Yugoslav Journal of Operations Research
(20##), Number #, #-#
DOI: https://doi.org/10.2298/YJOR240615010I

Research article

OBJECT DETECTION IN VIDEO SUMMARIZATION FOR VIDEO SURVEILLANCE APPLICATIONS

Mohammed Inayathulla*

Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India
183030016.phd@gmail.com, ORCID: 0000-0001-9358-3687

Karthikeyan C

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India
ckarthik2k@gmail.com, ORCID: 0000-0002-8129-2687

Received: June 2024 / Accepted: November 2024

Abstract: For effective Data Extraction (DE) and Data Analysis (DA), the constant flow of visual information offers unique techniques in the Video Surveillance (VS) domain. This VS application demands the significance of advanced Object Detection (OD) for obtaining Video summarization in this study. Then, the accurate detection and the location of objects in the video frames are known as OD, as it is crucial for DE in the VS. To improve OD, the research offers advanced techniques like Faster R- (Convolutional Neural Networks) CNN or FRCNN using Inception ResNet V2 (IR-V2) via the application of CNN, Region Proposal Networks (RPN) and Deep Learning (DL). The empirical outcomes indicate that this suggested framework delivers improved OD accuracy of 93.5% compared to other techniques. By overcoming Big Data (BD) in the modern VS, the combination of sophisticated Computer Vision (CV) techniques with inception modules and residual connections.

Keywords: Video Surveillance, Object Detection, Faster R-CNN with Inception ResNet V2, Convolutional Neural Networks, Region Proposal Networks.

MSC: 68T45

*Corresponding author

1. INTRODUCTION

The constant flow of visual data in the dynamic field of VS and the effective techniques that may obtain DR from massive video streams is very important. In the framework of video summarization, the important features of CV emphasize the advanced OD techniques. The study named "OD in video summarization for VS applications". The accurate detection and location of objects within video frames is known as OD [1]. The VS includes detecting individuals, automobiles, or other objects [2].

The basis for DE in the video stream is that OD is an important procedure in video summarization [3]. Then, the procedure of compressing long video clips into smaller clips and managing summaries is called video summarization. The summary of vital data, incidents or actions is captured in the VS, which is the main purpose of summarization [4].

In the VSS (VS Systems), the effective VS, detection of problems, and events are achieved via effective video summarization. Through the VSS, this study will effectively resolve the problems of massive amounts of data. Then, quick comprehension of vital data in the observed locations by the surveillance operators may help provide accurate DE by combining OD into video summarization.

A major challenge in effective surveillance and VS analysis is the constant development of visual information in the current VS field. For DE, the current techniques are essential and employed due to the widespread surveillance cameras, which will lead to the vital enhancement of video clips. The challenge is the OD in video summarization, which is greatly emphasized. This analysis is crucial for OD and emphasizes significant data in dynamic visual contexts.

The current statistical analysis emphasizes the demands for OD enhancement methods to enhance OD techniques. The algorithms are necessary for delivering more accurate and adaptable data, as the widespread VSS applications produce it. The inconsistency was increased among the conventional VSS techniques, and adequate OD accuracy was revealed by statistical analysis. The increasing error rates and decreasing video summarization efficiency will also emphasize the necessity of OD development. Thus, the sophisticated technique is essential for OD enhancement [5].

Many challenges arise due to complicated real-life situations, and OD inconsistency will demand effective VSS. Then, accurate OD is very important due to the varying light conditions, challenges, and constantly changing surveillance settings [6]. The lack of current standard techniques will lead to the presence of the issues, and a major shift in the methodology is essential in the video summarization. An effective technique is essential for overcoming the OD's complicated technical features and managing unpredictable problems in surveillance settings.

To overcome those challenges, an effective solution must employ CE (Cutting Edge) techniques. Through the application of advanced CV techniques, ML (Machine Learning) techniques and NN frameworks, for exceeding the conventional OD techniques [7]. Then, video summarization in the VSS applications is obtained by combining these sophisticated aspects. The overall efficiency of VSS and accurate OD can be achieved by integrating techniques [8].

An enormous quantity of video data necessitating effective and precise analysis has resulted from the fast expansion in the installation of video surveillance systems. Due to inefficiencies caused by data complexity and volume, traditional video summarizing algorithms may miss important details regarding security monitoring. To promote fast and accurate analysis, there is an urgent need for sophisticated approaches that can consistently

recognize objects automatically and summarize films. The growing number of video surveillance systems and the resulting deluge of video data have prompted the need for more effective and precise analytic methods, which motivated this study. There may be security monitoring gaps since traditional video summarizing approaches can't handle the complexity and amount of this data. More effective monitoring of large-scale surveillance networks is possible with the help of automated object recognition and video summarizing, which ease the pressure on human operators. Improving object identification accuracy is critical for spotting suspicious activity and securing objects. In the framework of video summarization, this study is overcoming those issues. It also emphasizes OD. The suggested technique will support the promotion of the VSS. Then, the current statistical and in-depth analysis of current techniques will contribute to this study.

The main contribution of the paper is

- Designing Faster R- (Convolutional Neural Networks) CNN or FRCNN using Inception ResNet V2 (IR-V2) via the application of CNN for improving object detection,
- The research offers advanced techniques like Region Proposal Networks (RPN) and Deep Learning (DL),
- Experimental results have been performed, and the suggested model increases the accuracy and efficiency of object detection and reduces the error rate compared to other existing models.

The rest of the paper is prearranged as follows: Section 2 deliberates the literature survey, Section 3 proposes the FRCNN and Inception ResNet V2 model, Section 4 discusses the results and discussion, and Section 5 concludes the research paper.

2. LITERATURE SURVEY

An approach for weather-aware OD called Weather-OD has been suggested by Chen et al. [9]. An on-shore cloud-based system and an on-board edge system enhance marine surveillance. This approach utilizes ML frameworks for OD. These models are updated in real-time based on weather conditions to allow for accurate detection in maritime environments with minimal latency. To continuously improve Weather-OD training models using new data gathered from travel. In addition, it considers EC constraints and oversees the development of various OD frameworks. To expand training data diversity, Weather-OD includes synthetic image data, including weather noise, to simulate different weather situations.

Habib Khan et al. introduced a novel deep pyramidal refinement network backed by a Vision Transformer (ViT) [10]. The network predicts a crucial frame score and extracts and enhances multi-scale features. The suggested network has four primary parts, each with multiple stages.

By employing ViT backbone and an extensive predictive transformer, optimal representations from the input frames are extracted by this unique technique. Feature Maps (FM) from different layers are individually processed for merging as well as enhancing features of various sizes prior to proceeding to the final prediction module. Then, multi-level feature set is enhanced by a unique pyramidal refinement block and it has been priorly refined for predicting the critical outcomes. Then, one of the steps in developing the video summarization is known as Predictive frame selection.

A new OD technique was introduced, named as You Only Look Once, and None Left (YOLO-NL) and it has been suggested by Zhou et al. [11]. Novel dynamic label assigning

method was implemented, as it strikes a balance among the demand for more precise detection and more precise localization through assigning labels to those particular features. Thus, with the support of ViT backbone and an extensive prediction transformer, a unique technique has been employed for extracting such optimal representations form the input frames. Feature Maps (FM) from different layers are individually processed for merging as well as enhancing features of various sizes prior to proceeding to the final prediction module.

