# Towards Scalable Video Analytics at the Edge


**Nguyen Van Dien**

Supervisor: Prof. Jaehyuk Choi

Prof. Eun-Seok Ryu

Department of IT Convergence Engineering

Gachon University

A dissertation submitted to the department of IT Convergence Engineering and the committee on graduate studies of Gachon University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

January 2021

# Acknowledgements

First, I would like to express my gratitude and appreciation to my major advisor, Prof. Jaehyuk Choi's mentorship and support, I was able to get access to numerous opportunities, develop myself to be a better researcher, and I hope our cooperation will continue in the future. Thanks to Prof. Eun-Seok Ryu for letting me join the Graduate School to study and research at Gachon University. I also want to thank my committee members: Prof. Myung-Mook Han, Prof. Yong Ju Jung and Prof. Joohyung Lee for supporting me complete the PhD program as well as all the valuable inputs to my dissertation. Additionally, the discussions with them helped me broaden my perspective on my research.

I am grateful to the support from my family in VietNam. I thank all my friends at Gachon and many others. Finally, a very special thanks to my wife, a woman who is excellent in all aspects, beautiful, smart, and generous. I am very grateful for her encouragement, her patience, and all her love and tenderness.

# Abstract

Video Analytics(VA) systems have an important role in many security areas such as public security, transportation and other industry fields. An important feature of these systems is that it is difficult and critical to analyze at real-time speed to generate timely event alerts on different scenario application. Because of the bandwidth and computation limitation, VA servers can not meet the real-time requirements when serving for many cameras simultaneously. In order to overcome these challenges, this dissertation propose a method to minimizes transfer of video data from the surveillance camera sources to a video analytics server on the cloud by using the edge-computing based system. The proposed method was conducted and published in our lab publication [46]. By utilizing the information from the encoded bitstream, the edge device can obtain low-complexity of the moving object tracking and filters non-motion frames from the list of frames which will be forwarded to the cloud server. In order to show the effectiveness of the proposed solution, we deployed a video surveillance system which consist of an edge device with low computational capacity and a hight computing GPU-enabled server. The performance evaluation results show that our method is compatible with the edge-to-cloud platform and can efficiently catch the motion of frame. The average inference time of this method is around 39 ms/frame with high definition(HD) resolution video, which is the state of art method. The experiment results show that the method helps the cloud server reduce 48% utilization of GPU, 48% load of CPU, and 51% of download network traffic while still maintaining the accuracy of real-time video analytics event alerts.

# Table of contents

# List of figures

# List of tables

# Nomenclature

**Acronyms / Abbreviations**

AVC Advanced Video Coding

CCTV Closed Circuit Television

CNN Convolutional Neural Networks

CPU Central Processing Unit

CUDA Compute Unified Device Architecture

FPS Frames Per Second

GPU Graphics Processing Unit

HEVC High Efficiency Video Coding

IoU Intersection Over Union

MCP Motion-Compensated Prediction

MV Motion Vector

R-CNN Regions with CNN

RAM Random Access Memory

RTSP Real-Time Streaming Protocol

SSD Single Shot Multi-box Detector

VA Video Analytics

VMS Video Management Software

YOLO You Only Look Once

# Chapter 1

# Introduction

Nowadays, the world is witnessing a very fast increment of camera deployment [10] as shown in Figure 1.1 with cities and organizations steadily increasing the size and reach of their deployments. For example, cities now deploy tens thousands of cameras, each continually collecting and streaming rich video data [1], [2], [3]. According to IHS Markit's annual report [50], as of 2018, "China has one camera for each 4.1 people in the country and the United State has a people-to-camera ratio of 4.6-to-1". The massive deployments of cameras are come mainly from the growth of the video surveillance industry because of increasing the concerns about public security and safety. With this trend, the intelligent video

(a) (b)

Fig. 1.1 Surveillance Camera Deployment and Application

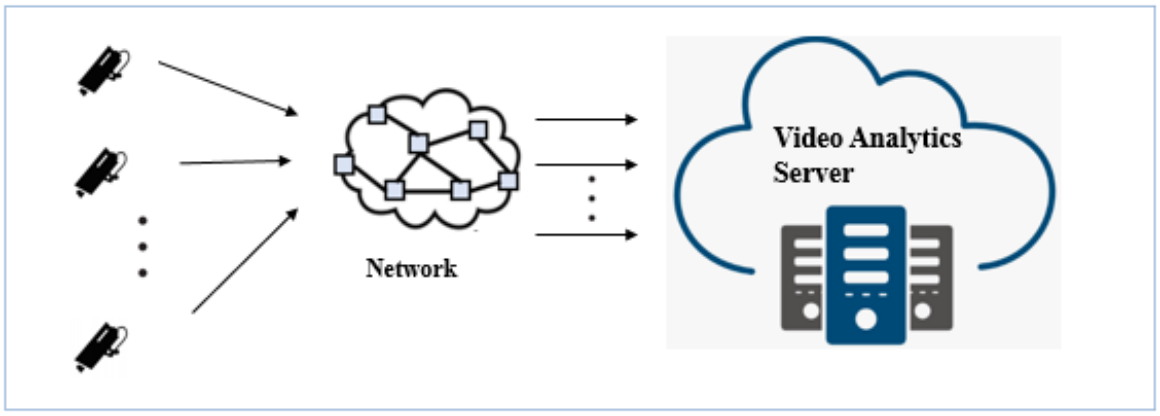



Fig. 1.2 Overall of cloud based video analytics server

analytics systems have been playing an essential and important role, performing main tasks in various fields including surveillance area, transportation monitoring, manufacturing, etc. Video analytics software analyzes videos in order to detect events the system is programmed to look for – when something is moving in front of the camera.These systems analyze video sources to handle long-running tasks such as traffic monitoring, customer tracking, and surveillance. The important key to the success of such applications has been recent advances in computer vision, particularly neural network based techniques for highly accurate object detection and recognition [15], [34], [41].

In a typical real-time video analytics pipeline as Figure 1.2 illustrates a traditional cloud-based video stream analytics platform. Video analytics is performed in a a central cloud and a large number of camera transfer video data to there. However, this traditional solution makes it difficult to perform real-time video analytics on live video streams from many cameras because the VA involves several computation-consuming tasks such as object detection, object tracking, object recognition and so on. Besides, streaming video from multiple camera to the cloud consumes a lot of network bandwidth over limited-bandwidth networks, which leads to high latency, causing significant challenge for real-time video analytics.

Many works has been conducted to improve the efficiency of video analytics platform [16], [18], [26], [29]. Accross these systems, a typical strategy is to improve efficiency by filtering

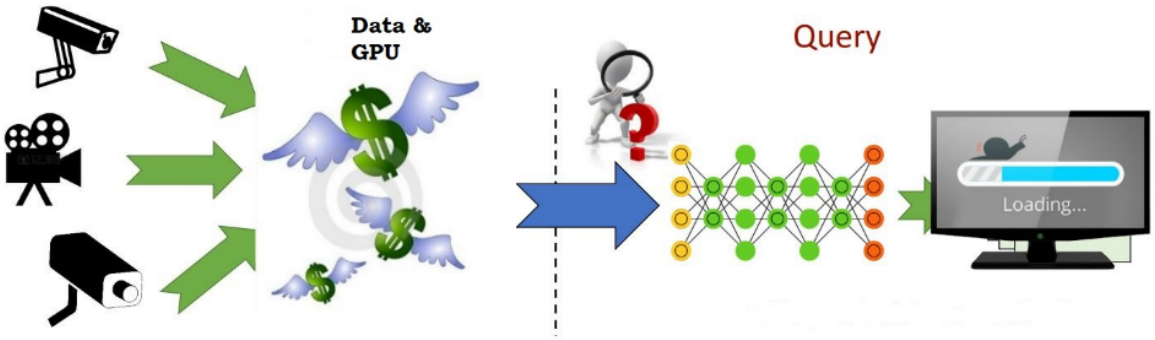



Fig. 1.3 Real-time Video Analytics Challenges

out frames that do not contain relevant information for the query using the additional edge devices. In this study, to address the question of how to address this network bottleneck and offload large volumes of data from a density camera deployment in real-time to a datacenter for further processing, we aim to develop a solution by harnessing edge computing. Video analytics at the edge has multiple benefits such as decreasing the response time, saving network bandwidth, and minimizing the peak workload to the cloud. In previous works, the solutions are based on either:

- Binary classification model[16][31]: remove frames that do not contain any object of interest.
- Pixel-based background modeling[18]: eliminate frames which have not changed substantially.

Typically, exist systems use deep-learning framework to understand video frame context. Video frame is analyzed on pixel domain which require fully video decoding. However, edge devices are normally much less powerful than the cloud server, with limited computing resources such as a few CPUs, GPUs and RAM capacities [56]. In the field of public security, the ability to simultaneously process multiple camera sources and provide real-time video analytics is critical. Thus, our study seeks to overcome how edge devices and the cloud can cooperate in an efficient manner to achieve real-time and scalable video analytics. Our

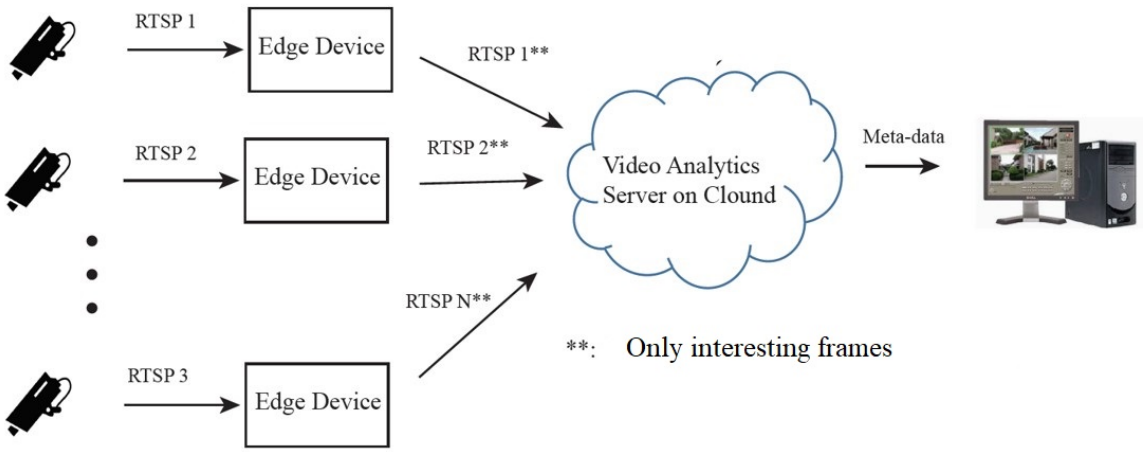



Fig. 1.4 The motivated system design

motivation is based on the observation that surveillance camera images are unchanged for a long time and video content is often unnecessary. So, it is undesirable to conduct analytics on redundant video frames from cameras, it will increase latency and system expense by network traffic congestion, wasting the computing resources and energy consumption in the cloud as described in Figure 1.3.

Motivated by this pupose, we present a pre-process module that acts as a video filter function at the edge device to remove out the redundant static images before forwarding video images to the cloud for further analysis as described in Figure 1.4. The proposed method is driven by the compressed-domain feature which doest not require GPU-based processing and fully video decoing. Therefore, it is light-weight method and cheap solution. The method run on an edge device and detect the motions (i.e., human walking) in the consecutive video frames. Depend on the motion detection result, the edge device decides whether to forward the video frame to the cloud or filter it out. As a result, the proposed filtering edge device can decrease not only computing resource of the cloud server but also reducing network traffic congestion to the cloud.

Motion detection solutions can be devided into two categories: (i) video pixel-domain based and (ii) video compressed-domain based approaches. In pixel-domain based approaches

[45], [35], [25], [61] video is completely decoded and then applying background modelling or CNN based deep learning framework to detect moving objects at the pixel level. The performance of video pixel-domain processing in large systems may be very challenged by the computing load of decoding multiple streams and the image pixel-based calculation. Therefores, pixel-based solution requires higher computational complexity and can make it difficult to fulfill the real-time requirement of edge devices. Compressed-domain approaches [22],[62],[20],[9] rely on video coding artifacts of compressed bitstreams such as motion vector (MV), macroblock partitions, and quantization coefficients for recognizing motion. Compared to pixel-domain based algorithms, compressed-domain methods generally require less computational resources because analyzing input information is already possible in the bitstream.

In this dissertation, we proposed a compressed-domain based moving objects detection method that applies a pixel clustering and an intersection of union (IoU) score-based object tracking technique of computer vision. Using only the MVs provided by video encoders in a compressed bitstream, our approach can efficiently detect moving objects. Furthermore, an experimental evaluation of our method is provided to compare it with the state-of-the-art compressed-domain tracking methods in term of processing time, and demonstrate its functionalities to evaluate the efficiency of the proposed edge device with a cloud video analytics server in a real-world scenario. In summary, this study makes the following contributions.

- Utilizing Video Coding Motion Vectors (MV) analysis to filter out the static frames from video stream.
- Present a compressed-domain based the moving objects detection method that is applied for many different surveillance applications.

- Deploy a novel edge-to-cloud computing system for surveillance video analytics that applies the proposed method at edge devices to minimize the data transfer from surveillance camera feeds to the cloud.

- Implementation and evaluation of the proposed method on cheap edge device (Pi4 ) and achieve real-time processing time of 39 ms/frame with high definition resolution video.

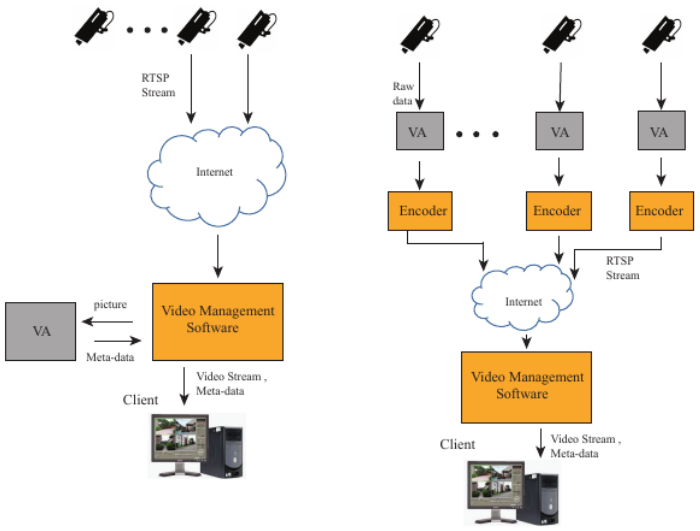
# Chapter 2

# Background

Object detection in images and videos has received a lot of attention of video analytics researchers in recent years. There are many different approaches to detect video moving object, both utilizing compressed and uncompressed video domain. In this chapter, we briefly introduce the current state-of-the-art approaches for moving object detection on both compressed domain and pixel domain. In addition, the background information necessary to understanding VA architecture is described briefly.

