IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 May 2024, accepted 4 June 2024, date of publication 6 June 2024, date of current version 14 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3410894

RESEARCH ARTICLE

Semantic Analysis System to Recognize Moving Objects by Using a Deep Learning Model

EMAD IBRAHIM[1], NIZAR ZAGHDEN[2], AND MAHMOUD MEJDOUB[3]
[1]National School of Electronics and Telecommunications of Sfax, University of Sfax, Sfax 3029, Tunisia
[2]Higher School of Business of Sfax, University of Sfax, Sfax 3029, Tunisia
[3]Faculty of Sciences of Sfax, University of Sfax, Sfax 3029, Tunisia

Corresponding author: Emad Ibrahim (emadmah236@gmail.com)

ABSTRACT This study focuses on enhancing the accuracy and efficiency of semantic analysis systems for recognizing moving objects within video sequences. The primary aim is to improve object detection capabilities in dynamic environments using a hybrid model that integrates Convolutional Neural Networks (CNNs) with Support Vector Machines (SVMs). Our contribution involves developing and testing an advanced detection algorithm that utilizes the Faster Region-based Convolutional Neural Network (R-CNN) framework combined with SVM classifiers for refined object recognition and interaction assessment in complex video scenes. We implemented the system using Python 3.7 and tested it on approximately 350 video frames. The findings demonstrate that our model significantly outperforms existing methods such as Scale-Invariant Feature Transform (SIFT), Centrifugal Compressor Performance (CCP), and Local Binary Pattern (LBP) in terms of detection accuracy. The proposed model consistently outperformed traditional methods such as SIFT, CCP, and LBP across various noise levels, maintaining higher accuracy, particularly in high-noise environments. At 80% noise, the proposed model demonstrated a marked advantage in detection accuracy compared to the baseline methods. Overall, the model showcased robust performance with less degradation in accuracy even under significant processing errors, validating its effectiveness in noisy and dynamic settings.

INDEX TERMS R-CNN algorithm, deep learning, semantic analysis, SVM classifier, synthesis technique.

I. INTRODUCTION

Detecting and recognizing moving objects is a critical area of study within computer vision [1], serving key functions in applications like intelligent video surveillance [2], [3], robotic vision navigation [4], virtual reality [5], and tracking cellular states in medical diagnostics [6]. The rise of unmanned aerial vehicles (UAVs) has heightened interest in this research due to their ability to capture video sequences [7]. UAVs, equipped with cameras capable of operating under varying degrees of movement and autonomy, face challenges such as motion blur and a dynamically moving background. Furthermore, outdoor environments introduce additional complexities like variable lighting, occlusions, and shadows, which can alter the appearance of moving objects and impact the accuracy of detection. Hence, there is a significant need for robust methods in motion detection and recognition. Unlike detection algorithms designed for static cameras, which include techniques like optical flow [8], inter-frame difference [9], and background modeling [10], the use of moving cameras often results in backgrounds that experience rotation, translation, and scaling.

By using semantic analysis, a computer can decipher the meaning or concept contained in any image or video, and then convert it into objects to complete its narrative. Developing an architecture for the semantic analysis of video sequences is one of the major obstacles to automating this process and analyzing specific videos efficiently. In applications such as search engines, video recommenders, and video summarizers, ontology methods, and approaches are crucial for finding the best answers. The concept of semantic analysis applies to the characteristics of videos and their associations, which include the colors, edges, and arcs. By utilizing

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

computer equipment to extract accurate information, this study aims to address the problem of semantic video analysis and processing. Semantic analysis creates a model that can diagnose moving objects and indicate their conditions, such as speed, emotions, and actions of people, as observed in real-time

In addressing the challenges of moving object detection and recognition, especially in dynamic environments like those encountered by UAVs, the adoption of advanced deep learning models such as Region-based Convolutional Neural Networks (R-CNN) offers a promising solution [11], [12]. R-CNN, known for its precision in detecting and classifying objects within images, can be particularly effective in scenarios where the background is constantly changing due to the motion of the camera. Convolutional Neural Networks (CNNs) and Region-based Convolutional Neural Networks (R-CNNs) play crucial roles due to their powerful feature extraction and object classification capabilities [13]. CNNs are deep learning architectures designed to automatically and adaptively learn spatial hierarchies of features through backpropagation. These networks are highly efficient at processing data that comes in the form of multiple arrays, such as images, where each spatial context of the pixel is relevant for understanding the content. This makes CNNs particularly useful for tasks like image classification, where identifying the presence of objects based on textural and boundary features is required.

Extending the capabilities of CNNs, R-CNN is used to accurately segment and identify objects despite the complications of motion blur, lighting variations, and occlusions that are typical in UAV-captured video footage. This model leverages a selective search to generate region proposals, which are then classified by convolutional neural networks, ensuring that even under significant environmental transformations, the detection and tracking of moving objects maintain high levels of accuracy. Therefore, integrating R-CNN into UAV systems could significantly enhance their capability for reliable surveillance and monitoring tasks in complex visual contexts.

The Fast R-CNN is similar to the R-CNN in its approach to object detection. Feature identification and regression are performed by a detector together with a trained Region Proposal Network (RPN) [14]. Data from the convolutional network for the full image is shared with the detection network. The RPN is a fully convolutional network that can generate better region suggestions after it undergoes end-to-end training. Object quality scores at each point are predicted simultaneously with the limits of the objects. Finally, a prediction of the bounding box of each image is generated. In response, two layers were created for bounding box regression and classification that receive region suggestions derived from the feature map. Detection speed and training time are the main drawbacks of R-CNN. The Faster R-CNN improves both the detection and training times. It is still important to keep in mind that the Faster R-CNN relies on the selective search to provide area suggestions, which could negatively impact its efficacy [15].

The Faster R-CNN uses shared convolution layers between RPNs and detectors, resulting in considerable savings in computing. Researchers have previously applied faster R-CNN to traffic sign detection. The detection of textures [16], the detection of moving vehicles in real-time [17], and the detection of pedestrians [18] are also included. As Figure 1 illustrates, the main goal of CNN is to extract meaning from an image. The image is analyzed using a variety of convolution kernels that are applied to numerous convolution layers. Due to the hierarchical structure of these kernels, CNN captures lower-level semantic information like edges and textures at its earliest levels, while higher-level semantic elements like objects, components, and shapes are captured in later layers. Using the hierarchical features obtained, RPN and Faster R-CNN detectors both work. As a result of these layers performing calculations for both jobs, processing time and memory usage are reduced significantly. To the best of our knowledge, to date, there are no many studies that use Faster R-CNN in this context.

In contrast, single-stage detectors process object detection in one step, eliminating the need for a separate region proposal phase. Examples of such detectors include SSD (Single Shot Multibox Detector), various YOLO (You Only Look Once) versions, RefineDet++, DSSD (Deconvolution Single Shot Detector), and RetinaNet. YOLOv1, introduced in 2016, was a significant development in single-shot object detection. Drawing inspiration from the GoogLeNet architecture [19], YOLOv1 replaced GoogLeNet inception modules with a combination of (1×1) and (3×3) convolutional filters. This model was evaluated using the VOC Pascal Dataset for the years 2007 and 2012 [20], and utilized the Darknet framework for training. It featured 24 convolutional layers, only four of which included max-pooling, and highlighted (1×1) convolutions and global average pooling as key features. Initially trained on the ImageNet dataset [21], the model underwent further fine-tuning by incorporating four additional convolutional layers and two fully connected layers with newly initialized weights. It used a Leaky Rectified Linear Unit (LReLU) for activation, except in the final layer which used a linear activation function. Despite its innovative approach, YOLOv1 had issues with large localization errors and a lower recall compared to two-stage detectors.

