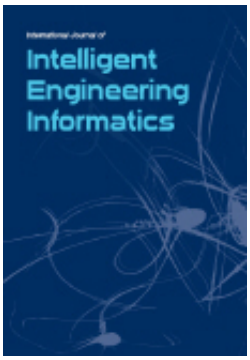


International Journal of Intelligent Engineering Informatics

ISSN online: 1758-8723 - ISSN print: 1758-8715
https://www.inderscience.com/ijiei

Dynamic video summarisation using stacked encoder-decoder architecture with residual learning network

M. Dhanushree, R. Priya, P. Aruna, R. Bhavani

DOI: 10.1504/IJIEI.2024.10062166

Article History:

Received:	02 September 2023
Last revised:	22 December 2023
Accepted:	29 December 2023
Published online:	02 April 2024

Dynamic video summarisation using stacked encoder-decoder architecture with residual learning network

M. Dhanushree*, R. Priya, P. Aruna and R. Bhavani

Department of Computer Science and Engineering,
Faculty of Engineering and Technology,
Annamalai University,
Annamalai Nagar, Tamil Nadu, India
Email: dhanushree269@gmail.com
Email: pryknd n@gmail.com
Email: arunapuvi95@gmail.com
Email: bhavaniaucse@gmail.com
*Corresponding author

Abstract: In the past decade, video summarisation has emerged as one of the most challenging research fields in video understanding. Video summarisation is abstracting an original video by extracting the most informative parts or key events. In particular, generic video summarisation is challenging as the key events do not contain specific activities. In such circumstances, extensive spatial features are needed to identify video events. Thus, a stacked encoder-decoder architecture with a residual learning network (SERNet) model is proposed for generating dynamic summaries of generic videos. GoogleNet characteristics are extracted for each frame in the proposed model. After the bi-directional gated recurrent unit encodes video features, the gated recurrent unit decodes them. Both the encoder and decoder architectures leverage residual learning to extract hierarchical dense spatial features to increase video summarisation F-scores. SumMe and TVSum are used for experiments. Experimental results demonstrate that the suggested SERNet model has an F-score of 55.6 and 64.23 for SumMe and TVSum. Comparing the proposed SERNet model against state-of-the-art approaches indicates its robustness.

Keywords: video abstraction; dynamic video summarisation; deep learning; residual learning; skip connections; GoogleNet; long-term memory; gated recurrent unit; stacked encoder; key shot selection; kernel temporal segmentation.

Reference to this paper should be made as follows: Dhanushree, M., Priya, R., Aruna, P. and Bhavani, R. (2024) 'Dynamic video summarisation using stacked encoder-decoder architecture with residual learning network', *Int. J. Intelligent Engineering Informatics*, Vol. 12, No. 1, pp.27–59.

Biographical notes: M. Dhanushree is a research scholar currently doing her Research Work in the Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India. Her major research interest includes video summarisation and apart from that she also has knowledge and interest in the emerging research areas such as image processing, computer vision, visual question answering and video understanding. She has published two papers in international journals and one paper in international conference.

R. Priya is a Professor in the Department of Computer Science and Engineering at Annamalai University. She has 24 years of teaching experience. She has published papers in 38 international and one national journal and ten national and 18 international conferences.

P. Aruna is a Professor in the Department of Computer Science and Engineering at Annamalai University. She has 32 years of teaching experience. She has published papers in 89 international and seven national journals and 25 national and 35 international conferences.

R. Bhavani is working as a Professor and the Head of the Department of Computer Science and Engineering at Annamalai University. She has 29 years of teaching experience. She has published papers in 58 international journals, 19 national and 57 international conferences and one book.

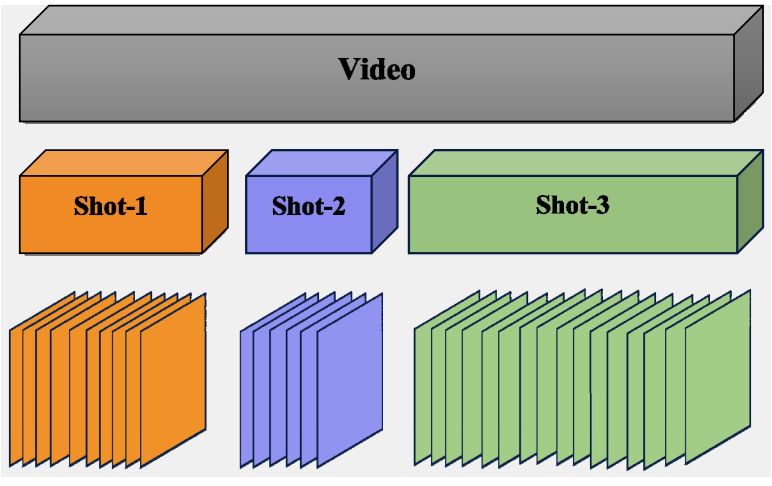
1 Introduction

In the ever-growing digital age world, multimedia content is increasing exponentially. Ever since the arrival of the internet, the world has shrunk in communication. While the internet reduces the communication gap worldwide, social media platforms allow the comfortable sharing of multimedia content such as images, videos, audio, text files, etc. Video capture and consumption have increased due to numerous devices, including smartphones, smart watches, portable cameras, and other wearable devices (Anand et al., 2023). YouTube, one of the most popular social media platforms, is witnessing more than 500 hours of video uploads every minute. Similarly, other social media platforms like Instagram and Facebook generate over a billion views on videos per day (Angeline et al., 2023). From the authenticated data, it is noticeable that people tend to upload and watch videos quite often. The exponential growth of video data increased the need for video summarisation, which helps in easy browsing, viewing, and storing. Video summarisation helps generate a shortened version of the original video without compromising its essence (Bose et al., 2023). It is also defined as generating a condensed version of the original video without compromising the underlying context for which the video is generated. Based on the output generated, video summarisation is classified into two types: static video summarisation (also called storyboard) and dynamic video summarisation. Storyboarding corresponds to the technique of finding keyframes, while dynamic video summarisation produces the key shots of the original video (Cirillo et al., 2023). A shot is a semantically similar sequence of frames. As with any other video analysis task, this also requires labelling to employ the full potential of deep architectures (Devi and Rajasekaran, 2023). The annotation for video summarisation is the importance score the annotators or humans give for a particular video. These importance scores may be for each frame in a video or for a particular segment (Gaayathri et al., 2023). Binary annotations are also used, wherein one keyframe or segment is labelled as one and the other as 0.

1.1 Types of videos

A video may be structured or unstructured based on the arrangement of frames. Structured videos contain a sequence of shots where each shot has a group of semantically meaningful frames. A pictorial representation of structured video is shown in Figure 1, in which a single video has three shots, each with a different colour, indicating semantically different shots (Joshi et al., 2023). Each shot contains a series of similar frames indicating a particular event (Kanyimama, 2023). Moreover, there are transition frames between the shots (Jeba et al., 2023). Using those transition frames, different shots are detected. Such videos include edited news, movies, interviews, and documentaries. Unstructured videos are those with no proper structure (Regin et al., 2023). These are unedited user-generated videos that are mostly blurry and shaky (Lodha et al., 2023). They do not have any transition between the shots and are continuous. Examples include those videos captured by people using digital cameras, smartphones, and egocentric videos.

Figure 1 Pictorial representation of a structured video (see online version for colours)



1.2 Motivation

Video summarisation aims to find shots worth watching or containing desired events, like sports, interviews, etc. Internet users find it difficult to watch a long video and prefer to be short. With the advent of short videos like reels and YouTube shots, it is evident that people do not want to spend too much time knowing the entire video content. A video is likelier to be enjoyable and watchable if it contains all the necessary context within a specified period, generally within two minutes. These paved the way for a new video analytics field called video summarisation.

1.3 Different techniques of video summarisation

Over the past two decades, video summarisation has emerged as a powerful and challenging field in video processing. Initial works of video summarisation include shot

boundary detection as a first step. Such works aim to find the transitions or cuts between the shots (Medentzidou and Kotropoulos, 2015). Handcrafted features such as motion, image entropy, and image differences are used to find the transition between the shots. Once the shots are segmented, desired frames or sequences are selected as keyframes or events. Then, dictionary learning-based approaches are used widely (Sivapriya et al., 2023). This method codes the features as dictionary codes, using which the summary is generated (Mademlis et al., 2018). Mathematical models like graphs and trees are other important approaches that proceed by converting a video into graphs. Different types of summaries are produced using the mathematical properties of graphs (Dos Santos Belo et al., 2016). Like other image and video processing tasks, video summarisation researchers also use its benefits as machine learning blooms. One such common technique is the use of clustering. After extracting the desired features, the clustering technique is used to aggregate those features. Each cluster represents a semantically similar frame, approximately representing a shot of the input video. Then, one frame from each cluster is chosen as the keyframe (Dhanushree et al., 2023). Supervised techniques are also used for video summarisation, in which a classifier is used to classify each frame as important (Rochan et al., 2018).

The advent of deep learning has positively impacted the efficiency of the video summarisation task. The major success of the deep learning technique is due to the learning ability of the deep models to learn and approximate any type of nonlinear function in the feature space. Some examples where deep learning is applied are medical image processing (Manju et al., 2022), video understanding, satellite image processing, cyber security (Arivukarasi et al., 2023; Mahasree et al., 2022), etc. Medical image processing includes the detection of breast cancer (Reshan et al., 2023), haematological disorders (Alshahrani et al., 2023), colon cancer (Azar et al., 2023c), diagnosis of gastro-intestinal diseases (Fati et al., 2022), parasite egg detection (Ray et al., 2021), paraquat-poisoned patient detection (Zhao et al., 2021), and corona disease detection (Waleed et al., 2022).