## 2.1 Surveillance Video Analytics Architectures and Challenges

Video surveillance systems are normally built using the following main components: surveillance IP cameras, VMS, device storage and VA modules (optional). VA can be implemented in two main configurations, as discussed below.




(a) (b)

Fig. 2.1 Video Analytics Implementation, (a) Video management server based implementation, (b) Edge camera based implementation

### 2.1.1 Edge Camera Based Deployment

In this approach, the VA is deployed through a camera device or video encoder [17] such as that shown in Figure 2.1b, which must have enough the processing power to run the VA functionality. Hence it is an expensive and challenging approach to wide-area video processing. This approach seems ideal, however it does not perform satisfactorily in many cases because of the limitations on the overall surveillance system design and performance. Most surveillance camera device still lack sufficient processing power for highend VA requirements, and therefore, this approach has many drawbacks in real-world deployment.

### 2.1.2 Video Management Server Based Implementation

In this approach, as shown in Figure 2.1a, VA is implemented through a dedicated server that pulls the video from the camera devices or from VMS, analyzes it, and issues analysis results. However, it has some challenges, which are listed below:

- The VA server requires the video to be transmitted and therefore causes an increase in network traffic congestion. In detail, running video analytics by forwarding all video to the cloud overloads with the bandwidth constraints of some deployments, which do uploading all camera data. Each camera's uplink bandwidth is limited, both by the physical constraints of network components such as: cable, switch, router.. Specifically, we consider large-scale deployments where each camera receives a bandwidth allocation of a few hundred kilobits per second, or less. For comparison, a low-quality H.264-encoded 1080p (1920×1080 pixels) stream is approximately 2 Mbits/s, an order of magnitude greater than our available uplink bandwidth. Yet, such low-quality data is often insufficient to perform accurate analysis: Modern 4K (3840×2160 pixels) cameras produce up to 30 to 40 Mbits/s. An edge-based filter answers this challenge with semantic filtering that uploads only frames that are relevant to applications.

- The video data quality analyzed by the VA server is usually degraded because of lossly compression and transmission effects, and therefore.

- The VA server is limited by its processing power, which makes it infeasible for large scale surveillance installations which deploy hundreds (and increasingly thousands) of cameras requiring a variety of VA functionalities. For example, a Full HD (1080P) stream at 30 FPS is $\approx$ 1.4 Gb/s when decompressed. Accomplishing VA at scale requires abundant compute, memory resources, so existing systems often perform this processing in the cloud, using GPUs.

However, this approach is independent with surveillance camera sources, and is therefore applicable to most types of surveillance systems, and recent technological developments can reduce the effect of the above drawbacks. For example, the release of high definition video surveillance footage with resolutions up to 1080p will decrease the impact of the image quality degradation during codec processing and by releasing the new video coding standard as well as the transcoder [58], high efficiency video coding (HEVC)[57] which

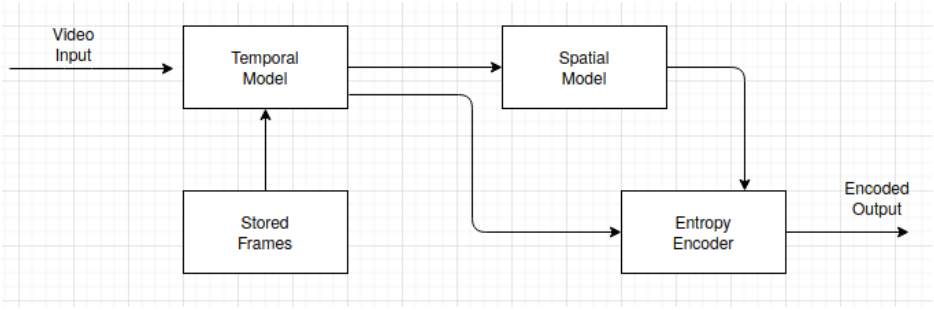



Fig. 2.2 Video encoder block diagram.

has achieved approximately twice the standard compression[47] will decrease traffic in the network and at the central server, where powerful devices will have sufficient processing power to handle hundreds of camera. Moreover, some edge-based filtering approaches that is designed to overcome these above issues [16][42][18]. These approaches use edge-compute resources allocate with the cameras to identify the video sequences that are necessary to cloud applications and forward only that data for further analysis. By this way, it helps to reduce the wide-area network links bandwidth. In this study, we present an edge-based filter, a system that offers the benefits of ultilizing the both edge and clound computing to video analytics processing.

### 2.1.3 Video Codec and Video Coding Motion Vectors

#### Video Codec

The previous video coding standards, such as MPEG-1, MPEG-2, H.264 and H.265, are based on block-based motion compensation, transform, quantisation and entropy coding. A video encoder as shown in Figure 2.2 consists of three main functional units: a temporal, a spatial and an entropy encoder. The input to the temporal model is an uncompressed video sequence. The temporal model tries to reduce temporal redundancy by finding the similarities between neighbouring video frames, normally by constructing a prediction of the current video frame. In MPEG-4 and H.264 codec, the prediction is formed from one or

more previous and future frames and is improved by compensating for differences between the frames (motion compressed prediction). The output of the temporal model is a residual frames (created by subtracting operation the prediction from the actual current frame) and a set of model parameters, typicially a set of MVs describing how the motion was compressed. The residual frame formes the input to the spatial model which makes use of similarities between neighbouring samples in the residual frame to reduce spatial redundancy. In MPEG-4 Visual and H.264 this is archieved by applying a transform to the residual samples and quantizing the results. The transform converts the samples into another domain in which they are represented by transform coefficients. The coefficiencents are quantised to remove insignificant values, leaving a small number of significant coefficients that provide a more compact representation of the residual frame. The output of the spatial model is a set of quantised transform coefficients.

The entropy encoder compressed the parameters of the temporal model (typically MVs) and the spatial model (coefficients). A compressed sequency consists of coded motion vector parameters, coded residual coefficients and header information.

At the decoder, a video frame is reconstructed from the compressed bitstream. The entropy decoder decoded the coefficients and motion vectors are decoded by an entropy decoder to reconstruct a version of the residual frame. The decoder uses the motion vector parameters, together with one or more previously decoded frames, to create a prediction of the current frame and the frame itself is reconstructed by adding the residual frame to this prediction.

#### Video Coding Motion Vectors and Blocked-based Motion Estimation

Changes between video frames may caused by object motion (i.e, a human running), camera motion (zoom, rotation) and lighting condition changes. These changes coresspond to pixel movements between frames. It is able to estimate the trajectory of each pixel through video frames, generating a field of pixel trajectories known as optical flows. Optical flow gives

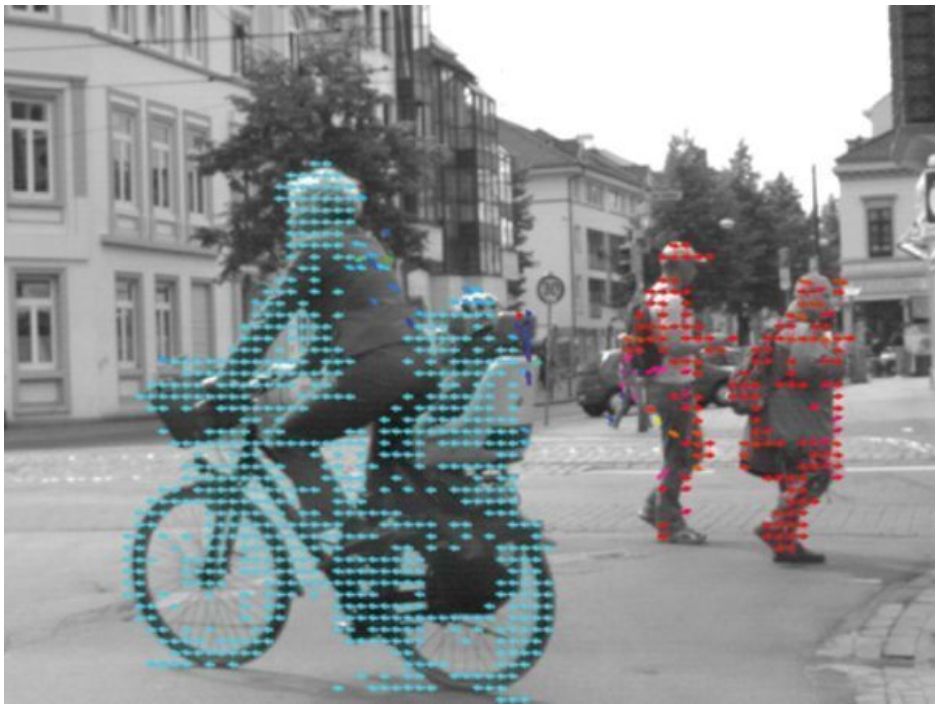


Fig. 2.3 Motion Vector in Video Codec.

the measure of movement of a pixel or a block in two consecutive frames. This measure of movement is presented by a vector where the magnitude of the vector represents the amount of motion and the angle of the vector specifies the direction of the motion. This vector is called motion vector. Figure 2.3 shows the optical flow (or motion vector) field of an image. Look at the objects in the image above and the motion vectors drawn on them. By analyzing the motion vectors, it is clear that a bicycle is moving towards the left side while two pedestrians are moving towards the right side. The speed of their motion is categorized by the magnitude of the motion vectors and their direction by the angel of the motion vectors. And therefore, it is necessary to send the motion vector for every pixel to the video decoder. In video coding standard, motion compensation is a practical method of motion compensate for movement of blocks of the current frame. The current frame is sub-devided in MxN blocks called macroblocks as shown in Figure 2.4. The following steps are carried out for each block MxN samples in the current frame:

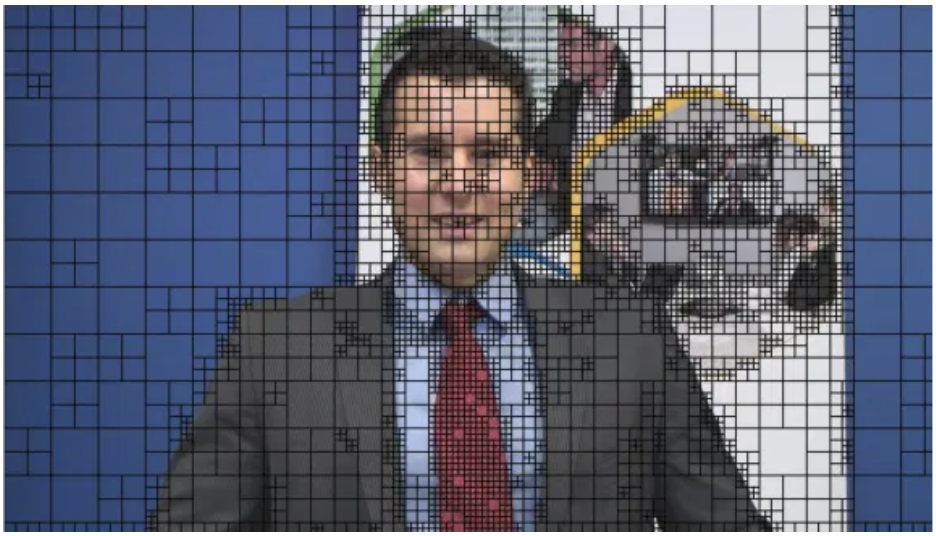


Fig. 2.4 Macroblock in Video Encoder.

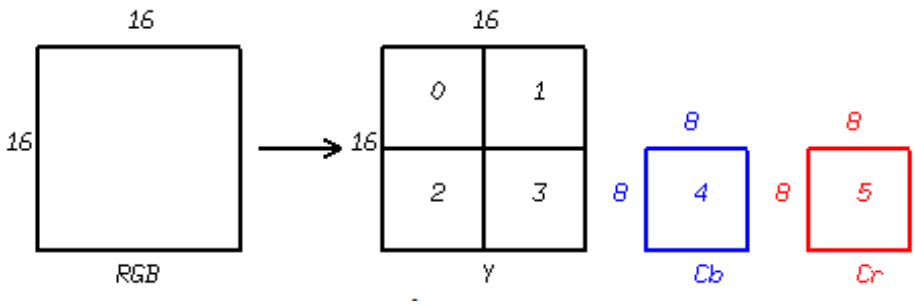



Fig. 2.5 Macroblock (4:2:0).

- Searching an area in the reference frame (past or future frame) to look after a 'matching' MxN-sample region. This process is to find the best match and is known as motion estimation.
- The selected candidatge region becomes the predictor for the current MxN block and is removed from the current block to form a residual MxN block (known as motion compensation).
- The residual block and the offset between the current block and the position of the candidate region (motion vector) is encoded and transmited.

In H264/AVC standard, the macroblock, coressponding to a block size of 16x16 pixel is the basic unit for motion compensated prediction. In the RGB colour space, each pixel is represented by three numbers indicating the relative proportions of Red, Green and Blue. However, human eyes are more sensitive to luminance ( brightness ) than to colour. Thus, data can be down sample by transforming the RGB to YCbCr which throw some color components without causing much visual distortion. In example, in the 4:2:0 sampling format, Y has double Cb and Cr each has half the horizontal and vertical resolution of Cb and cr. Figure 2.5 shows YCbCr 4:2:0 format with a 16x16 pixel macroblock. The macroblock is divided into 8x8 sample blocks. In the RGB space, a macroblock has a total of 3x4 = 12 (8x8 sample blocks) ( four for each of the colors Red, Green and Blue ). However, there are only six sample blocks in YCbCr 4:2:0 format, four for the Y component and 1 for each of Cb and Cr component.