YOLOv2 [22], inspired by the popular VGG architecture, utilized the darknet-19 framework with 19 convolutional layers and 5 max pooling layers. YOLOv3 [23] aimed to address earlier weaknesses by improving localization errors and detection efficiency, especially for smaller objects. Tested on the COCO dataset [24], YOLOv3 showed enhanced capability in detecting small objects, though it struggled with medium and large objects. Based on the Darknet-53 framework, it included 53 convolutional layers and employed 3×3 and 1×1 convolutional filters along with

skip connections. Notably, the Darknet-53 framework was twice as fast as ResNet-152 [25].

YOLOv4 [26] introduced numerous advanced techniques, establishing it as a quicker and more accurate detector suitable for production environments. It incorporated initial image processing, feature extraction through powerful networks like VGG16 [15], Darknet53, and ResNet50, and feature scaling through structures like Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) [27]. It also combined elements of single-stage and two-stage detectors for prediction. During their experiments, the creators favored CSPDarknet53, featuring 29 convolution layers with 3×3 filters and approximately 27.6 million parameters, using Cross-stage partial connections (CSP) to improve gradient combination efficiency at a reduced computational cost.

YOLOv5 [28], marking a shift to the PyTorch framework from Darknet, retained many of YOLOv4 enhancements while introducing significant changes. It began with a strided convolution layer to reduce memory and computational demands, followed by layers that extracted relevant image features. The SPPF (spatial pyramid pooling fast) layer and additional convolutional stages processed features at various scales, while upsampling layers improved the resolution of feature maps. The SPPF layer sped up computations by pooling features from different scales into a fixed-size feature map, and each convolutional layer was paired with batch normalization (BN) [29] and SiLU activation.

Introduced by the Meituan Vision AI Department in September 2022, YOLOv6 [30] featured innovative elements like RepVGG/CSPStackRep blocks, a PAN neck, and an efficient decoupled head with a hybrid-channel strategy [31]. The model employed advanced quantization strategies, such as post-training quantization and channel-wise distillation, making it faster and more accurate than its predecessors, notably YOLOv5, and surpassing previous models in speed and accuracy. YOLOv7 released the same year, was a groundbreaking advance in object detection [31]. Trained on the MS COCO dataset without using pre-trained backbones, it delivered an outstanding performance, achieving speeds ranging from 5 FPS to an impressive 160 FPS. In January 2023, Ultralytics unveiled YOLO-v8 [32], [33], a new iteration in the YOLO series that includes YOLO-v5. A formal paper on YOLO-v8 is forthcoming, and the model's repository will be further enhanced with additional features. Early comparisons indicate that YOLO-v8 surpasses its predecessors and establishes a new standard as the state-of-the-art in the YOLO series. The architecture of YOLOv8 retains a backbone similar to YOLOv5 but introduces significant modifications. The C3 module is replaced with the C2f module, which draws from the CSP structure and incorporates elements from YOLOv7's ELAN, blending C3 and ELAN concepts into the C2f module. This allows YOLOv8 to capture more extensive gradient flow information while maintaining a lightweight structure. At the end of the backbone, the widely used SPPF module continues to be employed, featuring three serially arranged Maxpools of size 5×5 followed by a concatenation of each layer. This design helps ensure accuracy across various object scales while keeping the model light.

The paper seeks to address the challenges inherent in the semantic analysis of video sequences, with a specific focus on recognizing moving objects without prior knowledge of their characteristics. The primary objective is to refine the accuracy and efficiency of object detection through advanced deep learning techniques, specifically employing CNN algorithms like Faster R-CNN in conjunction with SVM classifiers. The scope of this study is notably comprehensive, encompassing a wide array of applications from intelligent video surveillance systems to dynamic interaction environments such as those captured by UAVs. By implementing and refining these models on a robust platform, the research aims to significantly advance the field of computer vision, particularly in how moving objects are detected and analyzed within complex and continuously changing backgrounds. This approach not only aims to enhance the reliability of object tracking across diverse conditions but also strives to contribute substantively to the broader domain of automated video analysis.

A. CONTRIBUTIONS

The contributions of this research in the specified field are highlighted in the following key aspects:

- Enhanced detection of moving objects in complex videos: This study improves the detection capabilities for moving objects of any size using feature maps and introducing the Zeiler and Fergus method. It effectively reduces the probability and temporal dimension of feature fusion, leading to enhanced feature analysis. Consequently, it addresses prevalent issues such as misidentification and the omission of small objects in detection processes.
- Speed factor detection: Emphasizing the importance of speed in the integration of the proposed model, this research tackles it by analyzing the pixels related to specific moving objects within the videos. By increasing feature variation, facilitating the acquisition of long-distance feature information, and minimizing undue penalization of geometric elements, this approach aims to enhance the generalization performance of the system. This results in a model that not only improves accuracy but also reduces the number of parameters, thereby solving issues related to the loss of long-range information and challenges in achieving balance with anchor prediction.

II. RELATED WORKS

In recent literature on video processing and analysis, several studies have employed semantic algorithms to enhance the detection and retrieval of moving objects within video content. This review highlights notable contributions across diverse applications.

Due to constraints in data transmission bandwidth and storage capacity, classifying fast-moving objects over extended periods using high-speed photography poses significant challenges. In response, Zhu et al. introduced a novel single-pixel classification technique utilizing deep learning for fast-moving objects. This method involves modulating the scene image with orthogonal transform basis patterns, which are then detected by a single-pixel detector. Leveraging the sparsity of natural images in the orthogonal transform domain, they employ a limited number of discrete-sine-transform basis patterns to capture essential feature information for classification [34]. The designed neural network processes these single-pixel measurements as inputs, trained using simulated single-pixel measurements that reflect the physics of our measurement approach. To mitigate discrepancies between simulated and experimental data caused by slowly varying noise, they apply differential measuring techniques. Furthermore, to enhance the reliability of the classification results, they use a rolling utilization approach for measurement data, enabling repeated classification. They have successfully demonstrated the long-duration classification of fast-moving handwritten digits passing through the field sequentially, with results indicating the proposed method's superiority over human vision in classifying fast-moving digits. This approach not only provides a new avenue for classifying fast-moving objects but also holds promise for broad implementation.

Wang introduced a unique goalmouth detection method using the Hough transform to improve performance significantly in video analysis. The study provided a robust method for detecting features during the categorization stage, allowing for precise recovery and categorization of features within semantic boundaries. The effectiveness of the approach was demonstrated through real-world soccer videos, showcasing the capability of the proposed system to enhance video retrieval operations [35]. Similarly, Ammar et al. utilized the Hough transform in their study to develop a whistle-detection strategy that significantly boosts system performance. This research also emphasized the strong goalmouth recognition method that aids the feature categorization process. Through the application of semantic thresholds, the system ensures accurate feature recovery and categorization. The effectiveness of the proposed model was validated using real-world soccer films, reinforcing the utility of the proposed detection strategies [36].

Further, Isa et al. tackled the challenge of detecting offensive language in YouTube comments in Malay, employing a list of offensive words provided by the Malaysian Communications and Multimedia Commission. The study leveraged both Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) features in conjunction with random undersampling and oversampling techniques to address data imbalance. Support Vector Machines (SVMs) and Naive Bayes classifiers were used, with SVM showing a notable recall of 98.70% when weighted by TF-IDF. Both classifiers performed comparably, though Naive Bayes slightly outperformed SVM, suggesting that preprocessing data and fine-tuning classifiers could further enhance outcomes [37]. Setiawan et al. explored road surface classification using data from motion sensors in smartphones, utilizing a U-Net architecture combined with BiLSTM networks. Their method was validated with data collected from four different smartphones, demonstrating that BiLSTM could potentially enhance segmentation performance when integrated with U-Net. The study suggests that future improvements could include more reliable feature extraction methods and noise reduction techniques like signal decomposition to boost segmentation and classification performance [38].

