# Robust Virtual Content Integration and Occlusion Handling in 2D Video Streams

Aashka Desai(ad2280@rit.edu), Advisor: Dr. Fawad Ahmad
Rochester Institute of Technology
Department of Computer Science, GCCIS
Rochester, NY, USA

***Abstract*—This paper presents an integrated system for overlaying virtual content, like logos and scoreboards, in live sports with efforts on minimizing the impact caused by player occlusion, camera vibrations, and real-time processing demands. The proposed framework leverages state-of-the-art computer vision techniques to seamlessly integrate virtual content into live sports broadcasts. It uses static element detection to define the region of interest(ROI), real-time player detection and masking to handle occlusions, and displacement tracking to stabilize overlays against camera movement. The results demonstrate a system that keeps overlays stable and realistic, even in dynamic conditions, with low latency, high accuracy, and robust performance validated through quantitative analysis.**

***Index Terms*—Virtual Content Integration, Live Sports Broadcasting, YOLOv5, Camera Vibration Compensation, Computer Vision, Dynamic Masking, Overlay Stabilization.***

## 1 INTRODUCTION

Overlays like virtual logos and scoreboards have recently gained increased attention in enhancing viewer experiences and delivering up-to-date information in live sport broadcasts. However, seamless integration is not easy in dynamic environments where a lot of extraneous element movement is present. Occlusions are caused by moving players, the overlays are destabilized due to camera vibrations, and real-time processing is difficult to attain in synchrony with the live events.

This project aims at designing a system using the latest techniques in computer vision and artificial intelligence that is capable of embedding virtual content into live video streams in a manner that the final results appear natural and visually consistent. The project targets not only the challenge of overlaying content but also seamlessly integrating it with the scene, ensuring it adapts dynamically to movements and changes that may occur in the environment.

The system detects static elements, like stumps and pitch lines, through a state-of-the-art object detection model-YOLOv5-to identify a stable Region of Interest. Similarly, it handles seamless occlusion by dynamic player detection and masking in order to let virtual content respect real-world interactions. The system incorporates displacement tracking with smoothing algorithms that allow the overlay to remain stable even at times of high camera movement or vibration.

It therefore focuses on real-time performance, stability, and adaptability, and is a showcase for competencies in computer vision and AI on solving complex problems around live video processing. Outside sports, the developed methods have found useful applications in areas like augmented reality and interactive digital experience where there is a need for the integration of virtual content accurately.

## 2 BACKGROUND AND RELATED WORK

The integration of virtual content into live sporting broadcasts has grown remarkably over time. This development is based on increased usage of AR, VR, and computer vision. AR technologies utilize spatial mapping and depth sensors to overlay virtual elements into the real world while creating a complete virtual space for viewers. Computer vision techniques, object detection, and motion tracking enable the dynamic placement and interaction of virtual elements within live video streams. Applications of such technologies in real-world scenarios of fast-paced sports pose a number of challenges, such as occlusions by moving players, camera vibrations, and low-latency processing.

(1) **Occlusion Handling in Augmented Reality** [1] Guglielmo (2019) investigated occlusion handling in augmented reality (AR) by developing a system that utilizes depth information to manage occlusions between real and virtual objects. The approach involves creating depth maps to accurately render virtual elements in scenes with complex occlusions. While effective, this method requires additional hardware for depth sensing, which may not be feasible in all broadcasting scenarios.

(2) **Object Tracking-based Real-Time Occlusion Handling** [2] Tian et al. (2010) proposed a real-time occlusion handling method in AR that relies on object tracking. Their technique involves selecting occluding objects through interactive segmentation and tracking their contours across frames to manage occlusions effectively. This approach enhances the realism of AR applications but can be computationally intensive, potentially impacting real-time performance.

(3) **Virtual Reality in Sports Broadcasting** [3] Virtually Live has developed a system that captures live event telemetry to recreate sports events in a virtual environment, allowing fans to experience events in real-time through VR headsets. This method offers immersive experiences but requires specialized equipment and may not seamlessly integrate virtual content into traditional 2D broadcasts.

This project differentiates itself by focusing on seamless integration of virtual content into live sports broadcasts without the need

for additional hardware or specialized equipment. The developed approach is based on real-time players and umpire detection by YOLOv5, which dynamically masks moving elements such that the natural occlusions would be respected. Besides, the vibration of cameras is also compensated for by tracking displacement of markers within the ROI for the purposes of maintaining stable overlay placement. Both the approach and techniques have been modified and improved, taking into account the drawbacks of previous methods to the extent of being a scalable and live broadcasting environment compliant total solution.