In motion compensation, the MVs of other neighboring blocks in the current frame or in the earlier coded frames [36], [30] are correlated with MV of a current block. While intra-picture prediction uses correlations between spatially neighboring samples, then inter-picture prediction makes use of temporal correlation between frames to derive a motion-compensated prediction (MCP) for a block of frame samples [14]. For intra-prediction, we assume that neighboring blocks possibly correspond to the same moving object with similar motion and that motion is continuous to change over time. As the result, using MV in neighboring blocks as predictor can reduce the size of the signaled motion vector difference. For this block-based MCP, a video frame is divided into rectangular blocks. Figure 2.6 presents the novel concept of MCP with a translational motion model. Using this model, the position of the block in a previously decoded picture is represented by a motion vector $(\Delta x, \Delta y)$, where $\Delta x$ is the horizontal and $\Delta y$ the vertical displacement relative to the position of the current block. These motion vectors are coded by entropy coding and placed into a compressed bitstream to send to the video decoder side, and the decoder will use these MVs, along with

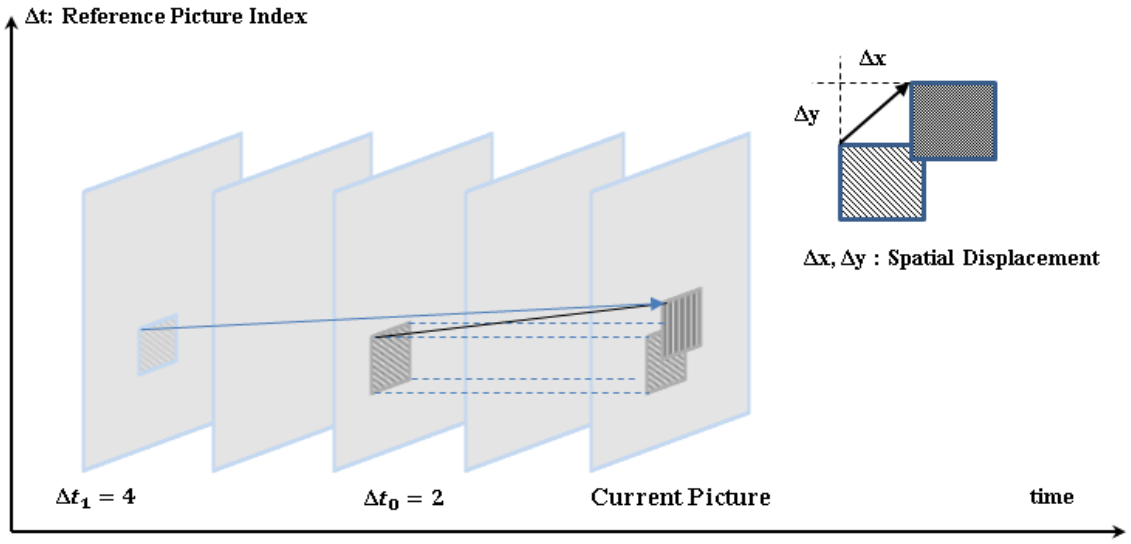
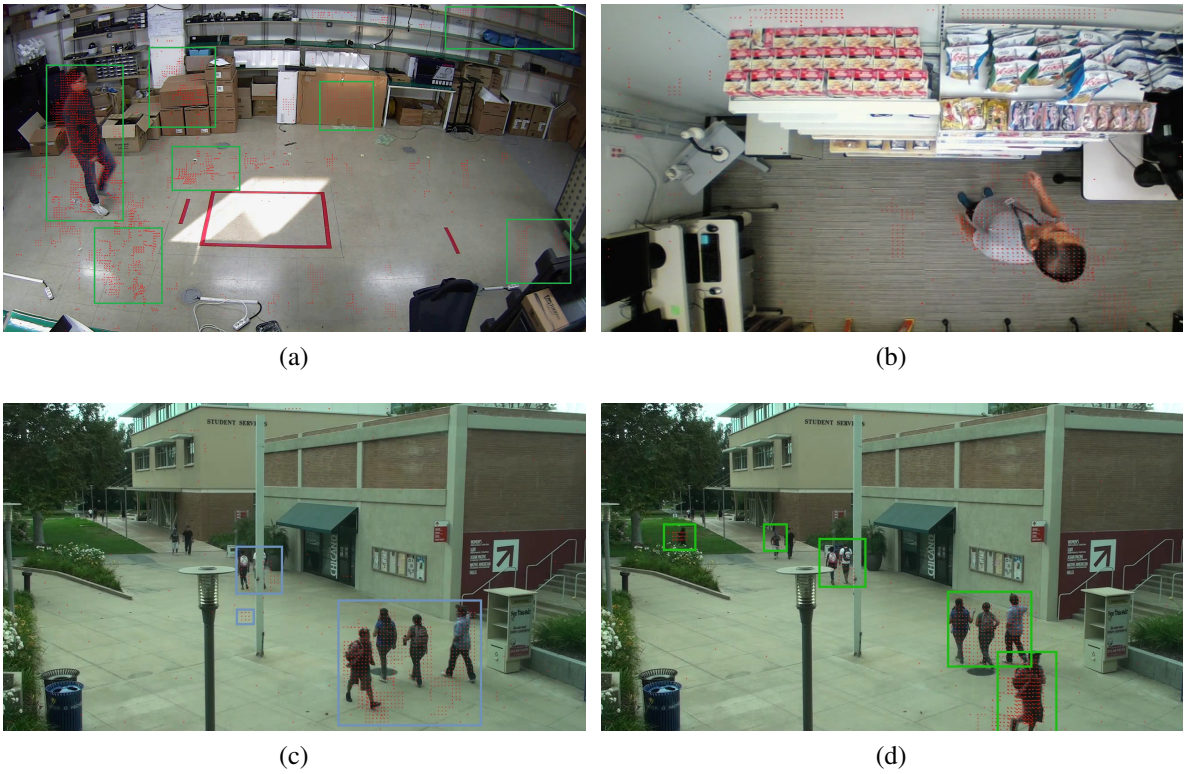



Fig. 2.6 The eeference MV model in video coding.

one or more previously decoded pictures, to reconstruct the current image. On the other hand, in H.264/AVC, each video sequence is devided into groups of pictures (GOP), comprising at least one intra code I frame, uni-directionally predicted P frames and bi-directionally predicted B frames. Normally, the first frame in a GOP is intra coded I frame and follows by P, B frames. I frame utilize raw data from camera, while other frames in a GOP use the predictive coding involved motion vectors displacement.

## 2.2 Compressed-Domain Based Moving Object Detection Using Motion Vector

As mentioned in last section, MVs are extracted for each motion block between the current and referenece frames. By minimising the prediction residual, the MVs present the temporal displacement between the two block in the process of motion compression. As the result, MV information will follow the real motion of objects and can be used for tracking purpose.



(a) (b)

(c) (d)

Fig. 2.7 The example of MV extraction.(a) Test video sequence from recorded camera, (b, c, d) Test video sequence from VIRAT dataset.

In recent researches, the compressed-domain based video analytics methods such as: [12],[32].. they are based on video coding features such as MVs, macroblock partition, and quantization coefficients, have been proposed. In [12], the authors applied a probabilistic technique of computer vision for image separation, known as Graph Cut [13], modeling with MVs rather than pixels and adapted to the additional temporal dimension of video signals. Using MVs and a spatio-temporal Markov random field (ST-MRF) model that naturally integrates the spatial and temporal aspects of the object's motion for tracking. In general, these approaches do not rely on pixels and studies by only using the codec's MVs and block coding modes extracted bitstream through inexpensive partial decoding. In this manner, computing and storage requirements have been significantly reduced compared to "pixel-domain".

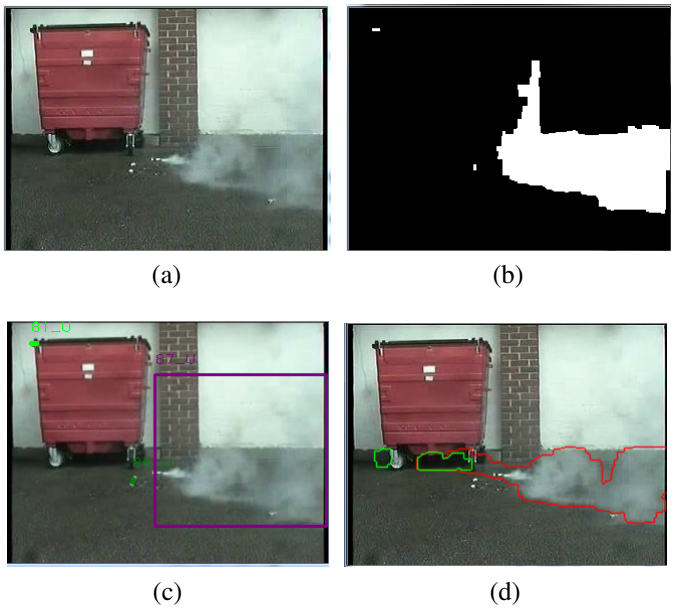
The primary limitation of this approach is that it may lead to a noisy MV field that does not necessarily correspond with actual object movement of the object in the scene, as shown in Figure 2.7. The noisy MVs fail to provide useful information such as those attributed to illumination changes and background movement. The amount of noise MVs is relatively reduced compared to the correctly estimated MVs because noisy vectors are continuous and similar MVs from real moving objects. Another challenge of this approach is the lack of information about the object's appearance such as color, edges and texture, because these features would require complete decoding of the compressed bitstream. In this study, our aim is to work in the "compressed domain" and uses only the MVs from the compressed bitstream to detect and track moving objects in video frames.

## 2.3 Pixels-Domain Based Moving Object Detection using Deep Learning

Because of more reliable features that can be extracted from the pixel data, the majority of moving object detections have been done in pixel domain, even it requires full decoding the video stream. In pixels domain, a VA server analyzes the R-G-B image to find the appearance objects and spatial events. The drawback and exciting research problem, and it is overcome by the recent advantages of hardware and deep learning [63],[19],[11],[60],[49],[48]. There are two main methods that are considered for moving object detection in pixels domain.

### 2.3.1 The combination of background modeling and CNN based object classification

The background subtraction approach [39][55] subtracts the current image and background image to eliminate the background, and then detects moving targets based on pixel clustering. This method is widely used for object detection. There are many methods that have been



(a) (b)

(c) (d)

Fig. 2.8 Smoke detection pipeline: (a) input frame, (b) the foreground subtraction, (c) the blob detection, (d) the smoke candidates classification

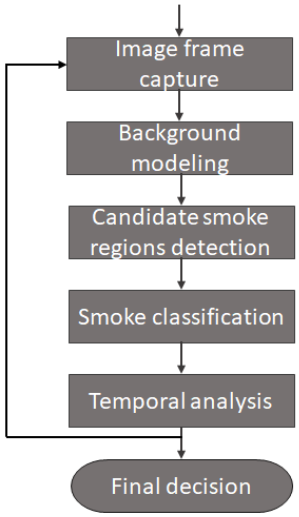



Fig. 2.9 The pipeline of video- based smoke detection algorithm.

purposed for building the background model, e.g., [44] presented a novel real-time motion detection method that integrates the temporal differencing, double background filtering, optical flow, and morphological processing methods to obtain excellent performance. Moreover, the authors of [55] discussed modeling each pixel as a grouping of Gaussians using an on-line approximation to renew the model. The Gaussian distributions of the adaptive mixture model were then assessed to determine which model can be presumed to be obtained from a background process. This method's advantage is its simple implementation, low computational resource requirements, robustness in the presence of environmental noise, and dynamic background. However, it has limitations with shadows, background changes, as well as object localization and classification. Furthermore, object detection based on the background model cannot detect the individual objects in a group or a block object. In general, the aim of background subtraction is to separate foreground images from background ones in the form of blobs, followed by an object classification process. The detected blobs then help classify each blob into subclasses, as shown in Figure 2.8, which shows the complete process of this approach is to perform smoke detection with the detail of flow-char is drawn in Figure 2.9. In this example, smoke classification was done by trained convolutional neural networks (CNN) model which was described in [34] for image classification. This a relatively new approach in machine learning, which ultilize CNNs to achive the very accurate classification of images. [37][28][38]. Figure 2.10 shows the original architecture of the trained model, first five layer are convolutional and some of them are followed by max pooling layer. The next are three fully-connected layer, the last fully-connected layer computes class score of trained class labels.

The accuracy of CNN based model can be changed by varying their depth and breadth. For example, by combining CNN and transfer learning, the authors of [27] achieved an average accuracy of 70.1% using the CIFAR-10 [33] dataset.

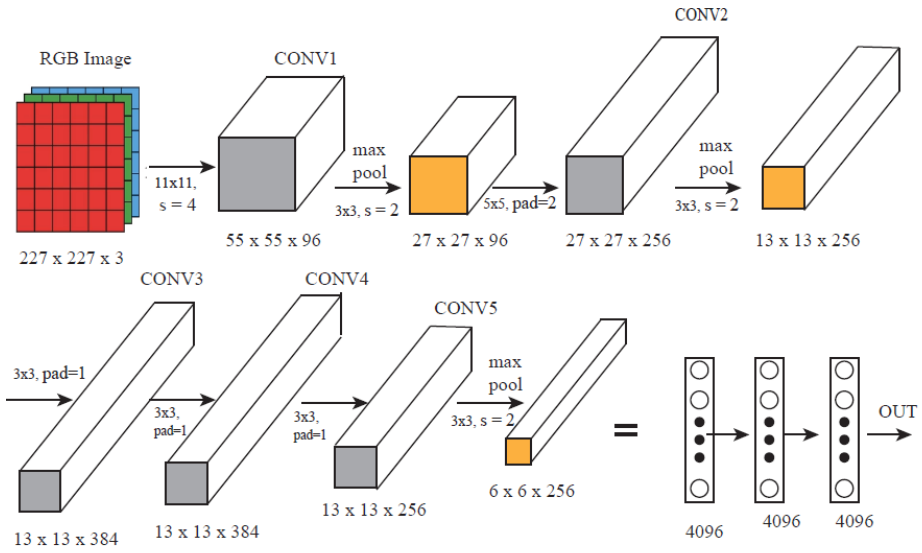


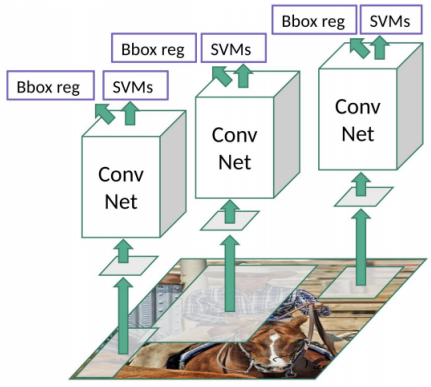

Fig. 2.10 Architecture of the Alexnet CNN model.

