Shane Rodricks and Ivan Orlovic
CS 615
Chapman University

Futbol/Soccer Formation Detection Using Digital Image Processing Techniques

Abstract:

Introduction:

The inspiration for this project came about as soccer (known as futbol/football across the world) is the world's most popular sport, with 3.5 billion fans as compared to its American football counterpart, which only has 410 million fans. This contrast in popularity is also evident when comparing the two major events in each sport: the World Cup Final and the Super Bowl. As reported by the Federation Internationale de Football Association (FIFA), the 2022 World Cup amassed a viewership of 1.5 billion viewers, while the National Football Association (NFL) reported that the event was watched by 113 million viewers. This key difference in popularity was one of the reasons why we chose to take on the challenge of applying Digital Image Processing and Machine Learning techniques to soccer.

The history of soccer as a sport dates back to 12th-century England, but it was not until 1848, when the widely accepted Cambridge rules were adopted, that soccer became the global sport we know today. One of the key factors that make the sport as popular as it is on a global level, with it being played in over 230 countries, is the simplicity of it. All anyone needs to play is a ball. In a more formal sense, the rules are simple: one grass pitch, two teams, 11 players on each team, two forty-five minutes halves with any additional time being added by the referee, up to five substitutions per squad, and disciplinary warning being issued by the referee such as a yellow card (warning) and red card (elimination). For simplicity purposes, we only named these rules are they will be crucial for our study.

The goal of this study is to use digital image processing techniques in order to be able to identify a team's formation and then employ data science and machine learning techniques to suggest formations for teams based on their performance. As this project is based on digital image processing, the majority of our study will focus on the image processing we used to achieve our goal, but the data science portion is also important to note as we will work to combine both disciplines as this has the great real-world and industry applications. Specifically, for the digital image processing, we will be using techniques such as image loading, noise reduction/removal, masking and filtering, marking and centroid detection, sectioning and counting, as well as thresholding. The data science portion will focus on starting formation, and the end result of the matches played by the team.

Methodology:

Digital Image Processing:

The study begins with us applying digital image processing techniques in order to process team lineups. In order for us to achieve this, we created six bird's eye view images that resembled team formation templates that were designed in order to replicate the six main base formations used in soccer. These base formations include: 3-5-2, 3-4-3, 4-4-2, 4-5-1, 4-3-3, 5-3-2. Take note of how each formation adds up to ten instead of eleven to match the number of players on the pitch. This is because it is common to omit the goalkeeper in position tactics, as each team is required to have a goalkeeper; therefore, the one missing is inferred as the goalkeeper. In addition to this, reading the formation goes from left to right in order of positioning on the field. The leftmost section is defenders, the middle is midfielders, and the rightmost section is attacking players. With this being noted, we can now go into the specific image processing techniques we incorporated to complete our study.

Beginning with image loading, processing, and noise reduction/removal. We begin by loading our image sample and processing the background of the image. As soccer is played on green grass pitches, we used the built-in MATLAB function of rgb2hsv in order to facilitate better color segmentation and define the hsv (hue, saturation, and value) ranges specific to green in order to identify the background of our images. We then created a mask for the green hsv components and non-green hsv components, which we combined in order to exclude the green color from the background of our images.

The next step in our digital image processing was masking and filtering to remove noise and fill gaps that may have occurred due to our initial steps. We begin this step by grayscaling the image and then applying a threshold to the image. Specifically, we used Otsu's method for applying a threshold, which includes using the graythresh() function that produces a threshold that is normalized to the range of [0,1] and then converts the image into a binary image using imbinarize() and the threshold value returned. In order to clean up this binary mask, we need to remove noise and fill in gaps that may have occurred. This is all for the purpose of making the player region more clear for further analysis and identification. In order to accomplish this, we employed the functions imopen() and strel(), which remove noise via erosion and dilation of the area. In addition to this, we also employed the functions imclose() and strel() to fill in the holes in our player region, making it a single-player unit. The strel() function we used creates a structuring element, a disk in our case. We displayed our results, and it returned us with a new image that just included the players with a black background.

After this has been accomplished, we go into the code phase, where we look at detecting centroids, which, in our case, are individual players. In order to accomplish the results we were looking for, we found a MATLAB function called regionprops(), which is often used to calculate centroids and superimpose locations on an image. A common purpose of this is to identify letters in text. In our case, the function was able to identify player locations, and we were able to visualize them by producing an image that built off of our last two steps that displayed the players, marked with a centroid and a black background.

Once this step has been completed and we have the player locations, it is time to section our image into four segments to replicate the four regions of the pitch. The four sections can be thought of as the goalkeeper, defender, midfielder, and attacker sections. We do this by taking in the image width and dividing it by four equal parts. Then, we iterate through each of the sections and count how many centroids are detected in each section. In order to visualize these steps, we create three lines that section the image into four parts, display the centroid count for each section, and print out the formation detected in the image.

