposal networks, and have successfully used anchor

boxes for object detection [16,18,19]. Though these models are very accurate, they are often slow. Because of this, this paper employed Yolo as the base to detect players accurately and quickly.

Previous research has proposed different methods for ball detection. For example, Reno et al. [20] trained CNNs on a dataset with two types of images, ball and no ball, to identify whether a given video frame contains a football. Doing this allows for single-frame analysis without the use of background models, which helps prevent misclassification due to changes in lighting. Alternatively, Zhang et al. [21] used both a CNN and a Kalman filter to detect and track the ball in real time. In this method, the CNN detects the ball and then the Kalman filter is exploited to predict the location of the ball. Vats et al. [22] developed another model for detecting both the players and the ball simultaneously based on semi-supervised learning. Their proposed system utilized an iterative teacher-student method alongside three loss parametrizations. These parametrizations took the detections and the confidence scores from the teacher and passed them onto the student to ensure a higher accuracy.

3.2 Team Assignment

Automatically labelling players according to their team is another important task in football video analysis. However, it can be difficult without prior information about each team's visual appearance [23]. Early works on this subfield [11,24] employed color information as a discriminant feature to separate players into two different teams based on their jersey colors. Other works [25,26] employed color histograms. These methods, though easy to implement, did not account for occluded image frames, lighting issues, or teams wearing similar jersey colors.

Recent research has demonstrated, however, that deep learning algorithms are very successful at team assignment tasks. Lu et al.'s work [27] used a supervised approach, employing a cascaded CNN and a large, labelled dataset for team classification. Unfortunately, this method is not generalizable because it requires fine-tuning on labelled samples. Istasse et al. [28] designed a CNN with descriptors for team classification in which teammates' pixels are close in the embedding space. Pixel clusters can then help identify players' teams. Again, however, this method requires a large, labelled dataset for training. Alternatively, Koshkina et al. [8] used contrastive learning and a triple loss function to generate features that were then used to cluster the players to their teams. Contrastive learning, however, is prone to mode collapse, in which all data maps to the same representation, which can make this method ineffective [29]. The research in [30] proposed TBE-Net to identify the identity using different views.

3.3 Player Jersey Number Recognition

Evaluating players' performances from a video is difficult without first identifying the players. Early research on player identification relied on hand-engineered features [31,32]. Though it is possible to identify players by their bodies, doing so by jersey number instead is more popular because players' numbers do not change, and are typically observable, throughout the game [33]. Multiple works have utilized CNNs for this purpose [1,34]. Gerke et al. [1] were one of the first to do so, and in their research deep learning methods performed better than hand-engineered features. Other researchers, such as Li et al. [34], utilized a spatial transformer network (STN) to: identify the location of jersey numbers using a manually labelled quadrangle; annotate the area surrounding the jersey number; and use semi-supervised learning to train the network.

Other work has researched the following alternatives. Chan et al. [35] proposed a ResNet + LSTM framework that combined residual networks [36] and long short-term memory [37] to identify spatio-temporal jersey features and long-term dependencies. They utilized a second, 1D CNN classifier to augment the initial framework. Liu and Bhanu [3] used a Faster RCNN network for jersey number detection. It incorporated human pose keypoint supervision to increase its success. Liu and Bhanu [3]

also utilized a pose-guided R-CNN alongside human keypoint prediction. To do this, they used a network branch and a regressor to create digit proposals. Vats et al. [39] proposed a multi-task learning loss-based approach to identify jersey numbers from static images. Their loss function comprised both holistic loss representations; they treated the numbers as separate classes and also accounted for "digit-wise" loss, which treated each digit in a number separately. Another study [40] applied weakly supervised learning to speed up the training by producing estimated frame labels to better track jerseys and numbers. Alternatively, Zhang et al. [41] used a multi-camera set up to track and identify players via a coarse-to-fine method, utilizing deep representations of players' identities.

4 Proposed System

As illustrated in Fig. 1, the proposed system is composed of pitch detection, player detection, ball detection, team assignment, and player jersey number identification.