All these and many more diseases detection and classification are possible because of the deep learning models. Deep models also have applications in intelligent video processing, like intrusion detection systems (Azar et al., 2023b), robot navigation (Azar et al., 2023a), recommendation systems (Dudekula et al., 2023), and helping visually impaired persons (Ganesan et al., 2022). Agricultural applications such as apple disease detection (Sulaiman et al., 2023), leaf disease detection, etc., also exploit the deep architectures. Not only image and video processing, but deep learning also performs well in another modality called audio and signal processing, such as bird sound classification (Boulmaiz et al., 2022), animal sound detection, and financial time series prediction (Nayak, 2021). Other applications in the field of automation and control include harvesting solar energy (Bhat, 2021), data-driven learning control (Dai et al., 2021), tracking the angular velocity of motors (Govindharaj and Mariappan, 2020), and trajectory tracking in an autonomous quadrotor helicopter system (Bellahcene et al., 2021). Owing to the vast problem-solving capability of deep architecture, this research work uses deep learning techniques to solve the problem of video summarisation.

1.4 Recent advancements in video summarisation

One of the recent advancements in video summarisation is multi-view video summarisation. For security purposes, sometimes more than one camera is installed in the

workplace, educational institutions, etc. Summarising such videos with multiple views is called multi-view video summarisation (Hussain et al., 2021). Various events are extracted from different views, and the correlation between the events of different views is identified and arranged temporally to generate a combined video summary (Singh et al., 2023). Other advancement is the summarisation of multi-stream videos, i.e., videos from different moving cameras that share their views occasionally (Elfeki et al., 2022). It differs from multi-view video summarisation in that the former videos are surveillance videos with a static background, while the latter have a moving background (Nagaraj and Subhashni, 2023). The summarisation of traffic surveillance videos is one of the vital real-time applications. By summarising the traffic videos, video segments containing the accident events or traffic rule violators are obtained, which helps in an intelligent traffic management system. Music video summarisation aims to summarise the musical videos, which predominantly use audio information for producing the summary. Besides audio and video features, textual features such as captions, video descriptions, and other metadata are utilised for video summarisation (Padmanabhan et al., 2023). Query-focused or user-desired personalised video summary generation is the current research topic. Summarising the videos requires understanding and retrieving the content, which ultimately contributes to video understanding. Semantic video understanding is the basis for futuristic technologies such as robot navigation and control. Visual video question answering is another emerging field similar to query-focused video summarisation. Whatever the emerging field of video processing is, video summarisation plays a vital role as a basic task.

1.5 Applications of video summarisation

The applications of dynamic video summarisation depend on the type of video that is being summarised. In the case of surveillance videos, long hours of video are summarised into desired events such as abnormal events (Kalaivani and Roomi, 2017; Mayya and Nayak, 2019; Tahir et al., 2023), people detection (Lai et al., 2016), etc. In the case of news videos, the primary task is to extract the highlights or headlines of a particular event from the video (Liang et al., 2021). For sports videos such as cricket, football, and tennis match videos, video summarisation aims to find the highlights by recognising the players' activities and/or detecting the object of interest (Rafiq et al., 2020). This greatly helps to reduce the manpower and time required for sports highlight generation. Others, such as the dynamic summary of medical videos, help medical students and professionals view shorter versions of the demo videos (Liu et al., 2020a). The summarisation of egocentric videos is of great help to people suffering from a medical condition called dementia (Ghozali et al., 2023). By summarising the 24-hour personal video, they are aware of their daily activities, thus helping them to be independent and lead their lives peacefully (Ghafoor et al., 2018; Sreeja and Kovoor, 2021).

Nowadays, online education platforms have grown enormously, and more lecture videos are available online to benefit students who wish to learn remotely. Even online degree programmes, which contain several hours of lecture videos, are available now. Summarising these lecture videos (Kota et al., 2021) eliminates the need to watch the complete video during revision (Obaid et al., 2023). Some people wish to learn only a particular topic from an hour-long lecture. In such cases, lecture video summarisation is a boon (Rajest et al., 2023). Video summarisation is not just limited to the mentioned

applications; it currently extends its branch to multi-modal video summarisation, which incorporates audio features, visual features, and video descriptions (Wei et al., 2018) for summary generation. Yet another example of multi-modal video summarisation is query-focused video summarisation (Xiao et al., 2020), which generates the output based on the user's natural language query (Sirajudheen et al., 2023). This query-based video summarisation helps generate personalised output according to the user's needs. It also has a real-time application for summarising the live streaming video, known as online video summarisation (Lal et al., 2019).

1.6 Problem statement

Video Summarisation is the process of finding events that are considered to be important from a lengthy video. Temporal dependencies of events in a video are important while finding those key events. Those temporal dependencies are handled using recurrent neural networks in the existing literature. Every activity and scenery is highlighted for a generic video summarisation (Ghozali et al., 2022). In such cases, spatial features are vital in identifying general events and activities. However, the literature does not properly handle the extraction of dense spatial features. Thus, in the proposed SERNet model, skip connections are used to extract hierarchical and dense spatial features both in the encoder and decoder to improve the efficiency of generic video summarisation.

The key objectives of this research work are:

- To develop a deep learning model for dynamic video summarisation of generic videos. The datasets for the experiment are taken in such a way that they contain different categories of videos, not just limiting them to a particular domain. Thus, we aim to provide a generic framework that works better for any type of video.
- Skip connections across the layers of bi-directional GRU to extract dense and hierarchical spatial features. Thus, spatio-temporal features are used to summarise the videos containing different categories.

2 Literature survey

Dynamic video summarisation is viewed as a sequence-to-sequence problem and is solved using a recurrent neural network. The input is a series of frames, and the output is a series of probability scores for each frame, showing the likelihood that it will be regarded as a keyframe. This section discusses the various literature that use recurrent neural networks for solving sequential problems, as follows:

Ji et al. (2019) proposed a summarisation model in which the bidirectional long short-term memory (LSTM) is used as an encoder while the decoder is LSTM with an attention mechanism. They proposed two types of attention mechanisms for computing attention scores based on addition and multiplication. They also demonstrated various comparative analyses, like results with and without attention and results with shallow and deep features. In the work proposed by Fajtl et al. (2019), a self-attention-based network is proposed that, during training, completes the whole sequence-to-sequence transformation at stretch. They used a regression network to find the importance score for each frame. This eliminates the need for computations in both directions, and they also claim that their proposed architecture does not suffer from any long-sequence

information loss. The authors exploit videos' hierarchical nature (Zhao et al., 2017), where two layers of LSTM are used, with the first being unidirectional and the second being bidirectional, producing a frame-level importance score. Previous works used LSTM, which lacks efficient summarisation of lengthy videos. Thus, they have utilised the hierarchical RNN for long-range video summarisation.

In the work proposed by Liu et al. (2020b), a new method for summarising videos uses several concepts and various contexts over temporal and spatial directions. They constructed a model that can learn from data with and without labels. This particularly makes it suitable for large-scale, unlabeled datasets. In a video, the architecture identifies concepts such as humans, vehicles, and food across temporal directions. These concepts are then used to summarise the video automatically.

The work proposed by Ghauri et al. (2021) is based on a multiple-feature set with parallel attention. Three types of features, namely image features, RGB motion features, and motion flow features, are extracted. These features are processed parallelly with the attention mechanism, concatenated at the end. They also presented a new evaluation metric for video summarisation. Feng et al. (2020) proposed a self-attention-based encoder-decoder architecture in which bi-directional GRU is used as an encoder, and GRU is used as a decoder and a regression network. A comparative analysis report was also discussed by comparing the LSTM and GRU for the encoder and the decoder. Zhang et al. (2018) proposed a retrospective encoder for video summarisation. Initially, hierarchical encoders extract frame- and shot-level features separately. These encoded features are then decoded using a decoder, whose output is given as input to the retrospective encoder. The predicted summary is embedded by the retrospective encoder into an abstract semantic space, and the embeddings of the actual video are compared.

Ji et al. (2020) proposed a deep learning model that primarily focuses on reducing the semantic gap between the original and shortened video. They used a semantic preserving loss in the deep seq2seq network. The major focus of this work is to reduce the semantic information loss between the original video and its corresponding summary. They introduced the use of Huber loss to effectively decrease the effect of outliers. Han et al. (2017) summarised unstructured videos. The input video is divided into equal-sized segments and summarised. Then, they reconstructed the video to compute the reconstruction error for predicting the interestingness factor and representativeness factor of each video segment in a video. Yuan et al. (2017) proposed a semantic embedding deep learning model that uses side information such as titles, comments, queries, descriptions, and visual features. Using the encoders, they measure the semantic relevance between the video and the side information. They used two types of losses for computing the relevance of semantic information; one is a feature reconstruction loss.

The visual and textual features are combined, and their correlation is learned during latent subspace learning. This is a new attempt by the authors, as it involves processing multiple modalities of data from multiple sources. The research work proposed by Yuan et al. (2019) uses a three-dimensional convolutional neural network to jointly obtain the spatio-toral features. These features are fused and given as input to a recurrent neural network for encoding spatial and temporal features. Finally, a multi-layer perceptron finds each frame's importance score. They defined a new loss function called Sobolev loss, a combination of the l2 norm and its derivatives, for their architecture to produce a dynamic video summary. Chen et al. (2022) proposed a transformer-based video summariser consisting of three components. The first is the embedding component, the

transformer component, and finally, the prediction block. In the transformer block, several layers of transformers are stacked to produce a U-shaped transformer to combine hierarchical video representations. In the work proposed by Nair and Mohan (2022), a multi-CNN model is used to split the input videos into meaningful shots. From these shots, features using several pre-trained models are obtained.