Data Science:

The data science portion of this study was designed to be a complement to the digital image processing portion and offer an extended amount of information about team tactics. Since the first part of our study on identifying team formations, this portion will study the performance of those team formations. We focused on six key features: most common formations by year, formations (home and away) and the average goals scored, formations (home and away) and the average possession time, formations (home and away) and the total shots, and looking at individual teams to see if their formation changes when they play home compared to away.

Gathering these metrics was quite simple, as we found a data base on Kaggle which supplied us with soccer match data from the top five european leagues and top 2 south american leagues from 2015 to 2023. The data set includes over 27,000 match reports and includes 228 features. The features that we found important to focus on were: ['Competition_Name', 'Season_End_Year', 'Wk', 'Round', 'Home', 'Away', 'HomeGoals', 'AwayGoals', 'manager_home', 'manager_away', 'formation_home', 'formation_away', 'possessiontime_home', 'possessiontime_away', 'shots_total_home', 'shots_total_away', 'yellow_cards_home', 'red_cards_home', 'yellow_cards_away', 'red_cards_away',].

The steps that we took in order to extract the relevant information was as follows: download the data set, load in the data, create a new data frame with the features that we selected, and then begin to calculate and visualize the metrics that we wanted. The common functions that we used were: mean(), sort_values(), and groupby().

We decided to separate home and away statistics, as in soccer the stadium atmosphere is something that should be taken into account as it may affect how a team sets up their formation in order to play. We gathered the 10 most frequent teams in our data, counted and analyzed their formations depending on if they were playing at home or away.

All in all, the data science portion of this study was used to provide us greater insight into the formations used by teams, and help up evaluate important metrics such as goal, shots, and possession time based on these formations.

Results:

Digital Image Processing:

Beginning with the digital image processing results, we used the script that we created to process six of our created base formations: 3-5-2 (figure 1), 3-4-3 (figure 2), 4-4-2 (figure 3), 4-5-1 (figure 4), 4-3-3 (figure 5), 5-3-2 (figure 6).