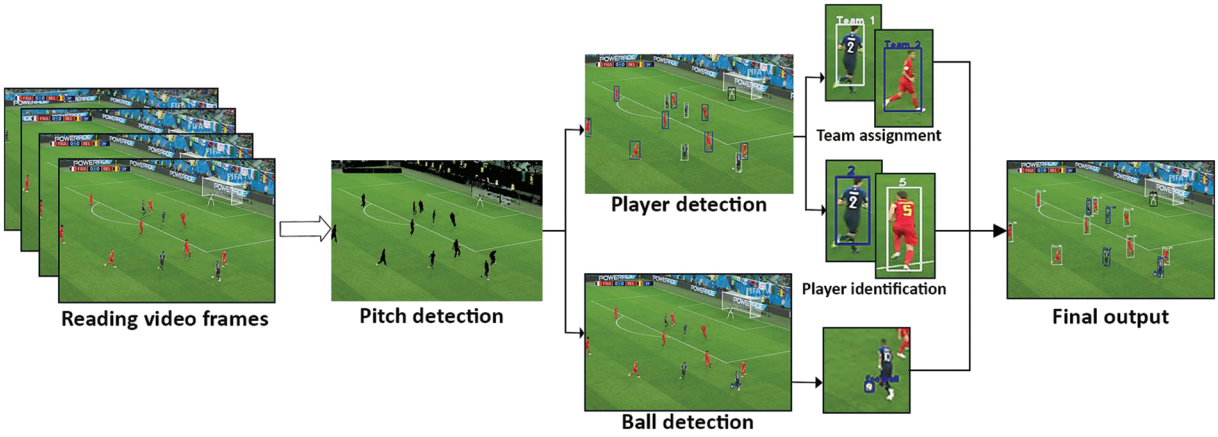

Figure 1: The architecture and workflow of the proposed system. The system includes the following phases: (1) The first phase is pitch detection, which segments the green (grass) region in each frame from surrounding areas, such as the stands. (2) Player and ball detection are then performed simultaneously. (3) After the detection phase, players are assigned to their corresponding teams. (4) Players' jersey numbers are identified. (5) Finally, all phases are combined in the last image (as the final output)

4.1 Football Pitch Detection

To best analyze football games, it is important to detect and track only the players on the pitch, not other people such as the spectators in the stands. Pitch segmentation can eliminate spectator regions and minimize false alarm errors in player detection. This segmentation is performed by selecting specific color ranges in the hue saturation value (HSV) color space, instead of in the red, green, and blue (RGB) color space. To do this, this paper proposed creating a mask for the color green. The green range was chosen empirically and was defined as a low bound $L = (36,\ 25,\ 25)$ and an upper bound $U = (86,\ 255,\ 255)$. A threshold was used with given upper (U) and lower (L) HSV boundaries to return a binary image, where areas that fall into the range limit appear white and areas that fall outside of that range are black. The proposed technique then constructed a mask to eliminate the area outside the pitch. The final output is the pitch. Fig. 2 illustrates this process.

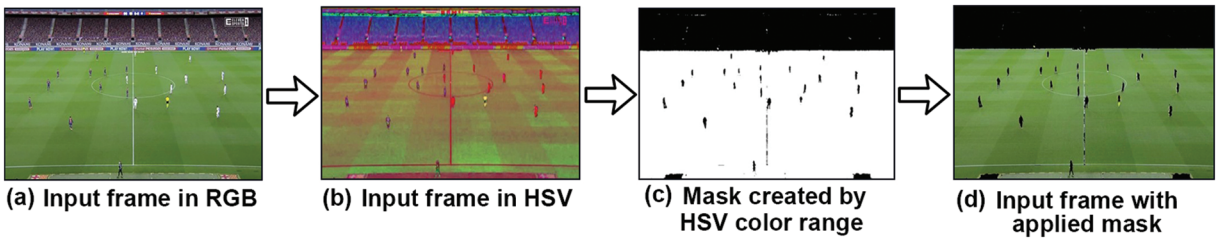

Figure 2: The process of pitch segmentation

4.2 Player Detection

Player detection requires the accurate processing of video frames, in real time. To complete this process automatically, computational complexity issues must first be mitigated. To do this, this research utilized the Yolov3 algorithm to perform player detections because it both outperforms different object detection algorithms and is very fast—it runs at 45 frames per second. The Yolo algorithm, introduced by Redmon et al. [15], is effective because it treats object detection as an issue of regression. By doing so, it yields both the location and class probabilities in a single step. The proposed Yolov3 model was trained on the Common Objects in Context (COCO) [42] dataset, which can detect 80 different everyday objects (e.g., persons, cars, cats, dogs, etc.). In every frame, the model recognizes the segmented pitch and detects any/all objects within it. When an object is recognized as a person, a bounding box is created and the player's team assignment and individual identity are both analyzed.

4.3 Team Assignment

To perform team assignment, recent research proposed the training of CNNs on labelled datasets. Unfortunately, this method has limited application because it necessitates regular fine-tuning, ideally before each game, in order to optimize the model's performance. By taking advantage of both unsupervised learning and unlabeled images, this paper utilized convolutional autoencoders (CAEs) instead to generalize this method to any football video, without relying on labelled data. The proposed method utilizes an unsupervised CAE to classify players via the extraction of a learned feature vector. Once the players have been detected, the system assigns each person to their corresponding team. To do this, this paper divided the bounding boxes into three groups: Team 1, Team 2, and others (a category that includes both goalkeepers and referees). As each team has a different color scheme and uniform, jersey color plays an important role in team assignment. This research used a CAE to extract useful representations from high-dimensional data (in this case, images). The CAE is a special type of CNN that replicates the input image into the output layer; it can then perform feature extractions on 2D input images.

