CHAPTER 1

INDTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Hypothesis (Optional); (6) Assumption (Optional); (7) Scope and Delimitation; and (8) Importance of the study.

1.1 Rationale

YOLO (You Only Look Once) is a real-time object detection algorithm developed by Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi in 2016. YOLO has become one of the most widely used object detection models in both academic research and industry applications. Its ability to detect and localize objects within images and videos makes it ideal for real-time systems such as surveillance, autonomous vehicles, and traffic monitoring. However, despite its strengths, YOLO shows significant limitations when applied to video streams. Because YOLO performs detection independently on each frame without incorporating temporal context, it often produces inconsistent results manifesting as flickering detections or missed objects especially when dealing with fast-moving, partially occluded, or blurred targets.

This issue becomes especially critical in real-time applications where detection reliability directly impacts safety and decision-making. For instance, in surveillance and intelligent transportation systems, momentary detection failures may result in delayed responses to critical events. According to a report by Grand View Research (2023), the global video surveillance market was valued at USD 73.75 billion in 2022 and is projected to surpass USD 130 billion by 2030, driven by the increasing demand for AI-powered video analytics. In Southeast Asia, and particularly Indonesia, government initiatives such as Electronic Traffic Law Enforcement (ETLE) and smart city programs are accelerating the adoption of real-time video detection systems, underscoring the need for stable and reliable performance.

Recent studies have highlighted that deep learning models, particularly those trained solely on still images, often experience performance degradation when applied directly to video streams [1] demonstrated that even in visually static scenes—where no changes are perceptible to the human eyeobject detection accuracy may still fluctuate significantly across video frames. These inconsistencies are primarily caused by automatic adjustments made by the camera, such as exposure or white balance changes, which subtly alter pixel values between frames. Although these variations may seem negligible, they can adversely impact the stability of detections produced by image-trained models like YOLO. This phenomenon illustrates the vulnerability of such models to dynamic video conditions and highlights the need for strategies that improve temporal consistency.

Empirical observations in our experiments confirm that YOLO-based detectors often suffer from frame-to-frame inconsistencies, leading to flickering detections. These issues become more pronounced when tracking partially occluded or fast-moving objects. While the severity may vary across datasets and scenarios, it underscores the necessity of temporal smoothing mechanisms to ensure output stability.

To mitigate these issues, this study proposes an enhanced YOLO-based approach that integrates Kalman Filtering and Polling (majority voting) as lightweight post-processing modules. Kalman Filtering enables temporal continuity by predicting object motion between frames, while Polling reduces classification jitter by aggregating label decisions over a buffer window. This hybrid method improves detection stability without modifying the core YOLO architecture, maintaining its real-time performance.

To address these challenges, this study proposes an enhanced YOLO-based approach that integrates Kalman Filtering and Polling (majority voting) as post-processing modules. Kalman Filtering helps maintain temporal continuity by predicting object positions across frames, while Polling aggregates detection results over a buffer of frames to smooth out inconsistencies. This combined method improves detection stability without altering the YOLO architecture.

By tackling this critical limitation, the proposed approach aims to enhance the robustness of YOLO in real-time video applications bridging the gap between high-speed inference and temporal consistency, and contributing to safer, smarter, and more reliable video-based systems both globally and in the rapidly developing context of Indonesia.

1.2 Theoretical Framework

1.2.1 Object Detection and YOLO Algorithm

Object detection is a fundamental task in computer vision that involves identifying and localizing objects within images or video frames. Among various detection algorithms, YOLO (You Only Look Once) has gained prominence due to its ability to perform fast and accurate detections in a single pass through the neural network. YOLO processes images in real time, making it highly suitable for applications such as surveillance, autonomous driving, and robotics. However, despite its advantages, YOLO processes each frame independently, which can lead to instability and inconsistency in sequential video detections.

1.2.2 Challenges of Temporal Consistency in Video Object Detection

In video-based object detection, maintaining temporal consistency is crucial to avoid flickering and sudden changes in detected objects across consecutive frames. Traditional YOLO implementations treat each frame as an isolated input without accounting for information from previous frames. This limitation results in noisy and unstable detection outputs, which can degrade the performance and reliability of downstream applications requiring smooth object tracking and decision-making.