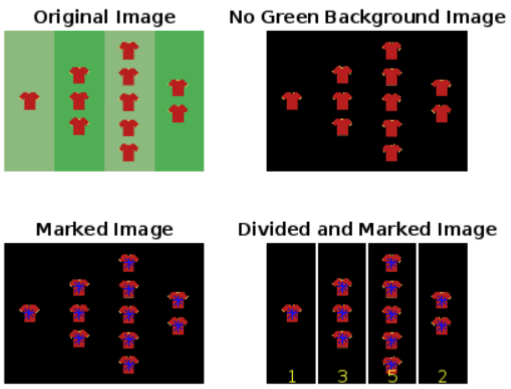


Figure 1

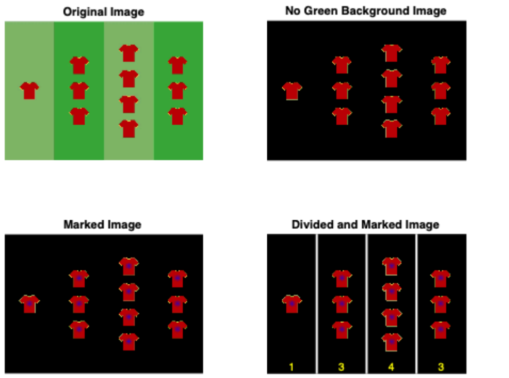


Figure 2

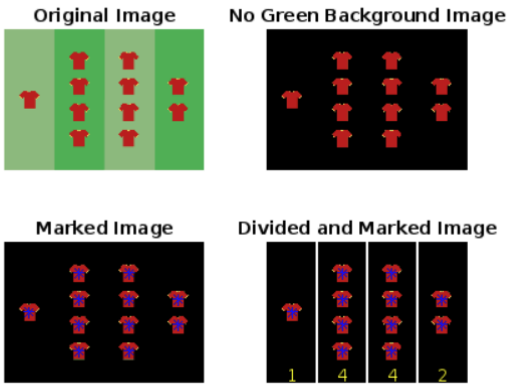


Figure 3

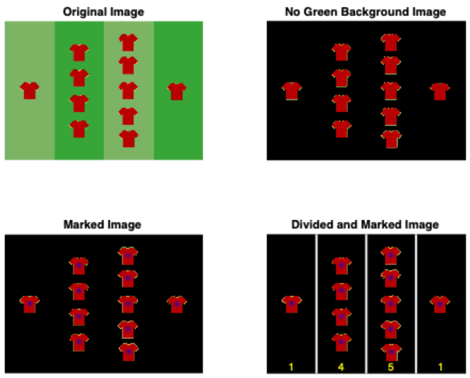


Figure 4

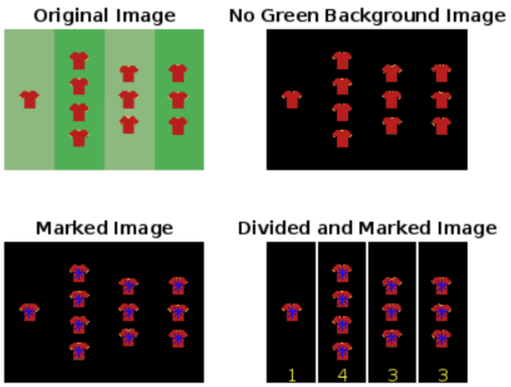


Figure 5

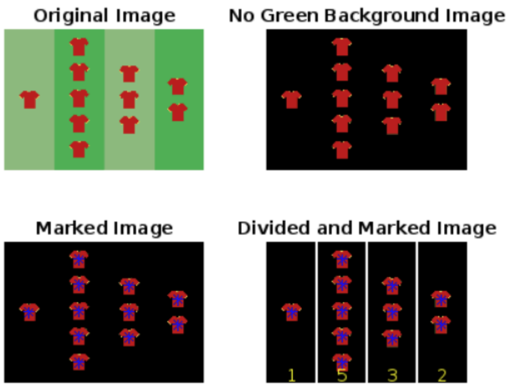
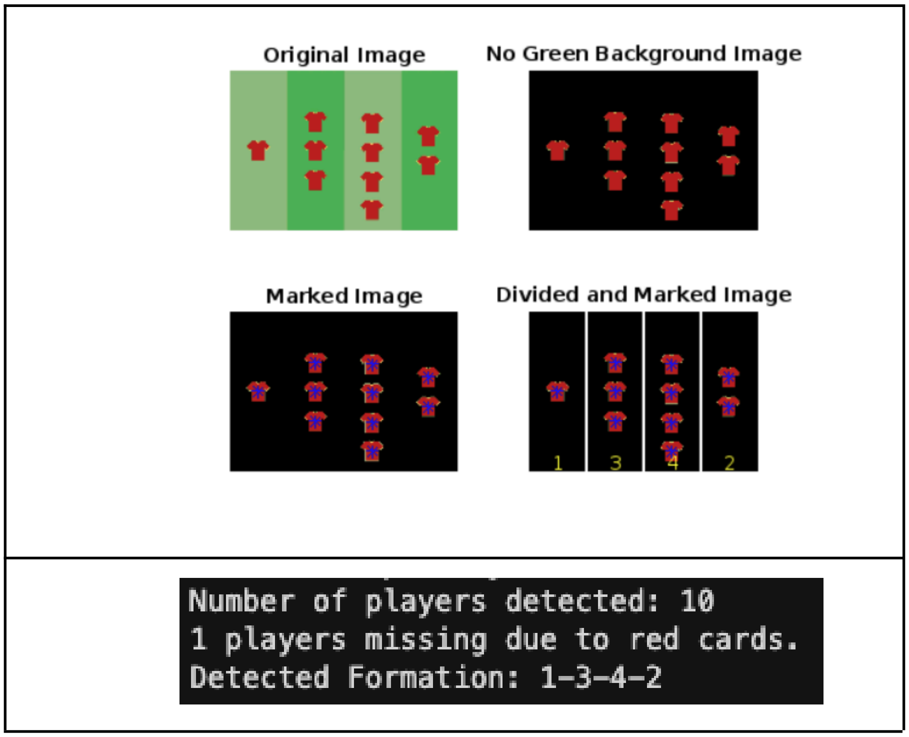


Figure 6

From our results, we can see that all of our formations were correctly identified. The script that we created was able to generate the four images that we wanted. Beginning with our original image, a no background image which shows that we were able to successfully remove the green grass background from our images, a marked image which shows that our centroid detection was working, and a final divided and marked image which was able to divide the image into four parts and then count the number of centroids in each part.

Our next challenge was being able to identify if a team was playing with a man down, which would mean that one of their players received a red card (a disqualification/sending-off). This process would work by adding an additional step to the end of our code which counted all of the centroids and determined if there were less than 11, then there must have been a sending-off. We loaded an image and obtained the results (figure 7). The results, were as expected that our image processing script was able to identify the players (and lack there of players) in the image we presented.



```
Number of players detected: 10
1 players missing due to red cards.
Detected Formation: 1-3-4-2
```

Figure 7

With this success in results, we decided to test our processing on an actual in game photo of a starting lineup. Unfortunately, due to copy rights we were unable to attain adequate photos for processing, but were able to find an image to use as a test. The results of the image were not what we were looking for as it was not able to detect an players in the image. This could have been due to the angle of the image we used, as games are filmed from a moving angle which requires much work on the part of editing the image to stretch it to be "straight". Building off of this issue, the division of the pitch also seemed to be a problem as the image we could find was looking left at the players from a center position and we were not able to adequately move the image in order to divide it as we did in the tests prior. Figure 8 shows our negative results.