After that, a custom pyramidal refinement block improves the multi-level feature set before predicting crucial scores. In addition, they expedite the process of extracting features by applying the sequential SSPP structure. The suggested approach exhibits resilience in identifying items that have been resized, even in difficult situations characterized by factors such as dust, density, ambiguity, and obscured sceneries.

Yunzuo et al. [12] provided a spatiotemporal interaction method specifically developed for creating object tube sets. This program does this by recognizing and tracking extracted object tubes. For the duration of the video, the tubes are categorized into sections according to their proximity in time and space and the correlation among pairs of tubes. As a result, relationships among moving objects in the summary films can be preserved. In addition, a fusion method for frame sequences is introduced that combines temporal and spatial constraints for the best possible rearrangement. The frame sequence and frame vessel variables are analysed by this technique for the purpose of determining the initial time label for all tube set frame sequence. In this analysis, the tube set frame's initial precise location in the frame vessel is considered. From the 3 publicly available datasets, the suggested technique is assessed, as it offers extensive test with testing videos. Finally, the outcomes of the test will indicate that tha suggested technique achieved superior performance when compared to other CE methods.

A systematic classification and methodological study for VS has been suggested by Shambharkar et al. [13]. This method emphasis on the real-time video summary (RVS) methods. This study offers easy accessibility, and basic comprehension. It also paves a way for future study. This study will give significant research findings and data. Then, video summarization was effectively employed in smart cities due to its valuable applications like VSS anomalies detection.

A thorough analysis for the most recent techniques in Video summarization has been conducted by Sabha et al. [14]. It also includes traditional and modern methods. A classification system is suggested to categorize video summarizing techniques according to various characteristics. The research also examines the assessment processes used in various techniques, using benchmark datasets and performance measures. For each sub-category of video summarization, they defined and outlined the precise research issues. The results suggest that contemporary deep learning techniques have more precision than conventional methods but need more training time. Furthermore, methods that rely on handmade techniques have restricted effectiveness in dynamic video situations, and discrepancies such as changes in size or rotation are seen under varying lighting circumstances. Then, additional analysis is essential for multi-criteria-based video summarization. So, the researchers must emphasize the analysis of multi-criteria-based video summarization. In the dynamic CV field, significant data for novice researchers was facilitated by this study.

In Video summarization, a new technique was established by Nair et al. [15] for determining a pivotal frames in VS. In this technique, FV (Feature Vectors) from several

pre-trained Convolutional Neural Network models are employed named as Multi-CNN. Then, pre-trained 4 CNN models have been employed in the FE (Feature Extraction) technique. The produced vectors are subsequently fed to a Sparse Autoencoder, which combines them into a single input model for the FV. Through the integration of FV, specific frames in the raw video can be identified by the support of Random Forest (RF) Classifier. The Comparison of user reports from the ground-truth dataset to VSUMM and OVP are considered to be the effective means of analysing methods performance.

During the COVID-19 pandemic, instances of non-compliance with face mask regulations can be detected by the support of automated method ad it has been suggested by Sabha et al. [16]. Then, the the congested Scene Video Summarization Network, or CoSumNet, a new technique established for summarizing COVID-19 protocol violations in congested environments. By differentiating peoples those wearing masks or not can automatically generates the video summarization that clearly shows the congested environments.

In the congested areas, implementing suitable measures and penalizing violators for protocols by the support authorities have made the CoSumNet more effective. Through the "Face Mask Detection ~12K Images Dataset" benchmark, the CoSumNet is trained, and it can be trained for the purpose of analysing the effectivenss of the approach. Also, efficiency can be verified by the application of multiple real-time CCTV recordings.

For the purpose of understanding Image-Based DL frameworks, the Explainable AI (XAI) was suggested by Abdullah et al. [17]. 9 SOTA methods like GradCAM, GradCAM++, GradCAMElementWise, HriesCAM, RespondCAM, ScoreCAM, SmoothGradCAM++, XgradCAM, and AblationCAM were included in this toolkit. The above-mentioned techniques will demonstrate the AI model's decision-making process particularly DL frameworks. When dealing with visual input, all tools highlight the DL frameworks in DM by facing multiple aspects of comprehension. Through case study, the effectiveness of toolkits can be proved, it supports in offering the openess and understanding to AI analyzers.

For 1 Human Activity Recognition from Inertial Sensors, the Machine Intelligence Approach was suggested by Ali and Abdelhafeez [18]. Apart from existing methods, the suggested strategic fusion method allows for robust feature propagation as well as enhancing the diagnostic performance regardless of computational costs. Then, the overfitting and perforamance saturation problems can be effectively reduced by the MFR-CNN. In this way, the suggested model outperforms the existing models. From the analysis, the results revaled that the suggested model achieved 94% accuracy. The suggested model surpasses the conventional trained DCNN in terms of accuracy and efficiency. Finally, the suggested model also has the ability to improve brain tumour diagnosis more cost-efficiently when compared to ensembles. Then, the suggested model overtakes the conventional pre-trained and fine-tuned DCNN. To improve the diagnosis of brain tumours, offering better cost-efficiency, the suggested model is more effective than existing methods.

For Diagnosing Brain Tumors from MRI images, the Multi-Fused CNN with Auxiliary Layers was suggested by Alkhatib et al. [19]. Apart from existing methods, the suggested strategic fusion method allows for robust feature propagation as well as enhancing the diagnostic performance regardless of computational costs. The occurrence of overfitting and performance saturation has been reduced by the MFR-CNN. During evaluation, the suggested model outperformed traditionally trained DCNNs in terms of efficiency and

accuracy. The suggested model achieved 94% accuracy. The MFR-CNN showed ability in improving brain tumour detection at a lower cost when compared to ensembles and more traditional pre-trained and fine-tuned DCNN. When compared to current approaches, the suggested MFR-CNN, including AuxFL and FuRB, has great potential in enhancing the accuracy and cost-effectiveness of brain tumour diagnosis.

Maria Lincy Jacquline and Sudha [20] discussed the Weighted Fuzzy C Means (WFCM) and Enhanced Adaptive Neuro-Fuzzy Inference System (EANFIS) for Chronic Kidney Disease (CKD) Classification. Any number of possible deviations from the expected data set will affect the classifier's precision. Misclassified results are still increasing for Multi-Kernel Support Vector Machines (MKSVM). These issues are addressed by applying a min-max normalization-based preprocessing step to the input CKD data values to standardize their scale. Next, the author will use the Improved Fruit Fly Optimization Algorithm (IFFOA) to identify important characteristics. The author will cluster the chosen features using WFCM to reduce misclassification and forecast the data sample's class label. Finally, the EANFIS will be used to classify CKD as normal or abnormal. The recall, accuracy, precision, and f-measure test results show that the proposed method works.

Nourkhah et al. [21] deliberated on the Role of Sensors in Smart Agriculture. When civilian usage of Global Positioning System (GPS) capabilities became possible in the 1980s, "smart agriculture" began to take root. After farmers perfected their field mapping techniques, they could selectively administer fertilizer and herbicide to specific regions. Early adopters of precision agriculture used crop yield monitoring in the 1990s to determine when and how much fertilizer and pH adjustments were needed. Better suggestions for watering, fertilizer application, and even peak yield harvesting might be made if additional factors could be monitored and input into a crop model. Good farming practices have grown in value for ranchers of all sizes over time.