Ultimately, these features are concatenated, and the combined features undergo dimensionality reduction to ease the computational complexity. A score is computed for each shot using the Chebyshev distance measure, wherein similar shots have a shorter distance and different shots have a longer distance. One shot is selected among the most similar shots to generate the dynamic video summary (Teng et al., 2022). A new model that combines spatio-temporal feature extraction and a cross-attention scheme was proposed using comparative learning. Four types of features, namely depth, breadth, local, and global, are extracted using deep architectures. These features are combined and undergo cross-attention to produce the desired summary.

Terbouche et al. (2023) proposed a method that uses multiple annotations for each video to train an attention-based recurrent neural network. Frame features are extracted using vision transformers. They utilised an expectation-maximisation algorithm for optimising the proposed attention-based model. Zhang et al. (2023a) introduced an integrated network combining the convolutional long- and short-term memory network and multi-scale transformer model. They employed contrastive learning in the spatio-temporal attention models for better representation of videos. The video under study is X-ray coronary angiography, which is useful in studying the diagnosis and treatment of cardiovascular diseases. In the work proposed by Psallidas and Spyrou (2023), the authors used a fusion of audio and video data to generate a meaningful video summary. Apart from benchmark datasets, they used their custom user-generated datasets to validate their proposed model. The segment-level feature fusion technique is employed in this work.

A multi-modal approach to video summarisation is used by Huang et al. (2023), where non-visual features such as interestingness and storyline consistency are utilised. A conditional attention module is used, which helps find the video's key components. Hsu et al. (2023) proposed a novel spatiotemporal vision transformer that helps obtain inter-frame and intra-frame attention. Both spatial and temporal attention heads are used for a better video summary during encoding and decoding. In this work (Khan et al., 2024), the authors introduced the transformer-based deep pyramidal refinement network that extracts multi-scale deep features and predicts an importance score for each frame. A progressive feature fusion technique is used for predicting the keyframes. Next, a self-supervised adversarial video summariser is used by Xu et al. (2023) for producing semantically meaningful video summaries. The authors proposed two new modules, namely clip consistency representation and hybrid feature refinement, to ensure the video clips are continuous and visually appealing.

In Su et al. (2023), a novel recurrent unit augmented memory network is proposed for generating a summary of long videos. This end-to-end memory network comprises an input module, a local and global sampling module, a memory module, and an output module. The memory module helps get the long-term memory information, which helps in efficient dynamic video summarisation. The work by Sreeja and Kovoor (2023) uses a generative adversarial network for modelling the video summarisation problem. The adversarial learning extracts diverse and good features from the video. The generator-discriminator model then utilises these to produce the summary of the video

(Suman et al., 2023). A convolutional recurrent auto-encoder is used as a generator with reconstruction loss. A discriminator tries to differentiate between the original video and the reconstructed video, thereby creating an appropriate condensed version of the original video. The key segments are selected using the knowledge distillation technique. In Zhang et al. (2023b), the researchers proposed a motion-assisted reconstruction network that extracts spatial and motion information to summarise video unsupervised. Their motion-assisted network includes a bidirectional modality encoder and a video context navigator, which ensures semantic consistency among the multi-modal features. A consistency loss term is used to remove the noise impact of the motion features.

The major challenge of existing literature is that it does not pay much attention to the spatial features. Since the recurrent neural networks handle the temporal features, the spatial features remain unattended. Dense spatial features are vital in identifying key events in a video. This work considers spatial and temporal features for a better generic video summarisation model. Thus, a deep encoder-decoder-based architecture with residual learning is proposed to improve the task of generic video summarisation.

This research uses skip connections in the encoder-decoder architecture for extracting dense spatiotemporal features, which helps identify the key events in a video. The encoder-decoder architecture is the backbone of the proposed SERNet model due to its capability of handling variable length input and output.

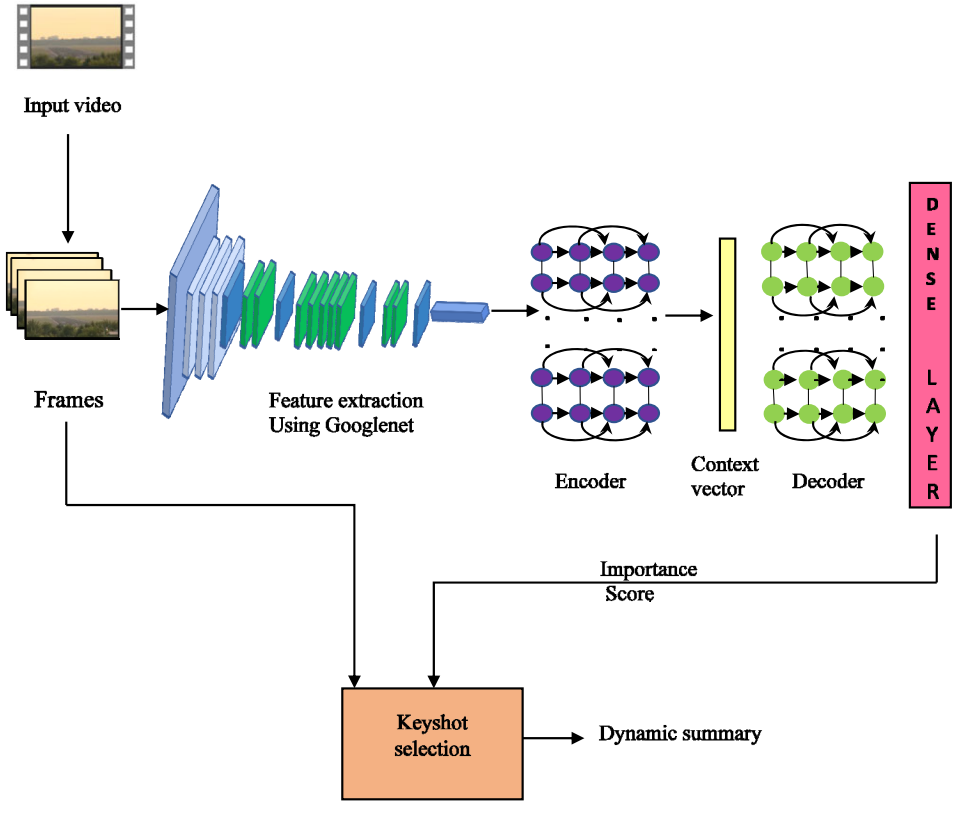
3 Proposed methodology

Figure 2 shows the block diagram of dynamic video summarisation using the proposed SERNet model. The input video is converted into a short, condensed version of the original video without compromising the important content. As the first step in the proposed work, the input video is converted into frames by a uniform sampling method, where every sixth frame is taken for further processing. Then, the visual features are extracted from the pre-trained GoogLeNet architecture. These visual features are passed through the encoder, which encodes the spatial and temporal features of the input frames. The encoded features thus obtained are sent to a decoder, which decodes the encoded features to produce an importance score for each frame. The key shots are selected from the frame level importance scores to generate a dynamic video summary.

3.1 Feature extraction

GoogLeNet features are the most commonly used feature extractor for the video summarisation task. Pre-trained weights on the ImageNet dataset are used, and the last two layers are truncated to extract a feature vector of size 1 × 1,024 from the average pooling layer. Since the GoogLeNet takes an input image of size 224 × 224, the input frames are resized to 224 × 224, which goes through a series of modules containing a convolutional layer, pooling layer and the inception layer to produce the final feature vector. Thus, a feature vector of size 1 × 1,024 is extracted from each sampled frame. These feature vectors are extracted for each frame as $FV = \{fv_1, fv_2, \ldots, fv_m\}$.

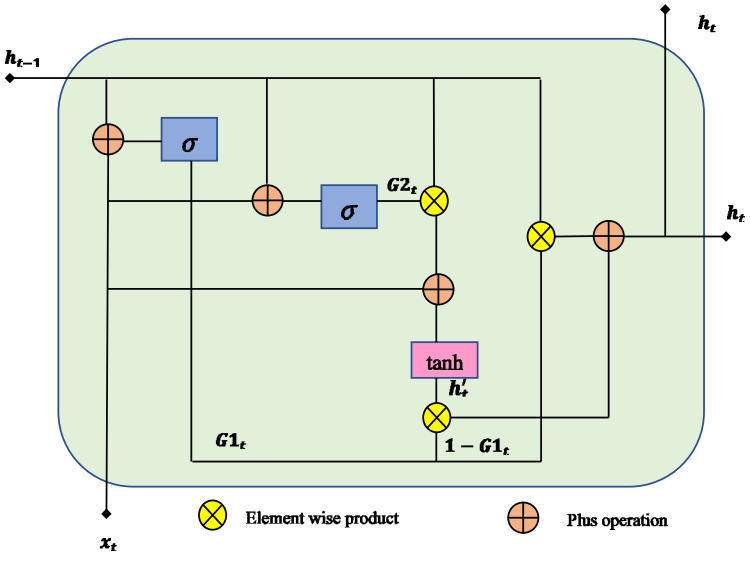
Figure 2 Block diagram of the proposed SERNet model (see online version for colours)



3.2 Stacked encoder architecture with residual learning

This paper modifies the video summarisation task as a sequence-to-sequence problem in which input and output are sequential data of variable length. Here, the input video and the output summary are sequential data consisting of a series of frames in temporal order. Consider an input video frame $F = \{f_1, f_2, \ldots, f_n\}$, where n is the length of the original video, which is converted into key shots $K = \{ks_1, ks_2, \ldots, ks_m\}$ where m is the number of key shots generated and $ks_m = \{f_1, f_2, \ldots, f_k\}$ where k is the total number of frames in shot ks_m. A video is a sequence of shots where a shot represents a sequence of frames representing a particular object or event in the video. Thus, encoder-decoder architecture is employed to find the dynamic video summary due to its efficiency in solving various sequence-to-sequence problems. The bidirectional gated recurrent unit (BiGRU) is an encoder to encode the video frames' semantic information from past and future frames. By this, the relation between frames in the temporal direction is well exploited. The objective of BiGRU is to extend the GRU cells in two directions. One is in a forward direction utilising the past information (forward encoded states), and the other is in a backward direction utilising the future information (backwards encoded states). The structure of a single GRU cell is shown in Figure 3.