## 3 SYSTEM OVERVIEW

This project develops a system that seamlessly integrates virtual content into live sports broadcasts-from logos to scoreboards. Taking into consideration challenges like occlusion of players, camera vibration, and performance in real time, it makes these virtual elements naturally blend and cohesive within dynamic video streams.

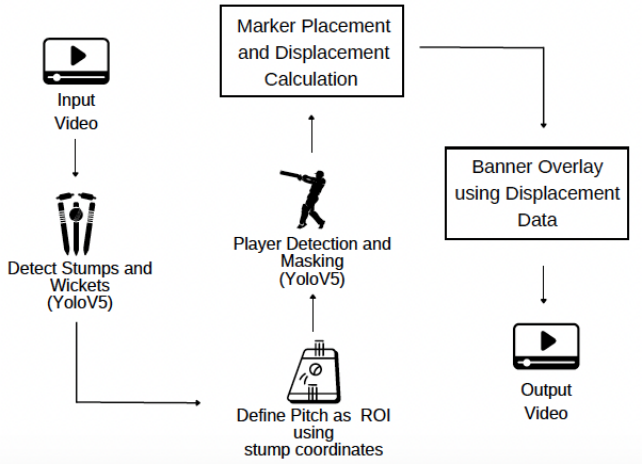


**Figure 1:** System Workflow

The workflow, briefly explained in Figure. 1 applies advanced computer vision techniques, which make it dynamically adapt to changing scene conditions. It first detects the static elements present in the video like the stumps and wickets using YOLOv5 on a custom dataset. It then identifies the pitch as the region of interest based on the coordinates of the stumps as we can assume that the pitch will always be present in between the stumps present on both the sides for any cricket play ground. The system detects player motions so overlays respect occlusions. Then, displacement tracking is used to compensate for vibrating cameras; refined masking and blending keep visuals consistent. These integrated processes yield a solid pipeline, where virtual content stays stable, looks real, and is reactive in a live broadcast environment.

Below explained are the steps involved in this structured workflow in detail, indicating the main stages to make the integration of virtual content both stable and realistic.

### 3.1 Stump and Pitch Detection

The static elements i.e., the stumps and wickets are detected in the video frames by YOLOv5 as seen in Figure. 2, which is immensely powerful and very recently developed object detection framework. The stumps and wickets are used because they are static elements and are always present for the entire duration of play in the game. The positioning of these objects makes them ideal reference points. Hence, once they are detected the coordinates of the bounding box are extracted and using them, we establish a Region of Interest (ROI) that covers the pitch area.


**Figure 2:** Stump and Wicket Detection using YoloV5

As the pitch always falls within the region of interest, there is no need to detect it in the earlier stages directly. This ROI ensures that subsequent stages focus only on the relevant portions of the inspected frames, reducing computational overhead by avoiding interference from superfluous elements such as players, shadows, or the boundaries of the pitch. The pitch region is highlighted in Figure. 3

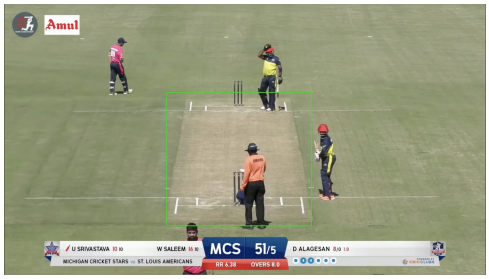


**Figure 3:** Pitch Detection using Stump Coordinates

*3.1.1* ***Edge Detection within the ROI.*** The identified ROI formed by the coordinates of the stumps and wickets is subjected to edge detection in the defined area in order to indicate the pitch lines. It first converts the cropped area to grayscale for easier image calculation. To emphasize the distinguished boundary, lines in the pitch, crease and other markings the Canny Edge Detection algorithm is applied. The edge detection done frame wise as seen in Figure. 4 provides additional visual information for aligning and determining where the overlay is stabilized or picked.

Additional precision is obtained by cleaning up these edges through morphological operations, such as dilation and erosion. This removes noise and helps sharpen them. The processed edges are further correlated onto the ROI for precise localization of pitch features. It is stored section-wise for subsequent processing.

