Autoref System Validation Methods Report

Deniz Akyazi

Executive Summary

This report outlines the validation methods for our robot soccer autoref system, designed to assist human referees in determining the last robot that touched the ball before it went out of the field

. The validation process aims to ensure the accuracy, reliability, and adaptability of the autoref system in various game scenarios, validating the functional requirements of the system. It is important to note that in this report main concern is the real-world performance and meeting user expectations.

1. Introduction

1.1. Background

The autoref system, addressed in this report, functions as an assistant for human referees, specifically in cases where it is tasked with determining which robot last touched the ball before it went out of play. The primary focus of the validation process is to assess the system's accuracy in identifying the robot that touched the ball last.

2. Validation Methods

2.1. Real-World Testing

2.1.1. Preparation

Before conducting the real-world testing, thorough preparation is essential to ensure the Autoref System's seamless integration and functionality. The following steps will be undertaken:

2.1.1.1. System Setup

The Autoref System will be configured on the Tech United field, ensured that all components are calibrated and operational for optimal performance.

2.1.1.2. Input Data Validation

Project team will initiate controlled scenarios to validate the Autoref System's ability to receive sensor/processed input data. This step ensures that the system can accurately interpret live game data.

2.1.1.3. Independence Verification

Validation will be performed to confirm that the Autoref System receives data only from independent sensors, avoiding interference from competing team robots. This verification guarantees the system's autonomy and impartiality during live gameplay.

2.1.1.4. Collaborative Testing

Team members will collaborate in simulated game scenarios to evaluate the Autoref System's responsiveness and decision-making capabilities. This collaborative testing phase aims to identify any potential issues and refine system performance.

Requirements Covered:

- ✓ System must receive sensor/processed input data of the “live” game.
- ✓ System must receive data from independent sensor (not from any sensor in competing team robots).
- ✓ System can distinguish which team touched the ball the last time.
- ✓ When there is change in game state, system can get game state.

2.1.2. Implementation

An example match of 2 vs. 2 robots will be played in the Tech United field located in TU Eindhoven. Human referees as well as the autoref system will be used during the match and the decisions made by both the human referee and autoref will be recorded.

Requirements Covered:

- ✓ System can distinguish which team touched the ball the last time.
- ✓ System can decide who is the last one to touch the ball at the current time.
- ✓ System can make combined decisions (out of field and last touch) in real-time.
- ✓ When there is change in game state, System can get game state.
- ✓ AR can communicate the decision it made through a medium when the ball is out of field.
- ✓ AR must be able to show its decision process when needed.

2.1.2.1. Edge Cases and Failure Analysis

During real-world testing, particular attention will be given to edge cases and scenarios that might pose challenges to the autoref system. This includes situations that are within the normal range of gameplay conditions but are considered challenging. Key aspects to consider:

- **Identification of edge cases:** Scenarios that are challenging (hard to be distinguished by the human referee, discussed in deliverables) but plausible within regular gameplay conditions will be identified by the teammates observing the match as well as by the feedback of the human referees.
- **Testing approach:** The match will be meticulously recorded from different angles to ensure comprehensive coverage. Post-match, decisions made by the autoref for identified edge cases will undergo thorough investigation. Confirmation of these decisions will be sought from multiple sources, including the perspective of human referees and a detailed analysis of the recorded match by expert reviewers.

Validation for requirements:

- ✓ Validate the ability to make combined decisions in complex scenarios.
- ✓ Confirm the system's resilience in challenging conditions.

2.1.3. Statistical Metrics and Benchmarking

In this section, statistical metrics associated with real-world testing will be discussed. This includes the data analysis process, scenario-specific thresholds, and benchmarking against decisions made by human referees.

2.1.3.1. Benchmarking Against Human Referees

Benchmarking against human referees involves comparing the decisions made by the autoref system with those made by human referees during the match. Two comparison metrics will be considered for validation, using the data analysis methods discussed below.