Pawar et al. proposed a deep learning-based method for detecting and localizing road accidents. Unique to this study is its training methodology, which involves the one-class learning paradigm, training solely on normal traffic events to detect anomalies as out-of-distribution samples. The performance of this model was evaluated against benchmarked data, highlighting its potential in real-world applications [39]. Lastly, Saad et al. presented a system for semantic annotation of video movements that integrates high-level movement concepts with low-level video features. The system uses temporal segmentation to extract movement objects from still scenes and employs SWRL rules and OWL ontologies to understand video movements. Based on the Benesh Movement Notation, the Video Movement Ontology (VMO) relates various movement features to improve the quality of video annotations through logical rules and reasoning processes [40].

This literature review on semantic analysis systems for recognizing moving objects reveals some gaps and areas for improvement that will be addressed in the proposed paper. Firstly, despite the advancements in deep learning models, such as the R-CNN framework combined with SVM classifiers, existing studies have highlighted the challenges faced in detecting moving objects under dynamic environmental conditions. These challenges include motion blur, varying degrees of movement and autonomy, and changes in outdoor lighting conditions which may affect the appearance of moving objects. This suggests a need for developing more robust detection and recognition methods that can adapt to varying environmental factors without compromising the accuracy of object detection. Secondly, while current models excel in extracting and analyzing semantic information from video sequences, there remains a notable difficulty in processing high-speed photography due to limited data transmission bandwidth and storage capacity. The literature cites innovative approaches like single-pixel classification methods; however, these are not yet widely adopted. This gap underscores the potential for exploring more efficient data processing techniques that could handle fast-moving objects over extended periods without requiring extensive bandwidth and storage. Lastly, the review points out that

most current systems rely heavily on predefined semantic models and do not adapt well to unanticipated scenarios that fall outside their trained datasets. This limitation calls for a more flexible semantic analysis framework capable of learning and adapting in real time to new or unexpected conditions in dynamic environments. Addressing these gaps could significantly enhance the practicality and effectiveness of semantic analysis systems in real-world applications, particularly in areas like surveillance, navigation, and automated monitoring, where the ability to accurately and efficiently recognize moving objects is crucial.

III. PROPOSED SCHEME

As previously stated, deep learning-driven object identification has made significant progress in recent years, boosting both speed and robustness in actual applications. Among the numerous methodologies, CNN has stood out due to its complete design, which enables nearly rapid processing while maintaining precision. Using this as a basis, an effective method for refining R-CNN in moving objects inside visual environments is described. This technique aims to improve the responsiveness of the model to moving elements by using personalized preprocessing steps and architectural changes. This guarantees that the model not only identifies objects but also monitors their movement in real-time. A refined and efficient identification of moving things across a range of contexts is expected via integrating CNNs' intrinsic processing speed with motion-focused upgrades.

To describe the suggested model shown in Figure 1, the particular processes of the proposed detection approach are as follows:

A. EXTRACTION FEATURES FROM OBJECTS OF VIDEO

In this article, we propose a semantic analyzer that generates semantic interpretations of observed scenes based on prior knowledge about the target domain and observed data (features extracted from movies). Interpreting a video means finding and understanding its items. It is important to note that the purpose of this video is not to label or categorize, but rather to describe the atomic activities and objects that occur. By detecting key semantic aspects from video data, such as actions and objects, and their relationships, we can interpret what is happening in the video [41]. The SVM model constructs connected structures to facilitate the semantic interpretation of videos. The nodes (generators) of the graph and their connections (graph edges) are represented as closed bonds. By observing the characteristics of objects and actions within a scene, their presence can be directly confirmed. Features observed are managed using priors like co-occurrence frequency tables. When certain items frequently appear together, semantically consistent interpretations are preferred. In this section, a highlighted object is produced, as detailed in [35]. A lack of evidence for expected actions or objects can also be captured with incomplete ties. There are no hanging edges in it, in contrast to standard graph architectures. Spatiotemporal regions of interest may be segmented, and identified, and motion salience extracted using preprocessing stages.

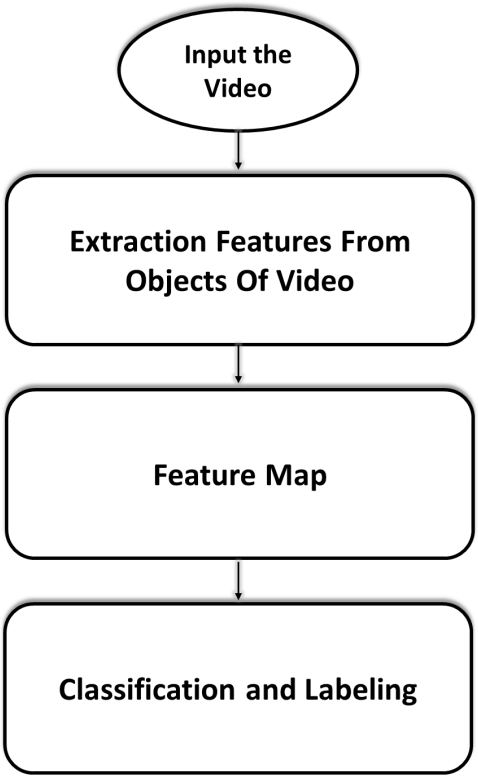


FIGURE 1. The proposed methodology.

Definitions:

1) **Definition 1.** Images environment and the relationship between its moving objects. These objects should be interpreted to certain semantic or clear titles of that environment (City, Super Market, Conference, etc.). So each object represents a variable (O_i).
2) **Definition 2.** The role of moving objects. This definition will represent the different roles (movement and the way of movements) of each object inside the environment. The variable of each moving object is $R_i(O_i)$.
3) **Definition 3.** The interpretation of role: The objects are the interpretation contents as a list of items, such as actions and objects. The items in a scene build a graph that specifies and describes the relationships between these parts. The combination of the two preceding definitions will be offered for the proper action or outcome of recognition for a certain moving item, such as the distinction between sport and dancing, etc.

In this segment, we introduce a refined semantic analysis approach designed to interpret and understand the dynamic scenes captured in video streams, leveraging a robust feature extraction process that emphasizes real-time detection and classification of moving objects.

- Dynamic feature extraction: Our proposed model employs an advanced dynamic feature extraction method that captures both spatial and temporal aspects of moving objects within videos. This is achieved through the continuous analysis of frames to detect changes and movement patterns. We implement a multi-layer convolutional network that processes video frames in sequence, extracting features such as edges, textures, and shapes which are critical for identifying and tracking objects.
- Adaptive learning from video data:
  Unlike traditional models that rely on pre-set feature libraries, our system uses an adaptive learning approach. By applying machine learning algorithms, specifically deep learning techniques, the system continuously updates its feature extraction capabilities based on new data encountered in the video streams. This adaptability allows it to improve its accuracy over time, particularly in handling diverse and complex scenarios such as varying lighting conditions, weather effects, and occlusions.
- Semantic feature integration:
  Key to our approach is the integration of semantic information into the feature extraction process. By employing semantic tagging and metadata, our model distinguishes between different types of movements and interactions within the video. For instance, it can differentiate between a person walking, running, or an object being moved by external forces. This semantic layer is crucial for applications requiring a detailed understanding and categorization of scene dynamics.
- Real-time processing and optimization:
  To ensure the system operates in real time, significant attention is given to optimizing the computational efficiency of the feature extraction process. Techniques such as parallel processing, hardware acceleration, and algorithmic optimizations are utilized to process high-resolution video without significant delays. This enables the application of our model in time-sensitive environments like traffic monitoring and emergency response systems.
- Enhanced object detection with machine learning:
  The extracted features are then processed by a hybrid model that combines CNNs and SVMs. This combination takes advantage of CNN's ability to hierarchically process visual data and SVM's effectiveness in classification tasks. The SVM classifier is fine-tuned to work with the high-dimensional data produced by CNN, providing a robust detection mechanism that significantly reduces false positives and improves object recognition accuracy. Feature Validation and Feedback Mechanism
- Feature validation and feedback mechanism:
  To further enhance the reliability of the feature extraction process, a validation mechanism is incorporated. This involves a feedback loop where the system's predictions are periodically reviewed and corrected by human supervisors. These corrections are then fed back into the model as additional training data, refining the feature extraction algorithms and adapting the model to new or unseen challenges in the video content.