### 2.3.2 Deep Learning based Moving Object Detection

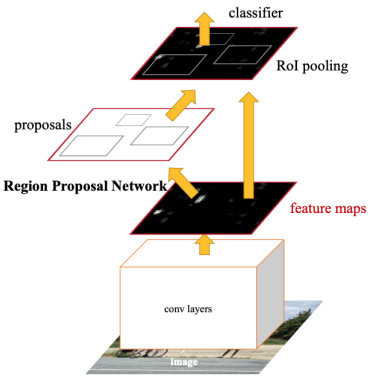
Hybrid background subtraction and deep learning classification enhance the system's accuracy; however, there are issues with detection of individual objects in a group or blocked objects and the background changes by the lighting condition and the environmental noise. Hence, the usage of deep learning is considered for robust object detection tasks. We have identified three primary object detection methods using deep learning:

- R-CNN model, Fast R-CNN model, Faster R-CNN model.
- You Only Look Once (YOLO)
- Single Shot Detectors (SSDs)

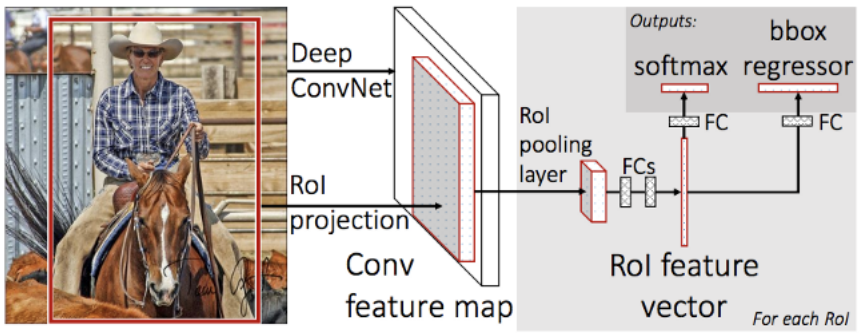
R-CNN [24] uses selective search [59] to extract several regions, called region proposals, from the image; it then attempts to classify a large number of regions. Each region proposal is placed into CNNs to extract features, which are fed to a support vector machine to classify the presence of object within the candidate region proposal as shown in Figure 2.11(a). The limitation of this method is that it requires a large amount of time to train and deploy networks




(a)




(b)




(c)

Fig. 2.11 CNN based Object Detection Framework Structure.(a) RCNN, (b) Faster-RCNN, (c)Fast-RCNN.

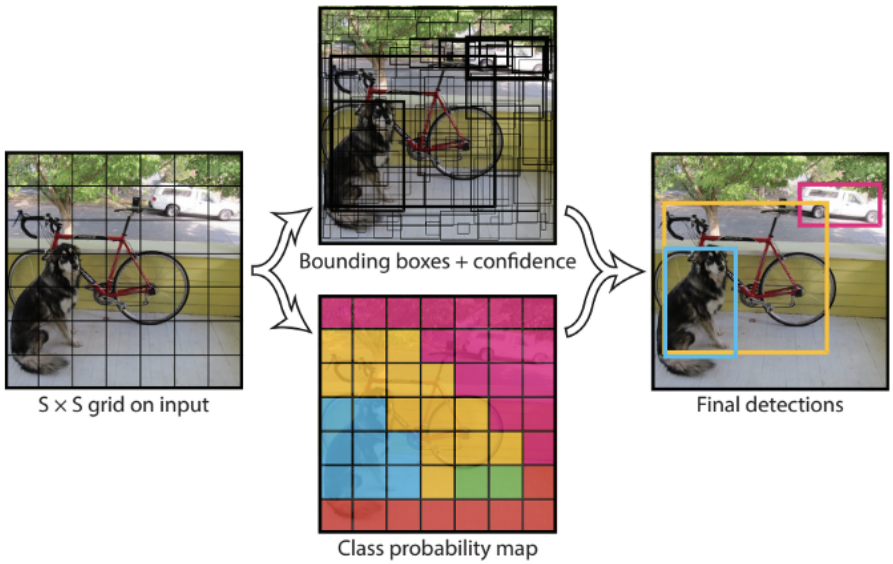



Fig. 2.12 YOLO Structer.

because it has to classify multiple region proposals per image. Fast R-CNN [23] solves the limitations of R-CNN by putting the entire image into the CNNs to generate a convolutional feature map and identify the region of proposals based on the map, thus reducing the number of classified region of proposals as drawn in Figure 2.11(c). Both R-CNN and Fast R-CNN use selective search to identify the region proposals. Selective search is a slow and tedious process that affects the network's performance. Faster R-CNN [53] was proposed to allow the network to learn the region proposals as shown in Figure 2.11(b); it uses a separate network to predict the region proposals rather than use a selection search algorithm on the feature maps output using the CNN layer. YOLO [51] is an object detection system to detect object at real-time speed. Unlike R-CNN, Fast R-CNN and Faster R-CNN, which use regions to localize the object within the image, YOLO ultilize a single neural network to predict the bounding boxes and the class probabilities for these boxes during training and test periods. Hence, YOLO only examines the input picture once to predict the presence and location of objects. YOLO divides the input image into an SxS grid, and multiple bounding boxes can exist within each grid. For each bounding box, the network outputs a class probability

and offset value for the bounding box. The bounding boxes with class probabilities above a certain threshold value are then selected and used to locate the object within the image as shown in Figure 2.12. Specially, YOLOv3 network architecture has 24 convolutional layers, followed by two fully connected layers as shown in Figure 2.13. The initial convolutional layers of the network extract feature from the image, while the fully connected layers predict the output probabilities and coordinates. According to performance comparison in [52], YOLOv3 is faster (45 frames per second (FPS)) than the other object detection algorithms. Faster R-CNN is more accurate than YOLOv3 (a mean average precision (mAP) of 73.2, as compared to 63.4); however, YOLOv3 is considerably faster than Faster R-CNN (FPS of 45, as compared to 7). Therefore, SSDs [43] were released as a balance between these two methods. Compared to YOLO, an SSD runs an input image through a convolutional network only once and computes a feature map. Then, a small $3 \times 3$ sized convolutional kernels are run on this feature map to predict the bounding boxes and categorization probability. Moreover, SSD uses anchor boxes at various aspect ratios comparable to Faster-RCNN and learns the off-set to a certain extent compared to learning the box. The SSD is able to detect objects of multiple scales because every convolutional layer function at a diverse scale. Compared to the YOLOv3 method [52], the SSDs attain similar accuracy but run slower. In this study, YOLOv3 was applied to robust human detection at cloud servers for our implementation.

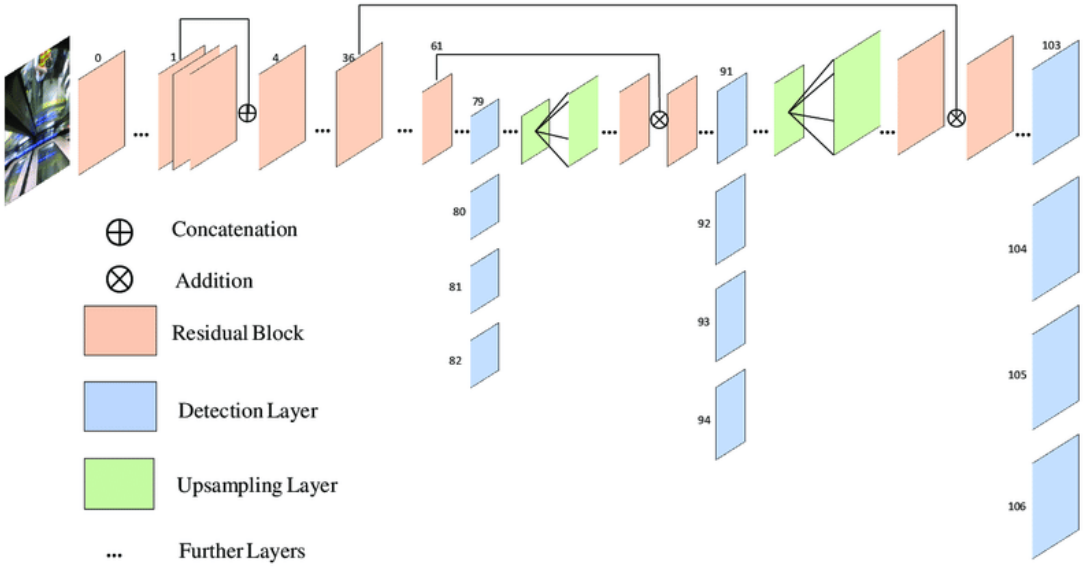



Fig. 2.13 YOLOv3 Network Model.

# Chapter 3

# Methodology

In this section, we will introduce the proposed edge-to-cloud sysem for surveillance video analytics applications in details. And then, the our approach for moving objects detection in compressed-video domain is presented. In addtion, we will introduce the performance evaluated model for video analytics platform used in this study.

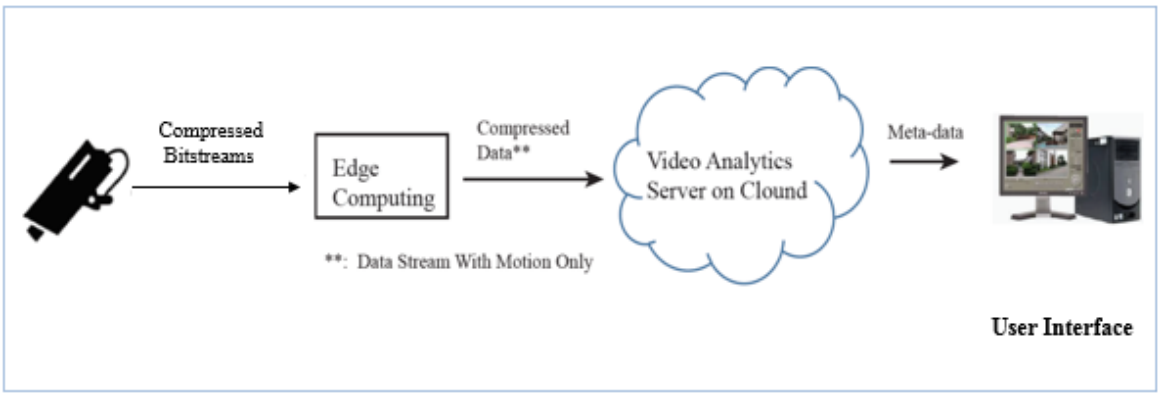



Fig. 3.1 Our proposed system.

## 3.1 The edge-to-clound architecture for Surveillance Applications

There is a trend to offload computation from the cloud to the edge where the data is generated. To minimize bandwidth consumption, and improving energy efficiency, edge computing has been used. “Edge computing performs data processing at the edge of the network, close to the data source. For network cameras, a camera source node can forward its video task to nearby edge nodes via local wireless/wire-link networks”. The camera captures the video sequences and divides each of them into multiple video trunks, compresses them and stream to edge nodes. Next, edge nodes implement video processing functions on the received video trunks and upload the results to a cloud server for further video analysis (such as object/event detection). In this study, as shown in Figure 3.1, we consider an edge computing network, which comprises three primary components, namely, camera source node, edge node, and cloud server.

- “Camera source node where the camera node periodically generates video tasks”: compresses video chunks at certain compression ratios, and then allocate compressed video chunks among all edge nodes according to specific scheduling policies.
- “Edge node where the edge helps preprocess video chunks”: moreover, edge nodes can cooperative groups based on specific group formation policy.
- “Cloud Server: cloud server collects the preprocessing results from edge nodes”: video analytics server has enough computational abilities, and performs additional video analysis.

During a sparse edge node deployment, an edge node connect to one camera at a certain location; however, in a dense deployment, an edge node may have multiple connection with multiple cameras. In this study, we focus on reducing processing load on cloud server by

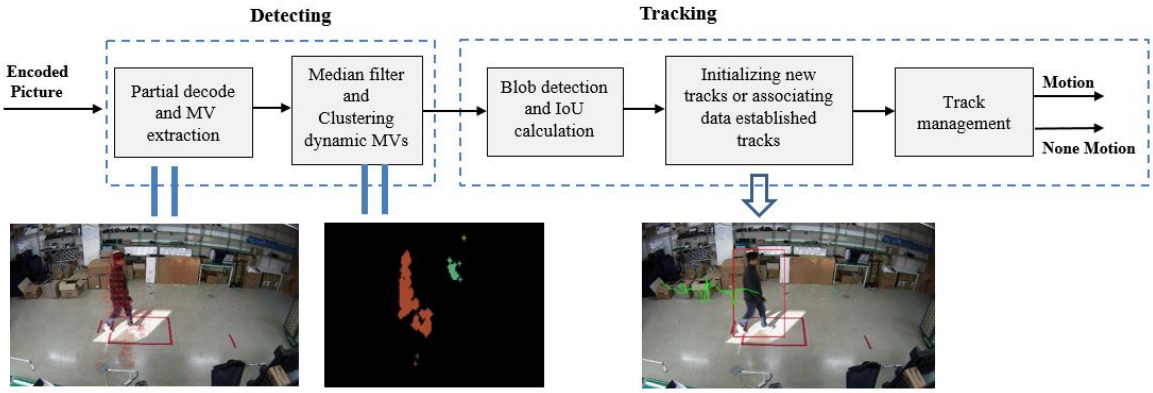



Fig. 3.2 The proposed method for moving object detection in compressed video domain.

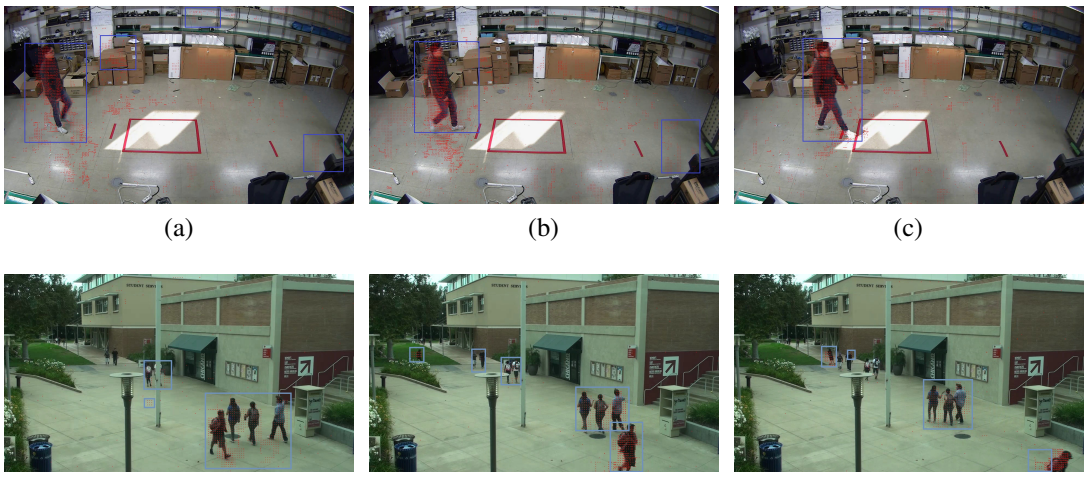
minimizing the video chunks that are forwarded from camera sources. Therefore, an edge device runs preprocessing tasks to filter uninteresting video chunks. The term “uninteresting video chunks“ is defined via scenario specification: In this study, it is static scenes without any objects and gradual changes.