1.2.3 Temporal Integration and Polling Mechanisms

To overcome these challenges, temporal integration methods have been proposed that leverage information from multiple frames to enhance detection stability. Techniques such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks incorporate temporal context but often increase model complexity and computational cost. An alternative is the use of a polling mechanism, a lightweight post-processing method that aggregates detection outputs across a sliding window of frames to smooth inconsistencies without modifying the underlying detection model. This approach offers a balance between computational efficiency and improved temporal coherence.

1.2.4 Post-Processing in Object Detection

Post-processing techniques play a significant role in refining raw detection outputs to meet application-specific needs. Beyond non-maximum suppression (NMS), temporal polling provides an additional layer of consistency by considering detection histories, thereby reducing false positives and flickering. Such methods contribute to more reliable object detection pipelines, especially in real-time systems where quick and stable inference is critical.

1.2.5 Proposed Model: YOLO Integrated with Temporal Polling System

This research aims to address the instability in YOLO's video-based object detection by proposing an enhanced model that integrates the YOLO algorithm with both a Kalman Filtering and a Polling. The Kalman Filtering functions as a predictive tracking mechanism, estimating the next position of detected objects and calculating the Euclidean distance between the predicted and detected bounding boxes. Kalman Filtering predicts similar data in subsequent frames based on the closeness of the distance. Meanwhile, the Polling acts as a lightweight post-processing module, aggregating detection results across consecutive frames to improve stability and reduce sudden fluctuations in detection outputs.

By leveraging both Kalman Filtering and Polling, the proposed model enhances detection reliability in video sequences without requiring modifications to the core YOLO architecture or additional training. This integration strikes a balance between maintaining real-time performance and improving temporal consistency, making the model ideal for applications that require both speed and stable detection.

1.3 Conceptual Framework/Paradigm

1.3.1 Identification of Variables

In this research, several key variables are identified that are crucial for improving the stability and accuracy of object detection in video sequences using YOLO and a polling system:

Independent Variables:

1. Animal Video:
   Raw input data in the form of sequential frames extracted from animal videos, serving as the test environment for object detection.
2. YOLO (You Only Look Once):
   A lightweight convolutional neural network used as the base object detection model, responsible for detecting and localizing objects within each video frame.
3. Kalman Filtering:
   A predictive tracking mechanism used to estimate the future position of detected objects. This mechanism matches the predicted bounding box with the currently detected bounding box by calculating the Euclidean distance. The class with the smallest distance will be considered the matching object and assigned to the relevant data category.
4. Euclidean Distance:
   Euclidean distance is a mathematical measure used to calculate the straight-line distance between two points in Euclidean space. It is one of the most commonly used distance metrics in geometry, data analysis, and computer vision.
5. Polling System:
   A temporal post-processing mechanism that aggregates detection outputs across multiple frames to improve the stability and consistency of YOLO's results.

Dependent Variables:

1. Detection Stability:
   The output measure that reflects the consistency of detected objects across video

frames. This is crucial for ensuring that the same object is consistently tracked and classified across multiple frames without flickering or mismatches.

2. Frames Per Second (FPS):
   The measure of the system's processing speed, indicating how many frames are processed per second. This is critical for real-time applications, balancing the accuracy of detection with the system's ability to maintain fast performance.

3. mAP (mean Average Precision):
   mAP is a performance metric used to evaluate the overall accuracy of object detection models by averaging the precision across all classes. It considers both the precision and recall of the model's predictions, providing a balanced measure of accuracy. A higher mAP indicates better model performance in detecting and classifying objects.

4. Precision:
   Precision measures the accuracy of the model's detections by calculating the ratio of true positives (correct detections) to the total number of predicted positives (true positives + false positives). Higher precision means the model makes fewer false positive errors, ensuring accurate object detection.

5. Recall:
   Recall measures the model's ability to identify all relevant objects in the dataset. It is calculated as the ratio of true positives (correct detections) to the total number of actual objects (true positives + false negatives). Higher recall means the model successfully detects more of the relevant objects, even at the cost of some false positives.

6. F1-Score:
   F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both false positives and false negatives. It is especially useful when you need a balance between precision and recall. A higher F1-Score indicates better overall performance of the detection system, particularly in scenarios where both false positives and false negatives are critical.

Moderating Variables:

1. Video Quality: The resolution, lighting, and clarity of the input video frames, which can affect detection performance.

2. Computational Resources: The available hardware and software infrastructure influencing model inference speed and feasibility of real-time processing.