B. CLASSIFICATION AND LABELING PROCESS

Using the affinity of the bond interaction between O_i and O_j, it was possible to measure the degree of acceptance between the two objects, O_i and O_j. A bond energy is calculated based on the logarithm of this bond affinity. It depends on the type of bond interaction that determines the form of affinity function $A(O_i, O_j)$. A bond interaction can be classified as either informational or support [42].

Informational bonds, which link informational generators, encapsulate contextually relevant data that supports the creation of semantically coherent interpretations. The affinities of these bonds are calculated based on the co-occurrence frequencies of actions and objects, as discussed in [43]. Additionally, the training dataset includes three types of conceptually appropriate labels, as noted in [38] and further elaborated in [39]. These reflect the frequency with which specific pairs of conceptually compatible labels co-occur.

$$f(O_i, O_j) = \sum_{k=1}^{|V|} \left[O_i \downarrow O_j \in V_k \right], \text{s.t. } i, j \tag{1}$$

The output of $\left[O_i \downarrow O_j \in V_k \right]$ is 1 if both concepts g_i and g_j co-occur in video clip k (or 0 otherwise), where V represents the training dataset annotations and V_k is the list of concepts in the video clip. As a result, b represents the affinity between ontological bonds. Equation 2 represents the affinity between ontological bonds as an exponential function of the hyperbolic tangent of the weighted sum of feature occurrences. Specifically, it is given by:

$$A(O_i, O_j) = \exp(\tanh(w_1 f(O_i, O_j))), \tag{2}$$

where $w_1 = 0.025$, a decision that is empirically determined and is based on the range of $f(O_i, O_j)$. Because we aimed to represent bond affinity as a monotonic, limited function, the hyperbolic function was an option. The energy function of an affinity bond is governed by its bound affinity, which prevents one bond from controlling it entirely. Actions and objects are grounded by support bonds. Objects and actions are linked by features. When actions and objects are classified using connected features by machine learning-based classifiers, the classification scores are used to calculate the support bond connections.

Equation 3 in the document represents the logarithmic transformation of bond energy, which is computed based on the affinity between two objects O_i and O_j. The mathematical expression for this is given by:

$$A(O_i, O_j) = \exp(\tanh(w_2 h(O_i, O_j))) \tag{3}$$

A confidence score for classification, denoted as C, is calculated for the concept "Suitable label" produced by

generator g_i when $w_2 = 2$. The output function $h(\cdot, \cdot)$ determines C, where C ranges from 0 to 1. A score of 1 represents a classifier with maximum confidence.

IV. PRACTICAL PARTS

This section describes the benchmark data set used to assess the performance of the proposed framework, as well as the Faster R-CNN approach utilized for a side-by-side comparison.

A. DATA

A video depicting traffic conditions was used to evaluate the proposed system. Various automotive types were filmed along with pedestrians, who were crossing the street, for instructional purposes in the videos. The film collection displays various challenging settings such as cluttered backgrounds, multiple themes, object occlusion, and multiple perspectives. The performance evaluation was limited to movies with accessible annotations. Video sequence annotations describe the spatiotemporal positioning of objects and events. Videos with annotations for 44 different types of videos were made accessible [44], [45]. The dataset is divided into a training group consisting of 22 movies and an assessment group consisting of 22 movies. Object interactions are extracted from training videos using annotations. The concept co-occurrence tables were generated using Equation 1.

The dataset includes interactions with objects encompassing things and activities. Eighteen items were examined and six types of behaviors were considered, categorized into actions (a car, bicycle, or person moving) and objects (a car, bicycle, or person). Due to the vast array of possible combinations of objects and actions, it would be computationally infeasible to create a classification model for every potential pairing. Configurations with three ideas have six times as many interpretations ($6 \times 18 \times 18$) as configurations with 18 ideas alone. Rhythmic movements are depicted in each video. The videos were automatically segmented into shorter video segments due to the sequential occurrences of activities across the records, creating temporal segmentation. This temporal video segmentation resulted in 8 video clips for evaluation purposes and 5 video clips for training purposes. A brief movie illustrates an interaction described by an action carried out on an object, serving as a unit of interpretation for performance assessment [41].

B. OBJECT FEATURE

Histograms of optical flow (HOFs) illustrate the motion dynamics of each video clip by using sequential frames combined into a temporal series. This sequence of HOFs is composed of three sets arranged chronologically, with clusters represented by the sum of their histograms. The motion feature histogram, a composite of these three histograms, reflects the dynamics of motion in each video clip. Motion dynamics of video clips are displayed using motion feature histograms generated by a feature generator. There will be a feature generator instance associated with each detected item

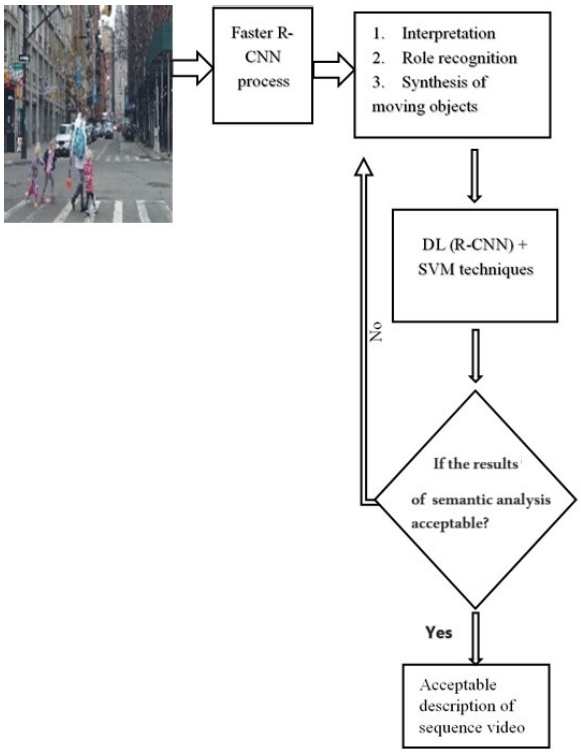


FIGURE 2. A scheme of the semantic analyzer.

bounding box. Objects' bounding boxes are characterized by Histograms of Directional Gradients (HOGs). Each object track has one HOG for each bounding box, thus each object track has a collection of HOGs. A bounding box can only be assigned to an object track once per frame. A video object track is represented as a collection of bounding boxes in an object feature generator [46].

One of the most challenging computer vision tasks is identifying and localizing objects in images or video streams [47]. Significant progress has been made in this field since the introduction of deep learning. Real-world object recognition and classification have been revolutionized by Faster R-CNN. The functions of the R-CNN family are divided into three stages:

- Region proposal networks suggest potential regions in an image where items may be present.
- CNNs extract key characteristics.
- Classification and regression are used to forecast the class and fine-tune the coordinates of an object bounding box.

C. FEATURE MAPPING

Space and bond strengths are generated by an ontology specification during the mapping process. Unlike graphical models, there is no need for training on possible interpretation structures. Based on the video annotations from the training dataset, the specific generator space is determined from the types of actions and objects present. Each action and

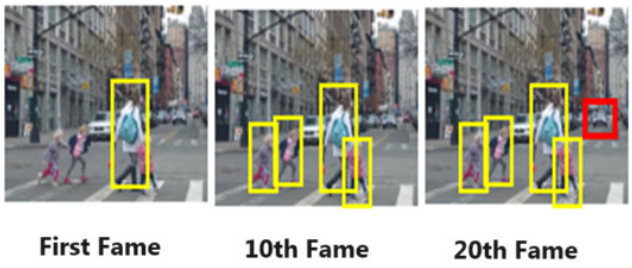


FIGURE 3. The interpretation steps.

object generator is predetermined by domain information. Classifiers were created according to predefined scenarios or object attributes. This method utilizes training data, and with classification models, bond interaction affinities can be calculated between feature generators and action/object generators (the support bonds). Actions and objects can be classified using a multi-class classification model based on LibSVM [48]. The authors used SMOTE [49] to generate synthetic samples for the minority categories, such as season and flip [50], due to the unequal distribution of training instances. This illustrates the challenges of detecting moving object sequences, which often require multiple repetitions for accurate labeling. The ontological bond affinities between actions and objects are learned from training data, and an ontology constraint can be used to remediate classification errors based on the co-occurrence of objects and activities [51].

