ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 5, May 2022

Multi-View Video Summarization Using RNN and SURF Based High Level Moving Object Feature Frames

[1] Vinsent Paramanantham*, [2] Dr. S. SureshKumar

[1] *Research Scholar, Sathyabama Institute of Science and Technology (Deemed University), Tamil Nadu, India.
[2] Principal at Swarnandhra College of Engineering Technology, Narasapur, Andhra Pradesh, India.
Corresponding Author Email: [1] *vinsent.storage@gmail.com

Abstract— ***Multi-View Video summarization is a process to ease the storage consumption that facilitates organized storage, and perform other mainline videos analytical task. This in-turn helps quick search or browse and retrieve the video data with minimum time and without losing crucial data. In static video summarization, there is less challenge in time and sequence issues to rearrange the video-synopsis. The low-level features are easy to compute and retrieve. But for high-level features like event detection, emotion detection, object recognition, face detection, gesture detection, and others requires the comprehension of the video content. This research is to propose an approach to over- come the difficulties in handling the high-level features. The distinguishable contents from the videos are identified by object detection and feature-based area strategy. The major aspect of the proposed solution is to retrieve the attributes of a motion source from a video frame. By dividing the details of the object that are available in the video frame wavelet decomposition are achieved. The motion frequency scoring method records the time of motions in the video. The frequency motion feature of video usage is a challenge given the continuous change of objects shape. Therefore, the object position and corner points are spotted using Speeded Up Robust Features (SURF) feature points. Support vector machine clustering extracts keyframes. The memory-based re- current neural network (RNN) recognizes the object in the video frame and remembers a long sequence. RNN is an artificial neural network where nodes form a temporal relationship. The attention layer in the proposed RNN network extracts the details about the objects in motion. The motion objects identified using the three video clippings is finally summarized using video summarization algorithm. To perform the simulation, MATLAB R 2014b software was used.***

Keywords – Joint multi-view video summarization high level features moving object detection SURF feature points

I. INTRODUCTION

Multi-view summarization has gained much traction as in today's video summarization era. As in real- time, there is always more than a single camera in surveillance. This approach also helps to build a framework with a few limited classes for classification and obtain high accuracy scores for the desired object detection in your video summary. The low-level visual key attributes such as colour, texture and motions form the prime fea- tures for low-level feature-based video summarization [1]. Colour is by far the most commonly used low-level feature. While comparing two frames by using colour histograms, One may encounter a situation in which two frames are entirely different, but, their histograms are similar because the video frames have the same colour distributions. TThe texture from the low-level feature for texture extraction, wavelet transforms like Discrete Haar Wavelet [2], and Daubechies wavelet [3, 4]. High-level feature-based summarization techniques involve high-time and computational costs, as discussed in [5, 6]. [7] views summarization as a structured prediction problem, those techniques model the long-range dependency in video using the popular long short-term mem- ory units (LSTM) and its variants [7].The prime insight is to optimize the accuracy of which frames or subshots to be in the summary. The initial stage computes the frame-level scores, and keyframes are selected depending on the combined scores in the next stage. Lastly, the elimination in the third stage of duplicate frames. Many variables, including acquisition, processing, compression, transmission, display and video reproduction sys- tems can affect the quality of visual media [8].

The rich feature quality of the individual frames forms the building blocks of the video [9]. A large, smooth region divided by sharp edges is component of the feature extraction, which depends on the particulars of the image. The colour image otherwise consists of continuous texture and colour information. [10] discussed that the summarization algorithm consists of two essential procedures: 1) searching the matched patches with per- ceptual similarity information for the hole patch, 2) fusing the matched patches to obtain the restored patch for hole regions, here the matching patches focusing on the feature extraction allow pattern detection for the sum- marization of media content, for which there are specific methods for the treatment of the audio plus image, feature extraction via artificial Intelligence with neural network and image processing like optical character recognition to identify desired portions or shapes in the video as discussed by [11] .

VS process is the detection and extraction of notable objects and their movements in the video content and to obtain condensed keyframes. Pixel intensities of the objects and non-objects are similar, thereby the existing methods fail to detect the objects in motion in low colour and contrast video frames. Interestingly, the edges of the artifacts are prominent in low contrast area connected to the lines arcs and other geometric primitives in high-level shape descriptors [12]. The curves obtained through traditional-features exhibits the cut effect phenomenon, as the feature curve changes dramatically between frames. The crop cut effect denotes a sudden change, but it needs more time to transform into another shot, this forms a violent vibration on the curve. The dissolve and fade effects are linear interpolation between shots [13]. The problems depend on feature extrac- tion, a type of dimensionality reduction that efficiently represents interesting-parts of an image in a compact feature vector. Reduced feature representation simplifies the tasks of image matching and retrieval.

The next session 2 discusses the literature survey, and session 3 discusses the feature extraction of moving objects in videos. Session 4 discusses the network analysis of both the finite impulse and infinite impulse recurrent networks can have additional stored state, and the storage can be under direct control by the neural network. Session 5 presents a feature matching of different video and the joint video summarization algorithm. Session 6 discusses the result analysis and follows the conclusion.

II. RELATED WORKS

[14]For accurate local object movement recognition, a video copy-move forgery technique is sug- gested. video copy-moves is a difficult problem to identify when uniform background are copied to the fore- ground object. A Patch Match algorithm determines the offset field to remove the false matches. False-match removal determines the efficiency of the algorithm.

[15]A novel detection method based on main component pursuit (PCP) called kinematic regularization with local null space pursuit (KRLNSP) was introduced that overcomes the challenge of false detections and com- putational loads.In aerial footage, KRLNSP models the setting as a subspace, and then the moving objects as sparse null spaces and the sparse objects depict unique kinematic properties. The author also covers the kine- matic regularization ’Γ’ [16]

[17] Proposed an objective function to identify the objects in motion involving complex background. This method solves using a linearize alternating direction method of multipliers (ADMM) based on batch optimiza- tion [18]. For real-time execution, the proposed solution works online. Moreover, an online optimization algorithm for real-time applications is proposed here. In different colour image backgrounds, a combined representation of red, green, blue colour characteristics, horizontal and vertical gradient characteristics, and motion function provides better accuracy.

[19] Instead of multi-frames, a local feature analysis route based on single-frame imagery was suggested, which can detect slow-moving target shadows in conventional ground moving targets (GMTI). Looking at the outcome of this process, The context model reconstructs using single-frame imagery to avoid the ”ghost” phenomenon in moving target detection algorithms involving multi-frame imagery.

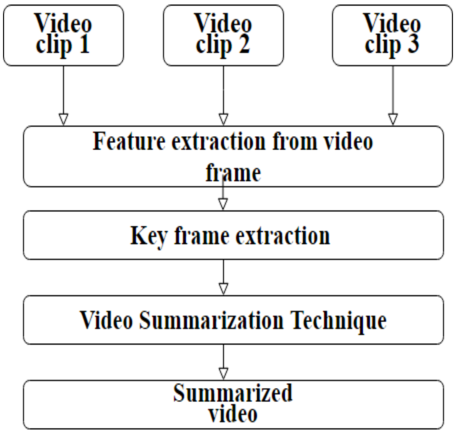


Figure 1. Summary of planned phases of the project

[20] Developed a mechanism for Unsupervised Online Video Object Segmentation (UOVOS) by ex- amining the motion properties of the mean moving of a universal object in segmented regions. The salient- motion and object map intersection defines the objects in motion. Salient-motion is measured using Mask R-CNN [21] from the optical flow data and object map.

[22] Using an adaptive blind update(random update) and RGB-D camera, to boost the accuracy in deep cam- ouflage scenarios, the implementation uses a bootstrapping segmentation and shadow detection system to dis- tinguish moving objects. The suggested technique, when referred with the original method [23] and the others implies that the segmentation findings are superior. The suggested solution achieves better efficiency in sta- tionary object detection and ghost phenomenon.

[24] In a high-speed environment, an algorithm was proposed to handle multiple object detection with limited hardware and tested the high-frame-rate method (HFR), high-data-throughput, and high-parallelism process- ing of video streams with low latency. Conventional Histograms of Directed Gradient (HOG) descriptor and Support Vector Machine (SVM) classifier algorithms are employed. The method provides low memory con- sumption by saving input pixel values and intermediate performance. Effective floating point to integer and integer truncation practices minimize the embedded memory usage.