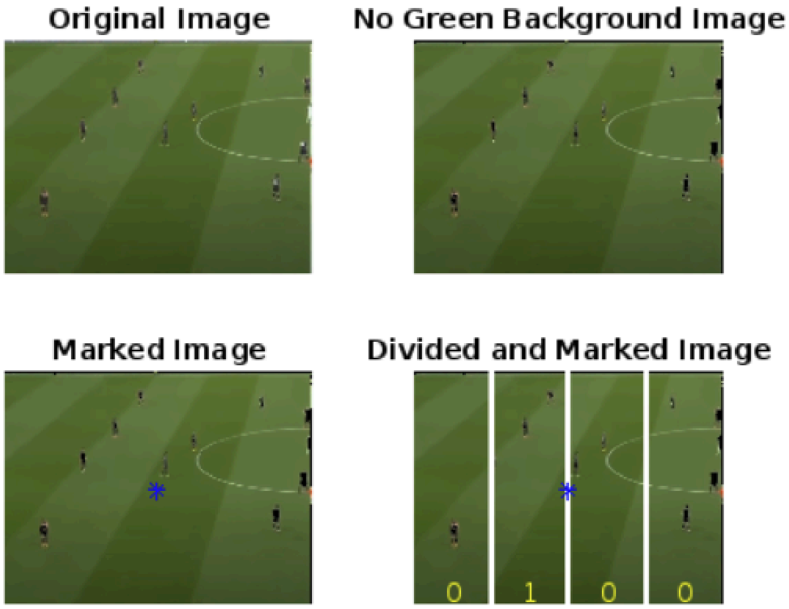


Figure 8

Data Science:

The results from the data science portion of our results were interesting and took some interpretation from our part. Luckily, we think that it is safe to assume that we have expert knowledge in this subject as both of us played soccer at a club level for 10 years. Shane's club team won the U17 Western Conference championship in Seattle hosted teams from six states, and his high school team won their open division conference championship. Ivan's club team was ranked top 3 in the nation for five consecutive years in their age group, and in addition, he played at the academy level for two years at the u19 level.

As mentioned before, focused on six key features: most common formations by year, formations (home and away) and the average goals scored, formations (home and away) and the average possession time, formations (home and away) and the total shots, and analyzing how a team's starting formation changes if they play at home compared to away.

Beginning with the most common formation by year (figure 9), we found that in all the years from 2015 to 2023, the 4-2-3-1 formation was the most common, which is an adaptation of the 4-5-1 base. With this formation, the team can have five midfielders on the pitch with a lone striker. As the game has become more and more athletic, the outside midfielders and outside backs play the entire length of the pitch, which allows teams to attack with more numbers. We noticed in our results that the 4-2-3-1 formation was most popular in 2017 and then slowly decayed in popularity until 2019, then had a resurgence until 2022, and then once again fell off as a popular formation. The second most popular formation was the 4-3–3, which grew in popularity from 2020 to 2023. This formation is a bit more offensive than the 4-2-3-1 as it has less midfielders on the pitch and primarily focuses on keeping possession in the opposing team's half as more attackers are in the game.