Figure 3 depicts results from the literature with the discovery that detecting moving object sequences was difficult to interpret successfully due to the multiple repetitions required to label appropriately. This suggests that trained models are useless when recognition is limited to labeling based on the model's best prediction scores. The overlap between overlapping groups is clear. In escaping action or athletic action, for example, there is a lot of running activity. Furthermore, training examples of things that are commonly used together confuse categorization models [50]. In an escape activity or sporting action, for example, there is a lot of rising motion. We hope that the incorporation of historical knowledge contained in the ontological connection affinities will lessen this sort of uncertainty. The ontological bond affinities between actions and objects are learned from training data. An ontology constraint can also be used to remediate classification errors based on the co-occurrence of objects and activities [51].

The combination of R-CNN and SVM methodologies allows the model to leverage the strengths of deep learning for feature extraction and the robustness of SVMs for classification. Faster R-CNN helps in quickly generating high-quality region proposals that are then refined by SVMs for accurate object classification, reducing the need for multiple labeling iterations.

Also, by implementing a selective search approach within the Faster R-CNN framework, the model can focus on probable object locations, reducing the computational burden and refining the process of object detection. This approach minimizes the instances where multiple detections are necessary by improving the accuracy of initial detection. Further, the model uses advanced preprocessing techniques to handle variations in object appearance and motion blur effectively. By enhancing the input data quality and the feature extraction process, the system reduces the dependency on multiple repetitions for label refinement.

In the subsequent sections, we will evaluate the effectiveness of the proposed methods by comparing them with the following established techniques:

1) SIFT

Developed by David Lowe in 1999, the Scale-Invariant Feature Transform (SIFT) is crucial for detecting, characterizing, and matching local features in images. It finds widespread applications in object recognition, robotic mapping, image stitching, 3D modeling, gesture recognition, video tracking, and identifying individual wildlife. SIFT keypoints are extracted from a database of reference images. Object identification in new images is accomplished by calculating the Euclidean distance between feature vectors. The consistency of object size, orientation, and location is verified through subsets of keypoints, employing the generalized Hough transform through a hash table for efficient determination of consistent clusters. Each cluster of three or more features that agrees on an object and its pose is subjected to model verification and outlier removal. The final validation involves assessing the accuracy of fit and the number of false matches to confirm the presence of an object.

2) CCP

This method analyzes pixel directions that converge at a common focal point to assess features.

3) LBP

A texture analysis technique, Local Binary Patterns (LBP), functions by thresholding a neighborhood with the grey level of the central pixel.

V. ENHANCEMENT TECHNIQUE

The default approach in Figure 2 simply returns object and action nodes associated with the detected attributes. However, a judgment cannot be made about what action is performed on which item because there is no link between the object and the action. Yellow blocks indicated that moving objects had been observed by the R-CNN previous recognition scheme. A feature map will then be used to examine these blocks. The nature of the map as analyzed by the scheme involves a critical component of Faster R-CNN known as the Region Proposal Network (RPN), which plays a vital role. The RPN generates region recommendations in images with objects using the following algorithmic components:

1) **Anchor boxes:** R-CNN Faster uses anchors in creating region recommendations. The system uses a predetermined number of anchor boxes in various sizes and aspect ratios. Many anchor boxes are positioned along the feature maps. It is important to understand two things about anchor boxes (scale and the aspect ratio). Throughout the input image, the anchor-generating layer disperses bounding boxes of different sizes and aspect ratios. These bounding boxes are independent of an image's content; they are the same for every image. While the majority of these bounding boxes do not encompass foreground objects, some do. The main objectives of the RPN network are learning which of these boxes are likely to contain a foreground object and generating target regression coefficients that improve the bounding box fit to the enclosed foreground object.
2) **Sliding Window Approach:** The feature map of the CNN backbone is displayed through the RPN. Using a tiny convolutional network (usually 3×3 convolutional layers), it processes the features in the sliding window receptive field. By combining these scores, we get scores indicating whether an item is likely to be present, as well as regression values for adjusting the anchor boxes.
3) **Objectness Score:** This score indicates how likely it is that an anchor box includes an object of interest rather than merely a background. Faster R-CNN forecasts a different score for each anchor. An objectness score of the anchor indicates whether it corresponds to an area of meaningful objects. This score is used to categorize anchors as either positive (object) or negative (background).
4) **IoU (Intersection over Union):** Overlap between two bounding boxes is measured by the IoU statistic. An area overlap with its union is calculated by dividing its overlap area by its union area.
5) **Non-Maximum Suppression (NMS):** To eliminate duplication and select the most appropriate suggestion, the objectness scores of overlapping proposals are compared and only the proposal with the highest score is kept.

Feature maps obtained from the CNN backbone are used by the RPN. Depending on the size and shape of the anchor boxes, RPN uses a sliding window approach to identify likely items on these feature maps. These anchor boxes are refined during the training process to more closely match the actual placements and sizes of items. The RPN predicts two parameters for each anchor:

- The likelihood that the anchor will contain an item (''objectness Score'').
- Modifications to the anchor coordinates to fit the geometry of the real item.

When there are a large number of ideas, the same area may be mentioned in several ideas. Based on their objectness likelihood, the NMS approach is used here to rank the anchor boxes and choose the top-N anchor boxes. By ensuring correctness and non-overlapping submissions, NMS ensures final proposals are chosen correctly. Possible region suggestions were selected from these anchor boxes.

VI. RESULTS

According to Table 1, the error rate for feature categorization is reported for each class of HumanEva video1 action, describing the scheme by its method. Compared with previous related articles, the total feature classification rate increased from 10% to 60% in one instance. According to the second scenario, there was a wide range of suitable label ranks for misclassified features, where the error rate was set at different levels of noise (20%, 40%), 60-60%, and 80-60% respectively. Specifically, these results discuss a testing environment where features were exposed to varying noise levels of 20%, 40%, 60-60%, and 80-60%. These noise levels represent artificial disturbances introduced into the data to simulate real-world inaccuracies or errors that might affect the model's performance. The range of suitable label ranks mentioned refers to the model's ability to still correctly identify or classify features despite the noise. This implies robustness in the model where it maintains a certain accuracy rate even as the noise level increases, demonstrating the model's capability to handle data corruption up to a certain extent. Further, the inclusion of two distinct percentages in ''60-60%'' and ''80-60%'' likely represents different experimental setups or thresholds in the study, possibly indicating varied conditions under which the model's performance was evaluated. For instance, the first number could denote the percentage of noise added to the dataset, while the second number might indicate a threshold percentage of acceptable classification accuracy or error rate under those conditions. This dual-percentage format highlights a more nuanced exploration of the model's performance across a spectrum of challenges. The study's focus on these specific noise levels and their impact on misclassified features suggests an in-depth investigation into the model's resilience and the effectiveness of its learning algorithm, crucial for applications where data integrity may be compromised.

In Figure 4, the proposed model shows how performance can be reduced only slightly and gradually when errors of low-level processing are encountered (for example, classification errors). There is a possibility that the method will maintain performance rates of 20% or more. An approach that heavily relies on concept classifiers functioning flawlessly, like the R-CNN implemented, might easily fail under these conditions. Multiple moving objects are detected by SVM in many videos (type MP4). It is possible to categorize and classify each object as a separate one based on its various characteristics. Figure 4 shows the accuracy of SVM. With the same datasets, the figure is compared with another method based on the same data set.