3. Object Motion: The speed and movement patterns of objects in the video, which may impact detection consistency.

1.3.2 Discussion of Relationships

Animal Video:
Serves as the primary input data for the object detection system. The quality and temporal continuity of video frames are crucial for effective feature extraction and temporal polling,

Yolo:
Functions as the core detection model, processing each frame independently to detect and localize objects. Its real-time performance is ideal for video analysis, but it lacks temporal awareness between frames

Polling System:
Enhances YOLO's outputs by aggregating detection results across multiple frames, smoothing out sudden fluctuations and inconsistencies. This temporal integration improves detection stability without altering YOLO's underlying architecture

Detection Stability and Accuracy:
The effectiveness of the combined YOLO and polling system approach is measured by how consistently and correctly objects are detected across frames. Improved stability reduces flickering and false positives, benefiting downstream applications such as tracking or behavior analysis.

Moderating Variables:

1. Video Quality: Higher quality videos provide clearer features for YOLO, aiding in accurate detection and improving the efficacy of the polling system.

2. Computational Resources: Sufficient resources are necessary to process frames and perform polling in real time, especially for high-resolution or high-frame-rate videos.

3. Object Motion: Rapid or erratic movement can challenge both YOLO and polling systems, potentially causing reduced detection stability.

1.4 Statement of the Problem

Real-time object detection in video sequences is critical for many applications, including surveillance, autonomous driving, and robotics. YOLO (You Only Look Once), a popular deep learning-based object detection algorithm, is widely used due to its high speed and accuracy. Despite YOLO's strengths in speed and accuracy, its frame-independent inference often causes unstable detections in video sequences, such as flickering and misclassification, which negatively affect real-time performance. There is a lack of empirical analysis on the extent of this instability and how post-processing mechanisms such as Kalman Filtering and Polling Blocks can mitigate these effects. This study explores how YOLO's architecture affects detection stability in video, and how adding post-processing methods like Kalman Filtering and Polling can make the results more stable without slowing down performance.

1.5 Objective

The objective of this study is to enhance the temporal stability of object detection results in video streams by integrating Kalman Filtering and Polling mechanisms into the YOLO object detection framework. This approach aims to reduce detection flickering, improve tracking consistency, and maintain high detection accuracy while balancing real-time performance.

1.6 Assumption

This research is based on the following assumptions:

1. **Frame Continuity Assumption**
   The video input consists of sequential frames with a reasonable frame rate (e.g., 15 FPS), such that temporal coherence between adjacent frames exists. The polling system relies on detection consistency across neighboring frames. If the frame rate is too low or the scene changes drastically, polling becomes less effective.

2. **Stationary or Moderately Moving Camera**
   The system assumes that the camera is either stationary or moves smoothly and predictably, avoiding abrupt scene transitions or camera shakes. Drastic changes in the field of view between frames would make it hard to match object positions across time, reducing the effectiveness of polling.

3. **Objects Appear Over Multiple Frames**
   Objects of interest remain in the video for at least several consecutive frames, allowing the polling system to observe and confirm their presence. Why: If objects appear for only one frame, the polling window may not collect enough data to stabilize the detection.

4. **YOLO's Detection Output is Sufficiently Accurate on Single Frames**
   The core YOLO model is reasonably accurate per frame, so that the polling system is aggregating mostly correct detections, not amplifying noise. The polling mechanism assumes that YOLO has a good baseline performance and focuses on temporal refinement, not fixing consistently wrong outputs.

5. **Object Motion Between Frames is Limited**
   Object movement between consecutive frames is assumed to be gradual, so the same object can be tracked and matched based on bounding box proximity or centroid. If objects jump too far from one frame to the next, polling may incorrectly associate or miss them.

6. **Polling Window Size is Properly Tuned**
   The number of frames used in the polling system (window size) is assumed to be appropriate for the scene's motion dynamics and frame rate. A window that is too short may not smooth out noise, while one that is too long could delay or suppress real-time responsiveness.

1.7 Scope and Delimitation

1.7.1 Scope

Principal Variables:

1. Independent Variables: The study focuses on video input, utilizing the YOLO object detection model, and integrates Kalman Filtering and Polling as temporal post-processing techniques to enhance detection stability. This approach leverages a custom YOLO model that has been specifically trained to detect cats, dogs, and bears. The combination of these methods aims to improve the accuracy and consistency of detecting these animals in dynamic video environments, reducing issues such as flickering or false positives that are commonly encountered in object detection tasks.

2. Dependent Variable: The primary outcome measured is the object detection tracking stability across consecutive video frames.

