
emerald.dev

From Keyframes to Narrative

A Multi-Stage AI Pipeline for Scene Understanding Using Object Detection and Large Language Models

MICAI-2025

GENERATIVE

GENERATIVE

TECHNOLOGIES

TECHNOLOGIES

TECHNOLOGIES

Institutions

- Emerald (ED)
- Tecnológico Superior de Purísima del Rincón (ITSPR)

Authors

- Valentín Calzada Ledesma (ITSPR)
- Victor Manuel Castillo Olivetto (ITSPR)
- Emmanuel Márquez Robles (ITSPR)
- Gabriela Alva Zavala (ED)
- Carlos Orozco Solis (ED)
- Faustino Neri Larios (ED)

Contents


CONNECTED
SYSTEMS
CONNECTED
SYSTEMS

Abstract

- We introduce a novel pipeline for real-time semantic scene understanding from video.
- The system efficiently selects representative frames, detects objects, and enriches each detection with semantic and contextual information.
- A narrative engine translates visual data into actionable summaries and alerts, supporting diverse applications in security, health, and IoT environments.

Our Goal

We aim to transform our **research** into an innovative **product** capable of analyzing video streams in near real time, automatically generating actionable narratives, key indicators, and timely alerts to support decision-making across diverse application areas.

Related Work

- Multi-object tracking in traffic environments: A systematic literature review.
- Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey.
- A Review of Deep Learning Applications for the Next Generation of Cognitive Networks.
- Temporal Scene Understanding using Contextually Unique Identification.

Methodology

- Preprocess video using grayscale conversion, downsampling, and smoothing to optimize performance.
- Select informative keyframes using motion-based analysis.
- Apply object detection models to extract entities and features from each keyframe.
- Generate a Visual Semantic Field (VSF) by adding semantic metadata and contextual relationships to each detection.
- Employ a Visual Narrative Engine based on a Large Language Model to interpret the VSF and produce narratives, indicators, and alerts.

Methodology

Keyframe Selection

1. Load the video and **reduce** each frame's size, convert it to **grayscale**, blur it, and skip frames to **speed up** processing.
2. For each **frame**, calculate **how much it changes** compared to the previous frame using a chosen method (like pixel differences or histogram comparison).
3. Normalize the change scores, then select the **frames** with the **highest scores**, making sure selected frames are not too close together in time.

Methodology

Object Detection

YOLO (You Only Look Once) is a real-time **object detection** system that processes the entire image with a **single neural network** pass.
It quickly identifies and classifies **multiple objects**, outputting both their locations and **categories** simultaneously.

Unlike traditional methods that scan images in parts, YOLO treats object detection as a **single regression** problem.
This enables fast and accurate predictions, making it suitable for tasks needing **real-time** results.

Methodology

Visual Semantic Field (VSF)

We define the VSF as a **structured scene** representation where each detected object is enriched with **categorical** semantic metadata derived from specialized models and geometric algorithms. In addition to attributes such as demographic features, sentiment, or identity for faces—and brand or physical features for various objects—the VSF encodes **contextual relationships**, including object neighborhoods and mutual positioning. All **features** are expressed categorically, enabling interpretable, **context-aware** scene understanding and supporting advanced reasoning across a wide range of object classes.

Methodology

Visual Narrative Engine (VNE)

We define the Visual Narrative Engine (VNE) as an intelligent agent that uses the structured outputs of the VSF to generate contextual narratives and actions within visual environments. The VNE, powered by large language models, continuously interprets and tracks events, responds to user queries, analyzes behaviors, searches for identities, and produces summaries or video reports, adapting its focus to the needs of each application—such as security, monitoring, or general IoT. By processing the history and structure of the VSF, the VNE enables real-time, context-aware visual understanding and decision-making across diverse deployment scenarios.


INTELLIGENCE
DATA
INTELLIGENCE
DATA
INTELLIGENCE
DATA

Experiments & Results

...

Conclusions

- We introduced a modular system for real-time video analysis and semantic scene understanding.
- The pipeline combines keyframe selection, object detection, semantic enrichment (VSF), and an LLM-powered narrative engine.
- Our approach enables automated event detection and interpretable reporting across multiple domains.
- Future work will expand specialist models and further adapt the system to new application areas.


emerald.dev