By using the stumps and wickets to define the Region of Interest (ROI) and applying edge detection within this area, the system can effectively identify pitch lines. This focused approach ensures

**Figure 4:** Edge Detection on the ROI

accurate preparation for subsequent steps, such as displacement tracking and overlay placement.

## 3.2 Player Detection and Masking

Apart from dynamically managing occlusions and providing a good contextual background for introducing virtual content, the system also uses real-time tool YOLOv5 to detect the players and umpires on the field in real-time. The development of YOLOv5 aims to enhance the robust object detection mode specified to detect humans (class ID 0) in the video frames. Using trained weights from the model, each frame from the input video is processed accordingly to localize the players and umpires as seen in Figure. 5 by creating bounding boxes around them with a high degree of confidence.

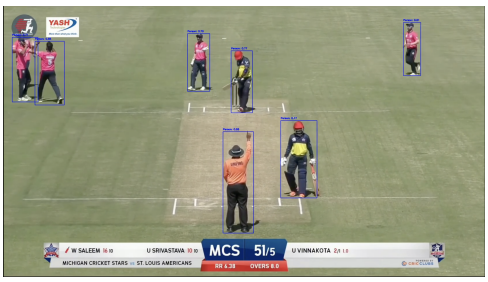

**Figure 5:** Person Detection using YoloV5

After detection, a dynamic masking script is executed. The bounding boxes for the detected humans in each frame are transformed into rectangular masks as visible in Figure: 6 so that the virtual content effectively masks out the area occupied by the players and umpires. Masking ensures that virtual content does not overlap players and respects natural occlusions.

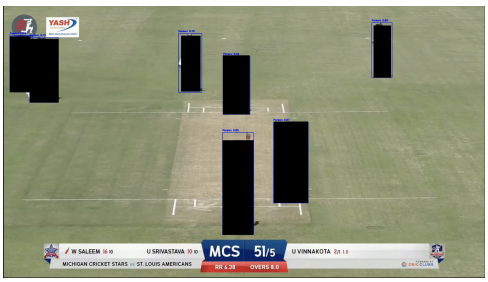

**Figure 6:** Masking Players using dynamic Masking Script

The extracted frames with masks are saved to an output video displaying the dynamic detection of players and umpires who have been removed from the overlay region. This is vital for making virtual content realistic, as the overlay will appear to be naturally part of the scene rather than an artifact.

## 3.3 Displacement Calculation

The system calculates displacements between successive frames to take into account any camera vibrational effects to maintain the alignment of virtual content, using markers that are positioned within the ROI identified in the earlier detection phase of the process and targeting the stable and more identifiable features of the pitch and stumps.

*3.3.1* ***Marker Placement.*** The contour analysis identifies the distinct points, like pitch lines or consistent features, in the ROI and places markers within them. These become reference locations from which movements are tracked. The coordinates of these markers in the first frame are recorded as the baseline against which all subsequent movements will be measured.

*3.3.2* ***Calculating Displacement.*** Position of the markers is calculated for every frame. From its initial and present position, the amount of shifting of the pitch and stumps can be calculated. The displacement can be calculated by the following relation:

$$dx = cx - ix, \quad dy = cy - iy$$

Here, $(ix, iy)$ are the initial marker coordinates, and $(cx, cy)$ are the coordinates in the current frame. These, therefore, represent the displacements arising due to camera motion or vibration.

*3.3.3* ***Smoothing of Displacement Data.*** With the following, the system may be able to refine data toward smooth tracking and eliminating noise from:

- Thresholding limits very rare outliers by restricting the displacement values within the normal bounds.
- Smoothing Algorithm filters minor variations in markers.

## 3.4 Dynamic Banner Placement


**Figure 7:** RIT logo Overlayed on a Frame

These refined displacement values dynamically create the position of the virtual content. This ensures that the overlays remain stable and perfectly aligned with the pitch, even when there is fast movement of the camera. It does this by interpolation and bounded movement parameters to prevent any jitter or misalignment of virtual content. This step is quite important because it keeps the natural feel of the scene intact while still managing to hold the virtual content in place. The overlayed banner looks like as shown in Figure. 7 where the RIT logo is placed on the pitch.

In summary, the system is designed to seamlessly integrate virtual content into live sports broadcasts by combining advanced computer vision techniques with dynamic adaptations. From detecting static elements to defining the Region of Interest, handling player occlusions, tracking displacements, and refining overlay placement, each stage of the workflow ensures stability, realism, and precision. Together, these interconnected processes form a robust pipeline that maintains the alignment and coherence of virtual content in dynamic and fast-paced environments.