CAEs consist of two components: an encoder and a decoder. The encoder contains convolutional and pooling layers that are trained to map the input x into a lower feature representation (z); this second image is comprised of the input image's content. The decoder, on the other hand, contains deconvolutional layers and an up-sampling layer that is trained to reconstruct the original images $\acute{x}$ from z. After being trained, researchers can utilize a CAE on any input to extract discriminative features so that the players on the same team map to the same feature space, while players on opposing teams map farther apart. The main advantage of unsupervised learning with CAEs is the method's generalizability: labelled data is not required, and the CAE needs minimal frames to correctly label each player.

Several types of conventional deep autoencoders—including sparse [43] and denoising [44] autoencoders—have been widely used to both reduce the dimensionality of inputs and learn robust feature representations via unsupervised pretraining. However, these techniques ignore the 2D image structure. The drawback to this method is its superfluous parameters, which then force the model to learn

global features. Compared to these conventional autoencoders, CAEs have been proposed as an alternative for image analysis tasks. CAEs have the same advantages as CNNs, which have demonstrated better performance on datasets with small changes, such as noise or translations. This is because the weights within the CAEs are shared, which maintains spatial locality [45]. They utilize reconstructions, which combine simple image patches via the latent space. In addition, max pooling is exploited to select the most important feature within a region in the convolved feature map. This facilitates improved learning [46].

Fig. 3 illustrates the overall architecture of this method. The CAE is a neural network that uses an encoder-decoder architecture. For the encoder, this paper used three convolutional layers, where the max-pooling layers were only implemented in the first two layers. For decoders, up-sampling layers and convolutional layers were added to return the data to its initial dimensions. The model was trained using Tiny ImageNet [47], without any label information. After training, the model could extract features for images from any football video. It used mean-square error (MSE) loss to minimize the difference between the original images and their reconstructions. MSE loss is defined as follows:

$$MSE = \frac{1}{2n} \sum_{i=1}^{n} (x_i - \acute{x}_i)^2 \tag{1}$$

where n represents the number of training images, x_i is the original image for image i, and $\acute{x}_i$ is the reconstructed image for image i.

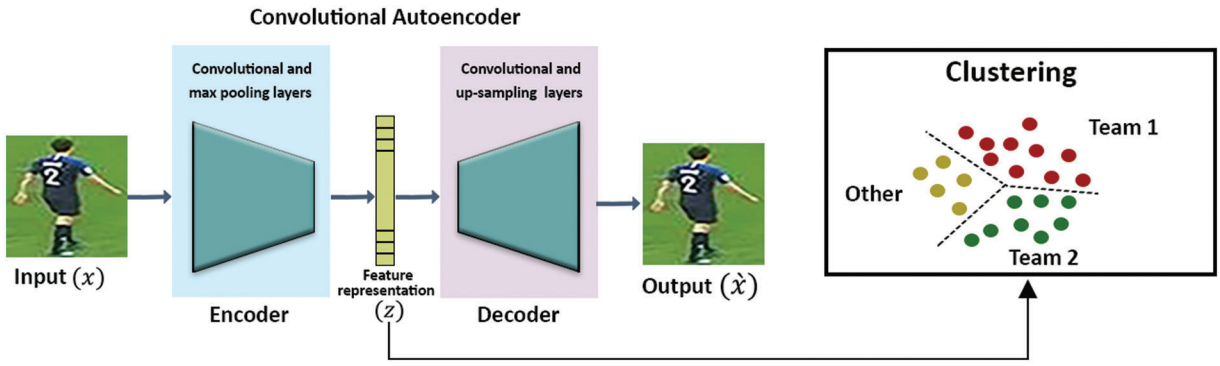

Figure 3: This approach's pipeline for classifying players, which contains both a CAE and K-means clustering

This research also applied unsupervised learning based on K-means to classify the teams based on their feature representations. *K*-means clustering [48] divides the data into clusters for classification. In this paper, after reading the video, 100 frames were chosen randomly and passed into the object detector to detect and retrieve all persons in the frames. Each was then resized to 64 × 64 pixels. These resized images were passed into the CAE to extract a feature presentation for each person. *K*-means was then used to estimate the cluster centers for all three categorization options: the two teams and the "other" category. To classify a new player, this research proposed computing the Euclidean distance between each feature representation and all three cluster centers, sorting the players into the category with the nearest center.

4.4 Ball Detection

Ball detection must be both efficient and accurate in order to successfully automate football video analyses. These qualities are especially important because the position of the ball provides the model with relevant information about players' actions. Ball detection is, therefore, important in collecting data about individual players' performances, such as the number of shots, passes, dribbles, and goals each player

attempts. However, detecting the ball in a video feed is very difficult because the ball is often small within the frame; furthermore, its size can vary greatly with respect to the proximity of the camera as well as the ball's speed and the possibility that it disappears from view, etc. To overcome these challenges, this research applied and modified a Yolov5 model to detect only the football. It then passed the colored images, whose sizes are all the same at 416×416, through the Yolov5 model. The output then estimated the ball's position and its confidence level on that position, which is presented alongside the bounding box. If there is no ball in the bounding box, the confidence value will be zero.