Mekawy [22] presented the Object Detection by Neural Network for Smart Home. Household object recognition is an innovative computer method that uses computer vision and image processing to identify things inside the home. This camera can find every item in the kitchen, bedroom, and other storage spaces. The term "object detection" describes the methodology used by low-end devices to identify humans in visual media. Video and image analysis have gotten us nowhere.

Khajehkhasan and Fakheri [23] introduced the Operational Strategies for Early Detection of Breast Cancers. The processing of mammographic pictures of patients' breasts began with their evaluation by medical professionals. So far, the author has been able to use fuzzy logic to speed up the cancer diagnosis and kind process. A novel method for distinguishing between benign and malignant malignancies is presented in this study. Tumours may be either benign (adenosis) or malignant (phyllodes tumour), with the former category including duct cancer and the latter including papillary cancer. The paper suggests a four-step process for detecting breast cancer. The procedure begins with preprocessing, then moves on to image analysis using wavelet transform, feature extraction based on the findings of wavelet transform, and finally, a fourth step. In other words, we classify pictures as benign or malignant using fuzzy logic.

Alzoubi et al. [24] investigated the Detection of Depression from Arabic Tweets Using Machine Learning. We classified whether the user is depressed or not. The author used machine learning methods, including Decision Tree (DT), Randon Forest (RF), Mutational Naïve Bayes, and AdaBoost, and employed feature extraction techniques such as BOW

and TF-IDF. Based on our studies, the most accurate method for grading tweets was Mutational Naïve Bayes with TF-IDF, which had an accuracy rate of 86%. Taking steps to ensure people's mental health throughout the early stages of an infection is crucial, so it's clear that caring for people's mental health is a top priority.

Wagner et al. [25] suggested the Automated Parameterization of a Sensorless Bearing Fault Detection Pipeline. To lessen the impact of human error caused by incorrect parameterizations, an AutoML pipeline search uses genetic optimization. To achieve this goal, a search space encompasses generic and domain-specific signal processing and modification approaches. The bearing failure detection use case assesses the proposed framework in real-world settings. Emphasis is placed on using the existing fault detection pipelines in a broader context. Evaluating the fault detection pipeline's resilience to changes in motor operating state parameters across the test and training domains was also a focus of experimental investigations.

Talouki et al. [26] proposed Image completion based on segmentation using neutrosophic sets. The exemplary-based image completion method begins by iterating from the outermost pixel with the highest priority until no pixel remains in the target area. To identify the patches most suitable for filling holes, our novel neutrosophic-based image segmentation method considers neighbourhood and similarity for the extended similarity measure. The strategy improved ASVS (Average Squared Visual Salience) by 18% compared to the previous approaches. The author achieved a PSNR of 38.96 and an MSSIM of 0.9919, while the best values for previous approaches were 0.9868 and 0.36.75, respectively, because of the neutrosophic segmentation effect.

Dirik [27] recommended the Fire Extinguishers Based on Acoustic Oscillations in Airflow Using Fuzzy Classification. This research takes advantage of a large dataset from many experiments that tested the efficacy of sound waves in firefighting. Using this large dataset with sound wave technology, the author created a fire suppression model. The model incorporates five distinct fuzzy logic algorithms: Fuzzy Rough Set (FRS), Vaguely Quantified K-Nearest Neighbors (VQNN), Fuzzy Rough K-Nearest Neighbors (FRNN), and Fuzzy Ownership K-Nearest Neighbors (FONN). The primary goal of these models is to differentiate between a flame's extinguished and non-extinguished states correctly. Several fundamental model characteristics, including fuel type, flame size, decibel level, frequency, airflow, and distance, contribute to this categorization.

Baig [28] discussed the Deep Attributes and Decisions Fusion for No-Reference Video Quality Analysis. This study developed a novel No Reference Video Quality Assessment (NR-VQA) architecture using features derived from pre-trained models of deep neural networks, transfer learning, periodic pooling, and regression. No manually generated features were used in our findings; instead, we relied only on dynamically pooled deep features. The authors of this article provide a new approach to NR-VQA based on deep learning. Their method involves using several pre-trained deep neural networks to identify potential distortions in images and videos simultaneously. Next, each pre-trained convolutional neural network (CNN) is mapped onto the subjective peer evaluations after extracting the features at the video level using spatial pooling and intensity adjustment. A video series' perceived quality is determined by summing the quality criteria from all the regressors. Experimental results on two massive baseline video quality analysis datasets with actual aberrations show that the proposed method establishes a new benchmark, according to many research efforts. The results demonstrate that NR-VQA may be substantially enhanced by integrating the judgments of many deep networks.

Several issues have befallen previous video summarization and object recognition approaches, including the trade-off between accuracy and speed. Fast models like YOLO have problems recognizing tiny or obscured objects, while more precise methods like Faster R-CNN are computationally intensive and not good for real-time use. Occlusions, noise, and objects of different sizes are additional challenges for conventional models, and they also struggle to generalize to new situations. Additional constraints on their efficacy include their significant resource consumption and reliance on high-quality training data. The suggested approach improves real-time performance and accuracy while reducing computational and generalizability difficulties by combining fast multi-scale feature extraction with robust detection capabilities. It does this by merging Faster R-CNN with Inception ResNet V2.

3. PROPOSED METHOD

In video summarizing, OD is essential because it helps with multimedia analytics, improving IR (Information Retrieval) and comprehension.

By employing the advanced DL algorithms like FR-CNN with IR- V2, the ability of OD in videos has greatly increased. For the purpose of making OD accurately and handling large amounts of video data quickly, Faster R-CNN Integration with Inception ResNet V2 is effective. It also facilitates for large scale surveillance networks. Then, this algorithm is designed especially for real-time processing, quick detection and response to significant cases, so, this algorithm is essential for security applications. Consistent performance across different surveillance situations are confirmed by its flexibility in multiple environmental circumstances. Operating expenses can be lowered by automated tasks like OD and video summarization. It will also eliminate the demand for human involvement. Inception ResNet V2's advanced feature extraction capabilities further enhance the model's ability to accurately represent and understand visual content for enhancing the overall effectiveness and reliability of surveillance operations. Thus, it will result in more effective and informative video summarization.

3.1. Faster R-CNN

To generate probability distribution, employing a vector with real values is the obejective. With the assurance that the sum of the probabilities equals 1, this function aims to transform a logit vector into a probability vector. When there is (MCC) Multi-Class Classification involved, Softmax is very effective.

An improved version of the R-CNN OD algorithm is called Faster R-CNN. It provides a standardized approach to identify and localize objects within images. The application of DL, CNNs, and RPN increases the efficacy and accuracy of the detection process. The Fast R-CNN detector and the RPN are critical components of the Faster R-CNN architecture.