Figure 3 Structure of GRU cell (see online version for colours)



There are two gates, namely update gate $G1$ and reset gate $G2$. The update gate takes input. x_t and previous hidden state h_{t-1} which are multiplied by their corresponding weights and added together, followed by a sigmoid operation to produce an output. The update gate determines how much past information should be considered for further processing.

$$G1_t = \sigma\left(W^{(u)}x_t + U^{(u)}h_{t-1}\right) \tag{1}$$

Next comes the reset gate, which determines what information needs to be removed from memory for an efficient computation.

$$G2_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right) \tag{2}$$

$$h'_t = \tanh\left(W^{(r)}x_t + G2_t \odot U^{(r)}h_{t-1}\right) \tag{3}$$

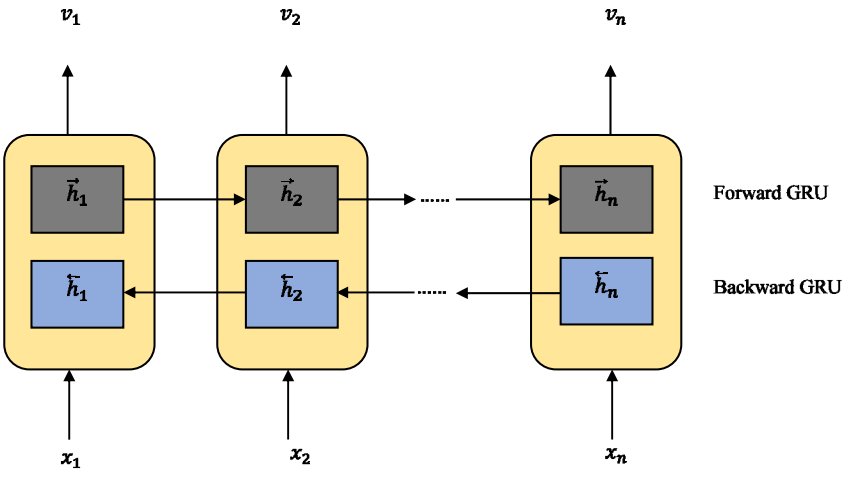
where W and U are the weights of input and the previous hidden state.

The final hidden state is calculated using the following equation:

$$h_t = G1_t \odot h_{t-1} + (1 - G1_t) \odot h'_t \tag{4}$$

The architecture of BiGRU is shown in Figure 4. From the input sequence, the forward states calculate the forward hidden states $(\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_n)$ by reading the input in the forward direction, i.e., from x_1 to x_n. Similarly, the backward states calculate the backward hidden states $(\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_n)$ by reading the input in the reverse direction, i.e., from x_n to x_1. The output vector is called the context vector (v_1, v_2, …, v_n) where each value v_t. It consists of information from both the past frames and the future frames. It is formed by the concatenation of both forward and backwards hidden states.

Figure 4 Architecture of BiGRU (see online version for colours)



Previous research work (Bengio, 2009) shows that a deep hierarchical model performs efficiently over shallow models. This motivated me to develop an efficient stacked BiGRU network for encoding the visual features of the frame. The stacked BiGRU in this proposed work consists of four layers where the output from the lower layers is fed as input to the higher layers. Each layer of BiGRU contains 128 unit cells. The working of the encoder is given as a series of equations as follows:

$$h_1 = encoder_1(FV) \tag{5}$$

$$h_2 = encoder_2(h_1) \tag{6}$$

$$h_3 = encoder_3(h_2) \tag{7}$$

$$V = encoder_4(h_3) \tag{8}$$

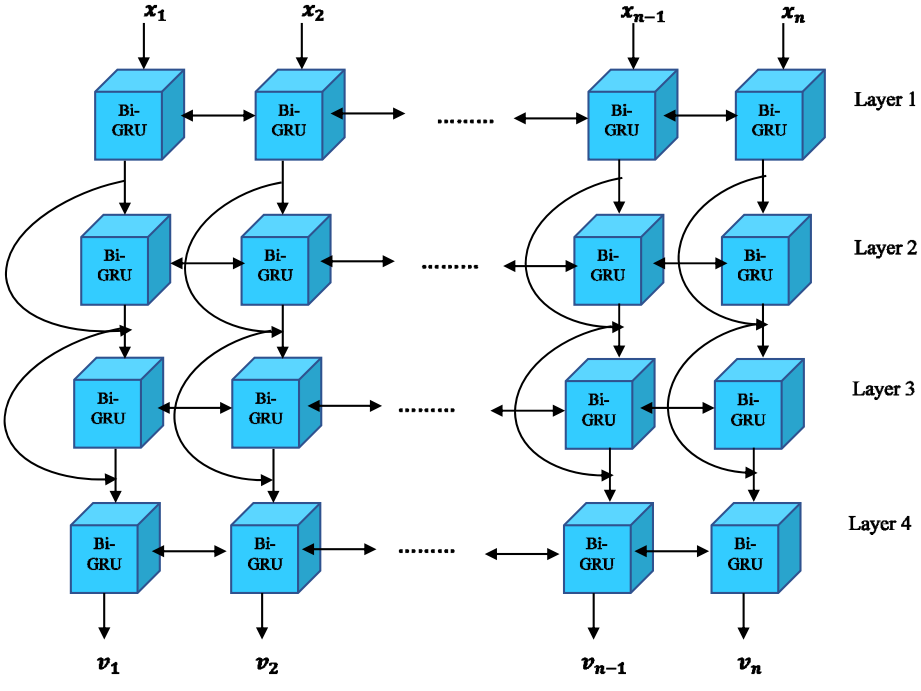
where FV is the input feature vector, the $encoder(x)$ represents the BiGRU, and the numbering from 1 to 4 indicates the layer, V is the context vector, h_n is the hidden state of the layer n where $n = 1$ to 3.

The input to the first layer of encoder i.e., BiGRU, which is denoted by $encoder_1()$ is the GoogLeNet features (FV) for each size 1 × 1,024 frame. The output of the first layer is h_1 which is obtained using equations (1) to (4). Here h_1 is the concatenation of both forward and backwards hidden states. Similarly, $encoder_2()$and $encoder_3()$ are the second and third layers of BiGRU, whose outputs or hidden states are h_2 and h_3, respectively. h_1 is the input for the second layer, while h_2 is the input for the third layer, as depicted in Figure 4. The final layer of the encoder is the $encoder_4()$ whose input is h_3 and output is the final encoded feature vector V.

From equations (5) to (8), it is evident that the hidden state of the final encoder is the context vector of the proposed encoder architecture. In addition to stacking the layers, the architecture introduces a skip connection, significantly preserving spatial information across layers. Residual learning is the term used when skip connections are utilised. As the name suggests, some connections across different layers are skipped to extract rich and dense features. The skip connections help combine the context vectors at different

levels to enrich the feature vectors. They also tend to reduce the problem of vanishing and exploding gradient problems of stacked networks. Skipping the connections across layers helps create additional paths for the gradients to flow. Here, the skipped connections are combined using an addition operation. The architecture of the proposed encoder is shown in Figure 5.

Figure 5 The architecture of the proposed encoder (see online version for colours)



3.3 Stacked decoder architecture with residual learning

The decoder uses the encoded context vector and generates a sequential output $Y = \{y_1, y_2, \ldots, y_n\}$ containing importance score for each frame. Hence, unidirectional GRU is used as a decoder. Like the encoder architecture, the decoder is also stacked as four layers, followed by the dense layer with a sigma activation function to generate the frame level importance score. Each layer contains 256 GRU cells. The decoder decodes the context vector obtained from the encoder. The input to the decoder is the output of the encoder ($v_1, v_2, \ldots, v_n$) and the last hidden state of the encoder. Since this research aims to automatically summarise a video, the output should be in probability, indicating the importance score for each frame. Thus, a dense layer is included as an ultimate layer of the decoder with a sigmoid activation function to output the importance or relevance score for each input frame. The working of the decoder is given as a series of equations as follows.

$$h_1' = decoder_1(FV) \tag{9}$$

$$h_2' = decoder_2(h_1') \tag{10}$$

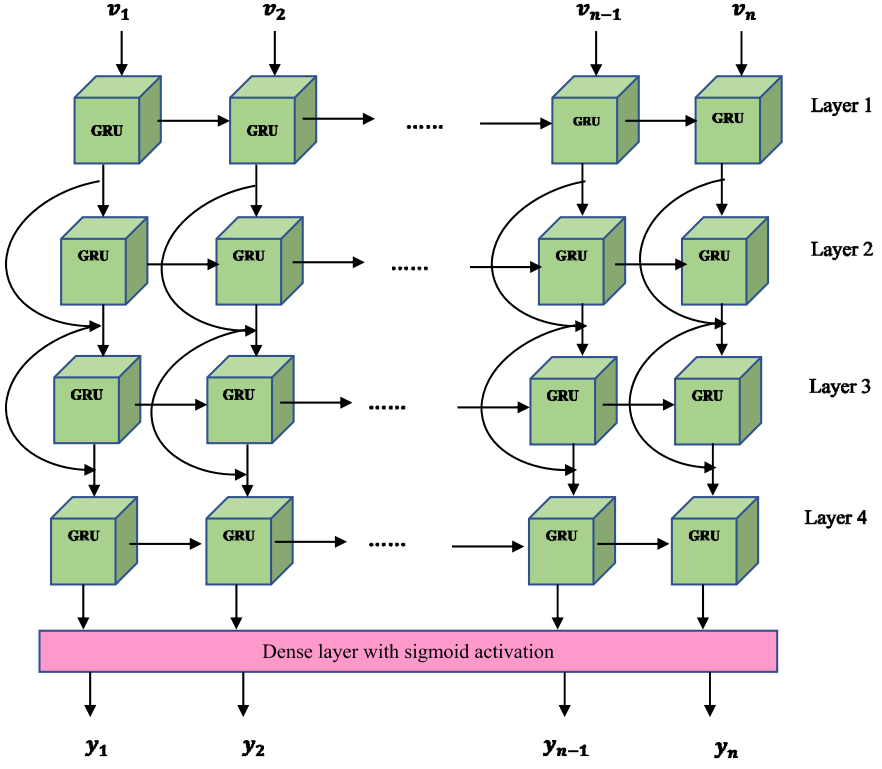
$$h_3' = decoder_3\left(h_2'\right) \tag{11}$$

$$Y = \sigma\left(decoder_4\left(h_3'\right)\right) \tag{11}$$

where h_n' is the hidden state of the decoder layer n, and n ranges from 1 to 3; Y is the output of the decoder architecture, which is the importance score for each frame in the original video.