## 3.2 The Proposed Method of Moving Object Detection in Compressed Domain

MV represents the changes between video frames may caused by:

- Object motion (car, human activity)
- Camera motion (panning, tilt, zoom, rotation).
- Lighting condition changes

The MV group generated by lighting change, usually randomly occur without following a certain flow and have various sizes. The MV group generated by human motion (walking, running), move in certain direction and It’s size is similar with object size in video frame. This observation is good factor to classify MVs into real-motion and noise motion group.



(a) (b) (c)

(d) (e) (f)

Fig. 3.3 The video coding MV extraction from consecutive frames with two different test sequences.

The aim of this method is to analyze only MVs at the edge device, recognize the real moving objects, and exclude the noised motions that are generated by environmental factors such as illumination changes and background movement, in the current video scene. For this purpose, we present a method that has a workflow shown in Figure 3.2 to analyze the video coding MVs. In order to extract MVs from compressed bitstream, the input video stream is partially decoded and the collected MVs will include both real object MVs and noise MVs. The example of extracted MVs are represented in Figure 3.3 in consecutive frames of two different video test sequences. To analyze these MVs, the following functionalities are involved.

### 3.2.1 Motion Vector Median Filter and The Moving Object Detection

The compressed-domain MVs may not present the actual of true motion due to the properties of motion estimation as shown in Figure 3.5(a,e). Therefore, it is desirable to remove motion vectors that can be categorised as false case for moving objects detection. We found that MVs whose magnitudes are very low, compared to the other MVs related to the real moving object,

(a) (b) (c) (d)

Fig. 3.4 Motion vector of a moving object in different video test sequences: (a, b) our record video, (c, d) Test video in VIRAT

has to be replaced with the median value of neighbouring MVs. Therefore, the application of vector median filter aims to eliminate isolated vector noises and smoothen the difference of MV between adjacent blocks.

Because H.264/AVC allows for variable-sized block partitioning, we construct a uniformly sample MV field by mapping all MVs to 4x4 blocks. Theoretically, the motion vector average (the normalized vector) among all elements $N_{mv}$ in NxN (N=4 is chosen in this study) window function is calculated by Equation 3.1:

$$N_{mv} = \frac{\sum_{i=1,j=1}^{N*N} MV_{ij}}{N*N} \tag{3.1}$$

Where: $MV_{ij}$ is the MV elements in the N×N window. Finally, the smoothened motion filter is presented, which is experimentally determined as shown in Figure 3.5(b,f) using the input MV, including isolated motion vectors noise, are smoothened in the areas that mostly correspond to the object boundaries. In order to detect the moving objects, a cluster detector is involved to define the number of moving objects and their positions in the image. Then, a density-based cluster technique is used to completely segment the moving objects. The first step, MV whose magnitudes can be used to easily extract moving objects from

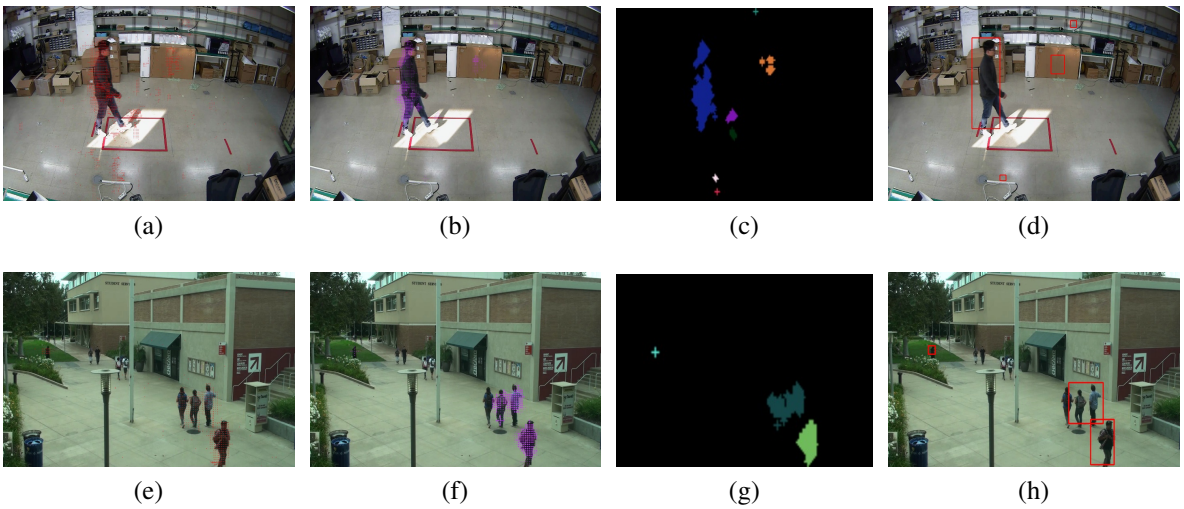
the background. However, it is difficult to directly and completely segment moving objects because not all MVs on the objects have the same motion state. While every part of a rigid body maintains nearly the same motion state, different parts of a non-rigid body can move in various ways (see Figure 3.4). In order to cluster detected MVs into dynamic clusters, a range search algorithm based on the euclidean distance of the MV point is used, under the presumption that dense points represents the same object. It requires a parameter $\varepsilon$ where $\varepsilon$ is the spatial distance threshold between density reachable MV. The goal is to partition $n$ MV points $\chi \subset \mathbb{R}^d$ into $k$ clusters using k-means filter. Each of the $n$ data points will be assigned to a cluster with the nearest mean. The mean of each cluster is called its "center". For the k-means problem, we wish to choose $k$ centers $\mathbb{C}$ so as to minimize the potential Equation 3.2:

$$\phi = \sum_{x \in \chi} \min_{c \in \mathbb{C}} \|x - c\|^2 \tag{3.2}$$

From these centers, we can define a clustering by grouping data points according to which center each point is assigned to. With MV points, following steps are performed frame by frame:

- For each $p_i \in \chi$, find all the neighboring points within spatial distance $\varepsilon$.
- Cluster all the MV points that are density reachable or density connecting [21] and label them.
- Terminate the process after all the MV points are checked. The output is a set of clusters of dynamic points.

The distance threshold $\varepsilon$ decides the range of density reachability [21]. After clustering, the label list includes all candidate objects with various size, represented by different colors. Typically, the size of candidate object is followed by object size in the video frame. In small object dataset, objects are small when they have mean relative overlap (the overlap area between bounding box area and the image is) from 0.08% to 0.58%. Thus, we used a cluster



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3.5 Moving objects detection by the proposed method with two different video test sequences:(a,e) MV extraction; (b,f) Apply median filter; (c,g) Clustering MVs; (d,f) Blob detection.

threshold size to filter out the small clustered to the candidate lists as shown in Figure 3.6.

After this process, the label list includes the moving objects and certain big MVs noises as shown in Figure 3.5(c,d). To eliminate these noises, we based the observation on the fact that noise motions because of lighting condition changes usually randomly occur without following a certain flow. Therefore, the tracking motion's trajectories length is then used to classify the detected motion into the real motion or noise group. If the object tracking trajectory length is larger than a certain threshold, it indicates that the objects move as a flow and the time is sufficiently long to consider it as a real motion. Moreover, I-Frame does not apply motion estimation process, therefore, MV analysis process will be skipped for I-Frames. In order to overcome the discontinuously object detection process, the object tracking algotithm will also be applied to derive the moving object's bounding box in I-frame based on the last states in P-frames. Because the real object moves with a certain flow, the tracking motion's trajectories length is then used to classify the candidate motions into the real motion or noise group, the IoU-based object tracking is applied to track motion because it is light-weight and widely tracking algorithm that calculates the overlap area between two

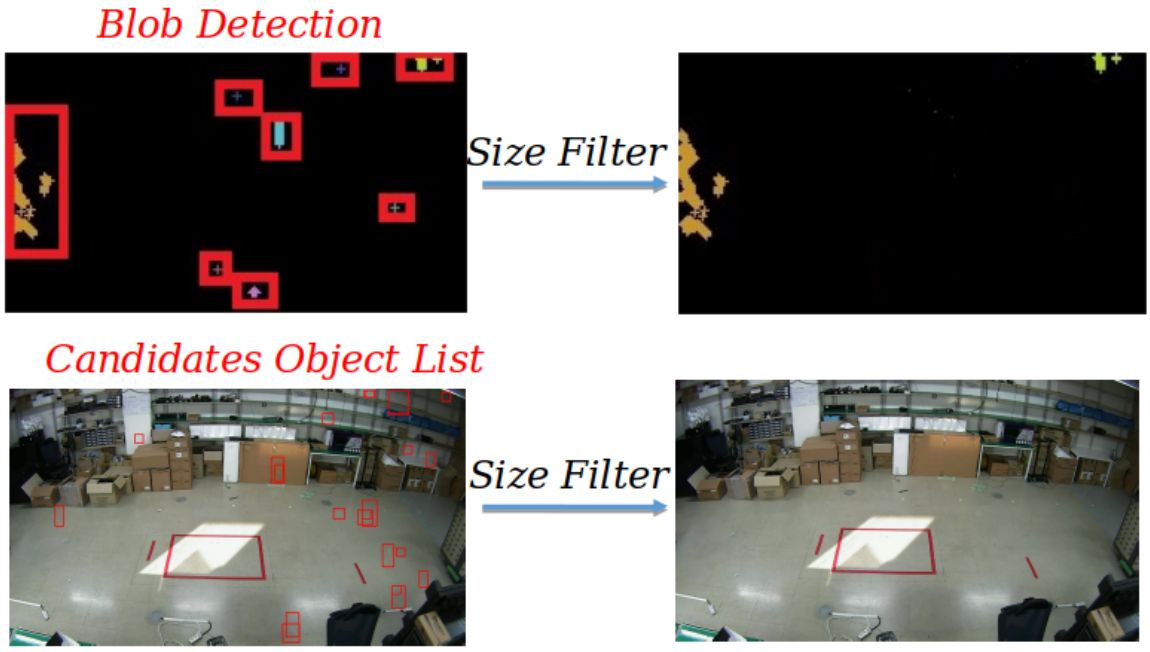



Fig. 3.6 Motion size based Filtering

bounding objects. Note that this tracking was implemented and evaluated in some previous studies [54], [40].

### 3.2.2 The low-complexity of overlapping based Moving Objects Tracking

To apply the IoU-based object tracking, the detected clusters are normalized with a rectangle bounding box according to the cluster's size as shown in Figure 3.5(d,h). The correlated regions are connected into blobs. Each blob is represented by its top-left and bottom-right corners, i.e., $(x_1, y_1, x_2, y_2)$ may include one or many moving objects. Because the moving object's bounding box size and shape can be different comparatively frame by frame depending on motion vector intensity. Therefore, to track the moving object, the object matching algorithm which is based on the overlapped area calculation is applied. For example, assuming that the existence of the real motion is continuous frame by frame. For each bounding box $B_1$ in the previous frame, we determine a bounding box $B_2$ at the current frame with the

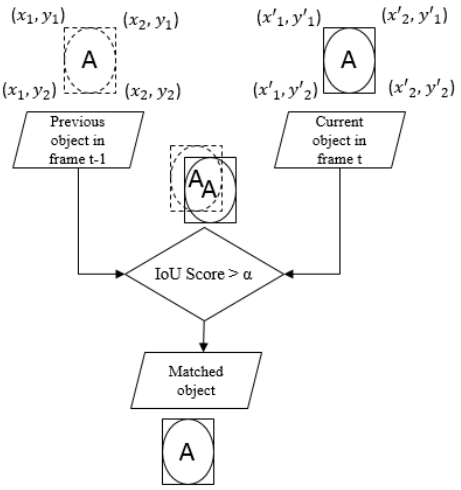



Fig. 3.7 Object matching method

highest IoU score. Note that IoU is attained by:

$$IoU(B_1 \bigcap B_2) = \frac{\|B_1 \bigcap B_2\|}{\|B_1 \bigcup B_2\|} = \frac{\|B_1 \bigcap B_2\|}{\||B_1| + |B_2| - B_1 \bigcap B_2\|} \tag{3.3}$$

As its definition in the Equation 3.3, IoU is invariant to the scale, indicating that the similarity between two arbitrary shapes A and B is independent from the scale of their space. The IoU computation's pseudo-code is given in Algorithm 1. If two bounding boxes do not overlap, the IoU value will be 0 and if the IoU score is greater than a detection score threshold ($\alpha$), two bounding boxes are considered in same account (Figure 3.7). The detection score threshold is determined through experiments and depend on the object velocity as well as the distance between objects and camera. Therefore, the method is run with multiple times with different detection score thresholds to tune the best value for each application scenarios.

## 3.3 Performance Evaluated Model

Computing Resources of VA server is affected by many explicit factors. For example: whether VA function is running, the complexity of VA function, video resolution, which kind of deep learning model used for VA function, how many cameras is serving. However, the

**Data:** Cordinates of the two bounding boxes.
- First box: $A1(x_1,y_1), B1(x_2,y_1), C1(x_2,y_2), D1(x_1,y_2)$
- Second box: $A2(x'_1,y'_1), B2(x'_2,y'_1), C2(x'_2,y'_2), D2(x'_1,y'_2)$

where $x_1 \leq x_2, y_2 \leq y_1 and x'_1 \leq x'_2, y'_2 \leq y'_1$.