The Faster R-CNN architecture's initial set of layers is called the CNN Backbone. The backbone creates FM by evaluating the input image; it can be created with networks like ResNet or VGG. The RPN and the Fast R-CNN detectors exploit the various levels of visual data captured by the FM. Examining the role of the CNN Backbone,

CNNs are particularly built to extract important features from input images. CNNs utilize many (CL) Convolutional Layers, each with a different kernel, to learn from the input data. These kernels are specifically made to extract the input image's structural representations. While the deeper layers of the CNN can recognize more complicated

structures like object components and structures, the initial levels recognize basic features like edges and textures.

The CNN Backbone retrieves similar hierarchical properties used by the RPN and the Fast R-CNN detectors. These layers improve throughput and memory utilization by performing calculations to classify inputs into different groups. This encoding of the hierarchical features provides the Softmax probability.

3.1.1. Region Proposal Network (RPN)

In earlier times, R-CNN and Fast R-CNN relied on Selective Search for region recommendations, which required the CPU and resulted in higher computation times because of the extra processing. However, Faster R-CNN addressed these issues with greater efficiency by employing RPN and NN, which significantly accelerated the detection of image regions, leading to an astounding 2-second reduction in processing time. Additionally, utilizing shared detection layers enhanced the feature representation.

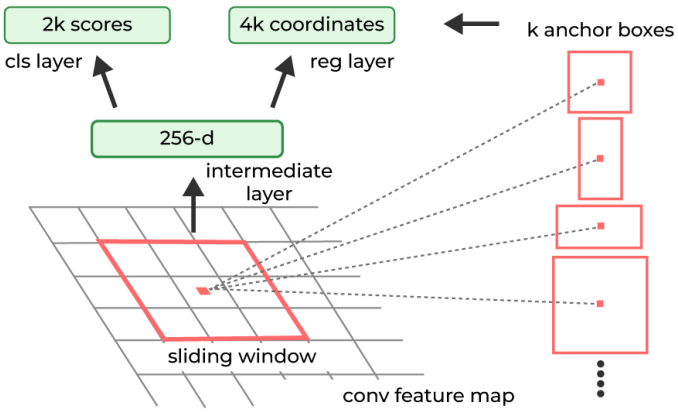

Figure 1: Region Proposal Network (RPN)

Within the Faster R-CNN model, the RPN is an essential component in Figure 1. The algorithm identifies possible Regions of Interest (ROI) in images that might contain objects. The NN directs the Fast R-CNN detector to the exact areas of the image where it is supposed to search for objects using an attention mechanism. The RPN's primary components are as follows:

Anchor boxes: Anchors are employed in the Faster R-CNN framework to help in RPN. The method uses a pre-made set of anchor boxes that differ in size and structure. On the FM, the anchor boxes are positioned at various scales.

An anchor box possesses two essential features:

- scale,
- aspect ratio.

The CNN backbone's FM is easily navigated using the RPN. An analysis of the (SW) sliding window's receptive field is done with a compact convolutional network, typically consisting of three layers. This method uses neural processing to produce regression values that enhance anchor boxes and scores that calculate the probability of OD.

The probability that an anchor box contains an important object rather than other noise is measured by its objectness. Every anchor receives a score from the RPN in the Faster R-CNN algorithm. This objectness score shows the probability that an anchor will be located in a location that has an important object. In the training process, this score determines the objectness of anchors by classifying them as either positive (to indicate the existence of an object) or negative (indicating background).

IoU (Intersection over Union): The term IoU is frequently used to describe the measurement of statistics for bounding box overlap. Concerning the two boxes' combined total area, the computation determines the percentage of the area that each box holds. It is stated mathematically as follows:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{1}$$

Non-Maximum Suppression (NMS): Utilizing overlapping suggestions' objectness scores, NMS is a strategy that removes redundancy and selects the most accurate concepts. The system eliminates all other suggestions and keeps only the one with the highest score.

The CNN backbone generates FM, which the RPN utilizes. Using anchor boxes of different sizes and shapes, the RPN uses a method known as SW on FM to identify possible object locations. The network modifies these anchor boxes to more accurately represent objects' actual locations and sizes throughout the training process.

For every anchor, the following 2 parameters was estimated by the Reverse Polish Notation (RPN) algorithm:

- The probability of the anchor containing an object (“objectness Score”),
- Modifications to the anchor's coordinates to match the object's precise shape.

Multiple area suggestions may intersect, suggesting the possibility of representing the same object or thing. NMS prioritizes anchor boxes with greater odds of containing an item by selecting the highest-scoring ones that meet a certain criterion. This technique ensures accuracy in the depiction of concepts and reduces unnecessary repetition. Here, Anchor boxes function as representations of region.

3.1.2. Fast R-CNN detector

For OD in proposals generated by RPN, faster R-CNN architecture relies on the Fast R-CNN detector.The region proposals generated by RPN can be combined by the process called ROI pooling. Before sent to next layer networks, this process normalizes the FM by converting RPN proposals of different sizes into uniform ones. Proposals are divided into equal-sized grids using ROI pooling, and max pooling is applied for each grid cell. Then, for every propisals, FM of predetermined dimensions can be generated by this method thus helps in adding network analysis in Figure 2.

- Feature Extraction: Similar to RPN technique, the important features of objects can be extracted with CNN backbone and RoI-pooled feature maps. The method extracts hierarchical data from predetermined locations, reducing complexity and maintaining geographic context to help the network comprehend proposal regions.
- Fully connected (FC) Layers: Newly generated RoI and FE can be analyzed by A sequence of FC layers. Object classification and adjusting the bounding box size are the functions included in this layer.
- Object Classification: To determine the possibility of each proposal area holding a certain kind of object for several classes, object classification was performed by the

network. This categorization combines region proposals' characteristics with the CNN backbone's learnt aspects.

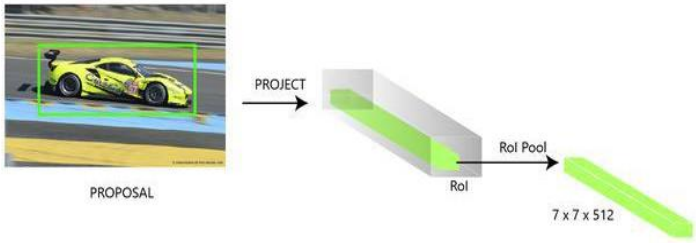

Figure 2: Region of interest pooling

- Box Regression: In addition to predicting the likelihood of each class, the network also predicts the necessary modifications to the bounding box for each suggested area. By adjusting the location and dimensions of the bounding box for the suggested area, the network enhances its precision and alignment with the item.
- The first layer consists of a softmax function with N+1 output, where N denotes the number of class labels plus the backdrop. This layer detects and classifies things inside the specified area. The next layer, responsible for bounding box regression, produces N output parameters, which properly predict the object's bounding box in the picture. The number of output parameters is four times N.
- Multi-task Loss Function: The Fast R-CNN detector utilizes a multi-task loss function that combines classification and regression losses. While the regression loss examines the variance in bounding box modifications among the actual and predicted values, the classification loss assesses the disparity among the actual and predicted probabilities for every class.
- Post-Processing: The network estimates class probabilities and modifies bounding boxes before using a post-processing approach to improve the final detection outcomes. In this stage, NMS is employed to reduce the number of duplicate detections and maintain the most dependable and non-overlapping detections.