The skip connections are introduced in decoder architecture, reducing the gradient problem for the proposed deep architecture. The skip connections across the layers for each time step help encode and decode only necessary spatial features. This helps generate an improved frame-level score for the dynamic video summarisation. The architecture of the proposed decoder is shown in Figure 6.

Figure 6 The architecture of the proposed decoder (see online version for colours)



3.4 Keyshot selection

The final step in generating the dynamic video summary is to select the key shots from the obtained importance score. For key shot selection, the kernel temporal segmentation (KTS) proposed by Potapov et al. (2014) is used. KTS is used to convert the original video frames into semantically meaningful shots. Then, a shot level relevance score is calculated by taking the mean value of the frame level importance score for each shot. Keyshots are identified by optimising the following equation:

$$\max \sum_{i=1}^{n} m_i s_i, \quad \text{s.t.} \sum_{i=1}^{n} m_i k_i \leq 1, m_i \in 0, 1 \tag{13}$$

where

$$s_i = \frac{1}{k_i} \sum_{i=1}^{k_i} y_i \tag{14}$$

where m is the cardinality denoting the actual number of shots, y_i is the frame level importance score, m_i is the relevance score of i^{th} shot and k_i is the length of the i^{th} shot. As it resembles the 0/1 knapsack problem, it is solved using dynamic programming. The final video summary is created by combining those shots. $m_i \neq 0$ in temporal order.

The algorithm of the proposed SERNet model is given as follows:

Algorithm 1 Proposed SERNet model

Input: Video

Output: Dynamic summary

Begin

Convert the input video I into frames $F = \{f_1, f_2, \ldots, f_m\}$

// Pre-processing and feature extraction

For each frame f_k

Resize the frame f_k to 224 × 224

FV = Googlenet(f_k)

End for

//encoding

$h_1 = encoder_1(FV)$

$h_2 = encoder_2(h_1)$

$h_3 = encoder_3(h_2)$

$V = encoder_4(h_3)$

//decoding

$h_1' = decoder_1(V)$

$h_2' = decoder_2(h_1')$

$h_3' = decoder_3(h_2')$

$h_4' = decoder_4(h_3')$

$Y = sigmoid(h_4')$

//keyshot selection

video_segments[]=KTS(I)

For each *video_segments vs*

score = 0

For each frame *Fvs*

score = *score* + *Y*[*Fvs*]

end for

s_score = *avg*(*score*)
end for
shots = argmax(*s_score*, *J*)
end

In the algorithm, *I* denote the input video, frames are denoted as $F = \{f_1, f_2, \ldots, f_m\}$, and *m* is the number of frames in the input video. *Googlenet*() function denotes the GoogLeNet pre-trained network used for feature extraction. *Encoder*() and *decoder*() functions denote the proposed encoder and decoder, respectively. *Y* is the frame level importance score for the input video. *KTS*()is the kernel temporal segmentation function that segments the input video *I* into small segments denoted by *video_segments* array. Score is the temporary variable to store the frame level importance score *F*[*Fvs*]. Here *Fvs* is the frame of the *video_segment vs*. *s_score* denotes the segment level score, the average frame level score in a particular segment. Finally, argmax(*s_score*, *J*) is the optimisation function which optimises the parameter or slack variable *J* such that the overall key shots denote the final dynamic video summary. Finally, the segments in the variable *shots* are concatenated to produce the final dynamic summary of the input video.

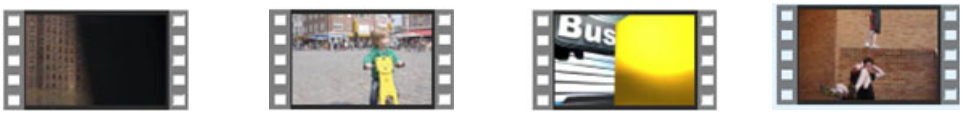
4 Experimental results

4.1 Dataset description

The proposed SERNet model is tested on two benchmark datasets: SumMe (Gygli et al., 2014) and TVSum (Song et al., 2015). The SumMe dataset consists of 25 videos from various events, such as holidays and sports. These videos contain 30 frames per second, and the average duration of each video is about 2–7 minutes. TVSum dataset consists of 50 videos under ten categories, including third-person and ego-centric videos. They contain an average of 30 frames per second, and the average duration of each video is about 2–10 minutes. Figure 7 shows the sample thumbnail images from SumMe and TVSum datasets. Both the datasets have the frame level importance score given to 15–20 people as ground truth annotation. During evaluation, the automated summary is compared against these ground truth annotations for each video, and the average performance measure is taken for result analysis.

Figure 7 Sample dataset, (a) SumMe dataset (b) TVSum dataset (see online version for colours)

(a)



(b)

4.2 Evaluation metrics

Precision, recall and F-score are the measures for evaluating the dynamic video summarisation. Let A be the automated summary created, and G be the ground truth summary. The precision P, recall R and F-score F are calculated using the following equations:

$$P = \frac{overlapped\ duration\ of\ \ A\ and\ G}{duration\ of\ \ A} \tag{15}$$

$$R = \frac{overlapped\ duration\ of\ \ A\ and\ G}{duration\ of\ \ G} \tag{16}$$

$$F = \frac{2 \times P \times R}{(P + R)} \times 100 \tag{17}$$

4.3 Experimental settings

The input videos are down-sampled into two frames per second to reduce the need for high-end computing devices. For a fair comparison, the output of the pool of five layers of GoogLeNet is used as a feature for the down-sampled frames. 80% of videos are taken for training, while the remaining 20% are taken for testing. The proposed SERNet model uses the gradient descent method for training with a learning rate 0.10. The training is stopped after five successive epochs with a decreasing F-score. The patience value for the SERNet model is set as 15, i.e., the training stops once there is no change in the F-score for about 15 epochs. The proposed SERNet model is run ten times, and the average performance is taken for fair result comparison.

The major assumptions considered in this study are that each video is independent and does not contain any correlation between them. Another important assumption is that the final summary represents the condensed version of the original video with full coverage of the context of the video. The final summary is not just any particular activity or event but all the events that humans would otherwise consider important. Finally, video summarisation is very subjective and varies according to every individual's point of view. This research work attempts to closely match the video summary with those generated by some selected human annotators.

4.4 Result analysis

Table 1 shows the precision, recall and F-score performance measures for the samples from SumMe and TVSum datasets.

Table 1 shows the sample output for the two datasets under study. Five sample videos are taken for each dataset, and the results of precision, recall and F-score are shown in Table 1.

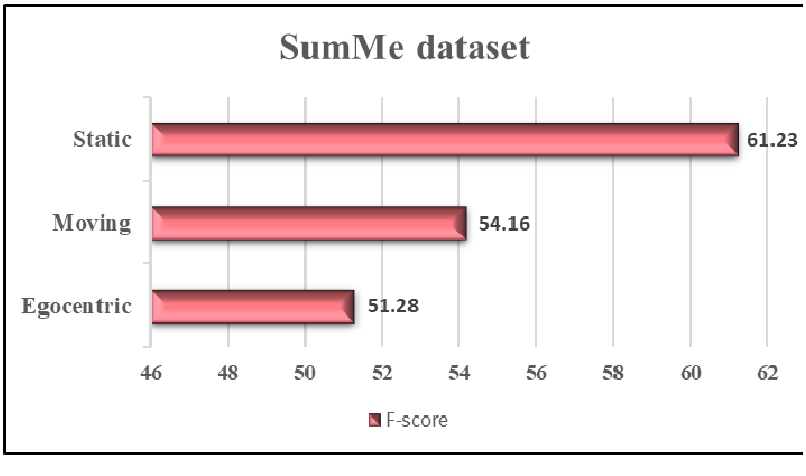
Table 1 Performance measures of the proposed SERNet model

Dataset	*Video name*	*Precision*	*Recall*	*F-score*
SumMe	Air_Force_one	62.03	64.10	63.04
	Base jumping	60.58	62.80	61.67
	Bearpark_climbing	58.31	59.48	58.88
	Bike Polo	61.45	63.24	62.33
	Bus_in_Rock _tunnel	54.20	55.32	54.75
TVSum	VT1	65.42	68.42	66.88
	VT2	66.21	68.00	67.09
	VT3	64.96	66.87	65.90
	VT4	59.45	60.28	59.86
	VT5	65.50	65.00	65.24

4.5 Result analysis on different categories of videos

Both SumMe and TVSum datasets contain different categories of videos. As the proposed SERNet model is for generic video summarisation, it is important to show the results for various categories. For the SumMe dataset, 25 videos are divided based on how they are captured, such as selfish, moving, and static. Egocentric videos are the ones that are taken by the first person in the video, while moving videos are captured by some portable capturing devices like smartphones and digital cameras by home users. Moving videos tend to be shaky and are difficult to summarise, which is a major challenge. Static videos are the ones in which the camera is not moving and is kept in a single point of view. Out of 25 videos, four are egocentric, 17 are moving, and the remaining four are static.