The training set for ball detection must be carefully designed to achieve a high detection rate. This paper proposed two steps to achieve high performance: (1) The proposed method used weights trained on the COCO dataset [42] as a starting point, which eliminated the need to train the model from scratch. This dataset has 123,000 images in the sports ball class, which means that these images' features are connected to training weights, thereby eliminating that step. (2) The model was trained and finetuned using a combination of footballs and cricket balls with different sizes to increase its accuracy.

4.5 Jersey Number Recognition

After assigning players to their corresponding teams, the model must identify each player in order to compute their individual statistics for any particular game. However, the automatic detection of jersey numbers is still challenging due to changing camera angles, low video resolution, the small size of jersey numbers in wide-range shots, and transient changes in a player's posture and movement. As a result, a player's jersey number may be difficult to see. To address these challenges, this research designed a deep neural network, similar to the Yolov5 model, that notes jersey numbers before attempting number recognition. The proposed model performs two tasks: detecting and locating the jersey number region, and then classifying the detected number.

The proposed model is shown in Fig. 4. This research first passed cropped player images from the player detection phase into the neural network for processing, as detailed below. The output first estimated the bounding box of the number, if it existed, and then the jersey number. The proposed model was composed of three major components, similar to Yolov5: the backbone, neck, and head models. First, the backbone utilized CSPDarknet to extract features from the input images, which each contained cross-stage partial networks. The neck then employed PANet, which created a feature pyramid network to aggregate the features and passed the data to the head model for predictions, which were obtained by analyzing the anchor boxes from the object detection process. To train and test the proposed model, this paper built a brand-new dataset with more than 6,500 images, which contained a total of 100 jersey numbers, ranging from 0 to 99. Since achieving a high performance requires a large training set, and the collected dataset was not enough, this research utilized transfer learning (TL) to pretrain the model on other datasets to improve its performance.

5 Experimental Results and Discussion

5.1 Team Assignment Evaluation

5.1.1 Datasets and Setup

To generalize this model, and to be able to extract discriminative features from any football video, this research trained a CAE on the Tiny ImageNet [47] dataset, which consists of 100,000 colored images with a size of 64×64 pixels. The CAE was implemented in TensorFlow and trained using an Adam optimizer with a learning rate of 0.001. The training ran for 355 total epochs, and the batch size was set to 64 images. MSE loss was then used to minimize the difference between the original images and their reconstructions.

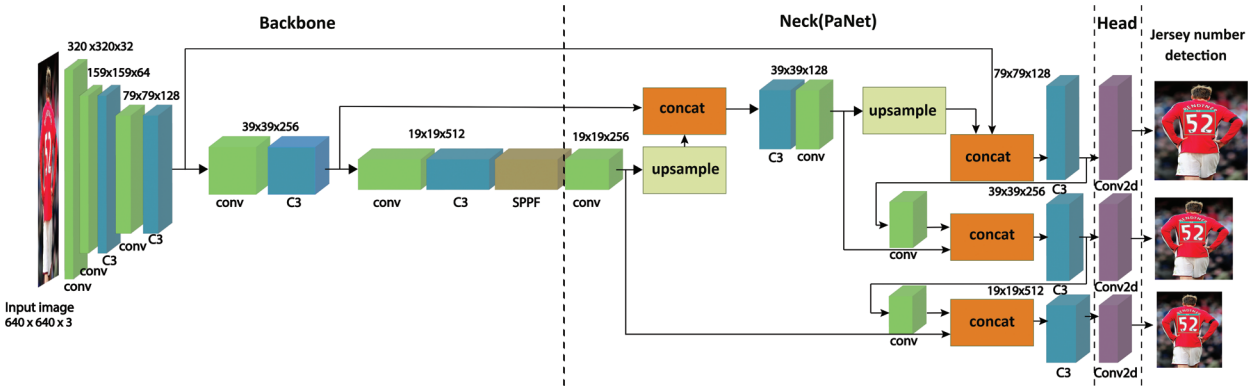

Figure 4: The proposed model's architecture for identifying jersey numbers. It combines number detection and classification into a single step

5.1.2 Results

For evaluation only, this research manually annotated the players in 100 randomly selected frames. Detected people were classified into: team 1, team 2, and other. The proposed model was then compared to other clustering methods, such as agglomerative [49] and BIRCH [50] methods in the embedding space. This comparison aims to identify the impact of utilizing CAEs in the image encoding process. The agglomerative method treats each object as a single cluster and then merges that cluster with the next object cluster based on the similarity between them; this process continues until all objects are clustered. The second method tested was BIRCH [50], which is another hierarchical clustering method. BIRCH grows dynamically in multiple dimensions, based on the similarity between points, to produce the best clusters.