3.2. Inception method

Google's research team developed the Inception model architecture, also called GoogLeNet. It is a sophisticated CNN. The idea was to solve the problem of maintaining high accuracy in DL models while efficiently using computational resources. Fundamentally, many convolutional processes of different sizes are used inside a single layer through inception modules.

The Inception framework consists of max-pooling layers and many CLs with varying filter sizes, including 1x1, 3x3, and 5x5 convolutions. The model can identify features at different spatial scales since the convolutional processes are combined simultaneously in inception modules. To reduce computational complexity and the number of parameters in the network, 1x1 convolutions are employed for DR (Dimensionality Reduction).

Regarding retaining computational economic growth, the Inception model architecture excels at delivering top performance in image classification tasks.

By employing inception modules, the local data and global data can be effectively extracted by the model. Thus, the inception modules can enhance the accuracy when compared to conventional CNN models.

The Inception model can be used by embedded systems and mobile phones with low resources because of its computational efficiency. This enables the application of DL in a variety of real-world scenarios with constrained computational resources.

The inception model was employed by many CV tasks like image segmentation, OD, and (Image Classifications) IC. Due to its versatility and efficiency, researchers and experts in DL field selects this method. The structure of the inception model influences the advancements in NN structure. It will results in more reliable and effective models for many applications.

3.3. ResNet V2

The efficacy and effectiveness of Deep NN (DNN) have been greatly enhanced by ResNet V2, an update of the original ResNet design created by Microsoft Research in 2015. The primary feature of ResNet V2 is the presence of residual connections, which allow deep network training without encountering issues such as fading gradients. The network learns residual mappings with the help of the residual connections, which makes training process optimization easier.

In architecture, ResNet V2 builds upon the core elements introduced in the previous ResNet. The structure includes a large number of residual blocks using CL, (BN) Batch Normalization, and (AF) Activation Functions like ReLU. ResNet V2 has bottleneck layers that employ 1x1 convolution to initially lessen and then maximizing the number of channels in every block for the purpose of reducing the computational cost of training deeper networks.

By employing ResNet V2, NN with a lot of layers can be effectively trained. ResNet V2 may be able to solve issues with deep network training, such the degradation problem, by implementing residual connections and other structural improvements. ResNet V2 is more effective than previous iterations, as bottleneck layers reduce the computing required for training deep networks.

ResNet V2 has been used in several domains, including speech recognition (SR), natural language processing (NLP), and CV. For semantic segmentation, OD, and IC tasks, ResNet V2 is widely employed in CV. Due to the ability of learning complex representations from visual input, it is suitable for tasks involving understanding and analyzing images. ResNet V2's capacity to produce hierarchical representations of text input has been used to adapt it for implementation in NLP applications, including text classification (TC), sentiment analysis (SA), and automatic translation.

By employing SR tasks like recognizing voices, ResNet V2 has demonstrated its versatility across multiple domains

From ResNet V2, many ML applications benefits significantly. Thus, ResNet V2 is a robust model and also effective in training procedure.

3.4. Inception ResNet V2 Architecture

Convolution Layer

The CL is crucial in domains like CV and NLP, and CNNs rely significantly on it. Using a procedure known as convolution for FE from the input data is the main objective in Figure 3. This layer uses a collection of adaptable filters called kernels or feature detectors to handle the new input. Each filter moves across the input volume by convolving the dot product of its weights and the specific area of the input area it is analyzing. The final product of this procedure is called as FM, often called Activation Maps (AM).

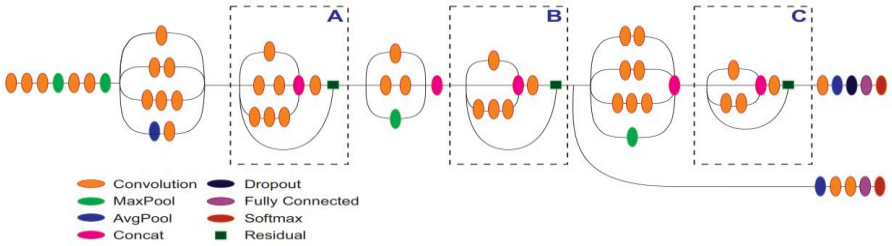

Figure 3: Inception ResNet V2 architecture

The spatial feature strutures of input data can be effectively captured by the convolution method. With time, the network learns filters that highlight particular components like edges, textures, or forms, enabling it to recognize ever-more-complex patterns. Development of a hierarchical representation of the input is the outcome of this. Some factors that impacts the output FM dimensions are the following: input volume size, filter size, stride, and (ZP) Zero-Padding value. For controlling the network's capacity, computational effectiveness, and receptive field, these parameters are crucial since they impact the output FM's spatial dimensions.

CL offers the advantages of weight sharing across several input data areas. To improve efficiency and facilitate learning from limited datasets, weight sharing reduces the number of parameters in the network. Through local connection and weight sharing, CL facilitates learning translation-invariant features essential for tasks like OD and IC.

MaxPooling Layer

A crucial component of many CNNs, particularly in (IR) Image Recognition applications, is the MaxPooling layer. The major goal is to use (DS) Downsampling to reduce the spatial dimensions of FM produced by CL while maintaining significant data. Reduced overfitting and increased computation speed result in DS, which helps control the number of parameters in the network and computational complexities. The MaxPooling layer operates in a simple yet effective manner. The procedure selects the highest value inside each window as it moves a window of a specific size, typically 2 x 2 or 3 x 3, over the input FM, acting on each FM independently. The remaining numbers in the window are eliminated, leaving only the highest value.

The layer filters out less significant data while keeping the most prominent features from the input.

MaxPooling is advantageous because it may bring translation invariance to the network. The layer only stores the highest values in each window, making it less affected by minor changes in feature positions in the input. This trait enhances the ability of Convolutional Neural Networks (CNNs) to perform well on new input and increases their resistance to translations, rotations, and other changes. MaxPooling may result in information loss by discarding non-maximum values in each window. In certain situations, the lack of knowledge may be harmful, particularly in jobs that need detailed geographical information. Several pooling methods, like AveragePooling or GlobalPooling, have been suggested to overcome this restriction. These approaches calculate the average value or a summary statistic throughout the window instead of selecting the largest value.

AvgPooling Layer

The Average Pooling layer is a key element in convolutional neural networks (CNNs) for extracting features and reducing dimensionality. This layer processes individual feature maps by decreasing their spatial dimensions while preserving crucial data. The main purpose of the Average Pooling layer is to decrease the input volume's size and complexity while retaining the most important attributes. The Average Pooling layer moves a constant-sized window (usually 2x2 or 3x3) over the input feature map during the forward pass, calculating the components' average inside each window. This procedure efficiently decreases the spatial dimensions of the feature map based on the pooling window size and the sliding operation stride. The layer summarizes local information by calculating the average values inside each window, which helps preserve a representation of the most prevalent aspects in the input.

The Average Pooling layer's main benefit is its capacity to provide translational invariance, reducing the network's sensitivity to minor changes or distortions in the input data. This characteristic enhances the network's ability to handle changes in input, which is beneficial in tasks like image classification, where item positions in an image might vary. The Average Pooling layer helps decrease the network's computational cost and memory needs by decreasing the parameters and operations in the following levels. The Average Pooling layer decreases the spatial resolution of the data by downsampling the feature maps, resulting in quicker computation and reduced memory use while maintaining performance.