The table 1 includes the literature survey that relates to low-level feature extraction process with various methods and techniques using artificial neural network. In the surveyed papers as in table 1, The authors propose video summarization using approaches compatible to handle large-scale video storage. The approach is ideal for computers in which the desired frame can be chosen from the video footage and transformed into static or dynamic content. To construct a chronological sequence of frames, the picked frames are then glued to Spatio-temporal orders. For any video search method, the proposed methodology were used to index backend video frames and to assist with the search results.

III. VIDEO PROCESSING FRAMEWORK

Figure 1 presents a project view covering three video clippings with video extraction and keyframe identification features and other processes for summarization.

In Joint video summarization process the different three video clipping taken for the method. The individual video clipping undergoes the pre-processing steps. The pre-processing steps as follows.

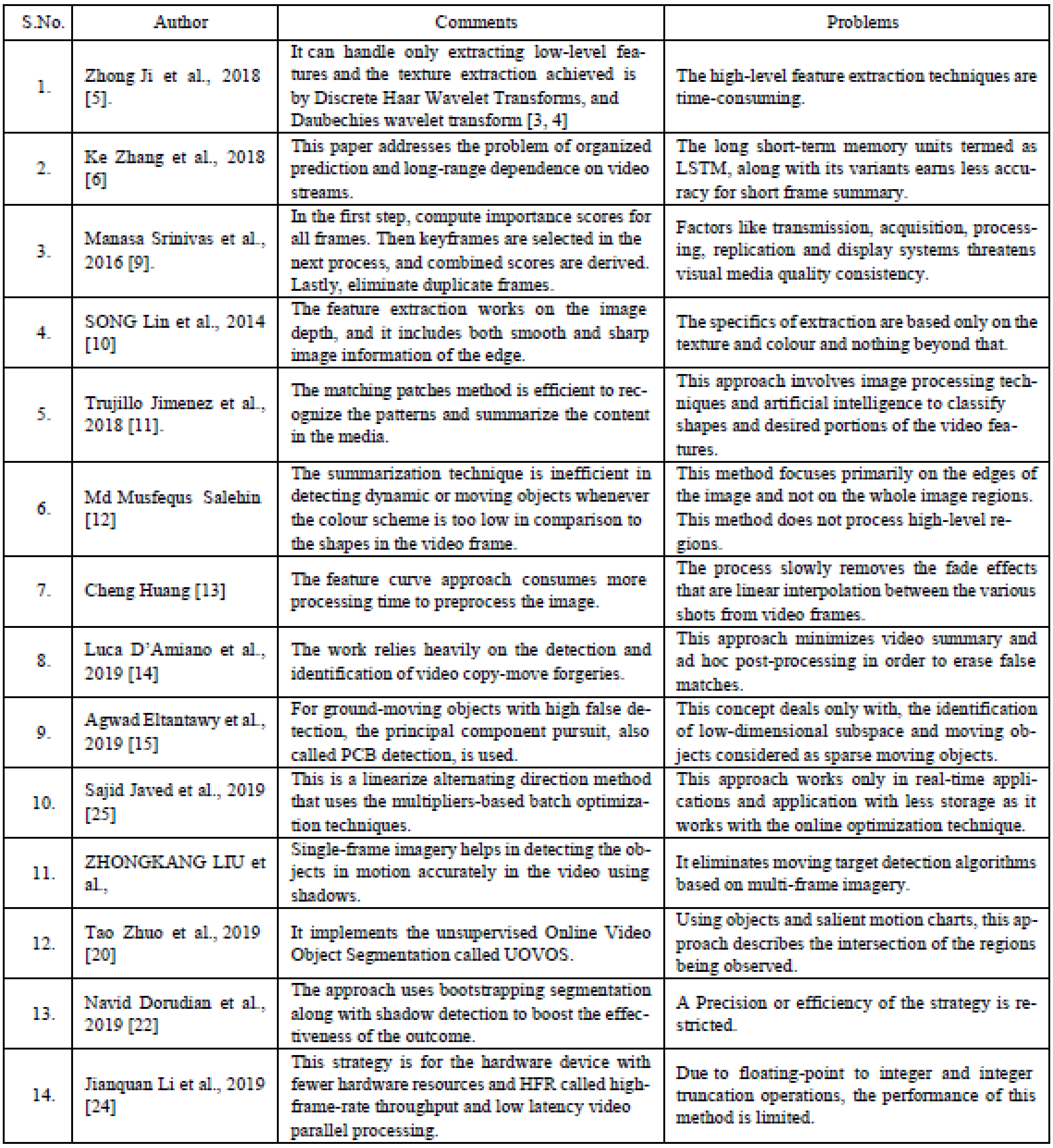
1. Video Frame Conversion
2. Gray Scale Conversion of individual frames
3. Frame Resizing

3.1. Video Frame Conversion

In general, when the frame rate is high, we will require more disk space as it consumes more space to store a quality image. The frame rate usually practised is 24 fps, as it gives a decent video quality. The video frame rate needs adjustment in this situation, and a reliable frame rate converter is required. The frames are created from the video and transformed into images, and then the luminance components are obtained. Later Kirsch edge detector is used to determine the edges, and finally, edge pixels counts are counted [26–28].

Table 1. Table view of introduction and literature survey.

The details of the literature review are given in this table.



S.No.	Author	Comments	Problems
1.	Zhong Ji et al., 2018 [5].	It can handle only extracting low-level features and the texture extraction achieved is by Discrete Haar Wavelet Transforms, and Daubechies wavelet transform [3, 4]	The high-level feature extraction techniques are time-consuming.
2.	Ke Zhang et al., 2018 [6]	This paper addresses the problem of organized prediction and long-range dependence on video streams.	The long short-term memory units termed as LSTM, along with its variants earns less accuracy for short frame summary.
3.	Manasa Srinivas et al., 2016 [9].	In the first step, compute importance scores for all frames. Then keyframes are selected in the next process, and combined scores are derived. Lastly, eliminate duplicate frames.	Factors like transmission, acquisition, processing, replication and display systems threatens visual media quality consistency.
4.	SONG Lin et al., 2014 [10]	The feature extraction works on the image depth, and it includes both smooth and sharp image information of the edge.	The specifics of extraction are based only on the texture and colour and nothing beyond that.
5.	Trujillo Jimenez et al., 2018 [11].	The matching patches method is efficient to recognize the patterns and summarize the content in the media.	This approach involves image processing techniques and artificial intelligence to classify shapes and desired portions of the video features.
6.	Md Musfequs Salehin [12]	The summarization technique is inefficient in detecting dynamic or moving objects whenever the colour scheme is too low in comparison to the shapes in the video frame.	This method focuses primarily on the edges of the image and not on the whole image regions. This method does not process high-level regions.
7.	Cheng Huang [13]	The feature curve approach consumes more processing time to preprocess the image.	The process slowly removes the fade effects that are linear interpolation between the various shots from video frames.
8.	Luca D'Amiano et al., 2019 [14]	The work relies heavily on the detection and identification of video copy-move forgeries.	This approach minimizes video summary and ad hoc post-processing in order to erase false matches.
9.	Agwad Eltantawy et al., 2019 [15]	For ground-moving objects with high false detection, the principal component pursuit, also called PCB detection, is used.	This concept deals only with, the identification of low-dimensional subspace and moving objects considered as sparse moving objects.
10.	Sajid Javed et al., 2019 [25]	This is a linearize alternating direction method that uses the multipliers-based batch optimization techniques.	This approach works only in real-time applications and application with less storage as it works with the online optimization technique.
11.	ZHONGKANG LIU et al.,	Single-frame imagery helps in detecting the objects in motion accurately in the video using shadows.	It eliminates moving target detection algorithms based on multi-frame imagery.
12.	Tao Zhuo et al., 2019 [20]	It implements the unsupervised Online Video Object Segmentation called UOVOS.	Using objects and salient motion charts, this approach describes the intersection of the regions being observed.
13.	Navid Dorudian et al., 2019 [22]	The approach uses bootstrapping segmentation along with shadow detection to boost the effectiveness of the outcome.	A Precision or efficiency of the strategy is restricted.
14.	Jianquan Li et al., 2019 [24]	This strategy is for the hardware device with fewer hardware resources and HFR called high-frame-rate throughput and low latency video parallel processing.	Due to floating-point to integer and integer truncation operations, the performance of this method is limited.