**Calculation**: IoU value
- The area of first bounding box : $Area_1 = (x_2 - x_1) \times (y_1 - x_2)$
- The area of second one : $Area_2 = (x'_2 - x'_1) \times (y'_1 - x'_2)$
- The area of overlap:
  $Area_{overlap} = (max(x_2, x'_2) - min(x_1, x'_1)) \times (max(y_2, y'_2) - min(y_1, y'_1))$
- $IoUvalue = \frac{Area_{overlap}}{Area_1 + Area_2 - Area_{overlap}}$

**Algorithm 1:** IoU calculation for two bounding boxes coordinates.

primary factor and biggest effect is number of serving cameras and whether VA function is running. Because if the VA server is in idle status, then other factors will be implicit. Let assume that we have a video test with N consecutive frames with K frames had the real motion (K<= N). For each frame, VA server cost S and T unit of average computing resource for processing and skip frame case respectively. For conventional method of video analytics server, where T = S because all frames are processed , the computing resource average is:

$$Comp_c = S \tag{3.4}$$

With our proposed method, the computing resource average during N frames is calculated as the following equation:

$$Comp_p = \frac{(K * S + (N - K) * T)}{N} \tag{3.5}$$

The performance ratio of two method is:

$$\frac{Comp_p}{Comp_c} = \frac{(K * S + (N - K) * T)}{N * S} = \frac{K}{N} + (1 - \frac{K}{N}) * \frac{T}{S} \tag{3.6}$$

if an object motion always appear in the video (K $\approx$ N), then:

$$\frac{Comp_p}{Comp_c} \approx 1. \tag{3.7}$$

In case of GPU consuming, when video analytic server skip a frame then T = 0 and:

$$\frac{Comp_p}{Comp_c} = \frac{K}{N} \tag{3.8}$$

If computing resource is CPU load and network throughput , T is very small. Becuase T is used only for listening new connection, then:

$$\frac{Comp_p}{Comp_c} \approx \frac{K}{N} \tag{3.9}$$

# Chapter 4

# Implementation And Performance Evaluation

In this section, the implementation and performance evaluation of the edge-to-clound system with proposed method is presented. Furthermore, the information of video datasets and the scenario setup are provided. In this implementation, VA server executes the application of intrusion detection. Although the evaluated platform integrates specific application, it is a general design and can be extended for other application with few modifications. The workflow of the evaluated platform is represented in Figure 4.1. It has two main components:

- Edge node implementation: The streaming data from camera sources are parsed and the proposed method is applied to detect moving objects in the current frame. If the encoded frame includes the motion, it will be forwarded to a cloud node using its own real-time streaming protocol (RTSP) server. To avoid decoding inaccuracies at the cloud node, all frames from starting time to ending time of the motion are continuously delivered in a connection session. Each session will start with an intra-coded frame.

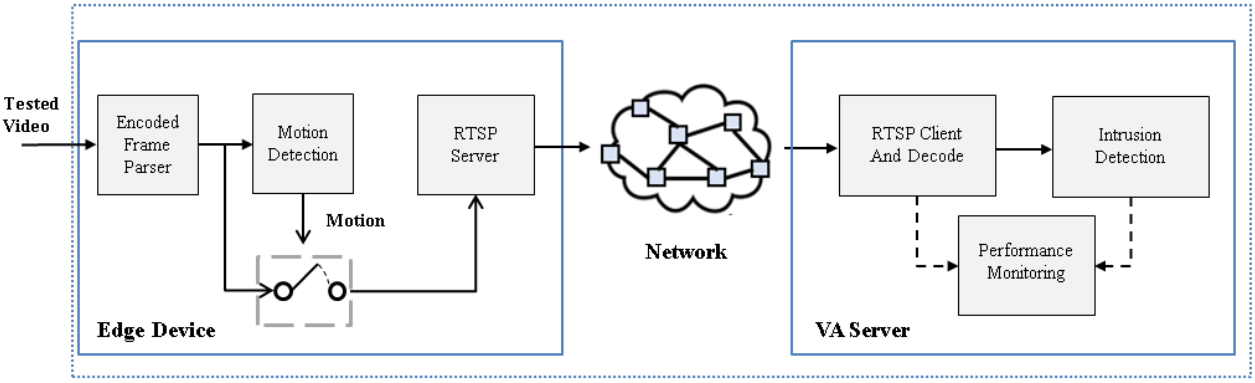



Fig. 4.1 Overview of System Design

Table 4.1 Hardware Specifications.

| Specifications | Edge Device | Video Analytics Server |
|---|---|---|
| Device Name | Raspberry Pi 4 | NVIDIA Jetson Xavier |
| Operating System | Ubuntu 18.04, 64 bits | Ubuntu 18.04, 64 bits |
| GPU | Not Supported | NVIDIA Maxwell architecture with NVIDIA CUDA c |
| CPU | Quad-core ARM Cortext-A72 | Quad-core ARM Cortext-A57 MPCore processor |
| RAM | 4 GB | 8 GB |

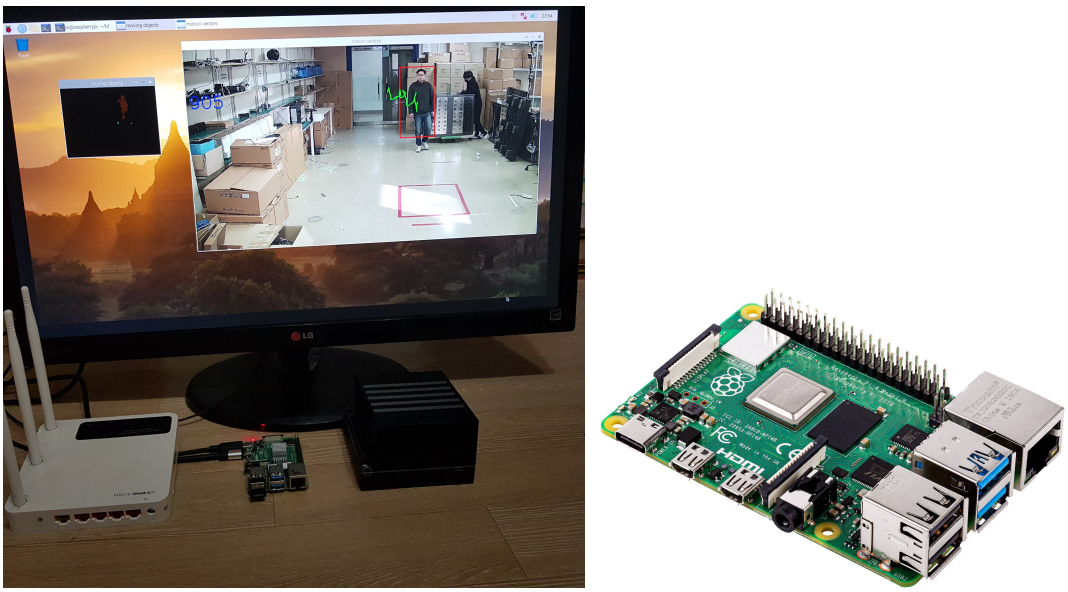
- Cloud node implementation: Receiving the forwarded encoded frame with the motion from the edge node over the network and then decoding and placing the output images into the intrusion detection module, which uses YOLO to detect humans.

### 4.0.1 Scenario Setup

- Testbed: We built a testbed comprising a single edge device node and a single video analytics server that runs as a cloud node, as shown in Figure 4.2. The edge device

Table 4.2 Video Test Squence.

| Tested Video Information | |
|---|---|
| Resolution | 1920x1080 |
| Length | 6 minutes |
| Codec | H264 |
| Group Of Picture (GOP) | 30 |
| Frame Rate | 25 |



(a) (b)

Fig. 4.2 Testbed: (a) Scenario Setup, (b) The implemendted edge device.

Table 4.3 Testbed: video sequence grouth-truth motion time.

| Time | Duration (Seconds) |
|---|---|
| 00:00:50 00:01:10 | 20 |
| 00:01:25 00:01:45 | 20 |
| 00:02:12 00:01:10 | 5 |
| 00:02:39 00:02:45 | 6 |
| 00:03:50 00:04:48 | 58 |
| 00:05:00 00:05:35 | 35 |
| 00:05:40 00:06:00 | 20 |
| Total | 164 |

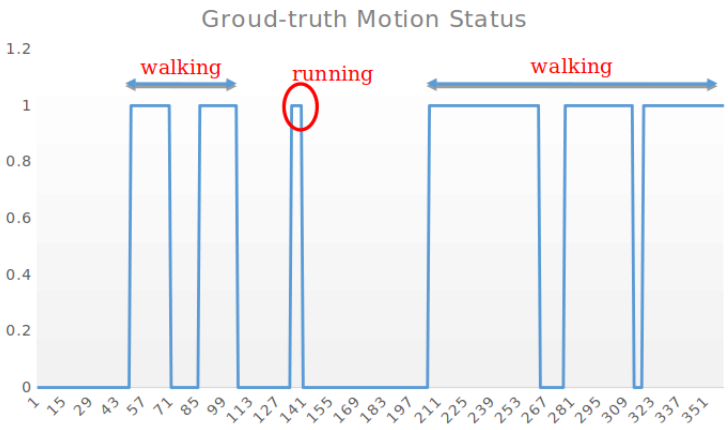



Fig. 4.3 The ground-truth motion time of our video test sequence.

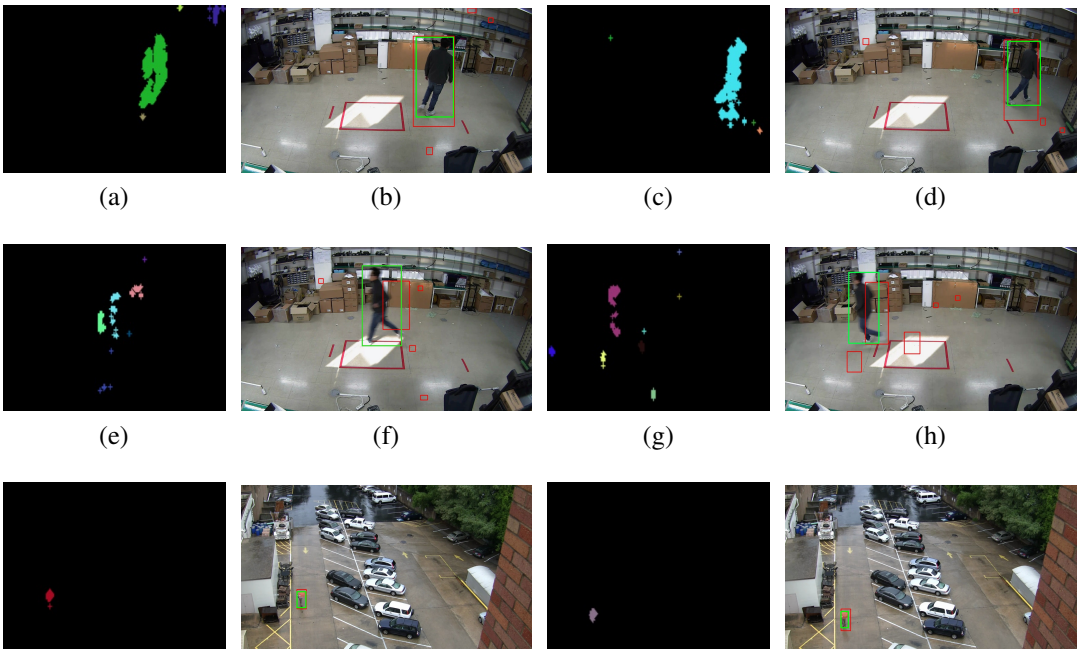
node involves the moving objects detection and runs on a low-computation device named *Raspberry Pi 4* and video analytics server is executed on *Nvidia- Jetson Xavier* because of a GPU that is supported to run YOLO. The hardware specifications of edge node and video analytics server are then listed in Table 4.1. Note that the two devices are directly connected to a router using a wired cable.

- Video Test Sequence: experiments have been conducted on the two video datasets. The VIRAT video dataset [4] was collected in natural scenes showing people performing normal actions for video surveillance domains. The second dataset is previously recorded from the our surveillance camera and uploaded [5]. The details of our video test sequence and the ground-truth motion time are shown in Figure 4.3 and listed in Tables 4.3,4.4.

We evaluate the performance of the moving object detection method and the proposed edge-to-cloud system separately.

## 4.1 The low-compelxity algorithm of moving object detection in Video Compressed Domain

To evaluate the quality of the proposed method for the moving object detection, we calculated the IoU score mettric of the detected object's bounding box with those of the ground-truth bounding box. Because the accuracy of the peoposed method is depend on the MV's density, we run the proposed method multiple times with different scenario application such as: different camera distances, different moving object speeds. We observe that except the I-Frame which does not apply motion estimation, the moving objects including sometimes big MV noises always be detected in other frames. Example of the moving object detection results are shown in Figure 4.4.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 4.4 The comparision of moving objects detection by the proposed method in different cases: (a, b, c, d) for human walking; (e, f, g, h) for human running; (i, j, k, l) when test with a far distance of camera.

Table 4.4 The average IoU calculation of the moving object detection in compressed-domain in different cases.

| *Test Sequence* | *Scenario* | | *IoU Average* |
|---|---|---|---|
| | Camera Position | Moving Speed | |
| Our recorded test video | Near | Normal | 0.75 |
| Our recorded test video | Near | Fast | 0.26 |
| Video Test from VIRAT | Far | Normal | 0.6 |

Table 4.5 Per-Frame Interference Time. Values are expressed in miliseconds (ms) and frame per second(FPS).

| *Frame Size* | *ST-MRF[16]* | *Graph Cuts[15]* | *Proposed Method* |
|---|---|---|---|
| 1280x720 | 64 ms (16 FPS) | 62 ms (17 FPS) | **39 ms (26 FPS)** |
| 1920x1080 | N/A | N/A | **69 ms (14 FPS)** |

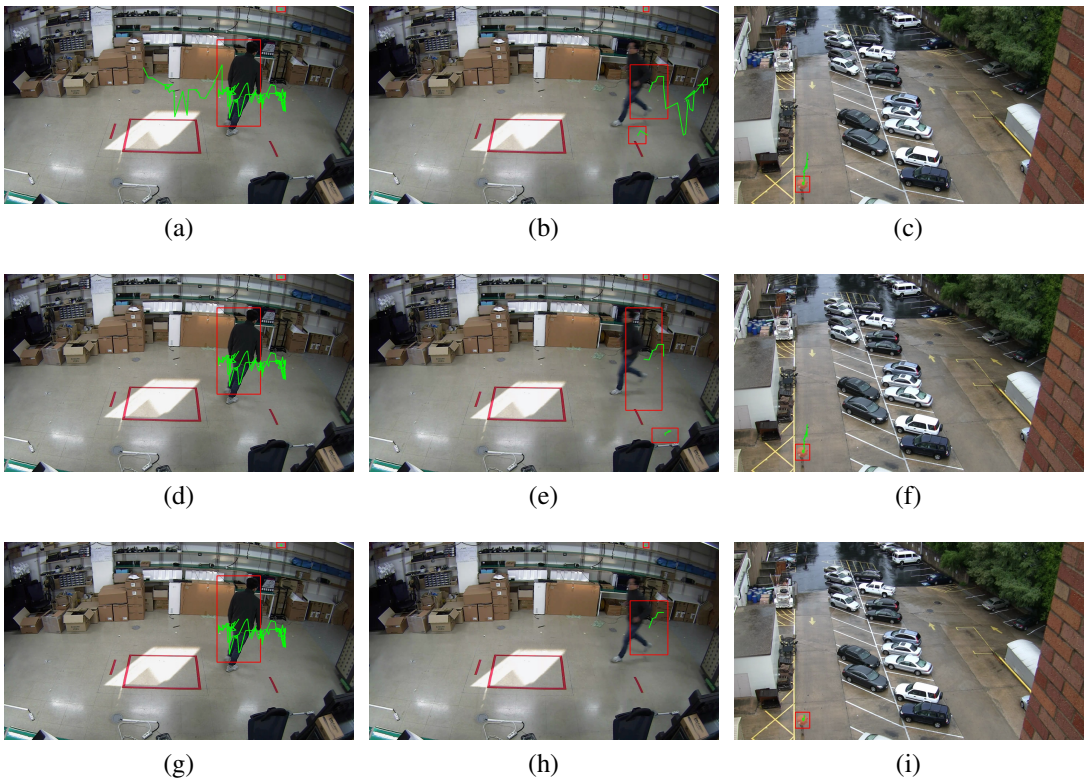


Fig. 4.5 The performance of moving objects tracking by the proposed method with different $\alpha$ value: for $\alpha = 0.25$ in (a, b, c), for $\alpha = 0.4$ in (d, e, f) and for $\alpha = 0.5$ in (d, e, f).