It is understood that the pooling procedure may result in information loss, particularly when the pooling window size or stride is very long. Important features may be overlooked in such instances, impacting the network's capacity to represent and categorize incoming data precisely. Hence, it is crucial to meticulously adjust the pooling parameters, such as window size and stride, according to the task's unique needs and the input data's properties.

Concatenation Layer

For processing sequences or multimodal input tasks, concatenation layer is crucial in NN structures. By combining or linking the results of many previous layers or branches of the network by which the layer operates. To enable the model to learn more complex representations and understand more complex relationships within the data, the concatenation layer's primary function is to combine features from distinct modalities or network segments.

In sequence-to-sequence models for tasks like machine translation and text summarization, concatenation layer is often employed in NLP. A sequence of hidden states representing the input text can be generated by the encoder and sequence of hidden states representing the output text can be generated by the decoder in these frameworks. At each time step, concatenation layer merges the hidden states from the encoder and decoder. It also facilitates in generating the output sequence while the decoder concentrates on relevant portions of the input sequence.

In multimodal DL, concatenation layer is crucial. Because the model analyzes data from many modalities, including pictures, text, and audio. From varied sources, model may integrate information. By combining FE from distinct modalities for enhancing predictions.

The concatenation layer improves sentiment classification accuracy in a multimodal sentiment analysis task by combining textual information from captions with visual features from images.

Dropout Layer

To prevent overfitting and improve generalization, the Dropout layer is a regularization technique used in NN, particularly DL models, during training, it works by randomly dropping out (i.e., setting to zero) a certain proportion of input units. In other words, for certain units, the layer's output will be zero, so "dropping out" some data from the network. By randomly dropping out units, Dropout introduces noise to the network during training, which helps to prevent the co-adaptation of neurons. Co-adaptation occurs when some neurons rely too heavily on the presence of other specific neurons, making the network less robust to variations in input data.

One of the key benefits of Dropout is that it acts as an ensemble method by training multiple subnetworks within the original network. Each subnetwork is trained to perform the task independently, with different subsets of neurons dropped out at each iteration. During inference (i.e., when making predictions), the Dropout layer is typically turned off, and the full network is used, but the weights are scaled to account for the dropout during training.

Fully Connected Layer

The FC layer is vital in several DL models, particularly in artificial NN. It is sometimes called the thick layer due to its compact network structure. Every neuron in this layer is linked to each in the preceding layer, making it known as "fully connected." This connectivity allows the layer to identify complex patterns in the input data, making it suitable for classification, regression, and feature learning tasks. FC layers are recognized for their ability to obtain hierarchical representations of the input data. Every neuron in the FC layer analyzes a mix of characteristics from the previous layer to generate a more complex representation of the input as data flows through the network. This hierarchical form enables the network to extract relevant components and make accurate predictions using the learned patterns.

The fully connected layer executes a sequence of matrix multiplications and non-linear transformations on the input data. Every individual neuron inside the layer calculates a weighted sum of its input values and then employs an activation function to induce nonlinearity. Some often used activation functions are sigmoid, tanh, and rectified linear unit (ReLU). The functions include non-linear elements in the model, enabling it to comprehend intricate connections between input and output variables. Training a fully connected layer entails optimizing the weights and biases to minimize a predetermined loss function. Backpropagation is a method that modifies the parameters of a neural network by using the gradients of the loss function concerning the model's parameters. The network improves its performance by using iterative optimization methods, such as stochastic gradient descent (SGD) and its variations, to comprehend the underlying structure of the data.

Layer implementing the softmax function:

Softmax

The softmax function is a common mathematical procedure in machine learning and statistics. The goal is to create a probability distribution from a vector of real numbers. An input vector, also called logits, is transformed by the function into a vector of probabilities that add up to 1. When assigning an input to one of many classes is the aim of a multi-class classification task, the softmax function is handy. The following formula is used to compute the softmax function:

$$Softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (2)$$

z, in this instance, symbolizes the input vector that contains logits. z_i is the i-th element of z, whereas K is the total number of classes. The function exponentiates each input vector element and then normalizes the results by dividing each exponentiated value by the total sum of all exponentiated values in the vector. Exponentiation amplifies differences in the logits by emphasizing larger values and diminishing the impact of smaller ones. Normalizing the vector by dividing each member by the total ensures that the resulting vector represents a valid probability distribution, with each element indicating the probability of its corresponding class.

Softmax is very sensitive to changes in the relative magnitudes of the logits. Small changes in the input may lead to significant changes in the output likelihood. Softmax is ideal for optimization jobs requiring distinct separation between classes, particularly in classification problems. The softmax function is frequently employed as the activation function for the result layer of a neural network in classification tasks. The predicted class is determined by picking the index in the softmax outcome vector with the greatest probability. Combining the softmax function with a loss function like cross-entropy is customary during training. The loss function quantifies the difference between the expected probability and the actual distribution of class labels. The softmax function and cross-entropy loss train the model and get precise probability estimates for each class.

Residual Layer

Residual layers are essential in deep neural networks, especially in structures such as ResNet. Residual layers were introduced to solve the vanishing gradient issue by improving the flow of gradients during backpropagation, allowing for the training of much deeper networks.

Residual layers use skip connections, sometimes called shortcut connections, to enable information from previous levels to bypass one or more layers and flow straight into deeper layers. This allows the network to learn residual functions rather than attempting to learn fundamental mappings directly, which facilitates the optimization and convergence of the network.

A residual layer processes the input through convolutional or fully connected layers and then applies a non-linear activation function like ReLU. The input to the layer is appended (element-wise) to the layer's output, establishing a residual link rather than transferring the output of these layers straight to the next layer. Mathematically, this may be expressed as output = input + F(input), where F (input) represents the residual function being learnt by the layer.

Deep networks may efficiently transmit gradients throughout the network by including residual connections, which helps alleviate disappearing gradients in deep topologies. This allows for training deeper neural networks, which have shown improved accuracy across many tasks such as image classification, object identification, and semantic segmentation.