- **Human-like decision making:** The decisions made by both the human referee and the autoref system will undergo a thorough comparative analysis. The focus will be on identifying areas of alignment and divergence between the autoref and human referee decisions. Emphasis will be placed on benchmarking the nuanced differences and similarities, providing insights into the autoref system's ability to replicate human-like decision making.
- **Performance:** A third layer of comparison will be introduced, incorporating an objective analysis conducted by expert viewers with reference to multiple videos. Expert viewers, possessing a high level of expertise in robot soccer dynamics, will provide an unbiased evaluation. The results of this expert analysis will serve as the base reference against which the decisions made by both the human referee and autoref system will be evaluated. Performance metrics will be calculated for both entities, facilitating a thorough comparison and revealing any distinctions in their decision-making accuracy.
  Requirements covered:
  - ✓ System can detect if the whole ball out of field with desired accuracy as good as human.
  - ✓ System can distinguish which team touched the ball the last time correctly.
  - ✓ System must be able to record the received data with timestamp.
  - ✓ System must be able to show its decision process when needed.

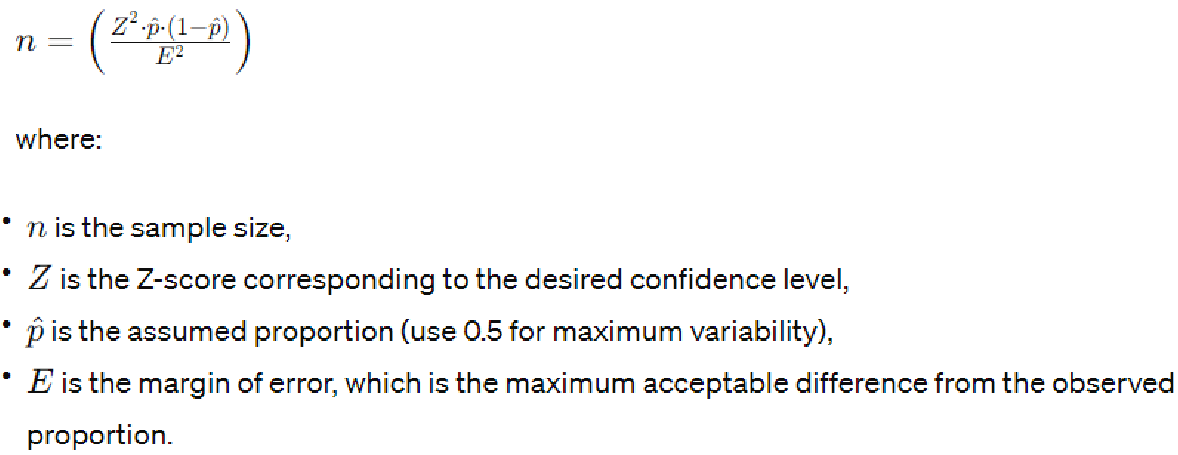
2.1.3.2. Data Analysis

The data analysis phase involves processing the decisions recorded during the match, applying statistical metrics to quantify the autoref system's accuracy, and drawing meaningful insights. Key components of data analysis include:

- **Confusion matrix:** It aids in understanding where the system excels and where it may need improvement.
- **Accuracy:** Overall accuracy, edge-cases accuracy and major cases accuracy need to be investigated separately since the dataset will be imbalanced.
- **F1 Score:** Harmonic mean of *precision* and *recall*. Particularly useful to measure overall performance when there is an uneven class distribution.

2.1.3.3. Confidence Interval and Required Number of Samples

To understand the required number of samples for the validation with a chosen confidence interval, some values were estimated. With a conservative estimation, the following formula was applied.



$$n = \left(\frac{Z^2 \cdot \hat{p} \cdot (1 - \hat{p})}{E^2}\right)$$

where:

- n is the sample size,
- Z is the Z-score corresponding to the desired confidence level,
- $\hat{p}$ is the assumed proportion (use 0.5 for maximum variability),
- E is the margin of error, which is the maximum acceptable difference from the observed proportion.

In the validation that will be carries, Z-score is chosen approximately 1.96, corresponding to 95% confidence level. p is chosen to be 0.7 due to requirements and margin of error is assumed as 10% of the observed samples. Carrying out this calculation, the number of samples required is 112. Corresponding to number of last touch decisions carried out due to ball out of play.

Validations for requirements:

✓ Validate the accuracy of decision-making in real-world scenarios.

3. Conclusion

Methods for rigorous testing of our Autoref System for robot soccer are proposed in this report. The upcoming steps involve the execution of the outlined methods, including real-world testing, edge case analysis, and benchmarking.

Moving forward, insights for improvement are aimed to be gained. The Autoref System, designed to assist referees in ball tracking, will undergo testing. Commitment to meeting user expectations remains, and refinements will be guided by the practical validation.