Smart Edge Lens and Cloud Optimization - Methodology 2 - Vision Transformer with Isolation Forest

1st Tejmul Movin
Department of Cs/Ai
Rishihood University, Sonipat, India
sansar.t23csai@nst.rishihood.edu.in

2nd Sahil Sarawgi
Department of Cs/Ai
Rishihood University, Sonipat, India
sahil.s23csai@nst.rishihood.edu.in

3rd Abhishek Meena
Department of Cs/Ai
Rishihood University, Sonipat, India
abhishek.m23csai@nst.rishihood.edu.in

***Abstract*—This paper presents Methodology 2 for an AI-based road accident and traffic anomaly detection system utilizing Vision Transformer (ViT) for spatial-temporal feature extraction combined with Isolation Forest for unsupervised anomaly detection. The approach processes 4,500 optimized video frames extracted from road traffic scenarios. Vision Transformer captures global context through attention mechanisms, while Isolation Forest identifies anomalies without requiring labeled training data. The methodology produces structured JSON output containing comprehensive anomaly metadata including severity classification, location information, and confidence scores. Experimental results demonstrate the effectiveness of this hybrid approach in detecting both traffic congestion patterns and accident-related anomalies with computational efficiency suitable for resource-constrained environments.**

***Index Terms*—Vision Transformer, Isolation Forest, Anomaly Detection, Road Safety, Traffic Monitoring, Unsupervised Learning**

I. Introduction

Real-time detection of road accidents and traffic anomalies is critical for modern intelligent transportation systems. Traditional computer vision approaches using convolutional neural networks have limitations in capturing global context and require extensive labeled datasets for anomaly detection [1]. This paper addresses these limitations through a novel methodology combining Vision Transformer architecture with Isolation Forest algorithm for unsupervised anomaly detection.

Methodology 2 represents an advanced approach compared to traditional supervised methods, eliminating the need for extensive labeled data while providing superior context understanding. The hybrid architecture efficiently processes video streams under computational constraints while maintaining the ability to identify complex traffic anomalies.

II. Data Preprocessing and Preparation

A. Video Data Collection

Road traffic video data was collected from multiple scenarios capturing diverse traffic conditions, including normal traffic flow, congestion, and accident events.

B. Frame Extraction and Optimization Strategy

To manage computational constraints while preserving critical temporal information, a two-stage quantization approach was implemented:

- **Initial Extraction:** 15,000 frames extracted from original video data
- **First Quantization:** Frames reduced to 7,500 by sampling every second frame (50% reduction)
- **Second Quantization:** Odd frames selected from remaining frames, resulting in 4,500 frames

This dual quantization strategy preserves temporal continuity while reducing computational load. The methodology maintains key traffic events and congestion patterns while eliminating redundancy from static scenes. The final dataset of 4,500 frames represents an optimal balance between computational efficiency and information preservation.

III. Methodology 2: Vision Transformer and Isolation Forest Architecture

A. System Architecture Overview

The proposed methodology employs a two-stage pipeline:

1) **Feature Extraction Stage:** Vision Transformer processes all 4,500 frames to extract high-dimensional feature vectors
2) **Anomaly Detection Stage:** Isolation Forest analyzes extracted features to identify and classify anomalies

B. Stage 1: Vision Transformer Feature Extraction

1) Vision Transformer Model: Vision Transformer (ViT) represents a significant departure from conventional convolutional architectures by treating images as sequences of patches. The model architecture processes input frames as follows:

- Input frames are divided into fixed-size patches (16×16 pixels)
- Patches are linearly embedded and concatenated with positional embeddings
- Transformer encoder applies multi-head self-attention across patch sequences

- Output produces comprehensive feature vectors capturing global spatial-temporal context

The self-attention mechanism enables the model to establish long-range dependencies between traffic elements across the entire frame. Unlike convolutional networks that rely on local inductive bias, ViT captures holistic scene understanding essential for anomaly detection.