TABLE 1. Performance evaluation of different models on video datasets with varying levels of noise.

Datasets	SIFT [52]	CCP [46]	LBP [53]	Proposed Model	Notes
Video1	100	100.00	100.00	100.00	No noise
Video2	83.3	82.50	80.95	85.41	20% noise
Video3	62.6	62.00	61.76	64.78	40% noise
Video4	40.1	40.50	40.78	40.23	60% noise
Video5	22.05	21.50	21.21	22.23	80% noise

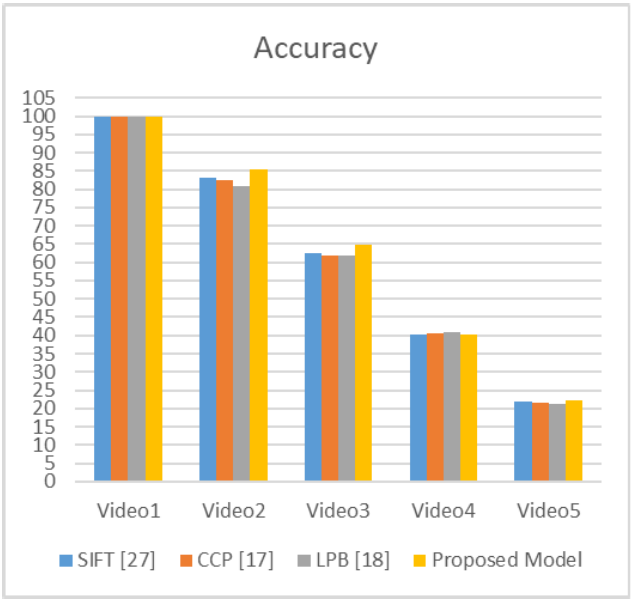


FIGURE 4. The accuracy of the applied techniques.

SVM iterations are critical in achieving high levels of segmentation or for making corrections to the boundaries. To examine the behavior of each algorithm, the videos include noise ratios. Due to its focus on identifying the best-moving objects, SVM produced an accepted result. Because the proposed model uses a different technique from Conventional Neural Networks (R-CNN) and LBP, its results are compared with those of the conventional neural network. The goal of this comparison is to determine the best results. It is very close to the results in video 1 because the noise level is very low. There are high levels of noise in the over videos, so the Gaussian algorithm explores the best detection method to come up with the best results. Within the first 150 epochs, the simulated data significantly improves both loss and accuracy. This achieves peak performance quickly. In practice, such rapid convergence indicates that it learns effectively, which is useful when dealing with large datasets generally involved with object recognition [54]. Both training and validation statistics appear to have stabilized after the 150th epoch. When encountered in real-life settings, this might imply:

- **Learning Limit of the Model:** Given the existing data, it is probable that the model has reached its learning limit. As a result, just increasing its complexity may not result in benefits unless there is an improvement in data quality or diversity.
- **Concerns about learning rate:** A high learning rate may cause the model to float about a local minimum [55], [56]. A pace that is too low, on the other hand, may result in stagnation. Adaptive learning rate techniques can correct this. The observation of little development beyond the 150th epoch suggests that using an early termination condition during training might save time and computer power.

VII. DISCUSSION

The performance results derived from the latest experiments conducted with the proposed machine learning model demonstrate substantial improvements over traditional methods such as SIFT, CCP, and LBP. The enhancements are particularly evident in environments with high levels of noise, showcasing the robustness and reliability of the proposed model under challenging conditions.

Firstly, the resilience of the proposed model is highlighted by its ability to maintain higher accuracy rates under increased noise levels, as shown in Table 1. While traditional methods show a marked decrease in performance as noise levels rise, the proposed model demonstrates a lesser decline in accuracy. For example, at an 80% noise level, the model maintains a detection accuracy significantly above that of traditional methods. This resilience is attributed to the model's sophisticated feature extraction capabilities, which leverage deep learning techniques to isolate and identify pertinent features even when noise corrupts the input data.

Moreover, the proposed model's superior performance can be partly attributed to its use of SVM for classification. Unlike conventional methods that might struggle with feature variance due to noise, the SVM component of the proposed model effectively categorizes and classifies each detected object based on its distinct characteristics, regardless of the environmental conditions. This approach ensures that each object is recognized and tracked consistently across the video sequence, enhancing the overall accuracy of the system.

The detailed performance analysis, as illustrated in Figure 4, further supports the effectiveness of the model. The graph indicates that even with errors introduced by low-level processing anomalies—such as classification discrepancies—the model's performance degrades only slightly, maintaining a baseline accuracy rate that exceeds 20%. This robust performance underscores the advanced error-handling capabilities of the model, which is crucial for practical applications where precision is critical, such as in surveillance or autonomous vehicle navigation.

Lastly, the rapid convergence of the model within the first 150 epochs of training, as indicated by the stability of both training and validation metrics, suggests a high level of learning efficiency. This rapid learning ability is essential for deploying the model in dynamic environments, where it needs to adapt quickly to new scenarios without extensive retraining. The model's performance in these experiments

suggests that it has reached an optimal balance between accuracy and learning speed, making it highly suitable for real-world applications where both factors are crucial for success. The ability of the model to quickly reach its learning limit and maintain performance with minimal further tuning highlights its practical utility and the effectiveness of its underlying architecture and training regimen.

VIII. LIMITATIONS

The proposed model, leveraging the Faster R-CNN combined with SVM classifiers, is designed to enhance the detection and analysis of moving objects within video sequences, particularly under dynamic conditions. However, like many sophisticated systems, it may reach a learning limit with the current dataset, potentially impeding further improvements in performance beyond a certain point. This plateau suggests that the model has maximized its understanding based on the available data and may not benefit from simply increasing model complexity or extending training duration without additional adjustments.

To address this issue and push the boundaries of the model's learning capabilities, several strategies can be implemented. First, expanding the diversity and volume of the training data can provide new patterns and scenarios for the model to learn from, thus enhancing its generalization ability. This could involve incorporating datasets from varied environments or scenarios not previously covered. Second, the implementation of advanced regularization techniques like dropout, batch normalization, or data augmentation could help prevent overfitting and encourage the model to develop a more generalized understanding of the features relevant to object detection.

Furthermore, exploring alternative neural network architectures or adjusting existing layers and their parameters could yield improvements. For instance, employing deeper or differently structured networks might extract more nuanced features from the data. Also, adjusting the learning rate adaptively during training could help in optimizing the convergence process, ensuring the model does not miss finer details in the data due to a suboptimal training pace.

Lastly, integrating feedback loops into the model where predictions can be manually corrected and reintroduced as training inputs could help in refining the model's accuracy. This approach, often referred to as active learning, allows the model to learn from its mistakes and adapt more effectively to complex or ambiguous detection scenarios.

By employing these methods, the proposed model can potentially move beyond its current limitations, enhancing its performance and applicability in real-world situations where dynamic object detection is critical.

The challenge of relying on predefined scenarios and object attributes for classifier creation, which may restrict adaptability to new or diverse datasets, is significant in dynamic environments where unexpected object behaviors or appearances can occur. To address this limitation, the proposed model could implement several strategies to enhance its adaptiveness and robustness:

- Transfer Learning: This technique involves using a model trained on one task as the starting point for training on a new task. By leveraging models pre-trained on large and diverse datasets, such as those available through ImageNet or COCO, the system can benefit from learning features that are generally applicable across various domains. This approach allows the model to adapt more effectively to new environments or object characteristics that were not part of the initial training data.
- Incremental Learning: This strategy involves continuously updating the model's knowledge without forgetting previously learned information. It is particularly useful in applications where new object types or scenarios are gradually introduced over time. Implementing methods like Elastic Weight Consolidation (EWC) can help the model maintain its performance on previously learned tasks while adapting to new data.
- Data Augmentation: Enhancing the training dataset with artificially modified copies of existing data can help improve the robustness and generalization of the model. Techniques such as rotation, scaling, cropping, and color modification introduce a variety of realistic scenarios that the model might face, reducing its reliance on the specifics of the predefined attributes and scenarios.
- Active Learning: This approach can be used to selectively acquire labels for the most informative data points. By integrating an active learning framework, the model can query the user or an expert for labels on new or ambiguous examples that are likely to be informative for learning. This method ensures efficient use of labeling efforts while continuously improving the model's adaptability to new conditions.
- Ensemble Techniques: Combining predictions from multiple models or configurations can enhance the robustness and accuracy of the system. By employing an ensemble of classifiers trained under different settings or on different subsets of the data, the system can better generalize across various conditions and reduce the risk of overfitting to predefined scenarios.
- Advanced Architectures: Exploring more complex neural network architectures that are inherently more adaptable, such as those involving attention mechanisms or transformers, could allow the model to focus on the most relevant features of an input irrespective of their position. This capability is particularly useful in unstructured environments where important features may not be consistently located.