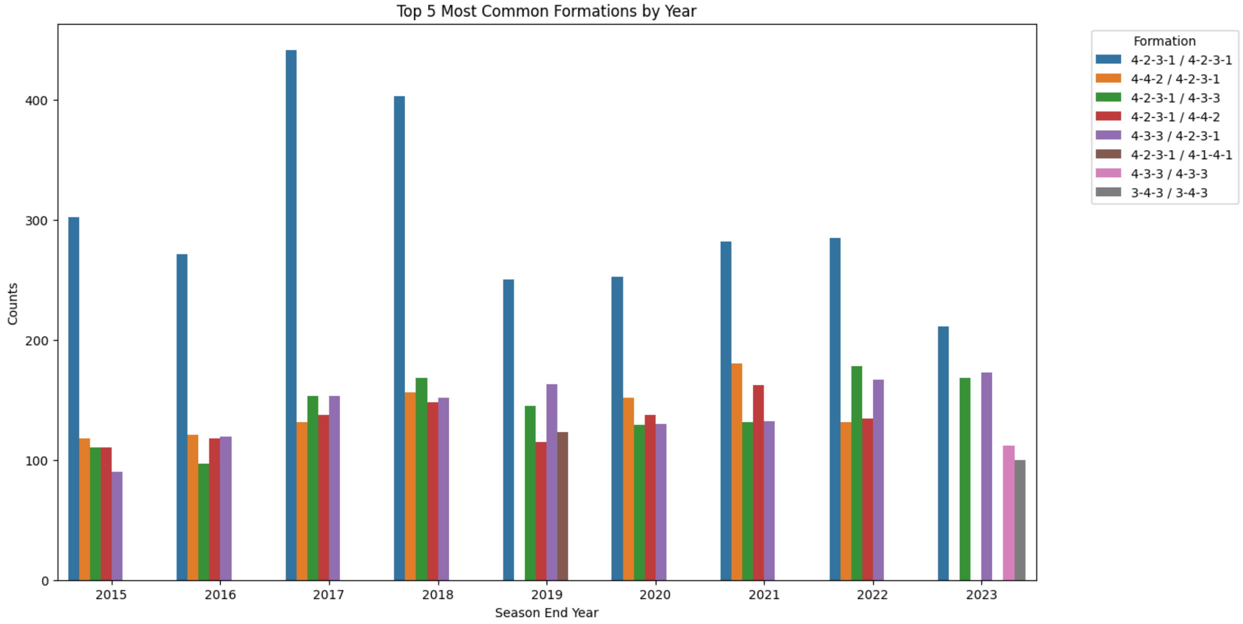


Figure 9

Our next metric that we extracted was formations (home and away) and the average goals scored by each formation. In this we captured that for home squads with a formation of 3-2-5-1 scored an average of just above 2.5 goals per match, and away squads with a formation of 5-1-2-1-1 scored an average of 6 goals per match. This away team statistic, seems to be off as scoring six goals in a match is quite difficult, but looking through the dataset at those matches, we noticed that they were blowout defeats from the away opposition when they were playing lower table standing teams. In addition to this, some of the formations were also quite unorthodox and this could be a result of teams lining up out of the ordinary in order to push for a result. An example of this is the home formation 3-3-3-1 which averaged just above 2 goals per match, and the away formation 3-3-3-1 formation which averaged about 1.5 goals per game. (figure 10)

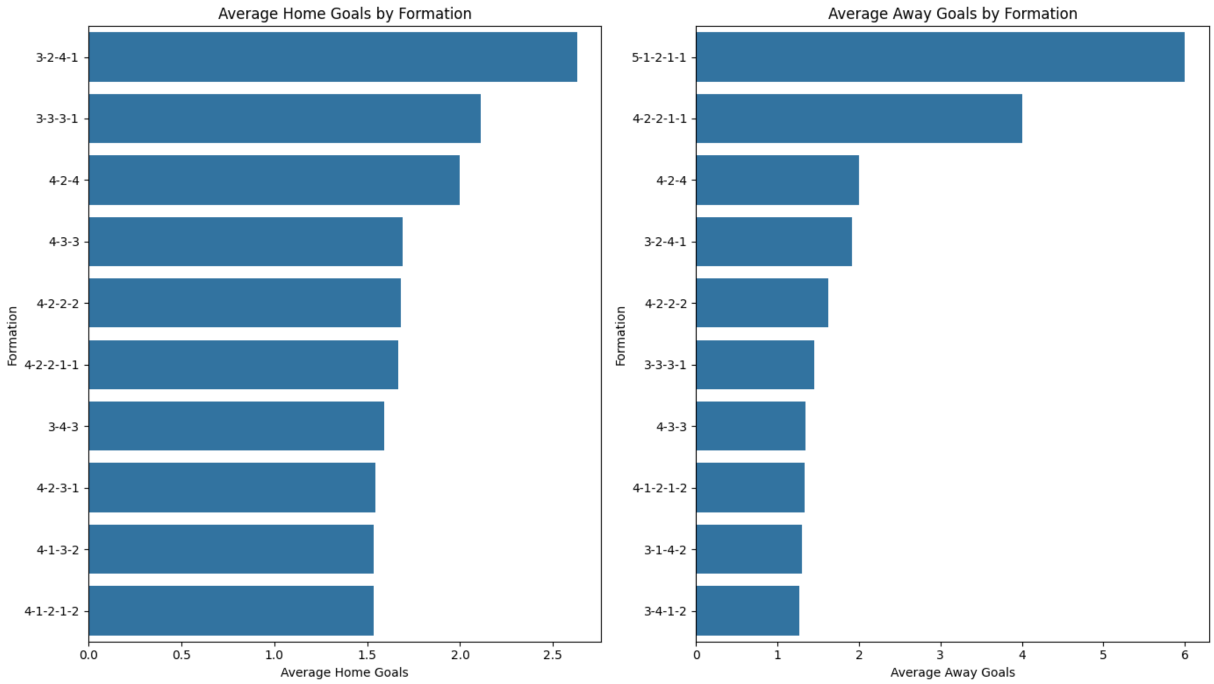


Figure 10

Our next result from average possession time by formation was interesting as we can see that for the 3-2-4-1 formation, home squads had about 55% and with the same formation away squads had about 60% of the possession. The best performing formation for home teams in terms of possession was 4-2-4 which averaged 60% possession, and the worst performing formation was 3-4-3 with about 51% possession. For away teams the best performing formation was 3-2-4-1 which is a base of 3-6-1, and averaged 60% possession, with the worst performing formation being 4-2-3-1 with 50% possession. (figure 11)