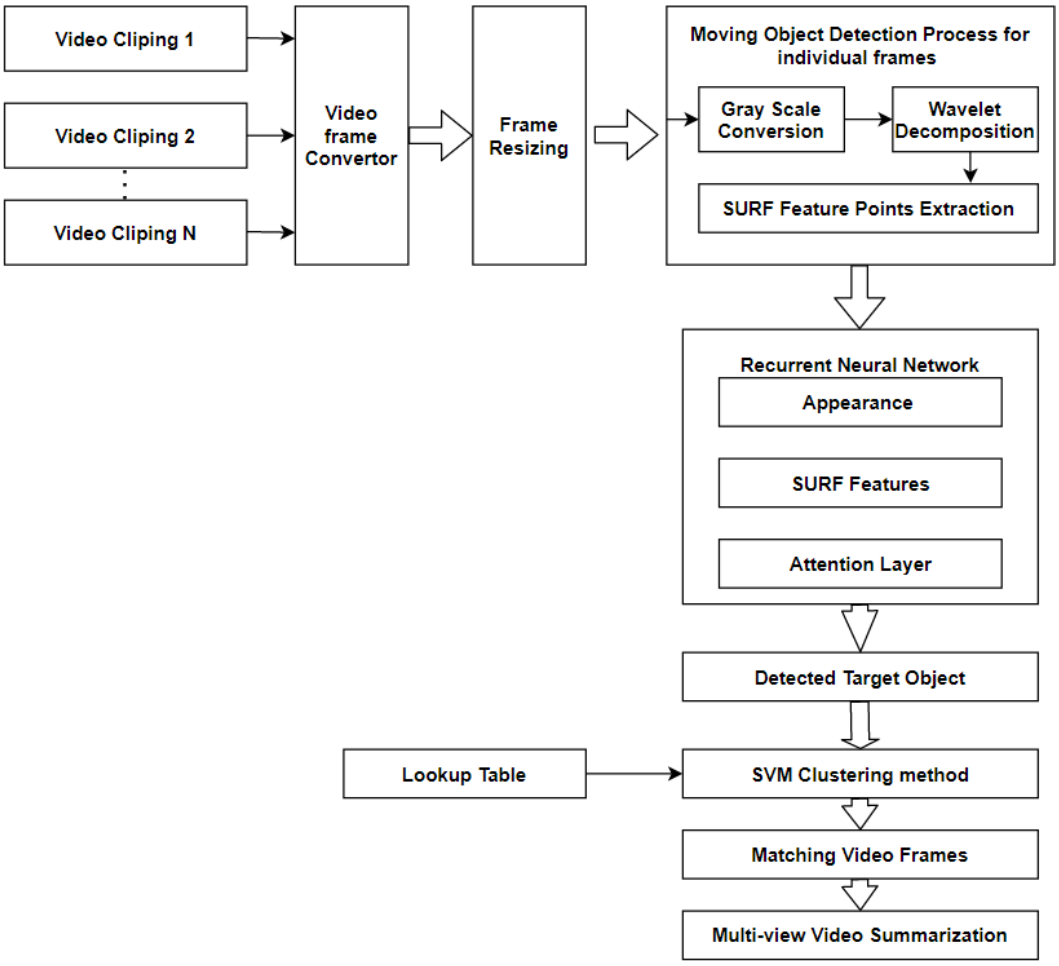


Figure 2. Multi-view video summarization processes

3.2. Gray Scale Conversion of individual frames

The luminance value in the grayscale image is 8-bit and ranges from 0 *to* 255 in the event of the RGB image, the values are 24 bits. To convert an RGB image into grayscale, the gamma compression function must remove the gamma expansion first and then transfer RGB into a linear colour space. For the linear colour components (*Rlinear*, *GLinear*, *Blinear*), the weighted sum is applied. When it is necessary to convert the grayscale to an encoded version and stored in nonlinear colour space, *Y linear* called linear luminance, the gamma-compressed image is restored. For the common *sRGB* color space, gamma expansion is defined as in equation 1

$$C_{linear} = \begin{cases} \frac{C_{srgb}}{12.92}, & C_{srgb} \leq 0.04045 \\ (\frac{C_{srgb}+0.055}{1.055})^{2.4}, & C_{srgb} > 0.04045 \end{cases} \tag{1}$$

In equation 1 C_{srgb} – three gamma compressed *sRGB* primaries, (R_{linear}, G_{Linear}, B_{linear}), each in range [0,1]) C_{linear} -linear-intensity value R_{linear}, G_{linear}, and B_{linear}, also in range [0, 1].

3.3. Video Frame Resizing

It is necessary to resize the picture based on the ANN network specifications and magnify the image known as resolution enhancement or scaling. Scaling of vector graphic images happens with a geometric transformation and validated image consistency. In scaling a raster graphic image, we need a new picture with lower or higher pixels. When the pixel number decreases, the consistency of the image tends to be lost [29]. In other phrases, using the Nyquist sampling theorem, the scaling of an image involves re-sampling or re- sampling or reconstructing an image. The theorem says that to prevent the aliasing artefacts a reduced sampling size and smaller image with high resolution and suitable 2D anti-aliasing filters are necessary. The reduced smaller size image carries the necessary data in it. Figure 1, depicts the pre-processing step involved in video processing. The figure 7,8,9 shows the video processes for resizing the frames and grayscale conversion [29].

Figure 2 depicts the process followed for the joint video summarization, where three videos are pro- cessed together and converted into a video frame. The individual video frame is resized through the resizing function, depending on the video length. Later the derived frames are pre-processed using filtering processes. In the developed framework, based on the length of the video, the three videos are regarded as input and later transformed into video frames, and the video frames are changed using the size filtering process. In the pre-process section, the grayscale conversion happens with the wavelet decomposition process with the help of Daubechies wavelets transform [3, 4]. In this step, the object and background are distinguished. Later the positions which were unable to recover are retrieved using *SURF* speeded up robust features.

The *RNN* forms the next session. The recurrent neural network that works on layers like appearance using attention layer and SURF feature points. The individual layers in the object are separated using feature points. At the end of this process, the object or person is identified in the context using the detected target object method. SVM classifies the video images and uses the lookup table or training dataset. Three video images are clubbed together using this feature point matching and produce the joint video summarization. The primary ideology is to suggest an unique approach to capturing the high-level features of objects in motion.

Keyframe extraction is accomplished by clustering with Support Vector Machine and recurrent neural network (RNN), along with attention layers. The attention layer function as the units/neurons in the hidden layers regarded as memory blocks, and each memory block can contain multiple memory cells. A video summary is a process that makes it easier to search video collections faster and also more effective in indexing and glancing the content. The summary can be former by choosing the keyframes that best indicates the actual video or through video skimming [30]. The keyframes extraction happens via change point, low-level features- based clustering, or clustering depending on the objects.

3.4. Video frame conversion

Frames are taken from videos and converted into an image. *Frame2im* function from MATLAB performs this transformation of video to an image. The .avi format format video is read using an *aviread* format function. To construct a high-resolution video frame, the spatial and temporal features of the image are needed. A novel framework is framed to resolve the problems of generating a video frame using motion-compensated sub-sampling [31]. The Bayesian framework deals with the challenge of multi-frame interpolation and video sequence alignment [32, 33]. The end design of the proposed algorithm satisfies to deliver high quality and usable image frame.

3.5. Frame Resizing

Image resizing and magnification of the image is termed as resolution enhancement or upscaling. In the vector graphic image, the images are scaled up or magnified using geometric transformations, and there is no compromise to the image quality. While handling a raster graphic image, the original image is transformed into a lower or higher pixel value. While lowering the pixel value, it will cause a loss in image quality. Image scaling is a type of resampling or rebuilding an image, according to the Nyquist sampling theorem. The Nyquist theorem states that an image can only be downgraded to a smaller scale using a 2*D* anti-aliasing filter to prevent the loss of aliasing artefacts. The picture is reduced to 100 100. The resized image holds necessary details for further process. Decreasing the size of the frame is pressing because they take additional space and helps computationally. In this process of minimizing the size, only the physical size and image resolution will change. The other attributes are unaltered [29].

3.6. Moving Object Detection Processes for Individual frames

To identify the objects in motion is the biggest challenge in video streaming surveillance applications. Object detection plays a critical role in video analysis and is usually performed by discarding the background image or using object detectors and manually marking the object [34, 35].