3. Moderating Variables: Factors such as video quality , computational resources (hardware performance), and object motion dynamics are considered to influence the detection performance and stability.

Locale:

This research utilizes image datasets from publicly available databases of cat, dog and bear. No specific geographical region is targeted for data collection beyond these public datasets.

Timeframe:

The research is conducted over a period of one year. This includes data collection, preprocessing, model development, training, evaluation, and analysis.

Justification:

1. Selection of Variables: The choice of independent variables such as YOLO is justified by their proven effectiveness in image classification tasks. The Polling System is also commonly used in decision making.

2. Locale: The use of publicly available databases ensures a diverse and comprehensive dataset, which is essential for developing a robust model. It also allows the study to be reproducible by other researchers.

3. Timeframe: One year period is deemed sufficient to carry out the various phases of the research, from initial data handling to final analysis, ensuring thorough and detailed investigation within a reasonable duration.

1.7.2 Delimitation

Limitation to YOLO Architecture:
This implementation is limited to YOLOv5. Other object detection algorithms, such as Faster R-CNN, SSD, or EfficientDet, are not included or compared in this study.

Limitation of Post-Processing Techniques:
This research exclusively employs the Kalman Filtering and Polling as post-processing techniques applied to the outputs of the YOLO object detection model.

Limitation of performance measurement
This research focuses solely on evaluating detection stability and improvements in precision, recall, F1-score, and mAP, and does not include tracking performance. **Limited Dataset/model Scope:**
The dataset utilized in this study consists of a custom dataset specifically created for the research. Unlike the default dataset that comes pre-packaged with the YOLO model, the custom dataset was built specifically for this research.

Hardware and Resource Constraints:
This research utilizes a CPU alongside an NVIDIA GTX 1070 GPU to conduct model training and inference. The GTX 1070, classified as a mid-range graphics processing unit, offers adequate computational capabilities to efficiently support the deep learning workloads associated with running the YOLO.

1.8 Significance of the Study

This study aims to contribute to the field of computer vision and object detection by proposing an enhanced YOLO algorithm integrated with a temporal polling mechanism to improve detection stability in video sequences. The significance of this study is multi-faceted, impacting both academic research and practical applications in real-time object detection systems.

1.8.1 Contributions as New Knowledge

Improved Detection Stability:

By incorporating a temporal polling system, the study introduces a novel post-processing approach that enhances the temporal consistency of YOLO detection outputs. This reduces frame-to-frame fluctuations and flickering, leading to more stable and reliable object detection in videos without modifying the core YOLO architecture.

Lightweight and Modular Enhancement:

The polling mechanism acts as a computationally efficient extension, enabling existing YOLO models to achieve better temporal coherence without the need for retraining or complex temporal models like RNNs or LSTMs. This makes it practical for real-world deployment in resource-constrained environments.

Broader Applicability in Video-Based Systems:

The proposed method benefits various applications including surveillance, autonomous driving, robotics, and traffic monitoring, where consistent object detection over time is critical for safety and performance. The research advances knowledge on how to improve video-based object detection through simple yet effective temporal integration techniques.

1.8.2 Usefulness to Specific Groups

Professionals:

Computer vision engineers, AI developers, and system integrators can benefit from the improved temporal stability of object detection results provided by the polling-enhanced YOLO system. This enhancement helps create more reliable and robust detection pipelines, which are essential for real-time applications such as surveillance, autonomous driving, and robotics. The modular nature of the polling system allows professionals to integrate this technique into existing workflows without extensive changes to the core model.

Researchers and Academics:

This study offers a novel yet practical approach to addressing temporal inconsistency in object detection, contributing valuable insights to the fields of computer vision and machine learning. Researchers can build upon this work to explore further improvements in post-processing techniques or combine polling with other temporal models, advancing the understanding of how to enhance video-based detection systems.

Industry and Technology Providers:

Companies developing video analytics, security systems, and autonomous solutions can implement the proposed polling mechanism to increase the reliability of object detection in their products. Improved detection stability can enhance user trust, reduce false alarms, and support better decision-making in safety-critical environments, providing competitive advantages in the market.

End-Users and Society:

End-users of applications relying on object detection, such as smart city surveillance, traffic monitoring, or automated vehicles, will benefit from smoother, more consistent detection results. This leads to safer and more efficient operation of these systems, which can positively impact public safety, traffic management, and overall quality of life.