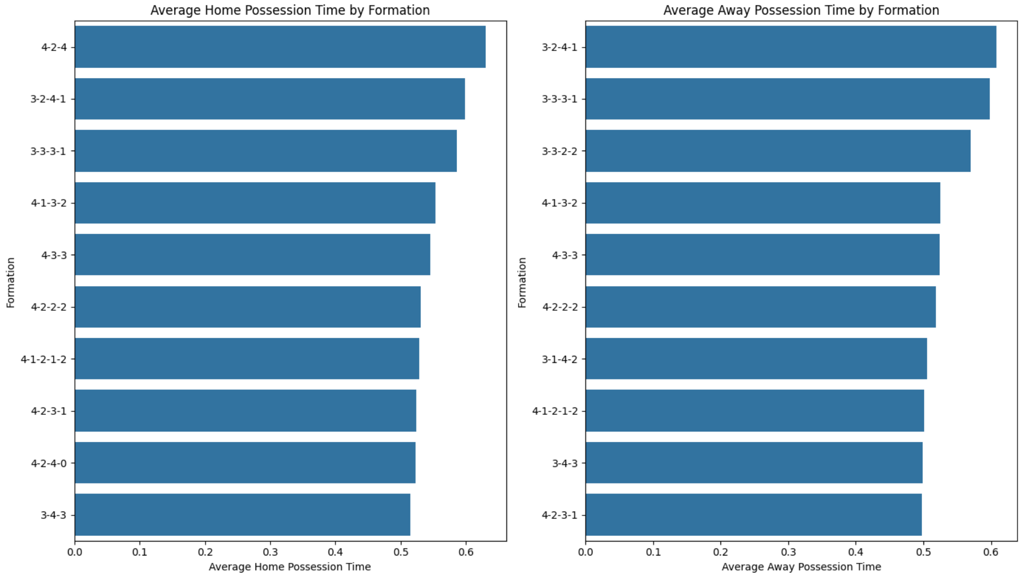


Figure 11

We then looked at the formations and their average total shots. The best performing home formations for shots on goal was 3-3-3-1 with 17 shots and 3-2-4-1 with about 15 shots. The best performing away formations were 3-3-2-2 with about 19 shots and 3-2-4-1 with about 17 shots. The worst performing home formation was 4-3-1-2 with just under 14 shots, and the worst performing away formation was 4-3-1-2 with about 11 shots. (Figure 12)