The temporal differencing [36] method implements the pixel-wise difference method where two or many frames are studied to spot the objects in motion. For dynamic frames in which the images keep changing, this method is well suited, and it is unsuccessful to retrieve all the relevant details from the object. In the scenario of a stagnant foreground object movement, the temporal differentiation strategy fails to act and record the object. Let Frame *i* represent the grey-level intensity value at pixel position *i* and at time instance *n* of video

image sequence *I*, which is within the threshold of [0, 255]. *T* is the threshold initially set to a pre-determined value.

3.7. Motion Frequency Scoring

The object in a frame that moves frequently becomes the prominent one. This motion frequency method is used to mark points for the most frequent objects in motion in the video. The results are calculated by first aligning the masks of the object to share all possible edges and then collecting all the masks of the object to get the frequency map M : [37]

$$M = \frac{1}{n} \sum_{j} O_j \tag{2}$$

Here, *oj* represents an aligned object mask, and *ok* represents a key object mask indicated as a mask with a maximum similar frequency map. *Ok* is achieved by reducing the inaccuracies between the aligned and motion frequency map M : [37]:

$$O_k = \underset{O}{\arg\min}\, j\{(O_j - M)^2\} \tag{3}$$

An innovative technique for evaluating a video's key-frame based on object motions. Using optical flow, We extract the foreground objects from the video and compute the motion frequency score to seek the most reflective movements. The key-frame contains the most frequent motion.

3.7.1. Gray Scale Conversion

Grayscale images are distinguished from black and white one-bit bi-tonal images, which are images of just two colours in the sense of imaging: black and white (also called bi-level or binary images. Grayscale frames have several shades of grey. Grayscale images are the outcome of weighted combination of the light intensity of every pixel. In some situations, monochromatic single frequency grayscale images are used. In the- ory, the frequencies can be from all over the electromagnetic spectrum (e.g. infrared, visible light, ultraviolet, etc.).

3.7.2. Wavelet Decomposition

To de-noise two-dimensional signals wavelets are used. Owing to the high contrast between neigh- bouring pixel intensity values the bi-orthogonal wavelets are extensively used in image processing to trace and filter white Gaussian noise.

Inspired by the work of Ingrid Daubechies, [3, 4], a family of orthogonal wavelets that define a discreet wavelet that transforms and is marked by the support of a maximum frequency of vanishing moments. There is a scaling function (called the *fatherwavelet*) for each wavelet form producing an orthogonal multi-resolution analysis. The transform of the wavelet is frequently contrasted with the transform of the Fourier. Here the signals are interpreted as a sum of sinusoids. With the choice of the mother wavelet $\psi(t) = e^{(-2\pi)}$, the Fourier transform can be regarded as a unique instance of a continual wavelet transform. The eminent distinction is that wavelets are located both in time and in frequency, while the regular Fourier transform is found only in frequency. The Short-Time Fourier Transform (STFT) is identical to the transformation of the wavelet being both located in time and frequency, however there are challenges with the frequency/time resolution trade-off.

In particular, the STFT can be thought of as a transform with a marginally different kernel, claiming a rectangular window field. Where $\psi(t) = g(t\text{-}u)e^{-2\pi}$ can be given as $rect(\frac{(t-u)}{\Delta t})$ where Δt and u denote the window function's length and time offset, respectively. Using the theorem of Parseval, one may define the energy of the wavelet as in equation 4

From this, u is supplied from the square of the temporal support window offset by the time.

$$E = \int_{-\infty}^{\infty} |\psi(t)^2| dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\widehat{\psi(\omega)^2}| \quad d\omega \tag{4}$$

and the square of the spectral support of the window over frequency is given below in the equation 5

$$\sigma_\omega^2 = \frac{1}{2\pi E}\int |\omega - \zeta|^2 |\widehat{\psi(\omega)^2}| \quad d\omega \tag{5}$$

Multiplication of time domain and a rectangular window leads to convolution in the frequency domain with a feature that results in unnatural ringing artefacts for short/localized temporal windows. Again with Fourier Transform continuous-time, and the convolution in Fourier space is a delta function, culminating in the sinc signal ($\Delta t\omega$) real Fourier transformation. The window function may be a Δt →∞ filter, for instance a Gaussian filter. The window function collection influences the prediction error compared to the real Fourier transformation. The time-bandwidth of the resolution cell should not be exceeded by the Short-Time-Fourier- Transform (STFT). For all temporal changes or offsets, all the STFT base components retain consistent spectral and temporal support. This approach ensures an equal temporal resolution for lower and higher frequencies. The sampling width specifies the image's resolution.

The edge is a digital image's most prominent high-frequency data. The conventional filter efficiently reduces the noise, yet leaving the video clip fuzzy. We can preserve the edge of the video image by noise suppression of the video image. The wavelet method analyzes time-frequency, which adaptively selects the appropriate frequency band based on the signal characteristics. Then the frequency band matches the spectrum, which improves the time-frequency resolution.

3.7.3. SURF Feature Points Extraction

The use of SURF [38] for object recognition is proposed using three steps: extraction, description and matching. The first and fundamental step in pursuing this approach is the extraction of features. The attributes that will decide the images are wisely chosen. The *Surf Detect.m* function acts as an entry-point for feature extraction. The input here is an 8-bit RGB or grayscale image, and the area of interest becomes the output from the array. The below-mentioned functions contain the computation for GPU parallelization.

Convert32bitFPGray.m Conversion of an 8-bit RGB image to an 8-bit grayscale image is the re- sponsibility of this function. This stage will be skipped for the images already with the 8-bit grayscale format. Once completed, The 8-bit grayscale image is transformed to a 32-bit floating-point for computation on the GPU.

MyIntegralImage.m This function calculates the 32-bit floating-point grayscale image that was de- rived at the previous stage. For any specified rectangular region of the image, the integral image uses the sum of the pixels. This function also estimates the rate of convolutions performed in the next iteration .

FastHessian.m This function performs image convolution with box filters of various sizes and stores the measured responses.

The extraction of features helps to reduce the amount of resources needed to analyze a broad data set. When dealing with complex data, the variety of factors considered becomes high. To function with regional, similarity invariant representation, and comparison of images SURF method is fast and robust. The main interest of the SURF method lies in its estimation of operators. Using box filters and SURF allows to capture real-time applications such as monitoring and object recognition [38].

IV. RECURRENT NEURAL NETWORK (RNN)

RNN or recurrent neural network is an artificial neural network designed to handle the inputs in a network structure to remember past information. They connect the nodes in a graph using their internal memory and process the sequences of inputs forming a directed graph in having temporal behaviour.

In the case of discrete-time settings,By considering one vector at a time, RNN sequences the vectors at the first layer and computes them as a non-linear form of the weighted sum. The supervised target activations are provided for certain output units at a specific time. The back-propagation training algorithm is used along with the recurrent neural network to sequence the given data using time series. In RNN one input is executed at a given instance of time and predicts only one output for a specific time. RNN is particularly useful in predicting the next scenario by understanding the given set of data.

4.1 Attention Layers

The outputs of the encoder and decoder in this network is joined to the context vector. In figure 3 the blue symbols denote encoder, and red ones the decoders. Context vectors obtain the output of the cells as input and assess the probability distribution, determining which word the decoder would produce. We have several variants in the various studies, and the choice makes it different in terms of score and attention function.

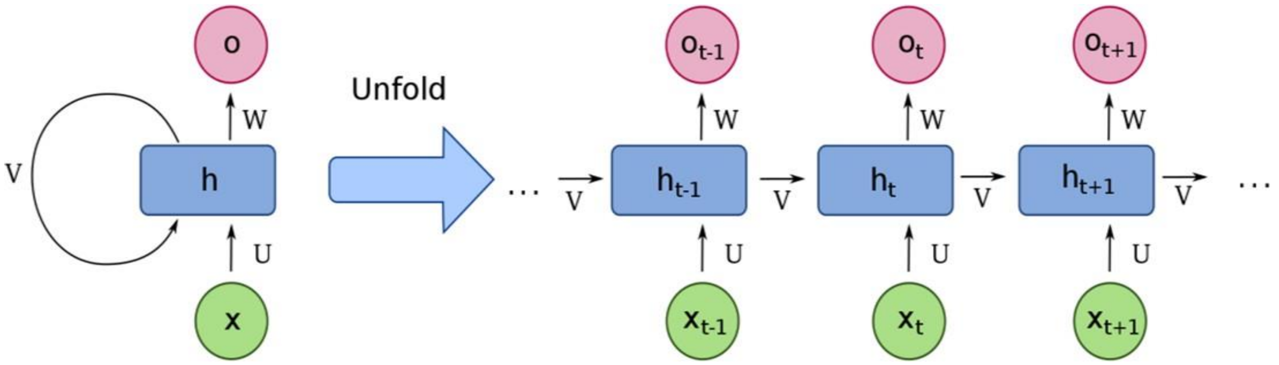


Figure 3. A basic representation of an unfolded recurrent neural network

Attention function can be of soft or hard attention. For the temporal attention process, the shift in the pixels of the adjacent frames is recorded as temporal changes and the temporal attention score is measured. You will find video summarization using ANN attention in the [7, 39] literature.