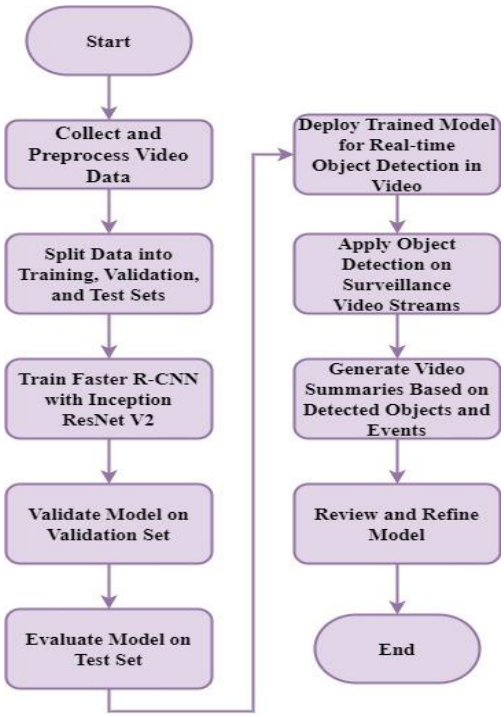

Figure 4: Overall flowchart of the proposed model

Figure 4 shows the overall flowchart of the proposed model. The Inception ResNet V2 enables rapid training with residual connections and advanced multi-scale feature extraction with its inception modules. Faster R-CNN is a strong framework for region proposal and object recognition. Because of this complementary relationship, the model can accurately recognize objects in various settings by capturing various features. Combined with these architectures, the algorithm can handle massive amounts of surveillance video data in real time with more accuracy, efficiency, and scalability. A major step forward in the area, this innovative integration solves important problems like computing complexity and resilience while increasing monitoring systems' operating efficiency and detection accuracy. The method uses the Faster R-CNN Inception ResNet V2 1024x1024 framework to identify objects in video summarization. Faster R-CNN is a sophisticated framework for object identification that integrates deep learning, CNNs, and region proposal networks (RPNs) to detect objects in images accurately. The system has two main components: the RPN and the Fast R-CNN detector. RPN utilizes anchor boxes and a sliding window approach to create probable regions of interest, while Fast R-CNN use methods such as RoI pooling and feature extraction to identify objects inside these areas. Inception ResNet V2 merges the efficiency of the Inception model with the depth of ResNet V2 by using inception modules for optimal use of computer resources and residual

connections to train deep networks successfully. The design includes convolutional layers, pooling layers, fully connected layers, and specialized layers such as concatenation, dropout, softmax, and residual layers. This combination enables strong object detection in video summarizing applications.

4. EXPERIMENTAL RESULTS

This part thoroughly examines the outcomes obtained from the simulations following the suggested approach. The data are taken from the YouTube video dataset [29]. Video summarization aims to extract the most relevant and useful information from a video and condense it into a summary. The final product is often a condensed version of the original video made from a collection of representative frames (also called video key-frames) or individual video clips (also called video key-fragments) stitched together in a specific sequence. The video storyboard describes the first video summary, whereas the video skim describes the second in Figure 5 (a,b,c,d,e) and Figure 6 (a,b,c,d,e,f).

Figure 5: Sample screenshots from the dataset (a)

Figure 5: Sample screenshots from the dataset (b)

Figure 5: Sample screenshots from the dataset (c)

Figure 5: Sample screenshots from the dataset (d)

Figure 5: Sample screenshots from the dataset (e)

Figure 6: Object detection results (a)

Figure 6: Object detection results (b)

Figure 6: Object detection results (c)

Figure 6: Object detection results (d)

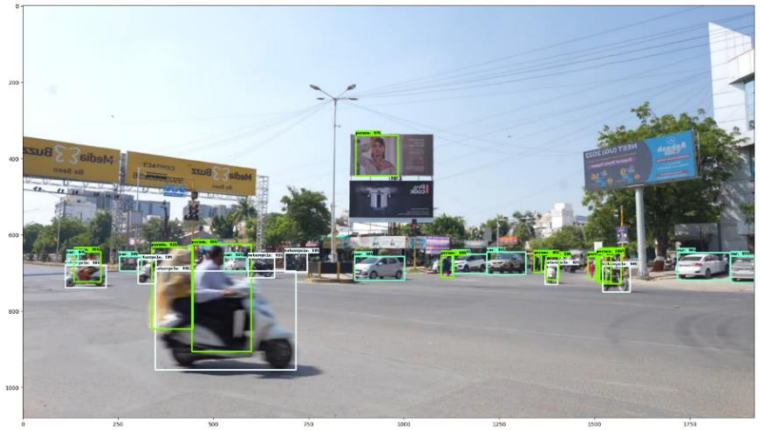

Figure 6: Object detection results (e)

Figure 6: Object detection results (f)

Table 1: Comparative analysis

Method	Accuracy (%)
CenterNet HourGlass104 1024x1024	58.5
EfficientDet D4 1024x1024	48.25
SSD ResNet50 V1 FPN 1024x1024	58.75
Faster R-CNN ResNet50 V1 640x640	74
Proposed Faster RCNN Inception ResNet V2 1024x1024	93.5

The Table 1, compares different object detection methods based on their accuracy percentages across various architectures and resolutions. CenterNet HourGlass104 at 1024x1024 achieves an accuracy of 58.5%, while EfficientDet D4 at the same resolution reaches 48.25%. SSD ResNet50 V1 FPN and Faster R-CNN ResNet50 V1 perform slightly better at 1024x1024, with 58.75% and 74% accuracy, respectively. Notably, the proposed Faster R-CNN with Inception ResNet V2 at 1024x1024 resolution outperforms all others significantly, achieving an accuracy of 93.5%, indicating its superiority in object detection tasks. Figure 7 shows the comparison of different methods for object detection.

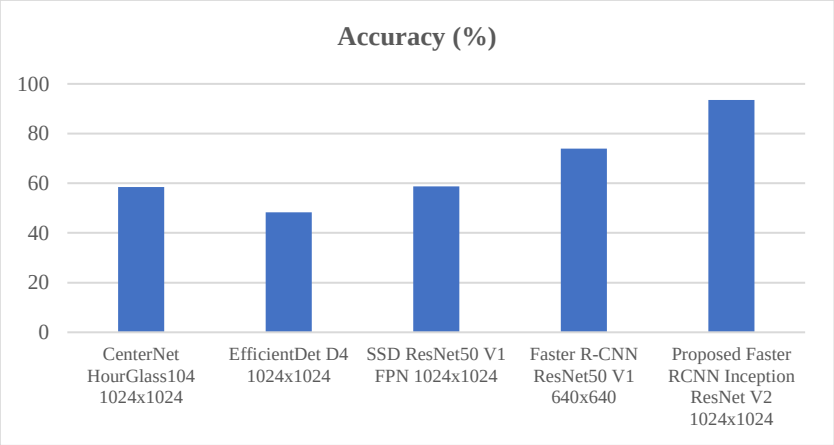

Figure 7: Comparison of different methods

5. CONCLUSION

The study on Object identification in Video Summarization for Video Surveillance Applications has highlighted the crucial role of advanced object identification methods in improving video summarization for surveillance purposes. The work improved object detection precision by merging modern deep learning models like Faster R-CNN with Inception ResNet V2, obtaining an accuracy of 93.5%. This research introduces a new method that combines Faster R-CNN with Inception ResNet V2 to improve object recognition in video summarization for video surveillance applications. Improving accuracy, efficiency, and resilience are three major issues with current surveillance systems that the suggested method aims to resolve. The algorithm outperforms the competition when extracting important events from surveillance film by combining the powerful object detection framework of Faster R-CNN with the sophisticated feature extraction capabilities of Inception ResNet V2. The empirical analysis demonstrates that the model is better than previous methods, showing a substantial improvement in addressing the constraints of

current surveillance technologies. The unique integration of inception modules and residual connections enhances computing efficiency and improves the training of deep networks, ultimately enhancing object recognition in video surveillance. The training and real-time processing of Faster R-CNN integrated with Inception ResNet V2 need substantial computing resources. This might be a problem for large-scale monitoring systems or places with limited computing capacity.

Funding: This research received no external funding.