By incorporating these strategies, the proposed model can significantly improve its ability to handle diverse and unexpected scenarios, reducing its dependence on predefined settings and enhancing its overall performance and utility in real-world applications.

IX. CONCLUSION

The task of improving the traditional object detection method to precisely recognize dynamic features in visual data streams was reported in this study. Our in-depth analysis of the Faster CNN foundation has confirmed its inherent qualities and strengths. However, there were clear chances for specialization and refining, just like with many broad-spectrum solutions. The reported improved Faster CNN model demonstrates an enhanced motion subtlety perception, resulting in a more advanced object detection system. The comparison analyses revealed that, even with the fundamental characteristics of speed and accuracy that are associated with Faster CNN, this model outperformed others, especially in highly dynamic circumstances. The suggested technique has higher accuracy for detecting moving objects and is faster than existing models since it combines an SVM model with a Faster CNN model. The Python 3.27 scheme platform made our task easier by testing about 650 videos from various datasets. This finding has wider implications across other fields. The augmented model puts the groundwork for increased real-time decision-making tools and deeper analytical views in domains where motion interpretation is critical, whether for security monitoring, controlling vehicular traffic, or analyzing motion in cinematic scenes. However, as is typical of the vast field of technology, the potential for additional improvement is boundless. Future studies could investigate combined model architectures, incorporate state-of-the-art motion forecasting methods, or adjust the model for specific use cases. This state-of-the-art Faster CNN acts as a lighthouse, demonstrating the vast opportunities that arise from customizing deep learning instruments for specific object identification applications.

Building on the foundations laid by this paper, several future research directions can enhance the practicality and effectiveness of semantic analysis systems for recognizing moving objects. Firstly, integrating additional sensory data such as audio, infrared, or radar with the current visual-based approach could significantly improve detection capabilities in environments with low visibility or high clutter. This multi-sensory approach could be particularly beneficial in complex dynamic environments such as urban areas or diverse weather conditions, where visual data alone may not be sufficient. Furthermore, optimizing the model for real-time processing could extend its applicability to immediate-response systems, such as autonomous driving and active surveillance. This could involve exploring more efficient computational methods or leveraging hardware acceleration techniques to enhance speed and efficiency. Another promising area of development is the advancement of the learning algorithms used. Investigating newer or more advanced neural network architectures could yield improvements in both accuracy and processing speed. Additionally, adopting unsupervised or semi-supervised learning methods could also enhance the model's ability to adapt to new or unseen environments without the need for extensive labeled datasets. Moreover, enhancing the model's capability to understand interactions between humans and objects within a scene could lead to deeper insights into the contextual dynamics of environments. This involves not just recognizing objects but interpreting human actions and predicting potential interactions or movements. Energy efficiency is also a critical factor, especially for deployments on mobile or edge devices. Research focused on developing energy-efficient neural networks or techniques for reducing computational load could make the systems more viable for widespread application. Lastly, the diversity and quality of training datasets play a crucial role in the performance of deep learning models. Efforts to expand dataset diversity and develop methods to reduce bias in model training and predictions could help in achieving more robust and equitable outcomes. By pursuing these avenues, future research can further the capabilities of semantic analysis systems, making them more adaptable and effective in real-world scenarios.

ACKNOWLEDGMENT

The authors express their gratitude to the University of Sfax in Tunisia for administrative and technical support.

REFERENCES

[1] M. Wu, C. Li, and Z. Yao, "Deep active learning for computer vision tasks: Methodologies, applications, and challenges," *Appl. Sci.*, vol. 12, no. 16, p. 8103, Aug. 2022.

[2] S.-H. Lee, G.-C. Lee, J. Yoo, and S. Kwon, "WisenetMD: Motion detection using dynamic background region analysis," *Symmetry*, vol. 11, no. 5, p. 621, May 2019.

[3] M. Ahmadi, W. Ouarda, and A. M. Alimi, "Efficient and fast objects detection technique for intelligent video surveillance using transfer learning and fine-tuning," *Arabian J. Sci. Eng.*, vol. 45, no. 3, pp. 1421–1433, Mar. 2020.

[4] G. Capi, "A vision-based approach for intelligent robot navigation," *Int. J. Intell. Syst. Technol. Appl.*, vol. 9, no. 2, pp. 97–107, 2010.

[5] G. Yin, Y. Li, and J. Zhang, "The research of video tracking system based on virtual reality," in *Proc. Int. Conf. Internet Comput. Sci. Eng.*, Jan. 2008, pp. 122–127.

[6] A. Singh and M. K. Dutta, "Imperceptible watermarking for security of fundus images in tele-ophthalmology applications and computer-aided diagnosis of retina diseases," *Int. J. Med. Informat.*, vol. 108, pp. 110–124, Dec. 2017.

[7] J. Cho, Y. Jung, D. Kim, S. Lee, and Y. Jung, "Design of moving object detector based on modified GMM algorithm for UAV collision avoidance," *J. Semiconductor Technol. Sci.*, vol. 18, no. 4, pp. 491–499, Aug. 2018.

[8] J. Shin, S. Kim, S. Kang, S.-W. Lee, J. Paik, B. Abidi, and M. Abidi, "Optical flow-based real-time object tracking using non-prior training active feature model," *Real-Time Imag.*, vol. 11, no. 3, pp. 204–218, Jun. 2005.

[9] X. Fan, Y. Cheng, and Q. Fu, "Moving target detection algorithm based on Susan edge detection and frame difference," in *Proc. 2nd Int. Conf. Inf. Sci. Control Eng.*, Apr. 2015, pp. 323–326.

[10] S. K. Jarraya, M. Hammami, and H. Ben-Abdallah, "Accurate background modeling for moving object detection in a dynamic scene," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Dec. 2010, pp. 52–57.

[11] S. Jung, Y. Cho, K. Lee, and M. Chang, "Moving object detection with single moving camera and IMU sensor using mask R-CNN instance image segmentation," *Int. J. Precis. Eng. Manuf.*, vol. 22, no. 6, pp. 1049–1059, Jun. 2021.

[12] E. Raja, G. Gandhimathi, A. Sriram, B. Ramasubramanian, and K. Priyadharshini, "A novel deep learning based approach for object detection using mask R-CNN in moving images," *AIP Conf. Proc.*, vol. 2946, no. 1, 2023, Art. no. 050004.

[13] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, Z. Yuan, and P. Luo, "Sparse R-CNN: An end-to-end framework for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15650–15664, Feb. 2023.

[14] S. Dhakal, Q. Chen, D. Qu, D. Carillo, Q. Yang, and S. Fu, "Sniffer faster R-CNN: A joint camera-LiDAR object detection framework with proposal refinement," in *Proc. IEEE Int. Conf. Mobility, Oper., Services Technol. (MOST)*, May 2023, pp. 1–10.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[16] A. Brodzicki, J. Jaworek-Korjakowska, P. Kleczek, M. Garland, and M. Bogyo, "Pre-trained deep convolutional neural network for clostridioides difficile bacteria cytotoxicity classification based on fluorescence images," *Sensors*, vol. 20, no. 23, p. 6713, Nov. 2020.