4.2 Feature Layers

Various features such as polygon lines, points, or regions Feature layers help in visualizing the base map details. Transparency, design, refresh interval, visible range and labels are fed to the input layers. Feature variable determines the lookup map. Feature layers help to analyze, view and edit and query against features and their attributes.

4.3 Appearance Layer

These attributes are effects, groups, layers, styles, objects with layers palette. When appearance layer applied to any object, it will hold that attribute. It is good to insert the layer of appearance to the object as a layer and not to the entire object.

V. VIDEO SUMMARIZATION ALGORITHM

The selected keyframe images in this segment take the requested information and output a summary of it. MVS(Multi-view video summarization) steps help to convert the video data into readable condensed infor- mation for a human to analyze large Videos. VSUMM method, a popular baseline method video summarization involves 1. Uniform Sampling, 2. Image histogram, 3. Scale Invariant Feature Transform

5.0.1. VSUMM

VSUMM [40] a basic and initial technique for video summarization as an unsupervised approach. To yield the clusters of the feature sets of each frame, the K-means algorithm is used.

5.1. SURF-RNN

RNN is a neural network wherein the input to the current layer originates from a consecutive output layer. In the conventional ANN network approach, output feedback is ignored. In this recommended RNN method, provided the previous term, RNN recollects the sequence of next probable words, RNN's memory network remembers the sequences.

The speeded-up robust features called SURF [38] are of the descriptor and a detector. The SURF feature is used primarily to recognize and record the object in the frame and to describe it using 3D reconstruction. We also have the SIFT scale -invariant descriptor along with this method. The traditional version of SURF is much more effective and quicker than all other image transformations. We presume that the frames form a cluster of' similarity classes' in this method. Two frames are then called identical only when they are of the identical group [41–43].

$$f_i \; and \;\; f_j \;\; Similar \Longleftarrow\!\!\Longrightarrow C(f_i) = C(f_i) \;\; (6)$$

Step 1: SURF Feature extraction is executed on the three different videos.

Step 2: To relate the involved object considering the key feature frames inferred from the three videos. For a collection of k classes C_1, C_2, ., C_k an optimum summary is obtained and holds on to the best summary solution Since the optimal summary is compute-intensive. Class selection involves caution, which contributes quickly to the right solution. The optimal summary can be reached by classifying all sets of k groups C_1, C_2,.C_k and the better class ones are retained. The optimal summary can be reached by classifying all sets of k groups C_1, C_2,.C_k and the better class is retained [41–43].

If a class C_m is being added to existing set C_1, C_2, ..., Cm-1, perhaps we can state the conditional as in equation 7

$$\begin{aligned} &Cov(C_m|C_1C_2...C_{(m-1)}) \\ &= Cov(C_1C_2...C_m) - Cov(C_1C_2...C_{(m-1)}) \\ &= Card\{i : \exists j, f_j \in E_i \;\; and \\ &C(f_i) = C_m \;\; and \;\; \forall f \in E_i, \\ &\forall r = 1, 2, ...m - 1, C(f) /= C_r\} \end{aligned} \tag{7}$$

Step 3:

We select each video v in turn and add to its current summary S_v, the one class C with maximal value.

$$value_v(C|\{S_v\}) = Cov_v(C|S) - \alpha \sum_{v\,v} Cov_v(C|S) \tag{8}$$

Here S represents the compilation of all classes that are already added in any of the compiled summary.

$$S = \bigcup_v S_v \tag{9}$$

If all the summaries reach the right size, we can iteratively substitute any chosen class only when we identify another better class [41–43].

Pseudocode :

```
Select the input video clipping
Frame conversion
Selection of high-level features such as the objects in motion
Select the keyframe k1,k2,k3 and Kn features
Train with key features, use a recurrent neural network,
   and moving objects are the source for the attenuation layer.
Video summarization algorithm is a
   process to give the ultimate summarized video output.
```

VI. RESULT AND DISCUSSION

The joint-video summarization method analyzes the three or more different videos and gives a sum- marized result. The individual video steps include the conversion of video files to individual frames. The figure 4, 5,6 shows the key video frame extraction from three videos. These steps are performed in MATLABR2014b version software. The choice representative frames is done using the moving sub shot summarization. To pick the optimal representative frame, an object-motion sub shot is partitioned into units. The threshold b_{err} param- eter computation is performed using successive frames representing the content diversity of motion prediction error. If the threshold of $berr$ is greater than the estimated value of Tb, then the unit boundary is picked from the recent picture. This is termed as "leaky bucket" algorithm [44–46]. Another threshold called T_f , is also mentioned here to eliminate successive selection in an instance of a highly active sub shot. In other terms, for each keyframe chosen, $KF_{i,k}(k = 0, ..., Mi)$, it satisfies the equation 10.

$$I(KF_{i,k}) - I(KF_{i,k-1}) \geq max(T_b, T_f) \tag{10}$$

Here $I(KF_{i,k})$ is the frame index of keyframe $KF_{i,k}$. For each keyframe, the timestamp and frame index are registered in the XML file [44–46].

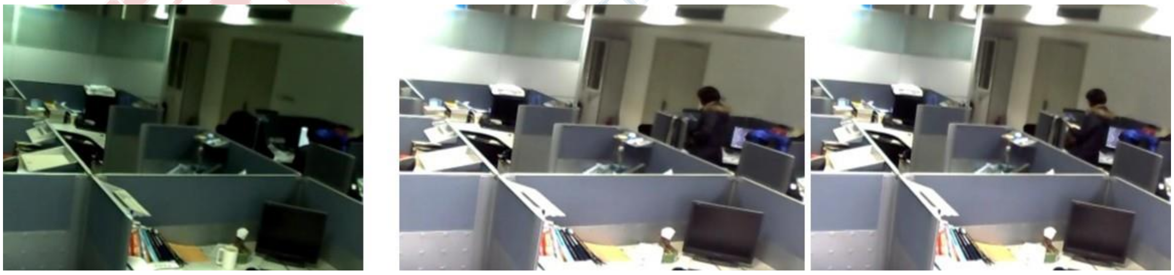


Figure 4. Representation of the Keyframes from Video 1.

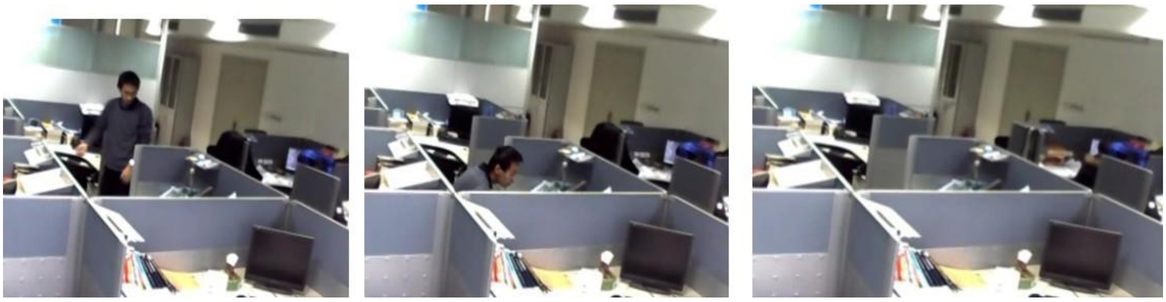


Figure 5. Representation of Keyframes from Video 2

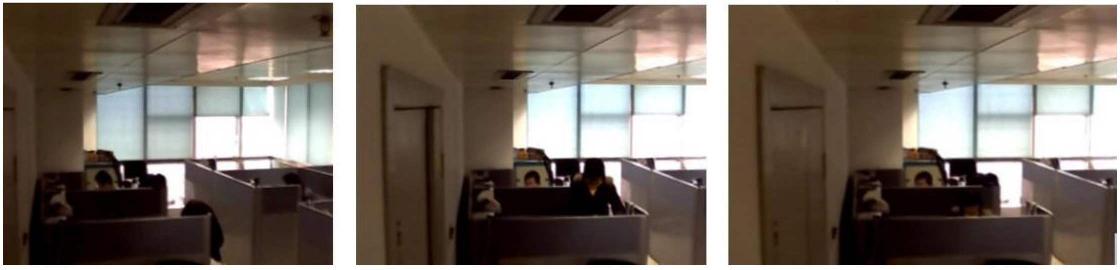


Figure 6. Representation of the Keyframes from Video 3.