## 4 RESULTS

The aim of the project was to create a powerful and flexible system for seamlessly inserting virtual content into sports broadcasting in a robust and real-time fashion. The basic principles were followed for dealing with most dynamic video issues, which are: stabilization regarding camera shake, occlusions naturally arising through moving players, and, obviously, the execution time required in real time.

- **Dynamic Occlusion Handling:** The system successfully handles occlusions as seen in Figure. 8 where we can see that when a player occludes the RIt logo, the logo behind is moved behind in the scene whereas the player remains in the front making the integration seem natural. In real time, the system also estimates these elements and applies a dynamic mask so that virtual content is not occluded and the natural occlusions are not violated. The system manages to maintain the illusion that virtual content is part of a scene. Moving players get excluded from the overlay region at the output video, therefore attesting to good masking behavior for a given sequence of fast-moving situations.


**Figure 8:** Handling Player Occlusion

- **Vibration Handling:** Vibration or camera motions are compensated by displacements of markers placed inside the region of interest. The system successfully estimates raw frame-to-frame displacement from which it refines this measurement further by interpolation in order to achieve smooth, jitter-free overlay positioning. For such applications where overlay positions need to align exactly with the pitch, it eliminates jitter even when fast-moving camera movements are involved. All this is clearly visible from the output video where the overlay is visually consistent due to its stability.

The system is developed so that overlays appear in place and visually consistent with fast movement by the camera and/or intricate player interactions. Using computer vision techniques such as advanced displacement tracking, dynamic masking, and refined blending ensures the virtual content not only closely aligns with the pitch but also respects the real-world context of the broadcast.

The results describe and show how well this system achieves these goals: low real-time latency, stable overlays, and perfect integration of augmented and real scene segments in motion. Visual outputs and quantitative analysis thus form a comprehensive backbone on which the response of this system to challenges in a live broadcast have been analyzed.

### 4.1 Average Confidence Scores per Frame Index Group

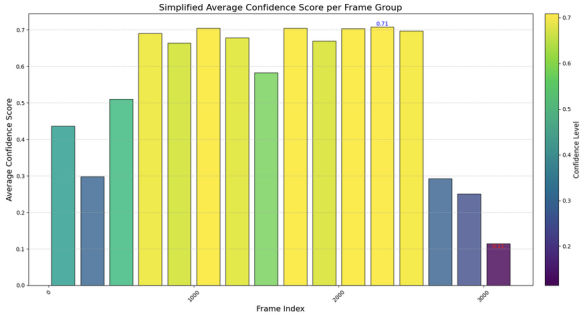


**Figure 9:** Average Confidence Scores per Frame Index Group

Here, graph present in Figure. 9 shows the confidence scores of YOLOv5 in detecting stumps, players, and umpires in a frame. It is observed that almost consistently, the confidence score is at 1.0 or very close to that value since YOLOv5 is able to detect the items quite reliably. However, small drops take place sometimes, maybe due to potential motion blur, partial occlusion, or variation in illumination which are negligible and almost will never affect overall accuracy.

It ensures reliable detection to have accurate ROI definition, player masking, and overlay alignment-things that are very crucial for stable virtual content integration.

Observations:

- Consistent high confidence scores validate YOLOv5's accuracy.
- Smaller dips suggest robustness, leaving little room for further improvements on the edge cases.

### 4.2 Latency Histogram for Rendering Frames

The histogram present in Figure. 10 shows the distribution of the frame processing time where for most of the frames, the time lies between 40 ms and 60 ms, whereas the average latency is 50.41 ms. Here the processing for a few frames extends 100 ms which result from increased computational complexity in challenging conditions.

The system provides latency performance that should be suitable for real-time applications because smooth adjustments of overlays can be done without noticeable delays.

Observations:

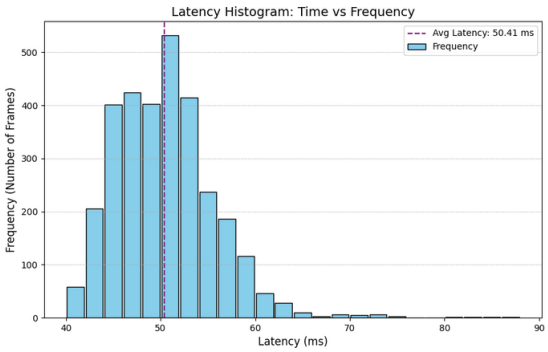


**Figure 10:** Latency Histogram for Rendering Frames

- Most of the frames fall within the efficient latency range, which confirms real-time suitability.
- Performance is hardly affected by rare outliers.