Table 1 summarizes the proposed method's team recognition evaluation results, including the accuracy rates, which are a standard metric in the assessment of team recognition approaches. The proposed method obtained the best accuracy, with an accuracy rate of 92%. In addition, this method performed better than those assigning players to their teams based on the closest cluster center and visual attributes like jersey color. Using color information with *K*-means achieved an overall accuracy rate of 86%. This research also verified that the proposed method of team assignment works well in real-world football videos, as shown in Fig. 5, which demonstrated that the proposed method achieved impressive results in labeling players to their teams.

Table 1: Accuracy via different methods of team assignment

Method	Accuracy (%)
CAE + Agglomerative	83%
CAE + BIRCH	83%
CAE + *K*-means (ours)	92%
Color space + *K*-means [23]	86%

5.2 Ball Detection Evaluation

5.2.1 Datasets and Setup

To detect the ball in the video stream, this paper modified the output layer of the Yolov5 model to detect only footballs. The modified Yolov5 was implemented in PyTorch. A stochastic gradient descent (SGD)

optimizer was employed to minimize the loss function with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0005. This research performed training phase runs for 100 epochs with a batch size of 64 images. In object detection systems, the dataset is very important to the success of the model. The dataset is comprised of both images and their labels. The labels contain references to the appropriate bounding boxes, including their locations (x, y), widths, heights, and labels. The modified model is first pre-trained on the COCO dataset. The pretrained model is then utilized to set the starting model weights. Upon completion, we trained and fine-tuned the model on the target dataset, which contained a total of 869 football images to help the model learn and detect the ball under various conditions.

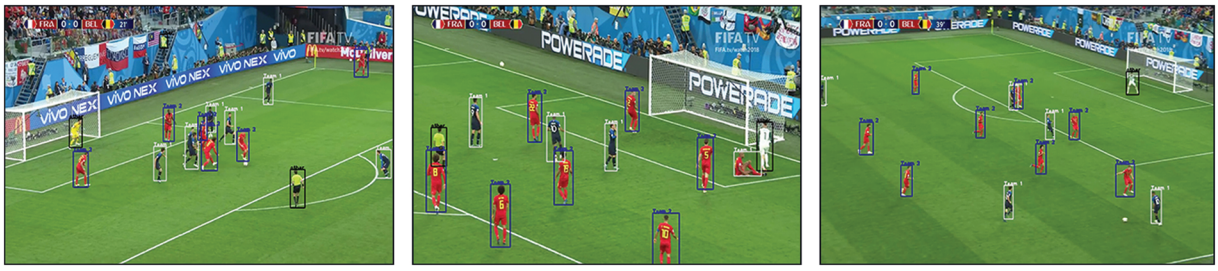

Figure 5: The results of the proposed method for team recognition. Notice that only the people on the pitch are detected and classified

5.2.2 Results

To evaluate the proposed model, this research compared the performance of the proposed model against that of a Yolov5 model trained from scratch with random initial weights. Note that this Yolov5 system was trained on the COCO [42] dataset, which contained football classes. This paper tested the proposed method for ball detection on 100 images collected from football video frames and the Internet. The images contained balls of different sizes and shapes under different conditions. These images were annotated in Yolo format, which were set manually using LabelImg. The models' performances were measured using mean Average Precision (mAP), which is an industry stand in the object detection field. The evaluation metrics were then presented in the form of Recall and F1 scores. The results shown in Table 2 indicate that the proposed model had higher Recall, F1, and mAP scores, in comparison to the original Yolov5 model, with one difference: the new model achieved a 99% detection rate using mAP; to the model that trained Yolov5 from scratch achieved a detection rate of only 80%. These results verified the effectiveness of the proposed method. They showed the benefits of utilizing transfer learning with limited data to achieve improved performance metrics. Fig. 6 shows some example images; the location of the ball has been marked exactly, and the model only detected the football.

Table 2: The results of the test sets in terms of mean average precision (mAP), recall, and *F1* scores computed for ball detection

Model	mAP	Recall	*F1* score
Pretrained Yolov5 + TL (ours)	99%	98%	99%
Yolov5 trained on COCO from scratch	80%	67%	76%

Figure 6: The results of the proposed football detection model. Note that the bounding boxes surround only the football in each image

5.3 Jersey Number Recognition Evaluation

5.3.1 Datasets and Setup

Because there are limited public datasets for jersey number identification, this research created its own, with more than 6,500 images. These images were collected from: open-source websites on the internet, data captured from soccer videos, and sample data from the popular Street View House Numbers (SVHN) dataset for detecting and classifying numbers. These images are annotated for numbers from 0 to 99, in Yolo format, as follows: class label, *x* center, *y* center, width, and height, all labelled manually using LabelImg. To evaluate the proposed method, this research compared this model with the model that currently achieves the best performance on this task, as proposed by [1]. To better understand the proposed method in context with other techniques, this research also compared our model against various CNN architectures, such as VGG16 [51] and MobileNetV2 [52], which have all achieved high accuracies on the ImageNet dataset challenge in the past.