REFERENCES

[1] A. Yali. P. Felzenszwalb and P. Girshick, Object detection. In *Computer Vision: A Reference Guide*, pp. 875-883, 2021.

[2] A. Karbalaie, F. Abtahi and M. Sjöström, Event detection in surveillance videos: a review. *Multimedia tools and applications*, vol. *81, no.* 24, pp. 35463-35501, 2022. doi: 10.1007/s11042-021-11864-2

[3] Y. Zhang, X. Liang, D. Zhang, M. Tan and E. P. Xing, Unsupervised object-level video summarization with online motion auto-encoder. *Pattern Recognition Letters*, vol. *130*, pp. 376-385, 2020. doi: 10.1016/j.patrec.2018.07.030

[4] V. Tiwari and C. Bhatnagar, A survey of recent work on video summarization: approaches and techniques. *Multimedia Tools and Applications*, vol. *80, no.* 18, pp. 27187-27221, 2021. doi: 10.1007/s11042-021-10977-y

[5] P. Pareek and A. Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, vol. *54, no.* 3, pp. 2259-2322, 2021. doi: 10.1007/s10462-020-09904-8

[6] J. Stjepandić and M. Sommer, Object recognition methods in a built environment. *DigiTwin: An Approach for Production Process Optimization in a Built Environment*, pp. 103-134, 2022. doi: 10.1007/978-3-030-77539-1_6

[7] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan and J. Walsh, Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1, no. 1*, pp. 128-144, 2020. doi: 10.1007/978-3-030-17795-9_10

[8] C. B. Murthy, M. F. Hashmi, N. D. Bokde and Z. W. Geem. Investigations of object detection in images/videos using various deep learning techniques and embedded platforms—A comprehensive review. *Applied sciences*, vol. *10, no.* 9, pp. 3280, 2020. doi: 10.3390/app10093280

[9] M. Chen, J. Sun, K. Aida and A. Takefusa, Weather-aware object detection method for maritime surveillance systems. *Future Generation Computer Systems*, vol. *151*, pp. 111-123, 2024. doi: 10.1016/j.future.2023.09.030

[10] H. Khan, T. Hussain, S. U. Khan, Z. A. Khan and S. W. Baik, Deep multi-scale pyramidal features network for supervised video summarization. *Expert Systems with Applications*, vol. *237*, pp. 121288, 2024. doi: 10.1016/j.eswa.2023.121288

[11] Y. Zhou, A YOLO-NL object detector for real-time detection. *Expert Systems with Applications*, vol. *238*, pp. 122256, 2024. doi: 10.1016/j.eswa.2023.122256

[12] Z. Yunzuo and Z. Tingting, Object interaction-based surveillance video synopsis. *Applied Intelligence*, vol. *53, no.* 4, pp. 4648-4664, 2023. doi: 10.1007/s10489-022-03477-5

[13] P. G. Shambharkar and R. Goel, From video summarization to real time video summarization in smart cities and beyond: A survey. *Frontiers in big Data*, vol. *5*, pp. 1106776, 2023. doi: 10.3389/fdata.2022.1106776

[14] A. Sabha and A. Selwal, Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions. *Multimedia Tools and Applications*, vol, 8*2, no.* 21, pp. 32635-32709, 2023. doi: 10.1007/s11042-023-14925-w

[15] M. S. Nair and J. Mohan, Static video summarization using multi-CNN with sparse autoencoder and random forest classifier. *Signal, Image and Video Processing*, vol. *15, no.* 4, pp. 735-742, 2021. doi: 10.1007/s11760-020-01791-4

[16] A. Sabha and A. Selwal, CoSumNet: A video summarization-based framework for COVID-19 monitoring in crowded scenes. *Artificial Intelligence in Medicine*, vol. *139*, pp. 102544, 2023. doi: 10.1016/j.artmed.2023.102544

[17] W. Abdullah, A. Tolba, A. Elmasry and N. N. Mostafa, VisionCam: A Comprehensive XAI Toolkit for Interpreting Image-Based Deep Learning Models. *Sustainable Machine Intelligence Journal*, vol. *8*, pp. 4-46, 2024. doi: 10.61356/SMIJ.2024.8290

[18] A. M. Ali and A. Abdelhafeez, DeepHAR-Net: a novel machine intelligence approach for human activity recognition from inertial sensors. *Sustainable Machine Intelligence Journal*, vol. *1*, pp. 1-1, 2022. doi: 10.61185/SMIJ.2022.8463

[19] A. J. Alkhatib, M. Alharoun, A. Alzoubi, E. Muqdadi and A. A. Aqoulah, Diagnosing Brain Tumors from MRI images through a Multi-Fused CNN with Auxiliary Layers. *Sustainable Machine Intelligence Journal*, vol. *6*, pp. 2-1, 2024. doi: 10.61356/SMIJ.2024.66102

[20] M. Lincy Jacquline and N. Sudha, weighted fuzzy C means and enhanced adaptive neuro-fuzzy inference based chronic kidney disease classification. *Journal of Fuzzy Extension and Applications*, vol. *5, no.* 1, pp. 100-115, 2024. doi: 10.22105/jfea.2024.437690.1376

[21] S. A. Nourkhah, G. Cirovic and S. A. Edalatpanah, The role of sensors in smart agriculture. *Computational Algorithms and Numerical Dimensions*, vol. 2, no. 4, pp. 210-215, 2023.

[22] I. Mekawy, Object Detection by Neural Network for Smart Home. *Big Data and Computing Visions*, vol. *2, no.* 4, pp. 143-148, 2022. doi: 10.22105/bdcv.2022.333756.1072

[23] S. Khajehkhasan and S. Fakheri, A new method based on operational strategies for early detection of breast cancers. *Innovation Management and Operational Strategies*, vol. *1, no.* 2, pp. 187-201, 2020. doi: 10.22105/imos.2021.266543.1027

[24] A. Alzoubi, A. Alaiad, K. Alkhattib, A. J. Alkhatib, A. A. Aqoulah and O. Hayajnah, Detection of Depression from Arabic Tweets Using Machine Learning. *Sustainable Machine Intelligence Journal*, vol. *6*, pp. 3-1, 2024. doi: 10.61356/SMIJ.2024.11103

[25] T. Wagner, A. Gepperth and E. Engels, A framework for the automated parameterization of a sensorless bearing fault detection pipeline, 2023. *arXiv preprint arXiv:2303.08858*. doi: 10.22105/jarie.2023.391005.1538

[26] A. G. Talouki, A. Koochari and S. A. Edalatpanah, Image completion based on segmentation using neutrosophic sets. *Expert systems with applications*, vol. *238*, pp. 121769, 2024. doi: 10.1016/j.eswa.2023.121769

[27] M. Dirik, Fire extinguishers based on acoustic oscillations in airflow using fuzzy classification. *Journal of fuzzy extension and applications*, vol. *4, no.* 3, pp. 217-234, 2023. doi: 10.22105/jfea.2023.401426.1291

[28] A. Baig, Deep attributes and decisions fusion for no-reference video quality analysis. *Big Data and Computing Visions*, vol. *3, no.* 3, pp. 91-103, 2023, doi 10.22105/bdcv.2023.415895.1165

[29] https://www.youtube.com/watch?v=4zfVFFw5nOg