Figure 8 Category-wise performance analysis for the SumMe dataset (see online version for colours)



The TVSum dataset contains various videos such as documentaries, vlogs, how-tos, news, cooking, user-generated, and egocentric videos. The 50 videos are distributed as follows according to the category. It contains 12 documentary videos, seven vlogs, 13 how-to videos, five news videos, four cooking videos, eight user-generated videos and one egocentric video.

Figure 8 shows that the proposed SERNet model performs comparatively well for static videos with an F-score of 61.23, which can be explained. For the moving videos, the F-score value is 54.16, while for egocentric videos, the F-score is the lowest value, 51.28. The results show that the proposed SERNet model performs less for moving and egocentric videos because they have too many blurred and shaky frames. Thus, this creates a need for further improvement in the robustness of the proposed model.

Figure 9 Category-wise performance analysis for the TVSum dataset (see online version for colours)

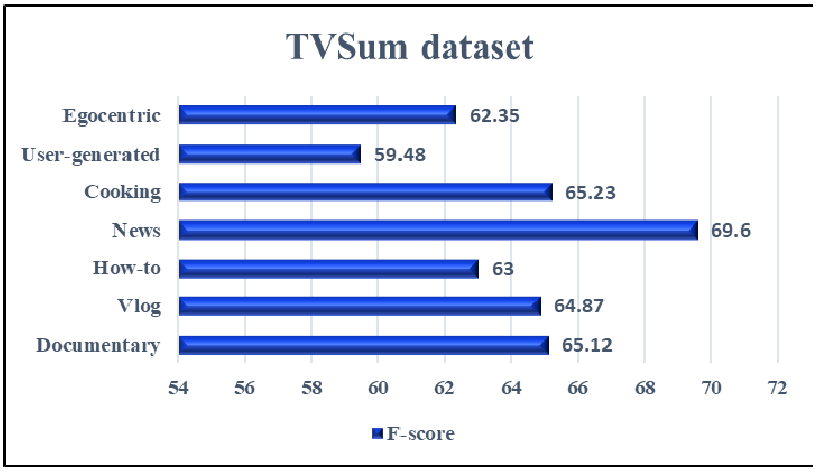


Figure 9 shows that the proposed SERNet model produces a robust performance against the different categories of videos, with the highest F-score of 69.6 for news videos. This is because the news videos have a particular structure with a considerable difference between the shots, which makes them favourable for finding the key events. The model produces an F-score of at least 59.48 for the user-generated videos. The comparatively poor performance of the proposed model is explained by the fact that the user-generated videos do not have any particular video structure. The videos are shaky, with minimum to no transition gaps between the scenes. Thus, there needs to be an improvement in the proposed SERNet model, particularly for user-generated videos with no transition gaps between frames. Overall, the proposed SERNet model effectively uses the encoder-decoder architecture and the spatio-temporal features to robustly predict the important events in a video.

4.6 Comparative analysis with state-of-the-art methods

The proposed SERNet model is compared against state-of-the-art methods and shown in Table 2.

Table 2 Comparison with state-of-the-art methods

Dataset	*Methodology*	*F-score*
SumMe	M-AVS (Ji et al., 2019)	44.40
	VASNet (Fajtl et al., 2019)	49.71
	MC-VSA (Liu et al., 2020a)	51.60
	AEDNet (Feng et al., 2020)	52.60
	Proposed SERNet	*55.60*
TVSum	M-AVS (Ji et al., 2019)	61.00
	VASNet (Fajtl et al., 2019)	61.42
	MC-VSA (Liu et al., 2020b)	63.70
	AEDNet (Feng et al., 2020)	62.70
	Proposed SERNet	*64.23*

Table 2 shows that the proposed SERNet produces an average F-score of 55.6 for the SumMe dataset and 64.23 for the TVSum dataset. Out of the research works taken for comparison, AEDNet has the highest F-score, 52 for the SumMe dataset and 62.7 for the TVSum dataset, which is the work done by Feng et al. (2020). The proposed SERNet model shows an increment of 3.8% and 2.4% for the SumMe and TVSum datasets, respectively, compared to AEDNet. This is justified by using a stacked encoder and residual learning, which learns better feature encoding along the spatial and temporal directions. Deeper stacked layers provide better feature representations and skip connections give a well-defined hierarchical spatial feature. Thus, better hierarchical features contribute to the model's efficiency in predicting a frame's importance score. This concludes that the proposed SERNet model outstands the other state-of-the-art methods.

4.7 Comparison based on different dataset settings

This work uses the supervised learning method, which utilises the annotations of each video in the dataset. Any deep learning model with a supervised learning strategy requires a large amount of data along with labels for learning. The task of summarising generic videos especially demands enormous data with labels. Labelling the dataset for video summarisation is a human resource-intensive task. The two datasets used in this work, namely SumMe and TVSum, contain only a few videos that are insufficient for training the complicated deep learning model. Hence, these datasets are augmented with two other datasets: YouTube and Open Video Project (OVP) (De Avila et al., 2011).

OVP contains 50 videos, which are mostly documentaries. The YouTube dataset contains 50 videos, such as news and sports, downloaded from the YouTube platform. These two datasets contain keyframes as ground truth annotations. For a fair comparison, the keyframes are converted into frame-level importance scores using the method described in Zhang et al. (2016). Three types of dataset settings, canonical, augmented and transfer, are utilised for experimentation. The canonical setting is the traditional way of splitting the dataset into training (80%) and testing (20%). In the case of the augmented dataset setting, 80 % of the original dataset is combined with the OVP and YouTube dataset, while the remaining 20% is used for testing. Two different datasets are used for training and testing in the transfer setting. Unlike the usual setting, this transfer

set does not split a single dataset into training and testing but uses different datasets separately for training and testing. Using this augmentation technique, the domain transferable ability of the proposed SERNet model is tested. Thus, OVP and YouTube datasets are used for training, while the original dataset is used for testing. The different dataset settings and the number of videos utilised for training the proposed model are shown in Table 3.

Table 3 Different dataset settings

Dataset	*Number of videos used for training*							
	Transfer settings				*Augmented settings*			
	OVP	*YouTube*	*SumMe*	*TVSum*	*OVP*	*YouTube*	*SumMe*	*TVSum*
SumMe	50	50	-	50	50	50	20	50
TVSum	50	50	25	-	50	50	25	40

Thus, in transfer settings for the SumMe dataset, 150 videos are used for training and 25 videos for testing. For the TVSum dataset, 125 videos are utilised for training and 50 videos for testing. In augmented settings, 170 and 165 videos are taken for training SumMe and TVSum datasets, respectively, while the remaining five videos and 20 videos are used for testing. The average F-score for both SumMe and TVSum datasets with different dataset settings are shown in Table 4.

Table 4 Result comparison for different dataset settings

Dataset	*Methodology*	*Canonical*	*Transfer*	*Augmented*
SumMe	VASNet (Fajtl et al., 2019)	49.71	-	51.09
	MC-VSA (Liu et al., 2020a)	51.60	48.10	53.00
	M-AVS (Ji et al., 2019)	44.40	-	46.10
	DR-DSN (Zhou et al., 2018)	42.10	42.60	43.90
	DPP-LSTM (Zhang et al., 2016)	38.60	41.80	42.90
	Proposed SERNet	*55.60*	*51.50*	*56.22*
TVSum	VASNet (Fajtl et al., 2019)	61.42	-	62.37
	MC-VSA (Liu et al., 2020b)	63.70	59.50	64.00
	M-AVS (Ji et al., 2019)	61.00	-	61.80
	DR-DSN (Zhou et al., 2018)	58.10	58.90	59.80
	DPP-LSTM (Zhang et al., 2016)	54.70	59.60	58.70
	Proposed SERNet	*64.23*	*63.20*	*65.00*

From Table 4, it is inferred that the augmented dataset increased the performance of video summarisation considerably. For the SumMe dataset, the performance is increased by 1.16 % with an average F-score of 56.22. In the TVSum dataset, the performance is increased by 1.18% with an average F-score of 65.00. This indicates the power of dataset augmentation, which helps improve the model's efficiency. It can also be seen that the performance measure of the transfer setting is slightly lesser when compared to the canonical setting. For the transfer setting, the F-score is 51.50 for the SumMe dataset and 63.20 for the TVSum dataset. The F-score decreases because the datasets used for

training and testing are completely different. This creates a need to improve the model, which is considered future work.

4.8 Evaluation of skip connections

For verification of the effectiveness of the skip connections, a baseline model without the skip connection, namely GRUvs, is created and compared with the proposed SERNet model. The results are shown in Figure 10.

Figure 10 Comparison of the proposed models with and without skip connections

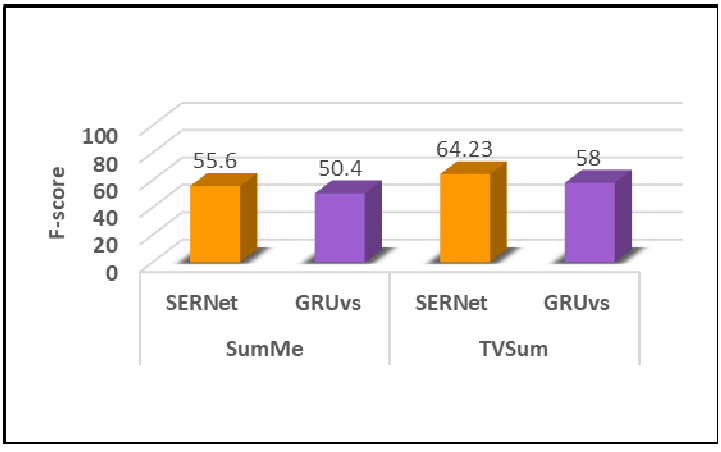


Figure 7 infers that the proposed model with a skip connection performs better than the model without a skip connection by an improvement of around 10.3% for the SumMe dataset and 10.7% for the TVSum dataset. This shows the significance of the skip connections. Skipping the layers creates an alternate path for the features to flow through the network. Thus, hierarchical features are extracted, contributing to the effectiveness of dynamic generic video summarisation.