The original Yolov5 model was pretrained on 80 classes using the COCO [42] dataset. Next, the output layer was removed and replaced with another softmax layer, which contained 100 neurons, representing the probability distribution of the 100 jersey number classes from 0 to 99. By reproducing the dataset in this way, this research could use the same training settings for the proposed model without access to the original dataset. In order to train the CNN [1], VGG16 [51], and MobileNetV2 [52] models, all players in the 6,500-image dataset were cropped and labelled according to their jersey numbers for manual classification. The dataset was then divided into three parts: 70% of the data for training, 10% for validation, and 20% for testing. The proposed model was pre-trained on the ImageNet dataset and then fine-tuned on the proposed jersey number dataset to create a baseline performance. The hyperparameter settings for each model are provided in Table 3.

Table 3: The hyperparameter settings for jersey number identification across different models

Hyperparameter	Modified Yolov5	MobileNetV2 [52]	VGG16 [51]	Gerke et al. model [1]
# of Epochs	100	100	100	100
Optimizer	SGD	Adam	Adam	Adam
Learning rate	1e-2	1e-3	1e-3	1e-3
Batch size	32	32	32	32
Framework	PyTorch	TensorFlow	TensorFlow	TensorFlow

5.3.2 Results

For consistency, every model was trained and then tested on identical data. Table 4 shows the performance of the MobileNetV2 [52], VGG16 [51], and Gerke et al. models [1] in comparison to the proposed model, in terms of accuracy on the test dataset. As shown in Table 4, the proposed model achieved the highest accuracy, at 99%. Conversely, the VGG16 model obtained the lowest accuracy (at 42%), likely because VGG16 is better suited for more complex problems. This research also evaluated an optical character recognition (OCR) method to identify jersey numbers. This technique achieved an accuracy rate of only 36%, the lowest tested. The advantage of the proposed model is that it guarantees that an image contains a number before moving into classification. By doing this, the proposed model detects numbers and recognizes them simultaneously, for faster results. Other models only recognize numbers.

Table 4: A comparison of results among approaches. The proposed model achieved the best accuracy

Model	Accuracy (%)
MobileNetV2 [52]	93%
Gerke et al. [1]	77%
VGG-16 [51]	42%
Proposed model	99%

As shown in Fig. 7, the proposed model recognized jersey numbers correctly under varying conditions. Moreover, this research examined the proposed model on football video frames, as shown in Fig. 8. The proposed model was most effective at identifying jersey numbers. In addition, this research found that training deep learning models to detect and classify jersey numbers simultaneously was better than training the models to learn detection and classification separately.

Figure 7: Examples of correctly labeled jersey numbers from the proposed model

Finally, once all phases were performed, this paper combined the results into one image, as shown in Fig. 8. White boxes represent team 1 (T1) players, blue boxes represent team 2 (T2) players, and black boxes represent goalkeepers and referees (other). The text in the boxes displays the results of both

automatic team assignment and jersey number recognition. If the player jersey was invisible, the bounding box would display “unknown” (UN) instead. Blue boxes can also represent the football.

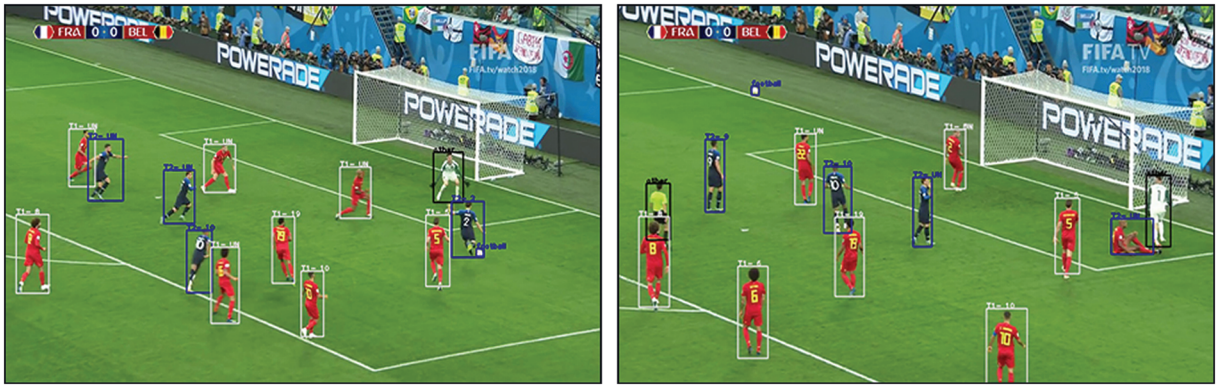

Figure 8: All phases have been combined into one output image. These images show the proposed system’s results for: player identification, football detection, team assignment, and player jersey number recognition

In building this system, the researchers learned a variety of lessons. For one, this paper identified that jersey number recognition was the most challenging task because players’ numbers were often occluded. This was particularly true when a player was passing the ball or otherwise moving. Even when jersey numbers were visible, they were often partially occluded or otherwise difficult to identify, which poses a second challenge for automatic video analysis systems. Future improvements should focus on these difficulties. Further research should also improve the tracking of multiple players at once, as this is critical in evaluating individual players’ performances. Doing so will become easier with a system that can identify players’ locations at consistent intervals, as this data can then be used to identify players when their jersey numbers are occluded. This automatic data analysis is most helpful in sports where players’ positions, relative to the ball’s position, are critical for better game strategy.