The green bounding boxes shown in the Figure are ground-truth bounding boxes, while the red bounding boxes are detected objects using the proposed method.The average IoU scores are shown in detail for each scenario in Table 4.4. The results show that when a camera is placed at a close distance and the human does not move fast (i.e, human walking), the proposed method detected well and achieved good average IoU score of 0.75. However, when speed of human is fast in case of running or the camera is in far distance, the MV's density is decreased, the detector return lower average IoU scores. For object tracking evaluation, the scenarios are run with multiple times with different detection score threshold $\alpha$ of 0.25, 0.4 and 0.5 with the results are shown in Figure 4.5. In this experiment, each $\alpha$ threshold was applied for three same scenarios of human walking in Figure 4.5 (a, d, g), human running in Figure 4.5 (b, e, h) and the far camera distance in Figure 4.5 (c, f, i). We see that with each

test scenario, with lower $\alpha$ threshold value,the tracking object capability is better with longer the object trajactories. However, it will increase the number of false alarm detection if the noise MVs apprear frequently in some specified areas.

To optimize the running time speed as well as the performance, the edge node and cloud node all are implemented in C++ using the multithread architecture. We observe that the proposed method does not indicate additional computational difficulty and achieve the approved evaluated results in terms of processing time when compared with previous studies [12] [32]. The average comsuming time is messasured with different video resolutions and shown in Table 4.5. Compare to other studies, the average processing time for high definition resolution video is approximately 39ms/frame, which outperforms most of the state of art method. This indicates that the proposed algorithm almost handles the data in real-time.

## 4.2 Performance Evaluation Results

In this experiment, the proposed edge-to-clound platform will be implemented and evaluated. Our recored video test sequence is selected because of including different the moving object speeds. Base on the last experiment result shown in Table 4.4, we choose $\alpha$ is 0.25 to cover both walking and running case and compared computing resources with the conventional method (no frame filter). The entire demonstration was recorded and uploaded [6]. The evaluated performances of the demonstration are presented in detail as follows.

### 4.2.1 Computing Resources Consumption

During demonstration of intrusion detection application, the clouds node's computing resources, including CPU, GPU utilization, and network download throughput, was monitored and recorded and shown in Figure 4.1. For comparison, the performance results are compared to that of the conventional method, which did not use edge node as shown in Figure 1.2. The

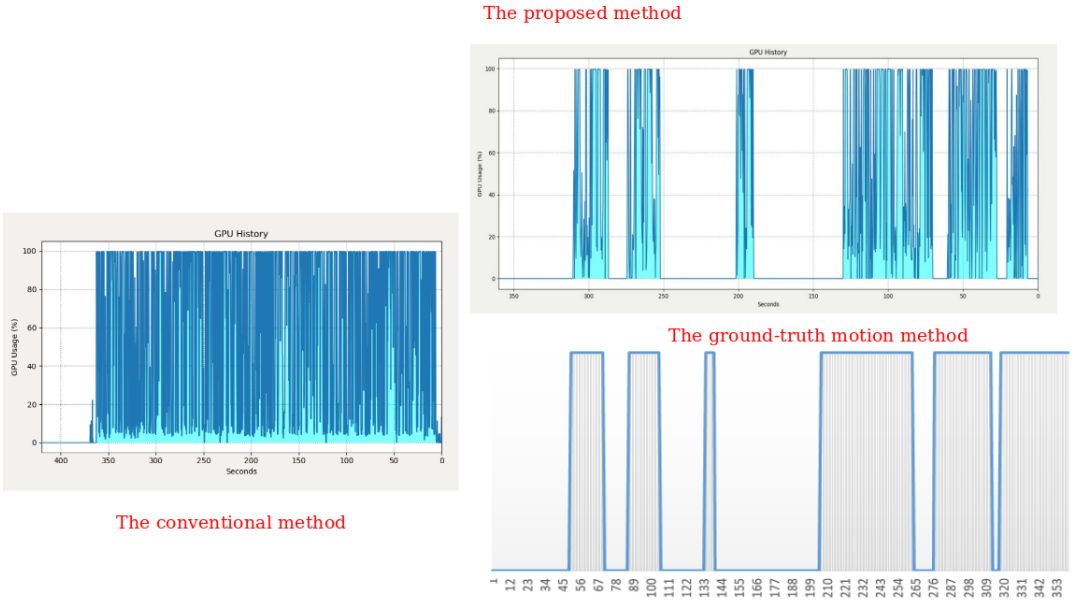



Fig. 4.6 GPU Load Comparision between the conventional method and the proposed method ($\alpha = 0.25$).

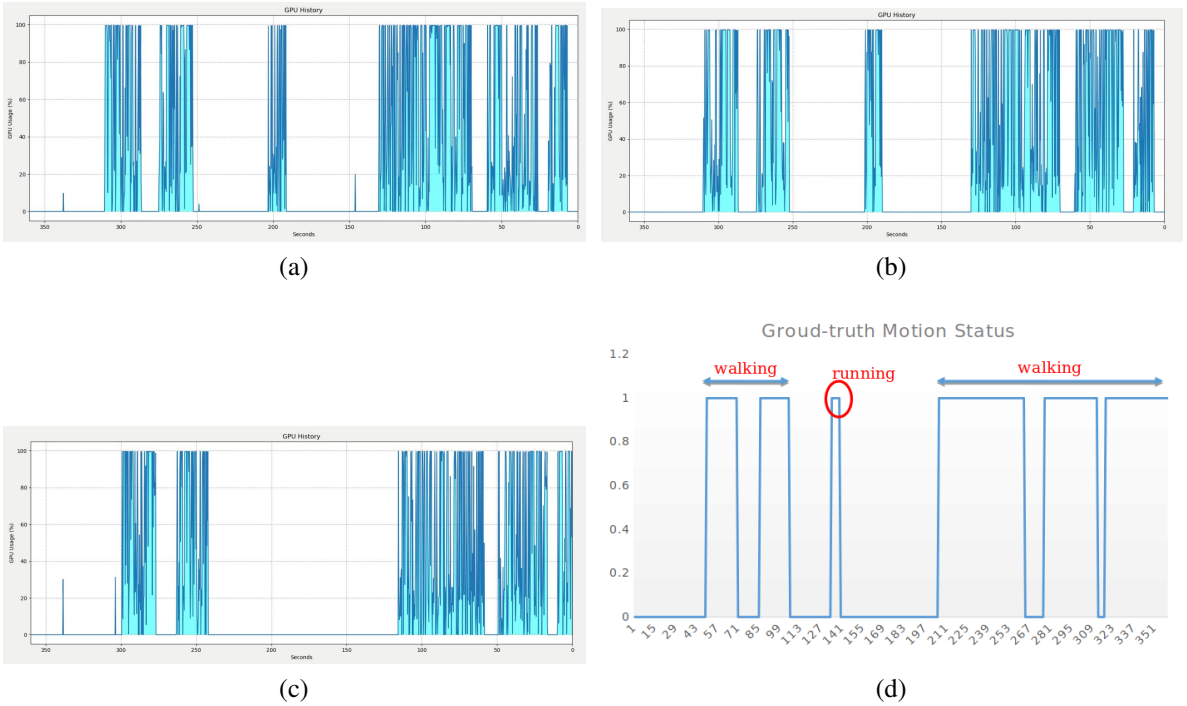



Fig. 4.7 GPU Monitoring: (a) $\alpha = 0.5$, (b) $\alpha = 0.25$, (c) $\alpha = 0.75$, (d) Ground Truth Motion Chart

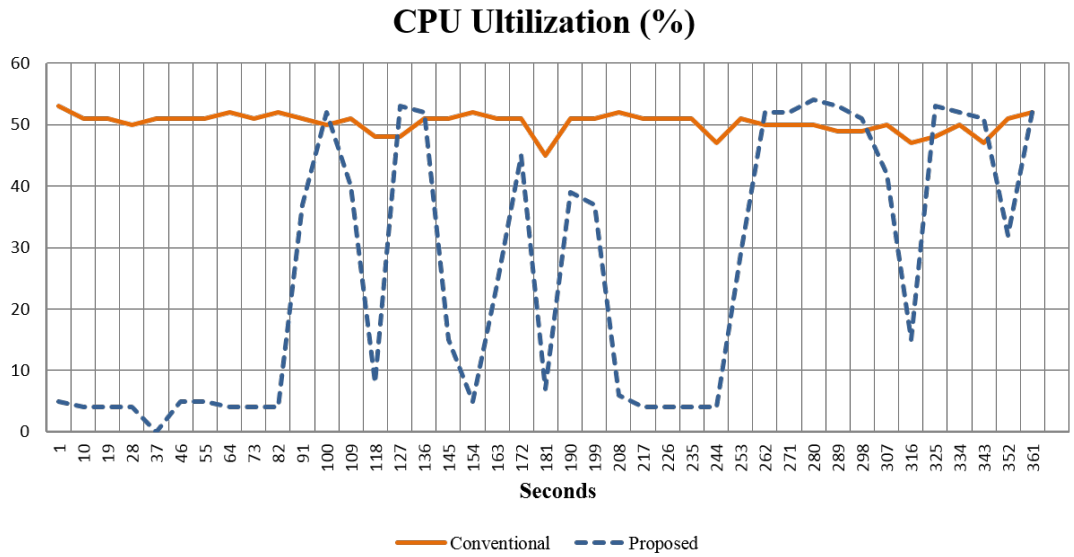



Fig. 4.8 CPU Load Comparision between the conventional method and the proposed method.

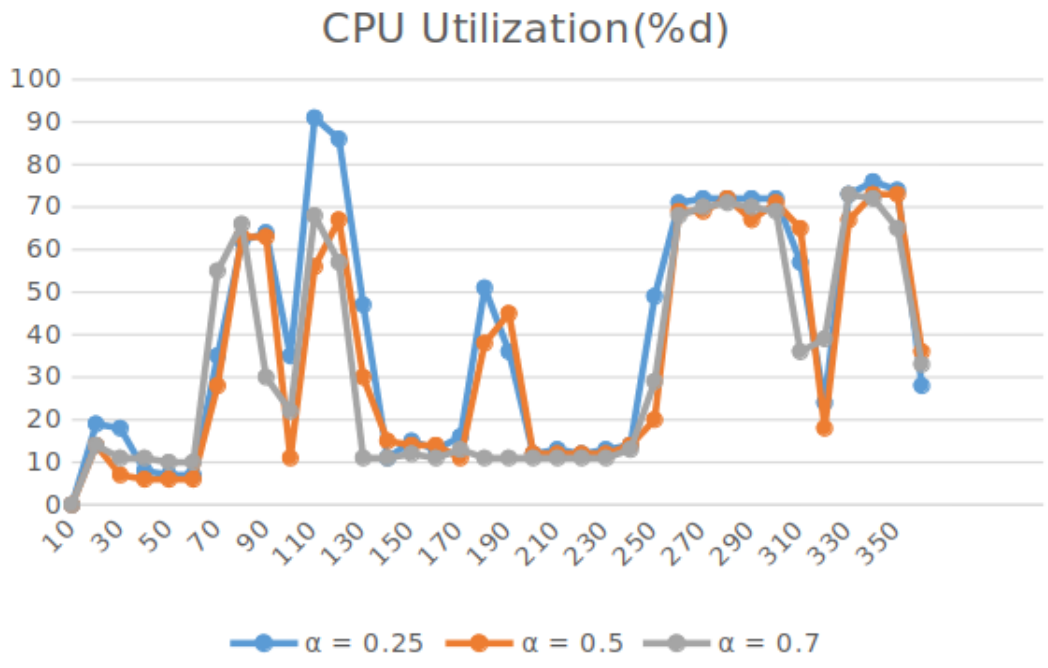



Fig. 4.9 CPU Ultilization with different alpha threshold.

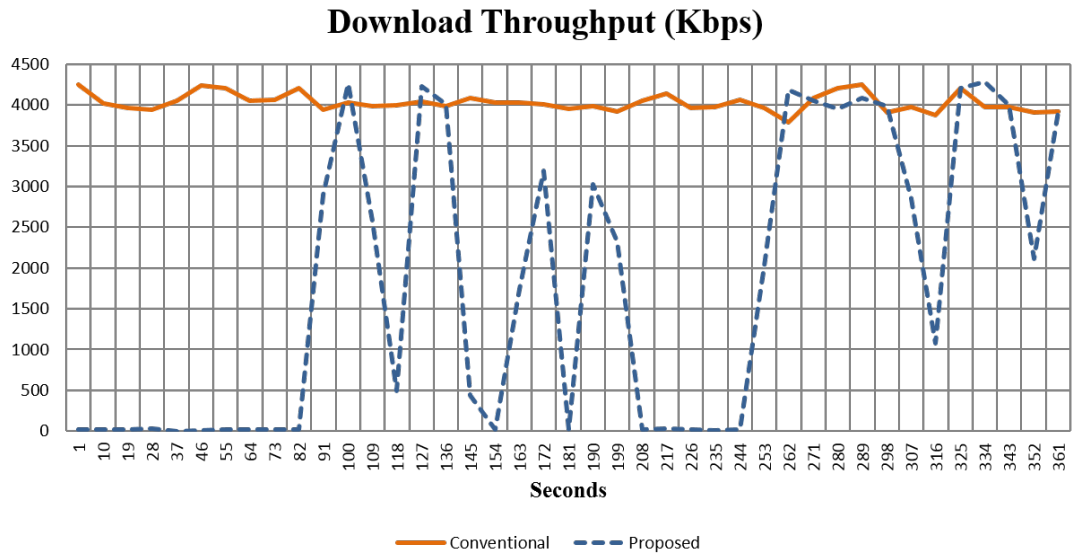



Fig. 4.10 Network traffic comparision between the conventional method and the proposed method.