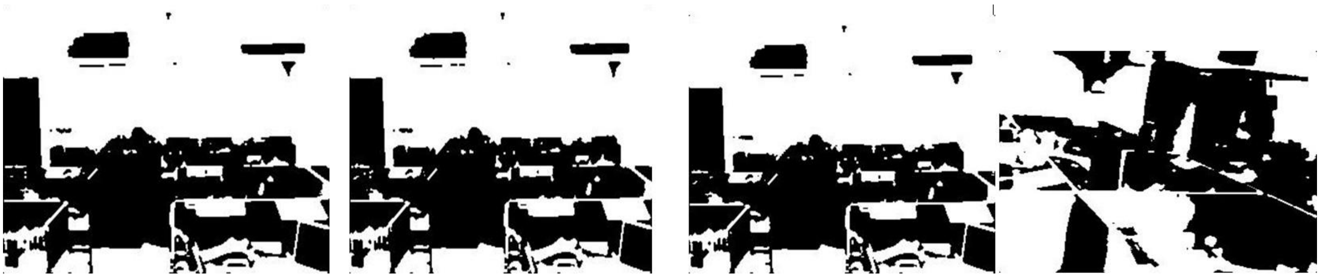


Figure 7. Resized grayscale frames in video 1

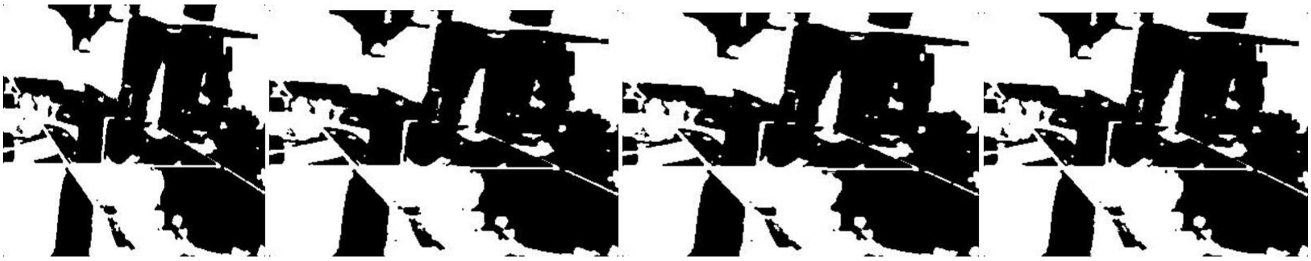


Figure 8. Resized grayscale frames in video 2

The figure 4,5,6 demonstrates the key feature frames from the three videos, and the object segmen- tation happens through the below process. The background and the person in the first video are tracked from frame 1. Similarly, the surroundings and the people are monitored in the second video and the third video.

6.1 Gray Scale Video Frames

In the figure 7, 8, 9 results, we see how the images are resized to $100X100$ and converted to a grayscale image to retrieve main frames.

6.2 Wavelet Decomposition Process

This approach helps to distinguish the objects that are contained in a given video frame. Now, the image is framed with horizontal, vertical, approximation and diagonal details for *LH*, *LL* as super, diagonal and decoded frames.

Figure 10 displays an input image decomposed into pixel processing that requires high-level segmen- tation and complex object recognition. This image is then processed to clear noise using a de-nosing algorithm, and other image reservations are created in this step. The wavelet coefficients of the specified images are higher than those of the scattered image, relying on the energy and small amplitude. The disturbing noises are eliminated, and the crucial information is recorded using an optimized threshold that processes the wavelet coefficients. The edge information is extracted and processed using the contrasts inside the video frame.

Figure 11, indicates the SURF features, SURF called as speeded up robust features [38] is a descriptor and detector, which performs tasks like image registration, object recognition, 3D reconstruction and classifi- cation. The figure 11 showcases the Object Recognition through SURF using three steps: feature generation, description and feature matching. The earliest step of the extraction function is the basic and crucial step in extracting useful information called features from the given input image. The features extracted must-have essential and unique attributes pertinent to the given image. The SURF calculation function considers images or input data with 8-bit RGB or grey-scale image. The output includes all points of interest.

The figure 12 depicts the summarized image of the three videos.

To analyze whether the ROC curve is in a perfect state, we need to evaluate the area under the curve called AUC and the confusion metrics. A confusion matrix is a table that describes the output of a classification model based on the test results. All values can be derived, except for AUC. Let's discuss the other evaluation parameters in detail.

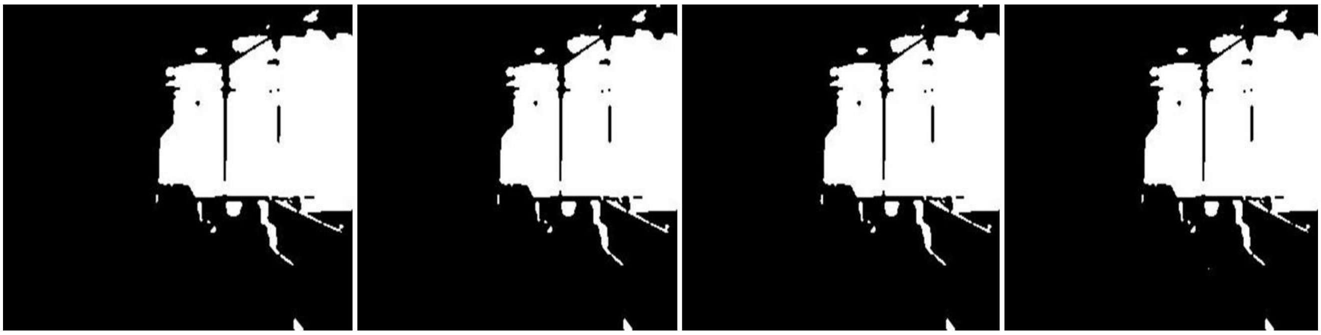


Figure 9. Resized grayscale frames in video 3

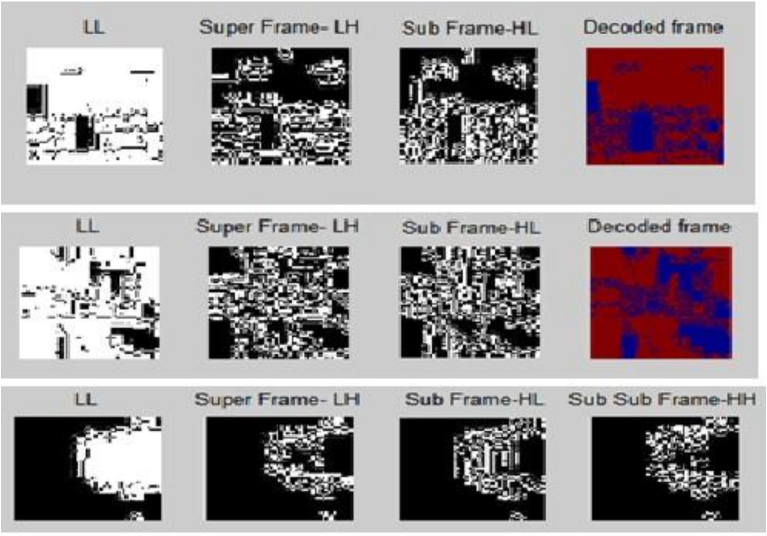


Figure 10. Wavelet Decomposition of the Video 1,2,3

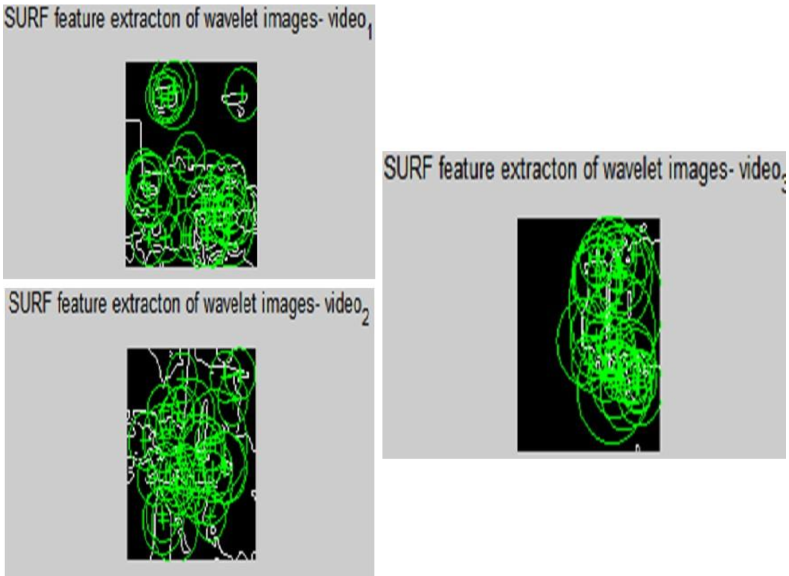

SURF feature extracton of wavelet images- video1
SURF feature extracton of wavelet images- video3
SURF feature extracton of wavelet images- video2