4.9 Comparison with LSTM-based encoder-decoder architecture

The proposed SERNet model uses GRU as an encoder and decoder. It is compared against another recurrent neural network (RNN) called long short-term memory (LSTM), a popular network for modelling sequence-to-sequence problems. LSTM is also an RNN which uses gate functions. It has three types of gates: input gate, forget gate and output gate. The input gate is used to retain the information for the long term. The forget gate decides whether the information stored by the input gate is worth storing. If it seems to be unworthy, the forget gate discards the information. The output gate provides a short-term memory to be saved based on the newly updated long-term information and previous output. GRU has fewer gates and thud fewer parameters when compared to LSTM. The bi-directional LSTM is used as an encoder, and the unidirectional LSTM is used as a decoder for comparison. The results are shown in Table 5.

From Table 5, it is observed that for the SumMe dataset, the average F-score is 53.20 and 55.6 for Bi-LSTM and Bi-GRU architecture and the TVSum dataset, the average F-score is 62.91 and 64.23 for Bi-LSTM and Bi-GRU architecture. The proposed

SERNet model shows an increment of 4.5% and 2.1% for the SumMe and TVSum datasets, respectively. Thus, the GRU outperforms the LSTM as encoder-decoder architecture for video summarisation.

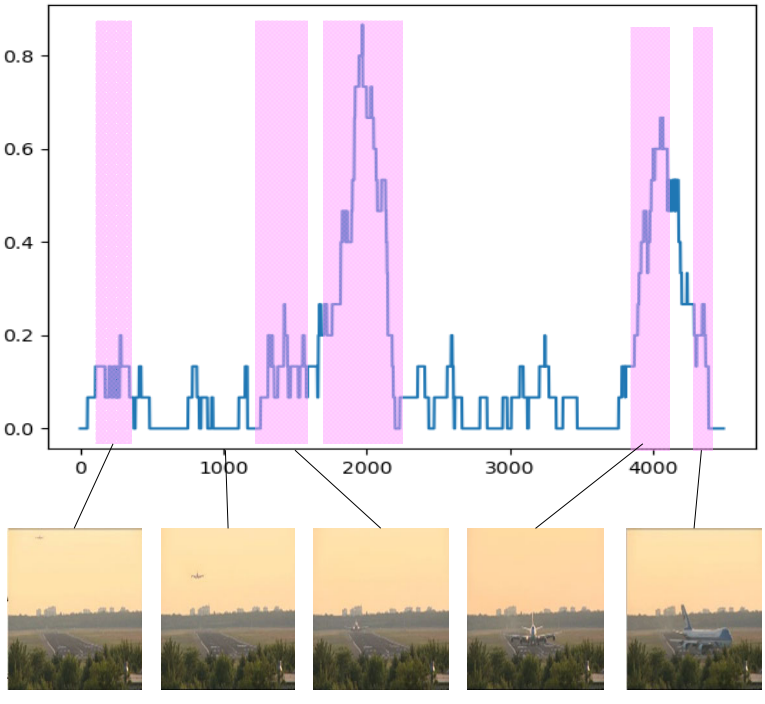
Table 5 Result comparison between the RNNs

Dataset	*Bi-LSTM*	*Bi-GRU*
SumMe	53.20	55.60
TVSum	62.91	64.23

4.10 Qualitative analysis

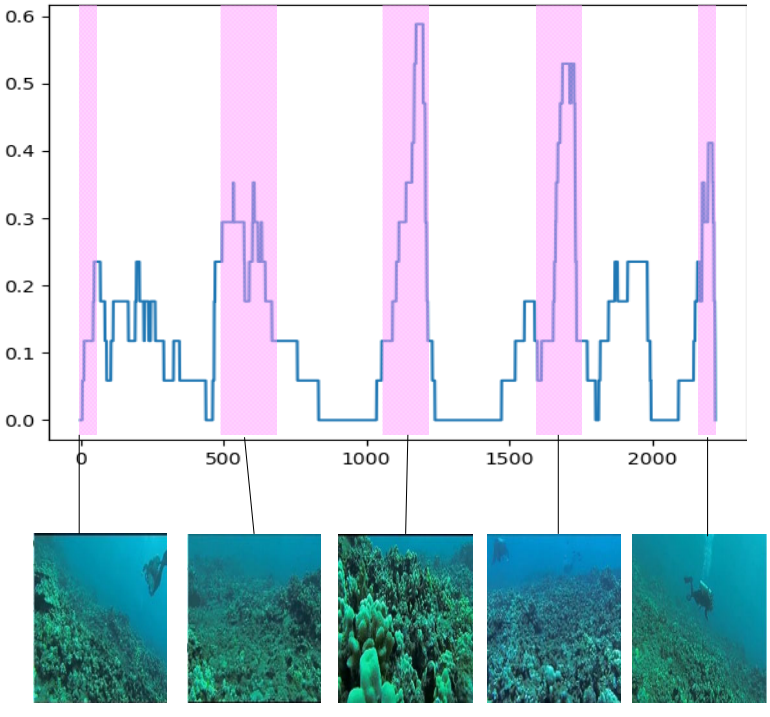
Qualitative analysis focuses on analysing the visual quality of the summary of a video. Visually appealing and continuous shots are necessary for a good video summarisation. As each video has ground truth, i.e., importance score for each frame, they are compared against the summary generated by the proposed SERNet model with visual representation. Figures 11 to13 show the qualitative analysis for the SumMe dataset under each category: static, moving and egocentric.

Figure 11 Sample qualitative analysis for the Static category from the SumMe dataset (see online version for colours)



The blue line in the graph indicates the ground truth for that particular video. The graph represents the importance score for each frame, with the x-axis denoting the number of frames while the y-axis denotes the importance score of each frame from 0 to 1.

Figure 12 Sample qualitative analysis for the moving category from the SumMe dataset (see online version for colours)



Figures 14 and 15 are sample qualitative analyses from two categories: cooking and news from the TVSum dataset. The above figures show that the proposed SERNet model captures most of the peaks shown in the graph, where the peaks indicate key events. The effective performance of the proposed SERNet model is due to the use of dense spatiotemporal features, which makes the model robust against various categories of videos. For representation purposes, the videos are chosen with only five shots, while in general, the model does not necessarily choose only five key shots.

An exclusive experimental analysis is made with different dataset settings and encoder-decoder architectures. Comparative analysis is given in which the proposed SERNet model is compared against the existing state-of-the-art methods. The analysis proves the efficiency of the proposed model quantitatively. Qualitative analysis is also made, which shows the visual quality of the automated video summary. By utilising the residual learning for dense feature extraction and modelling the video summarisation task as a sequence-to-sequence problem, this research contributes an efficient and robust deep learning model for generic video summarisation. Since this research focuses on producing automated generic video summarisation, it helps summarise major video categories like news and surveillance videos that are part of daily lives. In video storage and management, this study greatly helps store abstract forms of video for easy and quick access to videos in case of browsing and retrieval.

Figure 13 Sample qualitative analysis for the moving category from the SumMe dataset (see online version for colours)

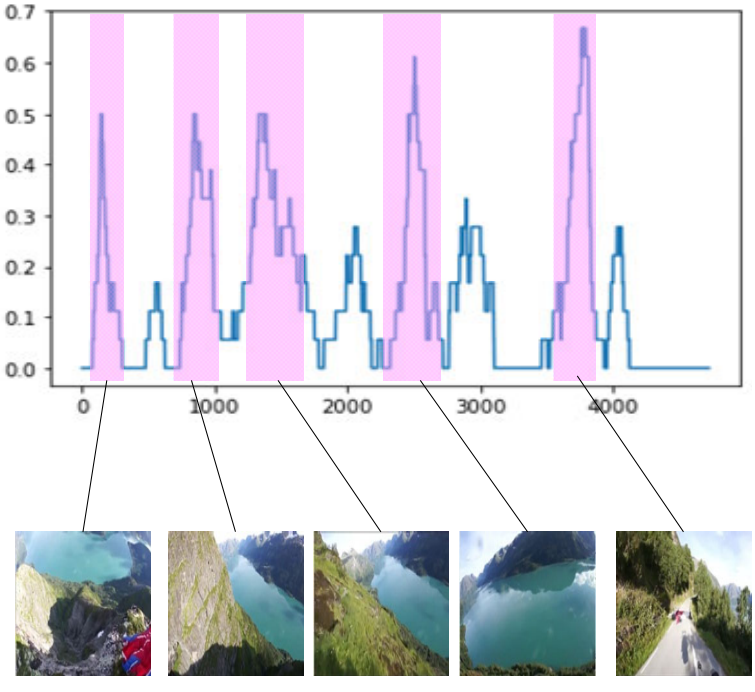


Figure 14 Sample qualitative analysis for the cooking category from the TVSum dataset (see online version for colours)

