Trajectory-Based Playstyle Identification

Thomas Premm
thomas.premm@gmail.com
Johannes Kepler Universität
Linz, Oberösterreich, Österreich

1 Introduction

The following project report, written in the scope of the practical work course at the *Institute of Computer Graphics* at JKU, delves into the fascinating world of gaming. Video games are a worldwide phenomenon that transcends age brackets and other social categories, making it such a versatile topic of investigation. Understanding playstyles and player behavior in video games is an essential aspect of game design and user experience. As humans, and especially as computer scientists, categorizing data helps to recognize complex patterns and gain valuable insights.

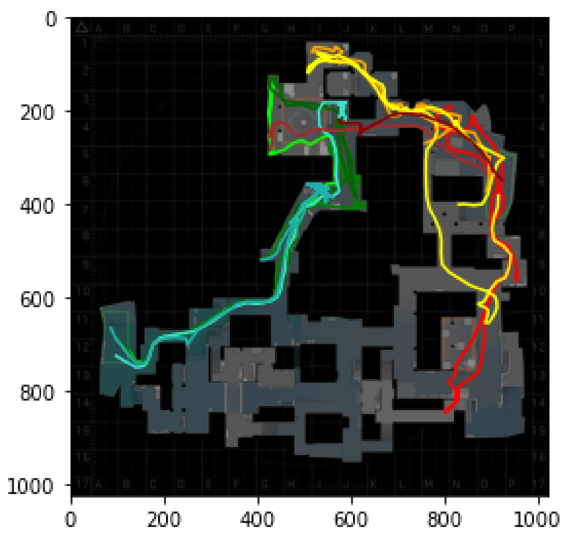

Figure 1: Trajectories of a single round

Through this project, I aim to explore the possibility of identifying different playstyles solely based on the trajectories of players within video games. In order to verify my claim, I computed various metrics from positional data and subsequently applied clustering techniques to distinguish between different archetypes of players.

2 Counter-Strike: Global Offensive (CS:GO)

For the purpose of this project I chose Counter-Strike: Global Offensive, commonly abbreviated as CS:GO, due to its meticulously documented and well-preserved in-game data. The game was developed by Valve and Hidden Path Entertainment and is a multiplayer tactical first-person shooter first released in 2012. At the start of each round, players get separated into 2 opposing teams, terrorists and counter-terrorists. They then have to collectively complete different objectives, such as planting or defusing a bomb, rescuing hostages, or eliminating all members of the opposing team. Despite its age, CS:GO continuously gets updated and, thus, maintains a highly renowned and competitive standard for the Esports community.

3 The Dataset

To access the trajectories of players in CS:GO, I utilized the *Esports Trajectories and Actions (ESTA)* dataset. The ESTA is a very large collection of parsed demo files from professional Esports tournaments spanning January 2021 to May 2022. It provides the viewer with compressed yet very detailed metadata of players' actions and spatio-temporal information, with updates occurring every half second (2Hz). Most notably the latter will be of great interest within this project since it offers a highly detailed snapshot of the trajectories each player follows. While the structure of the ESTA dataset files is nested and complicated, as shown in Figure 2. The the image illustrates the spatio-temporal data layout along with various other metrics captured in the dataset. In order to parse and analyze the chosen dataset, I used the *awpy* package, which was developed as a tool to handle such complex data files. This collaborative use significantly simplified the process of extracting and visualizing the data, therefore playing a vital role in the data preparation stage of the project. After having explored the nature of the dataset and the tools used to obtain and simplify it, I will now move on to address the main focus of my practical work: identifying different playstyles in CS:GO.


```
{:
  "parseKillFrame": true,
  "tick": 8174,
  "seconds": 3.71875,
  "clockTime": "01:52",
  "t": {
    "side": "T",
    "teamName": "",
    "teamEqVal": 3550,
    "alivePlayers": 5,
    "totalUtility": 1,
    "players": [
      {:
        "x": -1179.1435546875,
        "y": 483.21026611328125,
        "z": -55.96875,
        "steamID": 76561198035759667,
        "name": "thomas_the_tank_engine",
        "team": "",
        "side": "T",
        "velocityX": 109.83319854736328,
        "velocityY": 91.87308502197266,
        "velocityZ": 0,
        "viewX": 56.84326171875,
        "viewY": 3.33984375,
        "hp": 100,
        "isAlive": true,
        "isWalking": false
      }:
    ]:
  }:
}:
```