Figure 11. SURF feature frame for Video 1,2,3

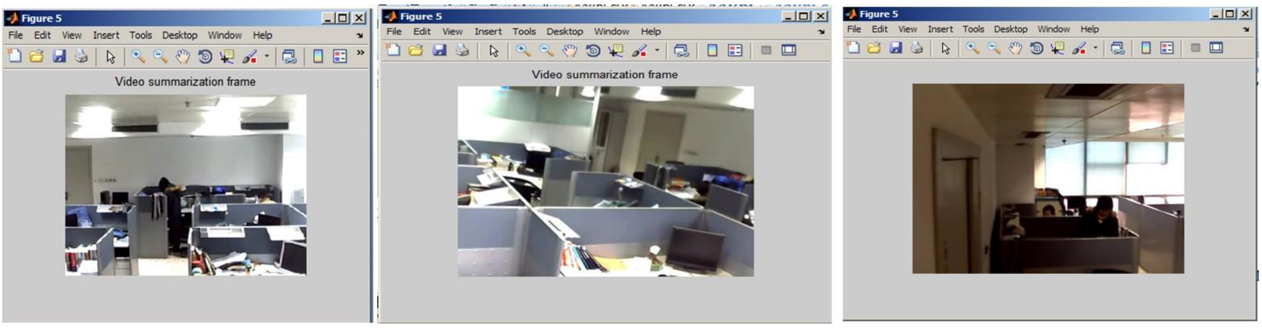
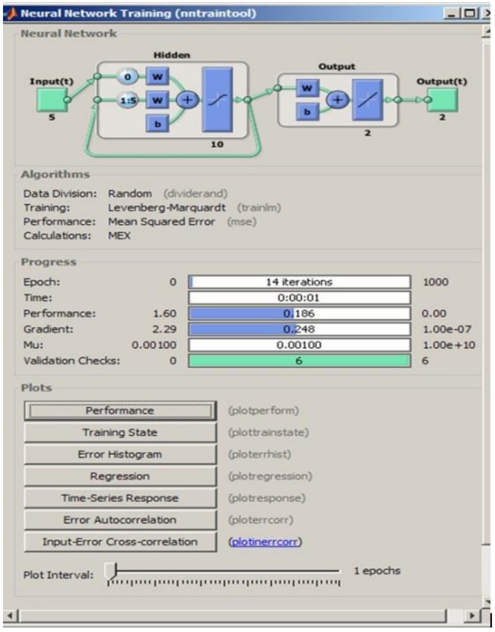


Figure 5
File Edit View Insert Tools Desktop Window Help
Video summarization frame
Figure 5
File Edit View Insert Tools Desktop Window Help
Video summarization frame
Figure 5
File Edit View Insert Tools Desktop Window Help

Figure 12. Multi-view Summarization frames



Algorithms

Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX

Progress

Epoch:	0	14 iterations	1000
Time:		0:00:01	
Performance:	1.60	0.186	0.00
Gradient:	2.29	0.248	1.00e-07
Mu:	0.00100	0.00100	1.00e+10
Validation Checks:	0	6	6

(plotperform)
(plottrainstate)
(ploterrhist)
(plotregression)
(plotresponse)
(ploterrcorr)
(plotinerrcorr)

Figure 13. The RNN neural network training screen shot using MATLAB.

Accuracy- This is the most significant parameter for predicting model efficiency using ratios that are rightly predicted.

Accuracy = TP + TN/TP + FP + FN + TN

Precision - This is the percentage of the cumulative positive observations predicted to the correctly predicted observations. When we witness a high precision, then it means a low false-positive rate.

Precision = TP/TP + FP

Recall- Recall also called sensitivity, a ratio to estimate the correctly predicted positive observations to the overall observations.

Recall = TP/TP + FN

F-score- A weighted average of recall and precision, and when this value reaches 1, it means the best score and 0 means the bad score. The relative contribution of recall and precision to $F\,1$ Score is equal and the formula is as below:

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

Figure 13 demonstrates the RNN neural network training for the performance. The layer consists of 1 : 10points.

Table 2 demonstrates a comparison analysis of the different video frames on the non-speed optimiza- tion of our code. The complex algorithm provides valuable information on the artefacts in the frame. he time spent on execution is calculated in conjunction with the complete automation of the summary process. For final post-processing, the desired frames with the highest likelihood in the video are considered. Office Dataset

[47] is the prominent and commonly used dataset in MVS research and used in simulations to validate the developed methods. The dataset used 4 asynchronous cameras in an office location. Each camera comes with varying lighting conditions, making it difficult to test this dataset using the MVS method leading to low ac- curacy. The summary derived here is taken from industrial surveillance, and no checking for truth is carried out. The participants of this survey are students in PhD and masters and are aware of the technology and the evaluation methods. Both local and cloud computing are identical, but the cloud implements GPU

computing, and the video mentioned here is Office-1 with a size of 235.8Mb with a time of 9 minutes and 4 seconds and 15fps.

The conclusions drawn by our method as shown in table 2 are in line with other reference works [48, 47], our high accuracy approach can be trained to perform video summarization for objects that need special attention. This research pays way to concentrate on target detection and tracking and tracking with the additional advantage of multi-view video summarization.

Table 2. Comparison of our evaluation scores with other references

Sl.No.	Video from Office [47]	Precision	Recall	Accuracy	F1 Score
1.	Office-1	0.9	0.93	98	0.91
2.	Office-2	0.95	0.98	97.9	0.96
3.	Office-3	0.93	0.89	95.6	0.90
4.	Refe [48]	0.93	0.86	93	0.90

VII. CONCLUSION

The developed system incorporates a multi-view video summarization technique that uses 3 videos of the target object and a wavelet of the target object using the SURF function extraction technique. The wavelet recognizes the person and objects in the screen. SURF takes care of feature extraction points in terms of the context object. Later, the derived information is processed in RNN networks to perform a better object classification for the person and other objects in the captured video frame. In this proposed method, heavyweight RNN model brings higher accuracy with the model. In the future, the system can be expanded with additional research to identify behaviour detection to provide immediate reporting on suspicious acts.

REFERENCES

[1]. J. Mas and G. Fernandez, "Video shot boundary detection based on color histogram." in TRECVID, 2003.

[2]. P. Porwik and A. Lisowska, "The haar-wavelet transform in digital image processing: its status and achievements," Machine graphics and vision, vol. 13, no. 1/2, pp. 79–98, 2004.

[3]. D. Popov, A. Gapochkin, and A. Nekrasov, "An algorithm of daubechies wavelet transform in the final field when processing speech signals," Electronics, vol. 7, no. 7, p. 120, 2018.

[4]. P. Lipinski and M. Yatsymirskyy, "Efficient 1d and 2d daubechies wavelet transforms with applica- tion to signal processing," in International Conference on Adaptive and Natural Computing Algorithms. Springer, 2007, pp. 391–398.

[5]. Z. Ji, Y. Zhang, Y. Pang, and X. Li, "Hypergraph dominant set based multi-video summarization," Signal Processing, vol. 148, pp. 114–123, 2018.

[6]. K. Zhang, K. Grauman, and F. Sha, "Retrospective encoders for video summarization," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 383–399.

[7]. Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder net- works," IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[8]. Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," IEEE Access, vol. 5, pp. 21 090–21 117, 2017.