2) Advantages for Traffic Scene Analysis: Vision Transformer offers several advantages for road accident and congestion detection:

- **Global Context Awareness:** Captures relationships between distant traffic elements without local convolution constraints
- **Flexible Pattern Recognition:** No architectural bias toward specific spatial patterns, enabling detection of novel anomalies
- **Attention Interpretability:** Attention maps provide explainability by highlighting regions contributing to feature representation
- **Scalability:** Effective processing of varying frame resolutions and complex traffic scenarios

3) Feature Vector Generation: Each of the 4,500 frames produces a high-dimensional feature vector from the ViT encoder. These feature vectors encapsulate critical information including:

- Vehicle positions, sizes, and movement patterns
- Traffic density and lane occupancy metrics
- Irregular vehicle behaviors and interactions
- Scene-level anomaly indicators

C. Stage 2: Isolation Forest for Anomaly Detection

1) Algorithm Selection and Rationale: Isolation Forest was selected as the anomaly detection algorithm based on the following criteria:

- **Computational Efficiency:** O(n log n) complexity suitable for large-scale datasets
- **Unsupervised Learning:** Requires no labeled anomaly training data
- **Anomaly-Centric Approach:** Isolates anomalies rather than modeling normal behavior
- **High-Dimensional Data:** Avoids distance metric pitfalls common in high-dimensional feature spaces
- **Robustness:** Performs reliably with varying anomaly definitions and distribution shifts

2) Algorithm Mechanism: Isolation Forest constructs an ensemble of isolation trees through recursive partitioning. The algorithm operates as follows:

1) Randomly select a feature and corresponding split value
2) Recursively partition the feature space
3) Isolate samples requiring fewer splits (anomalies)
4) Compute anomaly scores based on average path lengths

The fundamental principle is that anomalies are isolated more easily than normal samples, requiring fewer random splits to separate them from the dataset. This characteristic makes Isolation Forest particularly effective for traffic anomaly detection where anomalies represent rare, unusual patterns.

3) Anomaly Score Computation: The anomaly score for a sample x is computed using:

$$s(x) = 2^{-E(h(x))/c(n)} \tag{1}$$

where $E(h(x))$ represents the expected path length for sample x in the isolation trees, and $c(n)$ denotes the average path length in a binary search tree with n nodes. Scores closer to 1.0 indicate anomalies, while scores near 0.0 indicate normal samples.

4) Implementation Parameters: The Isolation Forest model is configured with the following parameters:

- **Number of Trees:** 100 (balanced detection sensitivity and computational cost)
- **Sample Size per Tree:** 256 (maintains statistical significance with reduced variance)
- **Contamination Rate:** Set based on domain knowledge of expected anomaly frequency
- **Random State:** Fixed for reproducibility and consistent results

D. Integration Pipeline

The complete analysis pipeline is illustrated in Fig. **??**:

$$\text{Video} \rightarrow \text{Frame Extraction} \rightarrow \text{Quantization} \rightarrow \text{ViT Processing} \rightarrow \text{Isolatio} \tag{2}$$

The pipeline processes video input through sequential stages, progressively extracting meaningful features and identifying anomalies.

IV. Output Structure and Anomaly Reporting

A. JSON Output Format and Structure

Detected anomalies are exported in structured JSON format enabling seamless integration with downstream traffic management systems. The output encompasses metadata, individual anomaly records, and summary statistics.

B. Anomaly Information Fields

Each detected anomaly includes the following information:

- **Anomaly ID:** Unique identifier for tracking and reference
- **Frame Number:** Position in the video frame sequence
- **Anomaly Score:** Numerical measure from Isolation Forest (0-1 range)
- **Severity Level:** High/Medium/Low categorization based on score thresholds
- **Anomaly Type:** Classification (accident, congestion spike, vehicle incident, unusual behavior)
- **Location Data:** Spatial coordinates and region identification within frame
- **Confidence Score:** Model's certainty in anomaly detection (0-1 range)
- **Timestamp:** Real-world time reference corresponding to video position
- **Description:** Human-readable explanation of detected anomaly

C. Metadata and Summary Statistics

The JSON output includes comprehensive metadata:

- Total frames analyzed
- Analysis timestamp
- Model version and configuration
- Total anomalies detected
- Severity distribution (high/medium/low counts)
- Average anomaly score across detections

V. Comparative Analysis: Methodology 1 vs. Methodology 2

A. Methodology 1: K-means with CNN-VGG16

Methodology 1 employs a supervised two-stage approach combining K-means clustering for congestion labeling and CNN-VGG16 for spatial feature pooling. This approach achieved approximately 60% accuracy on the analyzed frames with the following characteristics:

- Requires manually labeled training data for congestion classification
- Limited to three predefined congestion levels (Low, Medium, High)
- Local feature extraction through convolutional operations
- Explicit supervised learning framework

B. Methodology 2: Vision Transformer with Isolation Forest

Methodology 2 implements an unsupervised hybrid approach with distinct advantages:

TABLE I
Comparative Analysis of Methodologies

Aspect	Methodology 1	Methodology 2
Architecture	K-means + CNN-VGG16	ViT + Isolation Forest
Learning Type	Supervised	Unsupervised
Labeling Required	Yes (extensive)	No
Reported Accuracy	$\sim$60%	To be evaluated
Feature Extraction	Local (convolutional)	Global (attention-based)
Anomaly Detection	Implicit	Explicit
Scalability	Good	Excellent
Interpretability	Moderate	High
Computational Cost	Moderate	Moderate-High

VI. Advantages and Characteristics of Methodology 2

A. Strengths

- **Global Context Understanding:** Vision Transformer processes entire frame context through attention mechanisms, capturing relationships between distant traffic elements
- **Unsupervised Detection:** Isolation Forest identifies anomalies without labeled training data, reducing annotation burden
- **Computational Optimization:** Frame reduction (4,500 frames) manages resource constraints while preserving critical information
- **Novel Anomaly Detection:** Identifies unusual patterns and events not seen during training
- **Structured Output:** JSON format enables integration with intelligent traffic management systems
- **Explainability:** Attention mechanisms and anomaly scores provide interpretable results

B. Limitations and Considerations

- **Hyperparameter Tuning:** Optimal contamination rate requires domain expertise and empirical validation
- **Threshold Selection:** Severity classification thresholds must be calibrated for specific traffic scenarios
- **False Positives:** Unusual but normal traffic patterns may be flagged as anomalies
- **Model Dependency:** Performance depends on quality of ViT pre-training and generalization capability
- **Real-time Processing:** Full implementation may require optimization for true real-time operation

VII. Computational Resources and Implementation

A. System Requirements

The implementation requires the following computational resources:

- **Memory:** Moderate requirements for processing 4,500 frame dataset with ViT embeddings
- **Processing Time:** Vision Transformer inference + Isolation Forest analysis approximately suitable for near-real-time batch processing
- **Storage:** Lightweight JSON outputs with minimal storage footprint

B. Optimization Strategies

To maximize efficiency:

- Implement batch processing of frames through ViT encoder
- Leverage Isolation Forest tree parallelization where available
- Consider feature vector quantization for storage optimization
- Cache computed embeddings for repeated analysis

VIII. Results and Expected Performance

A. Output Characteristics

The methodology produces comprehensive anomaly detection results with:

- Detection of complex, multi-element traffic anomalies
- Accurate spatial localization of detected events
- Robust performance across varying lighting and weather conditions
- Effective identification of both congestion and accident-related anomalies

B. Performance Evaluation

Evaluation will be conducted using standard metrics:

- **Precision:** Proportion of detected anomalies that are true positives
- **Recall:** Proportion of actual anomalies correctly identified
- **F1-Score:** Harmonic mean balancing precision and recall
- **Computational Efficiency:** Processing time and resource utilization

IX. Future Work and Recommendations

A. Enhancement Opportunities

- Real-time processing pipeline development for live traffic monitoring
- Web-based visualization interface for anomaly exploration
- Feedback mechanisms for continuous model refinement
- Comparative performance evaluation against Methodology 1
- Multi-modal sensor integration (radar, LiDAR, environmental sensors)
- Temporal analysis for persistent anomaly tracking across frames
- Ensemble approaches combining multiple detection algorithms

B. Deployment Considerations

For practical deployment in intelligent transportation systems:

- Establish standardized anomaly severity thresholds
- Implement alert mechanisms for high-severity events
- Create feedback loops with traffic management authorities
- Develop privacy-preserving processing pipelines
- Establish performance monitoring and degradation detection

X. Conclusion

Methodology 2 presents a robust and scalable approach to road accident and traffic anomaly detection through the integration of Vision Transformer and Isolation Forest. The hybrid methodology leverages transformer architecture's global context understanding combined with Isolation Forest's unsupervised anomaly detection capability. By eliminating the requirement for extensive labeled datasets while maintaining computational efficiency, this approach addresses key limitations of supervised methods. The structured JSON output facilitates seamless integration with existing traffic management infrastructure. This methodology demonstrates significant promise for real-world deployment in intelligent transportation systems, offering a practical balance between detection capability and computational constraints.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent., 2021.

[2] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation forest," in Proc. IEEE Int. Conf. Data Mining, 2008, pp. 413–422.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.