6 Conclusion and Future Work

Football analysis requires the use of automatic tools for the best, and fastest, analysis and statistics. This paper presented one such tool for evaluating football videos, inspired by recent progress in deep learning and computer vision methods. All of these stages are sequential and interconnected, as well as extremely important for player-based analysis. This research demonstrated that the proposed approaches were effective in detecting the ball, assigning players to their teams, and identifying player numbers. These steps are critical to the design and development of automatic performance analysis. The proposed datasets and system implementations are publicly available for the benefit of future research.

Future work in this area should focus on tracking multiple players, estimating players’ positions, and action recognition. Doing this will allow researchers to design a fully automatic system that can analyze the skills of the individual players as well as team cooperation. The first iteration of this system will provide general information about the match, such as the number of passes, dribbles, interceptions, tackles, and shots. Then, as the system is further developed, it will eventually offer detailed reports about each player.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Gerke, K. Muller and R. Schafer, "Soccer jersey number recognition using convolutional neural networks," in *Proc. of the IEEE Int. Conf. on Computer Vision Workshops*, Santiago, Chile, pp. 734–741, 2015.

[2] A. M. Lopes and J. Tenreiro Machado, "Entropy analysis of soccer dynamics," *Entropy*, vol. 21, no. 2, pp. 187, 2019.

[3] H. Liu and B. Bhanu, "Pose-guided R-CNN for jersey number recognition in sports," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, pp. 2457–2466, 2019.

[4] R. Theagarajan, F. Pala, X. Zhang and B. Bhanu, "Soccer: Who has the ball? Generating visual analytics and player statistics," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, pp. 1749–1757, 2018.

[5] J. Komorowski, G. Kurzejamski and G. Sarwas, "Footandball: Integrated player and ball detector," arXiv Preprint arXiv:1912.05445, 2019.

[6] D. Bhargavi, E. P. Coyotl and S. Gholami, "Knock, knock. Who's there?–Identifying football player jersey numbers with synthetic data," arXiv Preprint arXiv:2203.00734, 2022.

[7] Y. Ohno, J. Miura and Y. Shirai, "Tracking players and estimation of the 3D position of a ball in soccer games," in *Proc. 15th Int. Conf. on Pattern Recognition, ICPR-2000*, Barcelona, Spain, IEEE, vol. 1, pp. 145–148, 2000.

[8] M. Koshkina, H. Pidaparthy and J. H. Elder, "Contrastive learning for sports video: Unsupervised player classification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 4528–4536, 2021.

[9] R. R. Nadikattu, "Implementation of new ways of artificial intelligence in sports," *Journal of Xidian University*, vol. 14, no. 5, pp. 5983–5997, 2020.

[10] J. Theiner, W. Gritz, E. Müller-Budack, R. Rein, D. Memmert and R. Ewerth, "Extraction of positional player data from broadcast soccer videos," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 823–833, 2022.

[11] P. L. Mazzeo, P. Spagnolo, M. Leo and T. D'Orazio, "Football players classification in a multi-camera environment," in *Int. Conf. on Advanced Concepts for Intelligent Vision Systems*, Berlin, Heidelberg, Springer, vol. 6475, pp. 143–154, 2010.

[12] A. Hiemann, T. Kautz, T. Zottmann and M. Hlawitschka, "Enhancement of speed and accuracy trade-off for sports ball detection in videos—Finding fast moving, small objects in real time," *Sensors*, vol. 21, no. 9, pp. 3214, 2021.

[13] F. Wu, Q. Wang, J. Bian, H. Xiong, N. Ding *et al.,* "A survey on video action recognition in sports: Datasets, methods and applications," arXiv Preprint arXiv:2206.01038, 2022.

[14] M. Carranza-García, J. Torres-Mateo, P. Lara-Benítez and J. García-Gutiérrez, "On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data," *Remote Sensing*, vol. 13, no. 1, pp. 89, 2021.

[15] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 779–788, 2016.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," *Computer Vision–ECCV 2016*, vol. 9905, pp. 21–37, 2016.

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, MIT Press, Montreal, Canada, vol. 39, pp. 91–99, 2015.

[18] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1134–1142, 2015.

[19] J. Dai, Y. Li, K. He and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, Curran Associates INC, Barcelona, Spain, pp. 379–387, 2016.

[20] V. Reno, N. Mosca, R. Marani, M. Nitti, T. D'Orazio *et al.,* "Convolutional neural networks based ball detection in tennis games," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, pp. 1758–1764, 2018.

[21] T. Zhang, X. Zhang, Y. Yang, Z. Wang and G. Wang, "Efficient golf ball detection and tracking based on convolutional neural networks and kalman filter," arXiv Preprint arXiv:2012.09393, 2020.