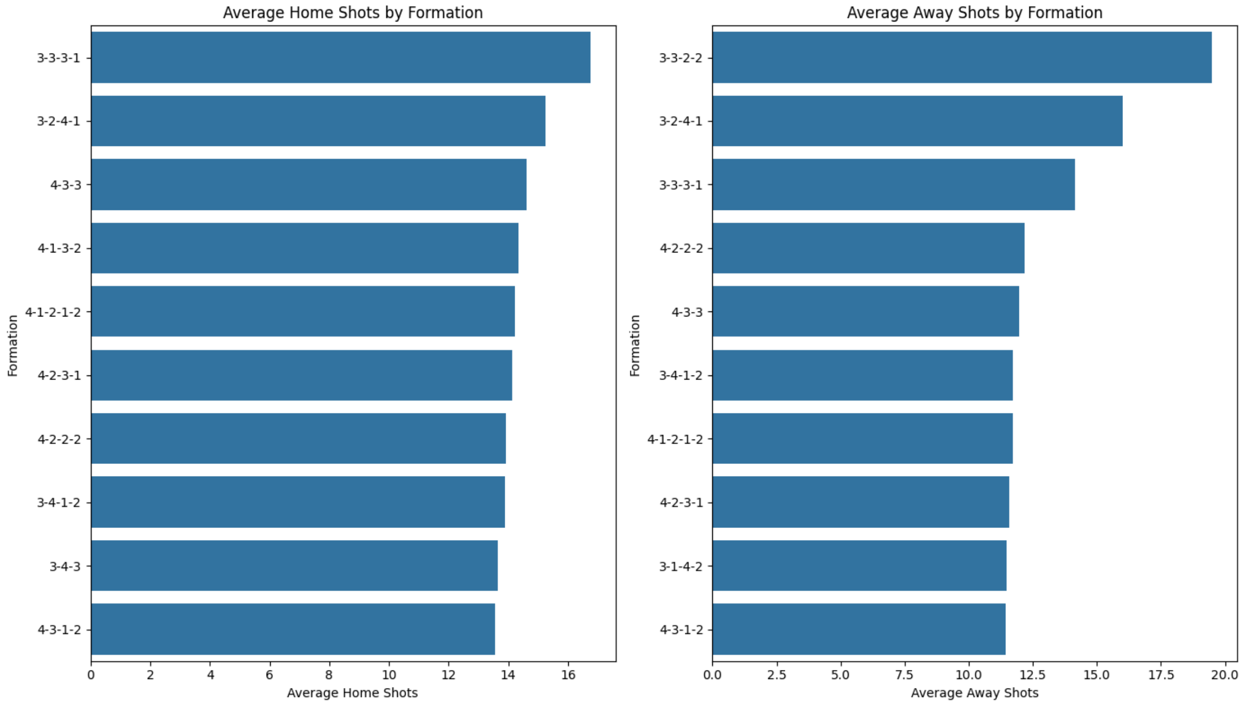


Figure 12

The final metric that we focused on was looking at the individual highest scoring games, and analyzing the formations used in those games. We extracted that teams for the most part, do not change their formation but, one did such as Juventus (Italy). The most popular home formations that Juventus used were 4-4-2 with about 48 games and 4-3-3 with about 44 games. Their away formation drastically changed with their most popula being 3-5-2 with about 51 games, and 4-3-3 with about 48 games. (figure 13)

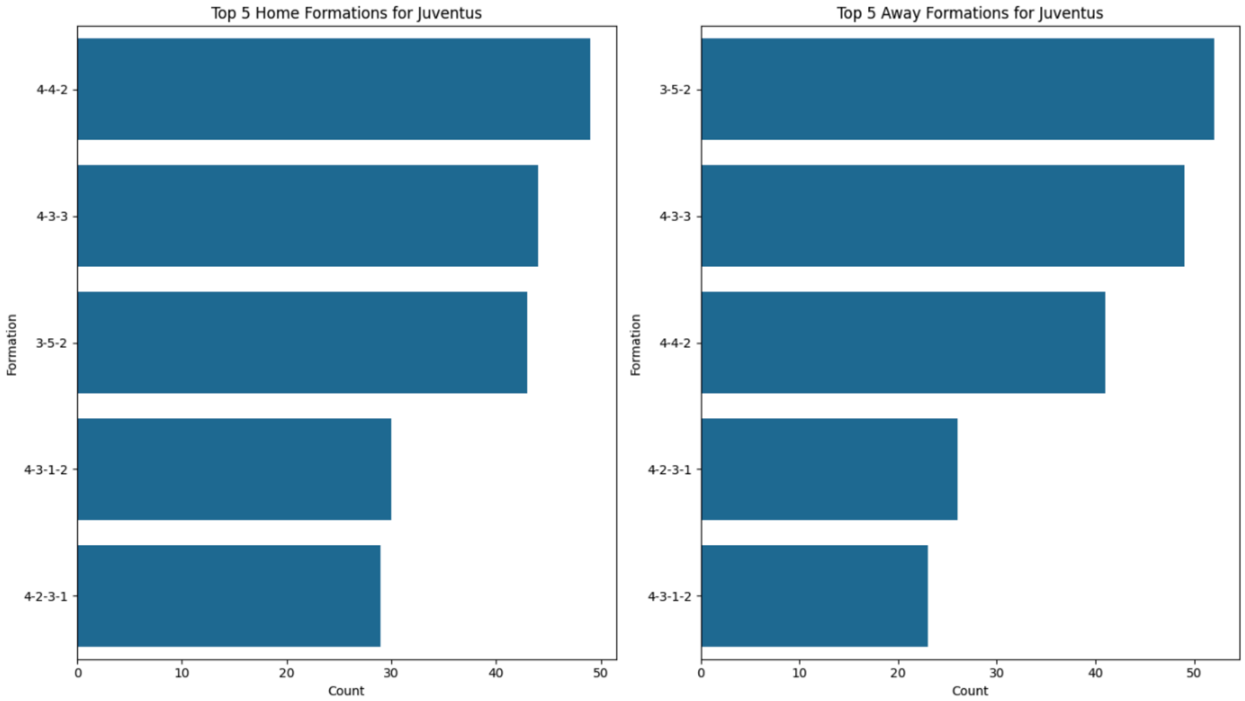


Figure 13

We also analyzed Real Madrid, which is commonly known as the best club in the world, and their main formation was 4-3-3 for both home and away. Their second most common formation changed from home and away, as for home their secondary formation was 4-4-2 and for away their formation was 4-2-3-1 which is a 4-5-1 base. (Figure 14)