[9]. M. Srinivas, M. M. Pai, and R. M. Pai, "An improved algorithm for video summarization–a rank based approach," Procedia Computer Science, vol. 89, pp. 812–819, 2016.

[10]. S. Lin, H. Ruimin, and Z. Rui, "Depth similarity enhanced image summarization algorithm for hole-filling in depth image-based rendering," China Communications, vol. 11, no. 11, pp. 60–68, 2014.

[11]. M. A. T. Jimenez, "Summarization of video from feature extraction method using image processing & artificial intelligence."

[12]. M. M. Salehin, M. Paul, and M. A. Kabir, "Video summarization using line segments, angles and conic parts," PloS one, vol. 12, no. 11, p. e0181636, 2017.

[13]. C. Huang and H. Wang, "A novel key-frames selection framework for comprehensive video summariza- tion," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577–589, 2019.

[14]. L. D'Amiano, D. Cozzolino, G. Poggi, and L. Verdoliva, "A patchmatch-based dense-field algorithm for video copy–move detection and localization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 3, pp. 669–682, 2018.

[15]. A. Eltantawy and M. S. Shehata, "An accelerated sequential pcp-based method for ground-moving objects detection from aerial videos," IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5991–6006, 2019.

[16]. A. ElTantawy and M. S. Shehata, "Krmaro: Aerial detection of small-size ground moving objects using kinematic regularization and matrix rank optimization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 6, pp. 1672–1686, 2018.

[17]. S. Javed, A. Mahmood, S. Al-Maadeed, T. Bouwmans, and S. K. Jung, "Moving object detection in complex scene using spatiotemporal structured-sparse rpca," IEEE Transactions on Image Processing, vol. 28, no. 2, pp. 1007–1022, 2018.

[18]. B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2513–2520.

[19]. Z. Liu, D. An, and X. Huang, "Moving target shadow detection and global background reconstruction for videosar based on single-frame imagery," IEEE Access, vol. 7, pp. 42 418–42 425, 2019.

[20]. T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Unsupervised online video object segmenta- tion with motion property understanding," IEEE Transactions on Image Processing, vol. 29, pp. 237–249, 2019.

[21]. K. He, G. Gkioxari, P. Dolla´r, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[22]. N. Dorudian, S. Lauria, and S. Swift, "Moving object

detection using adaptive blind update and rgb-d camera," IEEE Sensors Journal, vol. 19, no. 18, pp. 8191–8201, 2019.

[23]. Nonparametric background modelling and segmentation to detect micro air vehicles using rgb-d sensor," International Journal of Micro Air Vehicles, vol. 11, p. 1756829318822327, 2019.

[24]. J. Li, X. Liu, F. Liu, D. Xu, Q. Gu, and I. Ishii, "A hardware-oriented algorithm for ultra-high-speed object detection," IEEE Sensors Journal, vol. 19, no. 10, pp. 3818–3831, 2019.

[25]. C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," IEEE transactions on image processing, vol. 23, no. 7, pp. 2972–2982, 2014.

[26]. P. Gao, X. Sun, and W. Wang, "Moving object detection based on kirsch operator combined with optical flow," in 2010 International Conference on Image Analysis and Signal Processing. IEEE, 2010, pp. 620–624.

[27]. J. Lu, "An algorithm for edge extraction based on fast kirsch and the probability of being true edges,"

[28]. Computer Applications, vol. 21, pp. 33–35, 2001.

[29]. X. X. W. H. Y. Huisheng, "Application of an improved otsu method in the kirsch edge detection [j],"

[30]. Computer & Digital Engineering, vol. 3, 2009.

[31]. Wikipedia contributors, "Video frame resizing,image scaling," 15 October 2020, [Online; accessed 04-Nov-2020]. [Online]. Available: https://en.wikipedia.org/wiki/Image scaling

[32]. D. Sen, B. Raman et al., "Video skimming: taxonomy and comprehensive survey," ACM Computing Surveys (CSUR), vol. 52, no. 5, p. 106, 2019.

[33]. M. I. Sezan and R. L. Lagendijk, Motion analysis and image sequence processing. Springer Science & Business Media, 2012, vol. 220.

[34]. R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," IEEE transactions on image processing, vol. 5, no. 6, pp. 996–1011, 1996.

[35]. C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in CVPR 2011. IEEE, 2011, pp. 209–216.

[36]. Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212–3232, 2019.

[37]. L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietika¨inen, "Deep learning for generic object detection: A survey," International journal of computer vision, vol. 128, no. 2, pp. 261–318, 2020.

[38]. W. Shuigen, C. Zhen, and D. Hua, "Motion detection based on temporal difference method and optical flow field," in 2009 Second International Symposium on Electronic Commerce and Security, vol. 2. IEEE, 2009, pp. 85–88.

[39]. C. Jang and S. Lee, "Object motion based video key-frame extraction," in ACM SIGGRAPH ASIA 2010 Posters, 2010, pp. 1–1.

[40]. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," Computer vision and image understanding, vol. 110, no. 3, pp. 346–359, 2008.

[41]. J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," 2018.

[42]. S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Arau´jo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," Pattern Recognition Letters, vol. 32, no. 1, pp. 56–68, 2011.

[43]. I. Yahiaoui, B. Merialdo, and B. Huet, "Comparison of multiepisode video summarization algorithms,"

[44]. EURASIP Journal on Advances in Signal Processing, vol. 2003, no. 1, p. 613895, 2003.

[45]. "Automatic video summarization," in Proc. CBMIR Conf, 2001.

[46]. F. Chen, M. Cooper, and J. Adcock, "Video summarization preserving dynamic content," in Proceedings of the international workshop on TRECVID video summarization, 2007, pp. 40–44.

[47]. B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," ACM transactions on multimedia computing, communications, and applications (TOMM), vol. 3, no. 1, pp. 3–es, 2007.

[48]. T. Mei, L.-X. Tang, J. Tang, and X.-S. Hua, "Near-lossless semantic video summarization and its appli- cations to video analysis," ACM Transactions on Multimedia Computing, Communications, and Applica- tions (TOMM), vol. 9, no. 3, pp. 1–23, 2013.

[49]. C. Kim and J.-N. Hwang, "Object-based video abstraction for video surveillance systems," IEEE Trans- actions on Circuits and Systems for Video Technology, vol. 12, no. 12, pp. 1128–1138, 2002.

[50]. Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," IEEE Trans- actions on Multimedia, vol. 12, no. 7, pp. 717–729, 2010.

[51]. T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-assisted multiview video summarization using cnn and bidirectional lstm," IEEE Transactions on Industrial Infor- matics, vol. 16, no. 1, pp. 77–86, 2019.

BIOGRAPHIES OF AUTHORS

Vinsent Paramanantham Vinsent Paramanantham B.E., from Government College of Engineering; Tirunelveli, M.S. from BITS Pilani, has got around 15 years of industry and research experience, Cur- rently he is pursuing his M.Tech by research degree from Sathyabama Deemed University, Chennai. His core area of research includes Image processing and Video summarization. Other areas of interest include computer vision and Natural Language processing. An amalgamation of both Industry and academic keeps my research curiosity always high.

Dr. S. Suresh Kumar , Principal of Swarnandha College of Engineering ,Narasapur, Andhra Prad- hesh, India. His area of research is Smart Energy, Image Processing, Big Data, and Network Security. He has exemplary academic records. He completed Doctoral Degree (Ph.D) in the faculty of Infor- mation Communication Engineering from Anna University in 2009. And he obtained two Master Degrees in India from Premier institutions. M. Tech from Indian Institute of Technology, Kharagpur in 1999, M.S from Birla Institute of Technology and Science, Pilani, in 1993, and his bachelor de- gree(B.E.,) Computer Engineering from Madurai Kamaraj University in 1988.

He is a recognized Supervisor for Ph.D Programme in Anna University and Manonmaniam Sundara- nar University. Dr. Suresh Kumar has published 92 papers in various International and National Journals 61 papers in National and International Conference proceedings. He also published 6 books with co- authors. He has Thirthy one years of experience in the field of Engineering Education. He plays a vital role in many professional organizations. He is an active member in IET (UK) Insti- tution of Engineering Technology, UK. He was a Chairman for IET Chennai Local Networks during 2014-2016 , India. Also He was a elected council member of IET(UK) for the period 2013-2016. He is a fellow member in IE(I) and IETE. Also He is life member in CSI and ISTE , Senior member in IACSIT and member in IAENG,ICSES and ACEEE.