[22] K. Vats, W. McNally, P. Walters, D. A. Clausi and J. S. Zelek, "Ice Hockey Player Identification via Transformers and Weakly Supervised Learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 3451–3460, 2022. https://doi.org/10.1109/CVPRW56347.2022.00389

[23] M. Istasse, J. Moreau, and C. D. Vleeschouwer, "Associative embedding for game-agnostic team discrimination," arXiv, vol. abs/1907.01058, 2019.

[24] Z. Ivankovic, M. Rackovic and M. Ivkovic, "Automatic player position detection in basketball games," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2741–2767, 2014.

[25] X. Tong, J. Liu, T. Wang and Y. Zhang, "Automatic player labeling, tracking and field registration and trajectory mapping in broadcast soccer video," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 2, pp. 1–32, 2011.

[26] W. -L. Lu, J. -A. Ting, J. J. Little and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.

[27] K. Lu, J. Chen, J. J. Little and H. He, "Lightweight convolutional neural networks for player detection and classification," *Computer Vision and Image Understanding*, vol. 172, pp. 77–87, 2018.

[28] M. Istasse, J. Moreau and C. De Vleeschouwer, "Associative embedding for team discrimination," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, pp. 2477–2486, 2019.

[29] S. Albelwi, "Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging," *Entropy*, vol. 24, no. 4, pp. 551, 2022.

[30] W. Sun, G. Dai, X. Zhang, X. He and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14557–14569, 2022. https://doi.org/10.1109/TITS.2021.3130403

[31] W. -L. Lu, J. -A. Ting, K. P. Murphy and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *The Conf. on Computer Vision and Pattern Recognition 2011*, Colorado Springs, CO, USA, pp. 3249–3256, 2011.

[32] J. Poignant, L. Besacier, G. Quénot and F. Thollard, "From text detection in videos to person identification," in *2012 IEEE Int. Conf. on Multimedia and Expo*, Melbourne, VIC, Australia, pp. 854–859, 2012.

[33] K. Vats, P. Walters, M. Fani, D. A. Clausi and J. Zelek, "Player tracking and identification in ice hockey," arXiv Preprint arXiv:2110.03090, 2021.

[34] G. Li, S. Xu, X. Liu, L. Li and C. Wang, "Jersey number recognition with semi-supervised spatial transformer network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, pp. 1783–1790, 2018.

[35] A. Chan, M. D. Levine and M. Javan, "Player identification in hockey broadcast videos," *Expert Systems with Applications*, vol. 165, pp. 113891, 2021.

[36] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas Nevada, pp. 770–778, 2016.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] H. Liu and B. Bhanu, "Pose-guided R-CNN for jersey number recognition in sports," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2019.

[39] K. Vats, M. Fani, D. A. Clausi and J. Zelek, "Multi-task learning for jersey number recognition in ice hockey," in *Proc. of the 4th Int. Workshop on Multimedia Content Analysis in Sports*, New York, NY, USA, pp. 11–15, 2021.

[40] R. Vandeghen, A. Cioppa and M. Van Droogenbroeck, "Semi-supervised training to improve player and ball detection in soccer," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 3481–3490, 2022. https://doi.org/10.1109/CVPRW56347.2022.00392

[41] R. Zhang, L. Wu, Y. Yang, W. Wu, Y. Chen *et al.,* "Multi-camera multi-player tracking with deep player identification in sports video," *Pattern Recognition*, vol. 102, pp. 107260, 2020.

[42] T. -Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona *et al.,* "Microsoft COCO: Common objects in context," *Computer Vision—ECCV 2014*, vol. 8693, pp. 740–755, 2014.

[43] A. Ng, "Sparse autoencoder," *CS294A Lecture Notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[44] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. -A. Manzagol *et al.,* "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. 12, pp. 3371–3408, 2010.

[45] E. Blanco-Mallo, B. Remeseiro, V. Bolón-Canedo and A. Alonso-Betanzos, "On the effectiveness of convolutional autoencoders on image-based personalized recommender systems," *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 54, no. 1, pp. 11, 2020.

[46] Z. Cheng, H. Sun, M. Takeuchi and J. Katto, "Deep convolutional autoencoder-based lossy image compression," in *2018 Picture Coding Symp. (PCS)*, San Francisco, CA, USA, IEEE, pp. 253–257, 2018.

[47] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, pp. 3, 2015.

[48] J. MacQueen, "Classification and analysis of multivariate observations," in *Proc. of the Fifth Berkeley Symp. on Mathematical Statistics and Probability*, Oakland, CA, USA, pp. 281–297, 1967.

[49] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," arXiv Preprint arXiv:1109.2378, 2011.

[50] T. Zhang, R. Ramakrishnan and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM Sigmod Record*, vol. 25, no. 2, pp. 103–114, 1996.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv Preprint arXiv:1409.1556, 2014.

[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4510–4520, 2018.