Figure 2: Overview of one frame of the ESTA dataset

4 Preprocessing

The initial step in preparing the dataset for analysis was to decompress the *.xz* files using the *read_parsed_demo* function. The second function, which was of great interest in preprocessing, was *create_frame_row*. The combination of these functions searches for relevant metrics in each analyzed frame. These metrics include which team a player belongs to, their health points, and general spatio-temporal properties that are subsequently sorted into a dictionary. After establishing these two key functions, the following step included transforming the data. I later automated this process to create big data frames, consisting of thousands of rounds, which are much more easily usable compared to the dictionary-style beforehand

I noticed that comparing trajectories for different maps would not make sense and thus separated the different maps, which was an extremely important step in the preprocessing. As distinct maps call for different tactics, distinguishing between them is key to unlocking certain information about a player's style. Moreover, I divided the set into 2 teams, terrorists and counter-terrorists, because comparing the metrics and trajectories of them would not lead to promising results. After preparing the dataset and completing the preprocessing stage, I will now further address the methods used to categorize CS:GO players according to their playstyle.

5 Methods

The approach is based on a metric-based analysis. In order to find meaningful and relevant metrics for the task at hand, I used two different sources for my methodology. Firstly, I started researching the works of previous authors that might offer a new perspective. Secondly, a closer examination of the dataset gave me a good impression of which metrics should help achieve the purpose of my analysis.

5.1 Total Distance

Total Distance refers to the cumulative length of a player's path during a game round. It is calculated by adding up the differences in position of consecutive timestamps. This metric is highly relevant to provide insight into how much distinct playstyles move in general. Higher values might suggest a more aggressive archetype, while lower values could indicate a more defensive tactic.

5.2 Start-to-end distance

This metric measures the distance from a player's starting position to the end position. Players with a large start-to-end distance may be those, who are pushing into enemy territory, while players with a lower one might be more position bound to their own territory.

5.3 Linearity

Linearity assesses the directness of a player's trajectory, which is calculated as the ratio of the start-to-end distance to the total distance. This metric can help distinguish between different tactical approaches, such as flanking maneuvers and direct assaults. Additionally, this criterion could reveal a lot about a player's tactic by uncovering whether their tactic involves a rather direct approach or a more convoluted trajectory.

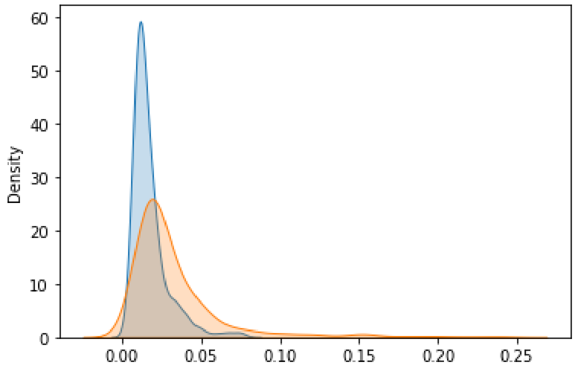

Figure 3: Differences in Linearity between the two teams

5.4 Number of turning points

Turning points are subsequent line segments exceeding a certain angle. Identifying these points can show patterns in a player's movement, such as evasive maneuvers or searching behavior. The number of turning points can be used to classify playstyles into either more dynamic or static categories.

5.5 Average Speed

Average speed is determined by dividing the total distance by the total time taken. This feature displays the mean velocity at which a player moves around the map. Analyzing the average speed can be indicative of playstyle, since a higher speed may indicate a more aggressive player, while lower velocities could suggest a more cautious approach.

5.6 Convex Hull

The Convex Hull of a set of points (the spatio-temporal data) is understood as the smallest convex polygon that completely surrounds all the points in this set. [1]

In the context of player trajectory analysis, this metric can be used for an approximation of the covered area by a player during a match. More specifically, it provides insights into the areal extension of a player's movement. Analyzing the area of a player's Convex Hull can show whether a player prefers to hold a tight defensive position (smaller Convex Hull) or move around across the map (larger Convex Hull), providing a metric to distinguish between defensive and aggressive in-game behavior.