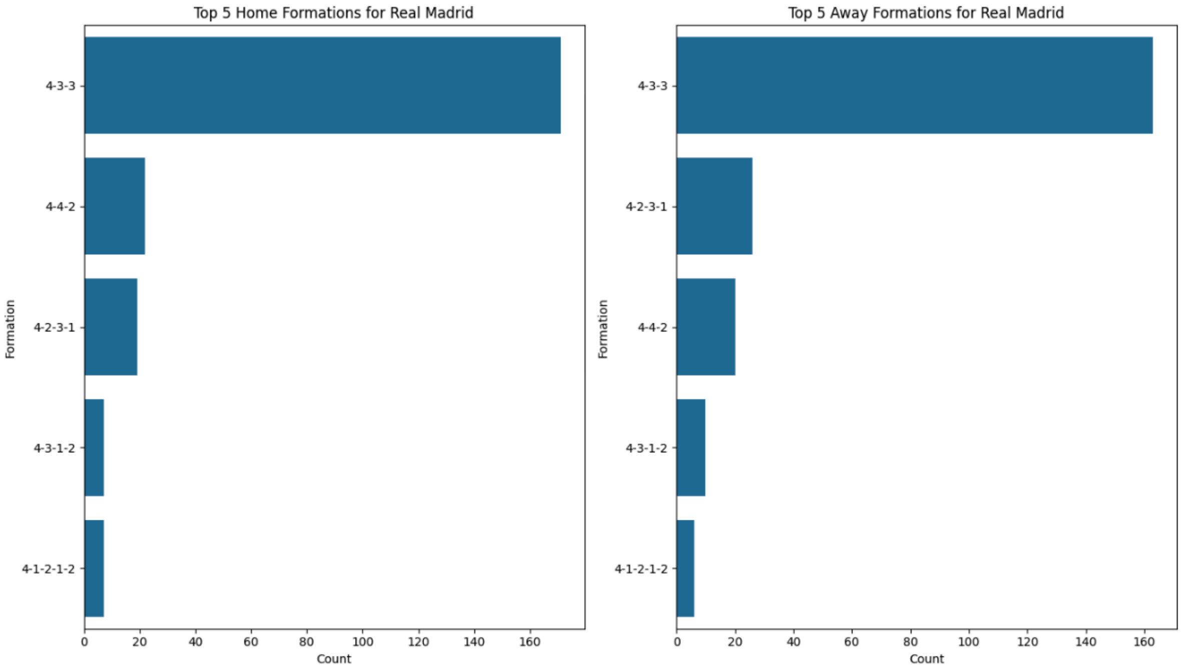


Figure 14

Conclusion:

Our image processing worked as intended for our generated images, we concluded that our project was successful in the digital image processing goals that we set out to accomplish. We were able to identify six out of six of the base formations that we generated. Incorporating the techniques that we used in class, into a Matlab script, to complete the project was a challenge that we believe we successfully implemented. We successfully imported an image, processed it using rgb2hsv() in order to better handle the color segmentation, got rid of the green grass background with a mask, then converted the image to grayscale in order to threshold the image to identify players using Otsu's method, and then found the centroids and identified the formation of the squad. The data science portion of the project was an extension that we decided to add to give us a deeper understanding of the formations teams used and the results that were obtained with those formations.

Discussion:

The goal of this project was to implement the techniques used in class to process images. However, this project can be expanded further to provide real-world and industry applications, not just for soccer/football but also for other sports. One key element that would have helped us analyze images further would be access to the copyright images from teams that they use to analyze their games. In addition to this, we would like to complete future work using video imaging techniques to analyze live video footage, which would allow great industry applications. We would like to focus on three topics in the future: formation tactics, player substitution suggestions, and adaptive formation recommendations. The formation tactics would be working with teams to analyze how they are performing in a match and provide data analysis for their performances. The player substitution suggestions would be a real-time substitution recommendation system that would analyze the current play of the squad and offer player suggestions to the managing crew of the team to suggest certain substitutions that may be beneficial for the squad. And the final future work would be to suggest real-time formation changes depending on how a match is going. As many players in the modern game can play many different positions, this would aid teams in changing their approach during the game to achieve the outcome of winning the game. All in all, this project was based on the image processing techniques that we learned in class but can be expanded greatly as sports is one of the biggest industries in the world.

References:

- Source Code Matlab
- Source Code GoogleColab
- https://www.fifplay.com/fc-24/formations/

- https://www.kaggle.com/datasets/ivanposinovec/football-matches-from-europe-and-south-america
- https://www.redbull.com/us-en/history-of-soccer