### 4.3 Cumulative Percentage of Frames Processed Within Latency Intervals

The graph present in Figure. 11 shows the percentage of frames within processed latency intervals: 90% are under 60ms, and for roughly 70ms, it's already close to 100% percent in accumulation. The slope is nearly straight, which suggests that well over most of the frames have performance similar in consistency and efficiency.

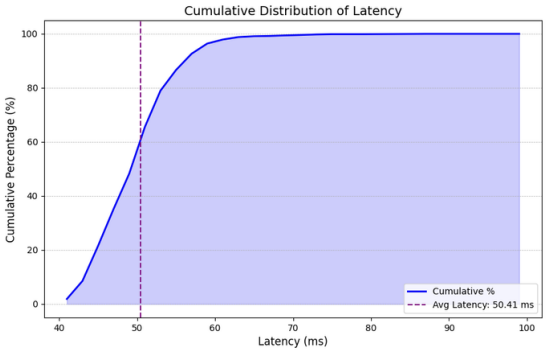


**Figure 11:** Frames Processed Within Latency Intervals

The system processes frames quite efficiently while predictably producing very few outliers, which does not critically affect its ability to maintain the correct sync between overlays and live events.

Observations:

- About 90% of the frames are processed within 60 ms, which ensures real-time capability.
- The graph confirms the reliability of the system in handling dynamic live sports environments.

## 5 CONCLUSION

The project showcases a very robust and scalable solution that integrates virtual content into live sport broadcasts, including but not limited to the embedding of logos and scoreboards. The system treats critical challenges such as occlusion handling, camera vibration compensation, and real-time processing requirements, seamlessly embedding virtual content into dynamic video environments.

The system detects key elements like stumps, players, and umpires reliably with the use of YOLOv5, thereby enabling the definition of ROI with high accuracy and its dynamic masking. It therefore allows the system to track displacements and refine the same for stable overlay placement under challenging conditions, such as fast camera movements. The real-time processing pipeline realizes low-latency performance, and this makes the system applicable to live sports broadcasting.

These results confirm the power of the system in bringing out the consistency and match between visual effects and live-action shots. This software-based solution will adapt easily to diverse broadcast environments without the addition of extra hardware. In conclusion, this project represents a significant advancement in virtual content integration for live broadcasts, offering a practical and efficient solution to enhance viewer experiences while setting a foundation for future innovations in the field

## 6 FUTURE WORK

While the project effectively deals with occlusion handling, camera vibration compensation, and real-time processing, it is always open for further modifications to increase the performance circle and expand its usability scope. Making the detection mechanism more robust for edge cases-a very large amount of motion blur, complicated occlusions, or light changes, for instance-will make a huge boost to the system. Further optimizations of the model can be done using pruning or hardware acceleration techniques which has great potential for making it computationally faster and easily scalable under any broadcasting scenario.

In future development, improvements can also be done to refine the overlay placement by incorporating adaptive blending methods for better visual realism. The ability of the system to expand on various sports, dynamic environments, such as augmented reality, film production, among others not necessarily relating to sport broadcasting, can add versatility to the system. This can render the system more powerful as well as be suitable for various real-time applications.

## REFERENCES

[1] S. Gugliermo, "Occlusion Handling in Augmented Reality Context", KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, Master's Thesis, 2019. [Online].
[2] Tian, Y.; Guan, T.; Wang, C. Real-Time Occlusion Handling in Augmented Reality Based on an Object Tracking Approach. Sensors 2010, 10, 2885-2900.
[3] Wikipedia contributors. (2022, August 11). Virtually Live. In Wikipedia, The Free Encyclopedia. Retrieved 20:37, December 10, 2024
[4] S. Biswas, A. Nandy, A. K. Naskar and R. Saw, "Real time Gesture Recognitionusing Improved YOLOv5 Model," 2024 11th International Conference on SignalProcessing and Integrated Networks (SPIN), Noida, India, 2024, pp. 328-333, doi:10.1109/SPIN60856.2024.10511787.
[5] Ping Ding and Yan Song, "Robust object tracking using color and depth images with a depth based occlusion handling and recovery," 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, 2015, pp. 930-935, doi: 10.1109/FSKD.2015.7382068.