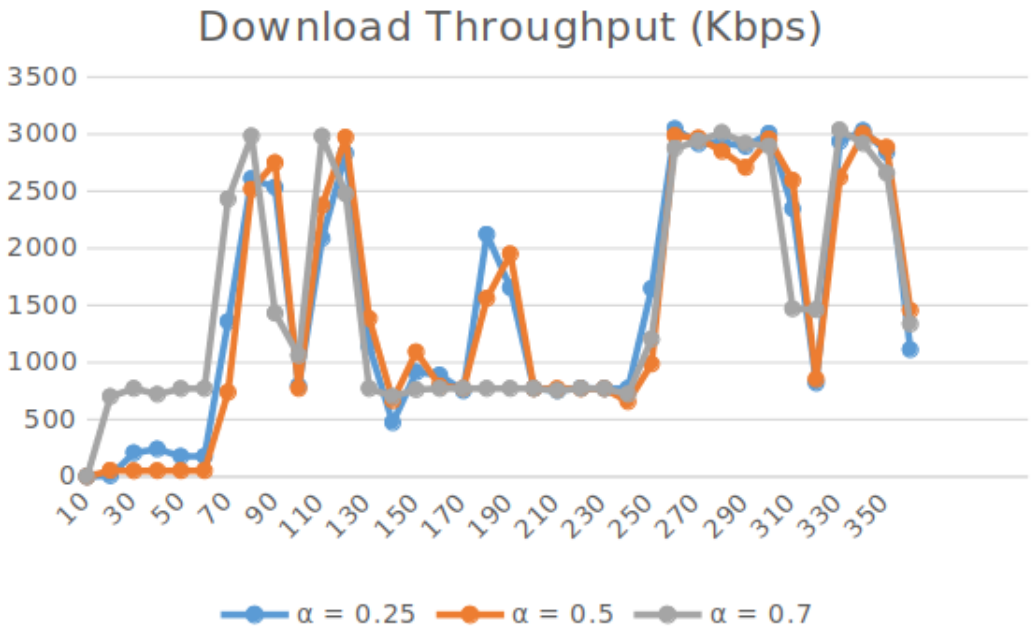



Fig. 4.11 Download Throughput with different alpha threshold.

Table 4.6 The average computing resources comparision between the conventional method and the proposed method with $\alpha$=0.25

| **Computing Resources** | **Conventional Method** | **Proposed Method** | **Performance Ratio** |
|---|---|---|---|
| GPU Utilization (%) | 65.61 | 30 | 0.45 |
| CPU Utilization (%) | 50.24 | 39.64 | 0.51 |
| Download Throughput (Kbps) | 4028 | 1509 | 0.35 |

Table 4.7 The average computing resources comparision the conventional method and the proposed method with $\alpha$=0.5

| **Computing Resources** | **Conventional Method** | **Proposed Method** | **Performance Ratio** |
|---|---|---|---|
| GPU Utilization (%) | 65.61 | 22.3 | 0.34 |
| CPU Utilization (%) | 50.24 | 34 | 0.45 |
| Download Throughput (Kbps) | 4028 | 1478 | 0.36 |

Table 4.8 The average computing resources comparision the conventional method and the proposed method with $\alpha$=0.7

| **Computing Resources** | **Conventional Method** | **Proposed Method** | **Performance Ratio** |
|---|---|---|---|
| GPU Utilization (%) | 65.61 | 33.73 | 0.33 |
| CPU Utilization (%) | 50.24 | 25.9 | 0.43 |
| Download Throughput (Kbps) | 4028 | 1808.2 | 0.35 |

results indicate that both methods achieve the same accuracy in terms of alert notification when humans entered the restricted area. The demonstration's record of both these methods are uploaded to [7], [8]. In terms of consumption of computational resources, GPU, CPU utilizations, and download throughput of both methods are measured. In figure 4.6, We measured the GPU computing resource of both the conventional method and proposed method. In conventional method, all frame is pushed into by deep learning model to detect object. So GPU was always at high loading status during video time. In contrast, the proposed method filters all static frame so GPU load status will change according to moving object status of the video. Compare the ground-truth motion chart and GPU load consumption chart, we can see the similarity between them. We also test the same experiment with other threshold alpha to evaluate the change of GPU load consumption graph. The results was shown in Figure 4.7. We found that when alpha value is big ($\alpha = 0.7$), the moving object at high speed will be missed. Because the IoU score is small when object move fast. Similar with GPU load consumption, CPU load and Download throughput have same story as presented in Figure 4.8 and 4.10 . With conventional method always be high consumption load and the proposed method, consumption load is varied following the ground- truth motion chart. And, alpha value have same effect to CPU and download throughput. When alpha is small, it can detect well for moving object at different speed (running and walking), however can be included false alarm. When alpha is too big value, some fast moving motion was missed detection. The figures clearly show the advantage of using the proposed method. While the conventional method of processing all video frames captured from camera leads to consumption of computational resources, our method only processes when there is motion, and therefore computational resources are dynamically allocated and economical. Another observation is that our proposed method is extremely effective for detecting motion of the scene compared to the ground-truth motion time in Table 4.3. In detail, the time for consuming and releasing computing resources of VA server was matched with the time for appearance and disappear-

ance of motion in ground-truth table.

In order to evaluate the efficiency of the proposed method, we calculate the performance ratio of the conventional method (without filtering function)and the proposed method (with filter function run on the edge device). Since the conventional method processes all video frames, we assume that the its computing resource average is S. According to the equation 3.6, the performance ratio of both methods in this video will be calculated as following:

$$\frac{Comp_p}{Comp_c} = \frac{(K*S+(N-K)*T)}{N*S} = \frac{K}{N}+(1-\frac{K}{N})*\frac{T}{S} = \frac{140}{360}+(1-\frac{140}{360})*\frac{T}{S} = 0.4+0.6*\frac{T}{S} \tag{4.1}$$

Compared with the performance ratio, which is calculated in real scenario test in Table 4.6, 4.7, 4.8, our performance theory model and real measurement are matching and reasonable.

# References


[1] (2020 (accessed September 3, 2020)a). *Available Online*. https://www.nytimes.com/2018/05/26/us/chicago-police-surveillance.html.

[2] (2020 (accessed September 3, 2020)b). *Available Online*. https://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html.

[3] (2020 (accessed September 3, 2020)c). *Available Online*. https://www.wsj.com/articles/BL-CJB-22562.

[4] (2020 (accessed September 3, 2020)). *Available Online*. https://viratdata.org.

[5] (2020 (accessed September 3, 2020)). *Available Online*. https://youtu.be/v24ldT1AGRw.

[6] (2020 (accessed September 3, 2020)). *Available Online*. https://github.com/diennv/MotionVectorAnalysis.

[7] (2020 (accessed September 3, 2020)). *Available Online*. https://www.youtube.com/watch?v=Cz_zxr_ElTU.

[8] (2020 (accessed September 3, 2020)). *Available Online*. https://www.youtube.com/watch?v=-fRc36HAduI&feature=youtu.b.

[9] Achanta, R., Kankanhalli, M., and Mulhem, P. (2002). Compressed domain object tracking for automatic indexing of objects in mpeg home video. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 2, pages 61–64. IEEE.

[10] Ananthanarayanan, G., Bahl, V., Cox, L., Crown, A., Nogbahi, S., and Shu, Y. (2019). Demo: Video analytics-killer app for edge computing. In *ACM MobiSys*.

[11] Babaee, M., Dinh, D. T., and Rigoll, G. (2018). A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76:635–649.

[12] Bombardelli, F., Gül, S., Becker, D., Schmidt, M., and Hellge, C. (2018). Efficient object tracking in compressed video streams with graph cuts. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.

[13] Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239.

[14] Bross, B., Helle, P., Lakshman, H., and Ugur, K. (2014). Inter-picture prediction in hevc. In *High Efficiency Video Coding (HEVC)*, pages 113–140. Springer.

[15] Cai, Z., Saberian, M., and Vasconcelos, N. (2015). Learning complexity-aware cascades for deep pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3361–3369.

[16] Canel, C., Kim, T., Zhou, G., Li, C., Lim, H., Andersen, D. G., Kaminsky, M., and Dulloor, S. R. (2019). Scaling video analytics on constrained edge nodes. *arXiv preprint arXiv:1905.13536*.

[17] Chen, N., Chen, Y., Ye, X., Ling, H., Song, S., and Huang, C.-T. (2017a). Smart city surveillance in fog computing. In *Advances in Mobile Cloud Computing and Big Data in the 5G Era*, pages 203–226. Springer.

[18] Chen, T. Y.-H., Ravindranath, L., Deng, S., Bahl, P., and Balakrishnan, H. (2015). Glimpse: Continuous, real-time object recognition on mobile devices. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pages 155–168.

[19] Chen, Y., Wang, J., Zhu, B., Tang, M., and Lu, H. (2017b). Pixel-wise deep sequence learning for moving object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

[20] Dong, L., Zoghlami, I., and Schwartz, S. C. (2006). Object tracking in compressed video with confidence measures. In *2006 IEEE International Conference on Multimedia and Expo*, pages 753–756. IEEE.

[21] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

[22] Favalli, L., Mecocci, A., and Moschetti, F. (2000). Object tracking for retrieval applications in mpeg-2. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(3):427–432.

[23] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

[24] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

[25] Gujrathi, P., Priya, R. A., and Malathi, P. (2014). Detecting moving object using background subtraction algorithm in fpga. In *2014 Fourth International Conference on Advances in Computing and Communications*, pages 117–120. IEEE.

[26] Hsieh, K., Ananthanarayanan, G., Bodik, P., Venkataraman, S., Bahl, P., Philipose, M., Gibbons, P. B., and Mutlu, O. (2018). Focus: Querying large video datasets with low latency and low cost. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pages 269–286.

[27] Hussain, M., Bird, J. J., and Faria, D. R. (2018). A study on cnn transfer learning for image classification. In *UK Workshop on Computational Intelligence*, pages 191–202. Springer.

[28] Jarrett, K., Kavukcuoglu, K., LeCun, Y., et al. (2009). What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 2146–2153. IEEE.

[29] Jiang, J., Ananthanarayanan, G., Bodik, P., Sen, S., and Stoica, I. (2018). Chameleon: scalable adaptation of video analytics. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 253–266.

[30] Jiang, X., Song, T., Katayama, T., and Leu, J.-S. (2019). Spatial correlation-based motion-vector prediction for video-coding efficiency improvement. *Symmetry*, 11(2):129.

[31] Kang, D., Emmons, J., Abuzaid, F., Bailis, P., and Zaharia, M. (2017). Noscope: optimizing neural network queries over video at scale. *arXiv preprint arXiv:1703.02529*.

[32] Khatoonabadi, S. H. and Bajic, I. V. (2012). Video object tracking in the compressed domain using spatio-temporal markov random fields. *IEEE transactions on image processing*, 22(1):300–313.

[33] Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

[34] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

[35] Kumar, S. and Yadav, J. S. (2016). Segmentation of moving objects using background subtraction method in complex environments. *Radioengineering*, 25(2):399–408.

[36] Laroche, G., Jung, J., and Pesquet-Popescu, B. (2008). Rd optimized coding for motion vector predictor selection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1247–1257.

[37] LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256. IEEE.

[38] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM.

[39] Lee, J. and Park, M. (2012). An adaptive background subtraction method based on kernel density estimation. *Sensors*, 12(9):12279–12300.

[40] Li, C., Xing, Q., and Ma, Z. (2020a). Hksiamfc: Visual-tracking framework using prior information provided by staple and kalman filter. *Sensors*, 20(7):2137.

[41] Li, H., Lin, Z., Shen, X., Brandt, J., and Hua, G. (2015). A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334.

[42] Li, Y., Padmanabhan, A., Zhao, P., Wang, Y., Xu, G. H., and Netravali, R. (2020b). Reducto: On-camera filtering for resource-efficient real-time video analytics. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 359–376.

[43] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

[44] Lu, N., Wang, J., Wu, Q., and Yang, L. (2008). An improved motion detection method for real-time surveillance. *IAENG International Journal of Computer Science*, 35(1).

[45] Lu, X., Izumi, T., Takahashi, T., and Wang, L. (2014). Moving vehicle detection based on fuzzy background subtraction. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 529–532. IEEE.

[46] Nguyen, D. V. and Choi, J. (2020). Toward scalable video analytics using compressed-domain features at the edge. *Applied Sciences*, 10(18):6391.

[47] Ohm, J.-R., Sullivan, G. J., Schwarz, H., Tan, T. K., and Wiegand, T. (2012). Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc). *IEEE Transactions on circuits and systems for video technology*, 22(12):1669–1684.

[48] Ou, X., Yan, P., Zhang, Y., Tu, B., Zhang, G., Wu, J., and Li, W. (2019). Moving object detection method via resnet-18 with encoder–decoder structure in complex scenes. *IEEE Access*, 7:108152–108160.

[49] Patil, P. W. and Murala, S. (2018). Msfgnet: A novel compact end-to-end deep network for moving object detection. *IEEE Transactions on Intelligent Transportation Systems*, 20(11):4066–4077.

[50] Philippou, O. (2020 (accessed September 3, 2020)). *Video Surveillance Installed Base Report – 2019*. https://technology.informa.com/607069/video-surveillance-installed-base-report-2019.

[51] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

[52] Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[53] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

[54] Sheu, R.-K., Pardeshi, M., Chen, L.-C., and Yuan, S.-M. (2019). Stam-ccf: Suspicious tracking across multiple camera based on correlation filters. *Sensors*, 19(13):3016.

[55] Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE.

[56] Stone, T., Stone, N., Jain, P., Jiang, Y., Kim, K.-H., and Nelakuditi, S. (2019). Towards scalable video analytics at the edge. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE.

[57] Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T. (2012). Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668.

[58] Thanh Le, T., Jeong, J., and Ryu, E.-S. (2019). Efficient transcoding and encryption for live 360 cctv system. *Applied Sciences*, 9(4):760.

[59] Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104(2):154–171.

[60] Wang, Y., Luo, Z., and Jodoin, P.-M. (2017). Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters*, 96:66–75.

[61] Wang, Z., Sun, X., Diao, W., Zhang, Y., Yan, M., and Lan, L. (2019). Ground moving target indication based on optical flow in single-channel sar. *IEEE Geoscience and Remote Sensing Letters*, 16(7):1051–1055.

[62] Yoneyama, A., Nakajima, Y., Yanagihara, H., and Sugano, M. (1999). Moving object detection and identification from mpeg coded data. In *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, volume 2, pages 934–938. IEEE.

[63] Zeng, D. and Zhu, M. (2018). Background subtraction using multiscale fully convolutional network. *IEEE Access*, 6:16010–16021.