5.7 Gyration Radius

The Radius of Gyration measures the dispersion of a set of points around their centroid, therefore computing how far the points spread out from their center of mass. [2]

The Gyration Radius serves as a metric to evaluate the variability in a player's movement. A larger Gyration Radius, on the one hand, indicates that a player's movements are widely dispersed around the mean position, suggesting that this player will change location frequently. On the other hand, a smaller Gyration Radius implies more positional concentrated movements, often associated with strategic positioning or guarding specific areas.

The Gyration Radius R_g can be calculated from the player's positional data as follows:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 + (y_i - \mu_y)^2}$$

where N is the number of positions recorded, x_i and y_i are the coordinates of the i-th position, and μ_x and μ_y are the coordinates of the centroid of a player.

This metric is particularly useful for identifying players who recurrently change their locations as opposed to those who tend to remain in a specific area. Therefore, the Gyration Radius offers knowledge about diverse strategic roles within the game.

6 Clustering Techniques

Before testing clustering techniques, it is vital to look at the correlation matrix of the already computed features, since highly correlated features may carry redundant information, which would lead to inefficiencies during clustering. By identifying and removing related features, the clustering process becomes more understandable and more effective. As illustrated in Figure 4 *total distances* and *turning points* have a correlation value of 0.79 which is extremely high. Since *total distances* also correlates significantly with other features, I decided to remove it from the feature matrix.

6.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a popular and important step in data analysis. Its overall goal is to down-project some high-dimensional data. This is achieved by transforming a set of possibly correlated variables into a smaller number of uncorrelated variables called principal components (PCs).

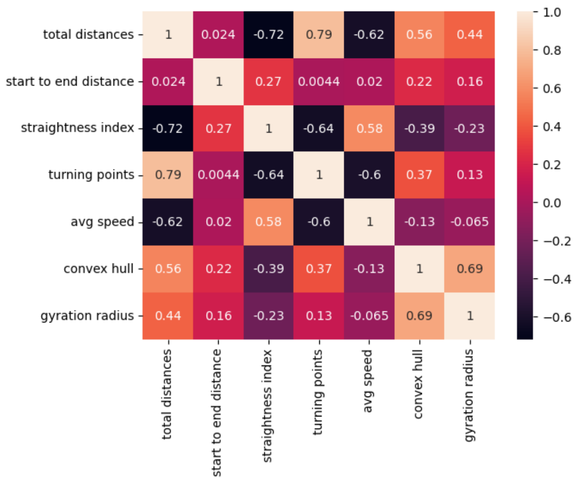

Figure 4: Correlation Matrix of the computed features

It is of the utmost importance to first standardize the data as it is very sensitive to the variances of the initial features. This process can be easily carried out by subtracting the mean and dividing by the standard deviation afterward for each value of each variable. Standardization ensures that all variables are on the same scale, preventing variables with larger ranges from dominating the outcome, hence it eliminates the possibility of a bias. As a next step, PCA calculates eigenvalues and eigenvectors of the covariance matrix to find the maximum variance within the data. The goal of this step is to find the principal components, which can be seen as a linear combination of the initial variables that capture the maximum variance. With this procedure at hand, it is possible to reduce the number of dimensions while keeping the majority of the information. The data will now be recast along the axes of the principal component, which completes the transformation into the PCA space[3].

6.2 K-Means Clustering

After successfully applying PCA and reducing the number of dimensions to 2, clustering the newly projected data will be the final step before identifying playstyles. For this I tried working with the popular K-Means and DB-Scan methods, however, in this paper, I will only consider K-Means, because it worked better.

K-Means clustering is an unsupervised machine learning technique that separates the points into k clusters in which each data point belongs to the cluster with the nearest mean. In the next step, the algorithm iteratively assigns each data point to the nearest cluster centroid and then updates the centroids based on the points assigned to each cluster. This process is repeated until no changes happen and as a result, the clusters are determined. K-Means is particularly useful for identifying distinct groups within a dataset and is widely used for cluster analysis in various fields[4].