[17] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, "Real-time vehicle detection based on improved YOLO v5," *Sustainability*, vol. 14, no. 19, p. 12274, Sep. 2022.

[18] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?" *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102524.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[23] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, Sep. 2014, pp. 740–755.

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

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[27] Z. Ma, J. Xuan, Y. G. Wang, M. Li, and P. Liò, "Path integral based convolution and pooling for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16421–16433.

[28] X. Yu, T. W. Kuan, Y. Zhang, and T. Yan, "YOLO v5 for SDSB distant tiny object detection," in *Proc. 10th Int. Conf. Orange Technol. (ICOT)*, Nov. 2022, pp. 1–4.

[29] Z. Yao, Y. Cao, S. Zheng, G. Huang, and S. Lin, "Cross-iteration batch normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12331–12340.

[30] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[31] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[32] M. Hussain, "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection," *Machines*, vol. 11, no. 7, p. 677, Jun. 2023.

[33] M. Safaldin, N. Zaghden, and M. Mejdoub, "An improved YOLOv8 to detect moving objects," *IEEE Access*, vol. 12, pp. 59782–59806, 2024.

[34] J. Zhu, Z. Wang, S. Wang, and S. Chen, "Moving object detection based on background compensation and deep learning," *Symmetry*, vol. 12, no. 12, p. 1965, Nov. 2020.

[35] Z. Wang, "Semantic analysis based on fusion of audio/visual features for soccer video," *Proc. Comput. Sci.*, vol. 183, pp. 563–571, Jan. 2021.

[36] S. Ammar, T. Bouwmans, N. Zaghden, and M. Neji, "Deep detector classifier (DeepDC) for moving objects segmentation and classification in video surveillance," *IET Image Process.*, vol. 14, no. 8, pp. 1490–1501, Jun. 2020.

[37] A. M. Isa, S. Ahmad, and N. M. Diah, "Detecting offensive Malay language comments on YouTube using Support Vector Machine (SVM) and Naive Bayes (NB) model," *J. Positive School Psychol.*, vol. 6, no. 3, pp. 8548–8560, 2022.

[38] B. D. Setiawan, M. Kovacs, U. Serdült, and V. Kryssanov, "Semantic segmentation on smartphone motion sensor data for road surface monitoring," *Proc. Comput. Sci.*, vol. 204, pp. 346–353, Jan. 2022.

[39] K. Pawar and V. Attar, "Deep learning based detection and localization of road accidents from traffic surveillance videos," *ICT Exp.*, vol. 8, no. 3, pp. 379–387, Sep. 2022.

[40] S. Saad, S. Mahmoudi, and P. Manneback, "Semantic analysis of human movements in videos," in *Proc. 8th Int. Conf. Semantic Syst.*, Sep. 2012, pp. 141–148.

[41] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614812.

[42] (May 2020). *Components Diagram*. [Online]. Available: https://www.tutorialspoint.com/uml/uml_component_diagram.htm

[43] F. D. M. de Souza, S. Sarkar, A. Srivastava, and J. Su, "Pattern theory for representation and inference of semantic structures in videos," *Pattern Recognit. Lett.*, vol. 72, pp. 41–51, Mar. 2016.

[44] (2024). *Humaneva Dataset*. Accessed: Feb. 2024. [Online]. Available: http://humaneva.is.tue.mpg.de/datasets_human_1

[45] (2024). *Urbanmotorbike*. Accessed: Feb. 2024. [Online]. Available: http://videodatasets.org/UrbanMotorbike

[46] E. Kurimo, L. Lepistö, J. Nikkanen, J. Grén, I. Kunttu, and J. Laaksonen, "The effect of motion blur and signal noise on image quality in low light imaging," in *Proc. 16th Scand. Conf. Image Anal. (SCIA)*. Oslo, Norway: Springer, Jun. 2009, pp. 81–90.

[47] M. A. A. Hammoudeh, M. Alsaykhan, R. Alsalameh, and N. Althwaibi, "Computer vision: A review of detecting objects in videos—Challenges and techniques," *Int. J. Online Biomed. Eng.*, vol. 18, no. 1, pp. 15–27, Jan. 2022.

[48] C.-C. Chang. (2001). *A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[49] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[50] T. Singh, S. Malik, and D. Sarkar, "E-commerce website quality assessment based on usability," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, Mali, Apr. 2016, pp. 101–105.

[51] R. Yazdanifard and A. Zargar, "Today need of e-commerce management to e-skill trainings," *Int. J. e-Educ., e-Bus., e-Manag. e-Learn.*, vol. 2, no. 1, p. 52, 2012.

[52] N. Liu, X. Xu, T. Celik, Z. Gan, and H.-C. Li, "Transformation-invariant network for few-shot object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5625314.

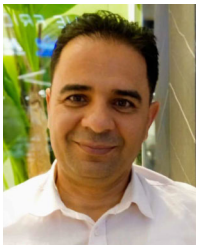
[53] J. Demagny, C. Roussel, M. Le Guyader, E. Guiheneuf, V. Harrivel, T. Boyer, M. Diouf, M. Dussiot, Y. Demont, and L. Garçon, "Combining imaging flow cytometry and machine learning for high-throughput schistocyte quantification: A SVM classifier development and external validation cohort," *eBioMedicine*, vol. 83, Sep. 2022, Art. no. 104209.

[54] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland. Springer, Sep. 2014, pp. 818–833.

[55] B. Ma, Y. Du, X. Zhou, and C. Yang, "A novel adaptive optimization method for deep learning with application to froth floatation monitoring," *Appl. Intell.*, vol. 53, no. 10, pp. 11820–11832, May 2023.

[56] X. Peng, Y. Chen, H. Liu, J. Xie, and C. Gu, "An online defect classification method for float glass fabrication," *Glass Technol.-Eur. J. Glass Sci. Technol. A*, vol. 52, no. 5, pp. 154–160, 2011.

EMAD IBRAHIM received the bachelor's degree in computer science from the University of Baghdad, Iraq, in 2005, and the master's degree in computer science from Acharya Nagarjuna University, India, in 2019. He is currently pursuing the Ph.D. degree in computer science with the University of Sfax, Tunisia, with a focus on information technology and artificial intelligence.



NIZAR ZAGHDEN received the master's degree in novel technologies in dedicated computer systems and the Ph.D. degree in computer system engineering from the National Engineering School of Sfax, Tunisia, in 2005 and 2013, respectively. His Ph.D. thesis is entitled characterization of the content of ancient document images. Concerning his professional career, he was recruited, in 2009, as an Assistant Professor with the Department of Computer Science, Superior Institute of Informatics in Medenine, Tunisia. In 2013, he was promoted to the rank of Assistant Professor. In 2014, he was moved to the Higher School of Business of Sfax, University of Sfax, Tunisia, as an Assistant Professor. He is currently working on intelligent applications for smart cities dealing with the classification of images and video from CCTV cameras.

MAHMOUD MEJDOUB received the engineering degree in computer engineering and the master's degree in novel technologies in dedicated computer systems from the National Engineering School of Sfax, Tunisia, in 2004 and 2005, respectively, the joint Ph.D. degree in computer system engineering from the National Engineering School of Sfax and in automatic, signal and image processing from the University of Nice Sophia Antipolis, France, in 2011, and the Habilitation (accreditation to supervise research) degree in computer system engineering from the National Engineering School of Sfax, in 2017. Regarding his professional career, he was recruited, in 2011, as an Assistant Professor with the Department of Computer Science and Communications, Faculty of Sciences of Sfax, Tunisia, where he was promoted to an Associate Professor, in 2018. He is currently a Research Member with the Research Unit Sciences and Technologies of Image and Telecommunications. His research interests include computer vision, image processing, artificial intelligence, and deep learning.