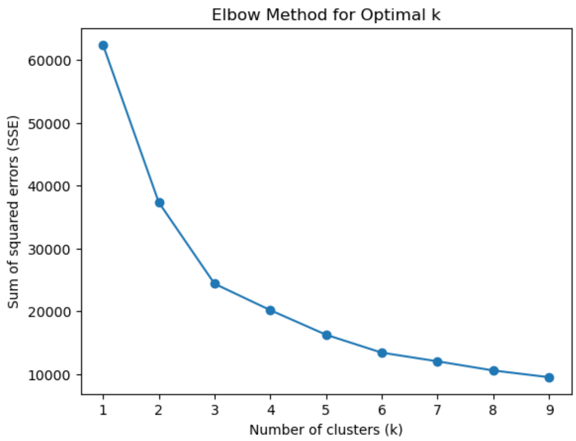

Figure 5: Elbow plot for optimal k

It is essential to determine the optimal number of k. This can be done by using methods like the elbow method, which I have applied. I let k be in the range 1 to 10, plotted the inertia for each k, and chose the *elbow* of the curve as my final k. In Figure 5 it is visible that the elbow is at k=3, thus 3 labels will be optimal.

7 Results

The analysis presented specifically focuses on the counter-terrorist team's movements on the map "Nuke". This decision was made to maintain consistency in the data analysis, as comparing different maps or opposing teams could introduce variability that is not beneficial to identifying distinct playstyles.

The application of K-Means clustering on the PCA-transformed dataset has yielded three distinct clusters. After adding the label to each player I decided to group them by their cluster. Subsequently, I created boxplots to visualize the comparative distribution of the calculated metrics, which is instrumental in distinguishing the characteristics and therefore the playstyles of each cluster.

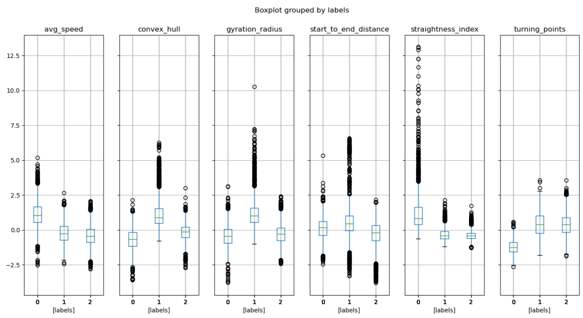

Figure 6: Boxplot comparison of player metrics across identified clusters

As depicted in Figure 6, Cluster 0 is characterized by the highest average speed coupled with the lowest convex hull, the highest straightness index, and the fewest turning points. This indicates a playstyle that is very fast and direct, with players likely engaging in swift, targeted movements toward objectives with minimal deviation.

Cluster 1 on the other hand is defined by the largest convex hull, the greatest gyration radius, the longest start-to-end distance, and plenty of turning points. Players in this cluster stand out due to their versatile playstyle that includes extensive area coverage and varied movement patterns, which indicates an archetype who excels in map control.

Cluster 2, in contrast, has the lowest average speed and start-to-end distance, although showing a tendency towards many turning points. This cluster suggests a more strategic playstyle, with players moving at a slower pace, possibly typical for a defender role. These players lay their focus on area denial and holding strategic positions.

8 Conclusion

This paper presented an exploratory analysis of player behavior in CS:GO through trajectory-based playstyle identification. The results illustrated distinct playstyles, which are characterized by various spatio-temporal metrics.

Potential future research in this area could include the application of sequence clustering techniques, which directly use the trajectories and perform an algorithm on them, without computing metrics beforehand. Such an approach would likely capture more nuanced patterns. While I have made efforts in this direction, I have not accomplished a method that worked out, thus this can be seen as an opportunity for further research. I suggest a more sophisticated approach capable of handling sequential data, like machine learning models, such as RNNs or hidden Markov Models.

It is important to acknowledge that the ESTA dataset used in this study comes from Esports tournaments. Hence, the playstyles identified may represent optimized gameplay, which could limit the generalizability of the findings to casual players. Future studies might consider including datasets from casual games to ascertain whether the identified playstyles hold across different levels of play.

References

[1] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. Springer New York, 1985.

[2] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” *Nature*, vol. 453, pp. 779–782, 2008.

[3] I. Jolliffe, *Principal Component Analysis*. Springer New York, 2002.

[4] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, pp. 281–297, 1967.