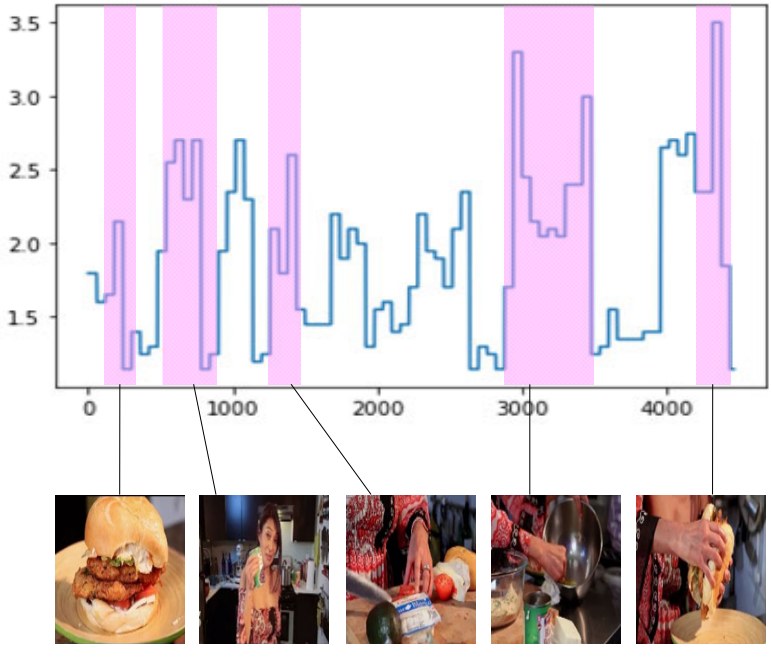
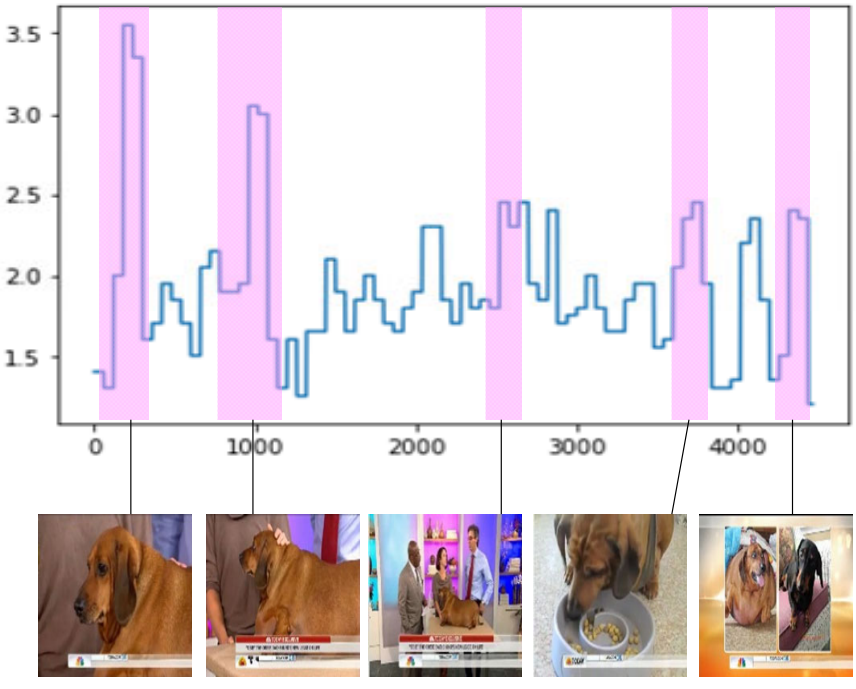


Figure 15 Sample qualitative analysis for the news category from the TVSum dataset (see online version for colours)



4.11 Outcomes of the proposed SERNet model

The outcomes of this research include

- effective modelling of video summarisation as a sequence-to-sequence conversion problem through deep stacked encoder-decoder architecture
- the summaries produced by the proposed SERNet model cover most of the important events of the video
- it helps manage a large video repository by storing the summaries of original-length videos as a preview for quick access and retrieval
- this generic video summarisation framework is robust and helps summarise various categories of videos efficiently.

5 Conclusions and future work

Dynamic video summarisation is a challenging field of video processing that generates a sparse representation of the original long video. The sparsity is maintained so that the overall meaning of the video content remains unchanged. It has various applications, depending on the type of video that is summarised. Some applications include news and sports video highlights, movie highlights, key events in a lecture video, and an abnormal

activity summary in a surveillance video. This work proposes an efficient SERNet model for summarising generic videos. The experiments are conducted in different settings, such as by varying the encoder-decoder architectures, with and without skip connections, and by varying the dataset's diversity. Finally, the obtained results are tabulated. The experimental analysis shows that the proposed SERNet model achieves a better F-score of 55.6 for the SumMe dataset and 64.23 for the TVSum dataset compared to other existing works. Improvements are still required, even if the proposed SERNet model performs effectively. Future work aims to increase the performance of the video summarisation task in the transfer settings so that video summarisation can be achieved for large datasets without being annotated explicitly.

Acknowledgements

This study was funded by the University Grants Commission (UGC) Of the Government of India.

Data availability

The datasets analysed during the current study are available in the following links.

- SumMe – https://gyglim.github.io/me/vsum/index.html
- TVSum – http://people.csail.mit.edu/yalesong/tvsum/
- OVP and Youtube – https://sites.google.com/site/vsummsite/download.

References

Alshahrani, H., Sharma, G., Anand, V., Gupta, S., Sulaiman, A., Elmagzoub, M.A. and Azar, A.T. (2023) 'An intelligent attention-based transfer learning model for accurate differentiation of bone marrow stains to diagnose haematological disorder', *Life*, Vol. 13, No. 10, p.2091.

Anand, P.P., Sulthan, N., Jayanth, P. and Deepika, A.A. (2023) 'A creating musical compositions through recurrent neural networks: an approach for generating melodic creations', *FMDB Transactions on Sustainable Computing Systems*, Vol. 1, No. 2, pp.54–64.

Angeline, R., Aarthi, S., Regin, R. and Rajest, S.S. (2023) 'Dynamic intelligence-driven engineering flooding attack prediction using ensemble learning', *Advances in Artificial and Human Intelligence in the Modern Era*, pp.109–124, IGI Global, USA.

Arivukarasi, M., Manju, A., Kaladevi, R., Hariharan, S., Mahasree, M. and Prasad, A.B. (2023) 'Efficient phishing detection and prevention using support vector machine (SVM) algorithm', *CSNT 2023: Proceedings of the 12th International Conference on Communication Systems and Network Technologies*, IEEE, Bhopal, India, pp.545–548.

Azar, A.T., Sardar, M.Z., Ahmed, S., Hassanien, A.E. and Kamal, N.A. (2023a) 'Autonomous robot navigation and exploration using deep reinforcement learning with Gazebo and ROS', *AISI 2023: International Conference on Advanced Intelligent Systems and Informatics,* Springer Nature, Switzerland, pp.287–299.

Azar, A.T., Shehab, E., Mattar, A.M., Hameed, I.A. and Elsaid, S.A. (2023b) 'Deep learning based hybrid intrusion detection systems to protect satellite networks', *Journal of Network and Systems Management*, Vol. 31, No. 4, p.82.

Azar, A.T., Tounsi, M., Fati, S.M., Javed, Y., Amin, S.U., Khan, Z.I. and Ganesan, J. (2023c) 'Automated system for colon cancer detection and segmentation based on deep learning techniques', *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, Vol. 15, No. 1, pp.1–28.

Bellahcene, Z., Bouhamida, M., Denai, M. and Assali, K. (2021) 'Adaptive neural network-based robust H∞ tracking control of a quadrotor UAV under wind disturbances', *International Journal of Automation and Control*, Vol. 15, No. 1, pp.28–57.

Bengio, Y. (2009) 'Learning deep architectures for AI', *Foundations and Trends® in Machine Learning*, Vol. 2, No. 1, pp.1–127.

Bhat, S. (2021) 'Machine learning-based novel DSP controller for PV systems', *International Journal of Automation and Control*, Vol. 15, No. 2, pp.226–239.

Bose, S.R., Singh, R., Joshi, Y., Marar, A., Regin, R. and Rajest, S.S. (2023) 'Light weight structure texture feature analysis for character recognition using progressive stochastic learning algorithm', *Advanced Applications of Generative AI and Natural Language Processing Models*, pp.144–158, IGI Global, USA.

Boulmaiz, A., Meghni, B., Redjati, A. and Azar, A.T. (2022) 'LiTasNeT: a bird sound separation algorithm based on deep learning', *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, Vol. 14, No. 1, pp.1–19.

Chen, Y., Guo, B., Shen, Y., Zhou, R., Lu, W., Wang, W. and Suo, X. (2022) 'Video summarization with u-shaped transformer', *Applied Intelligence*, Vol. 52, No. 15, pp.17864–17880.

Cirillo, S., Polese, G., Salerno, D., Simone, B. and Solimando, G. (2023) 'Towards flexible voice assistants: evaluating privacy and security needs in IoT-enabled smart homes', *FMDB Transactions on Sustainable Computer Letters*, Vol. 1, No. 1, pp.25–32.

Dai, X., Liu, L. and Deng, Z. (2021) 'Optimised data-driven terminal iterative learning control based on neural network for distributed parameter systems', *International Journal of Automation and Control*, Vol. 15, Nos. 4–5, pp.463–481.

De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A. and de Albuquerque Araújo, A. (2011) 'VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method', *Pattern Recognition Letters*, Vol. 32, No. 1, pp.56–68.

Devi, B.T. and Rajasekaran, R. (2023) 'A comprehensive review on deepfake detection on social media data', *FMDB Transactions on Sustainable Computing Systems*, Vol. 1, No. 1, pp.11–20.

Dhanushree, M., Priya, R., Aruna, P. and Bhavani, R. (2023) 'A keyframe extraction using HDBSCAN with particle swarm optimization', *SPIN 2023: Proceedings of the 10th International Conference on Signal Processing and Integrated Networks*, IEEE, Noida, India, pp.445–450.

Dos Santos Belo, L., Caetano Jr, C.A., do Patrocínio Jr, Z.K.G. and Guimaraes, S.J.F. (2016) 'Summarizing video sequence using a graph-based hierarchical approach', *Neurocomputing*, Vol. 173, No. 1, pp.1001–1016.

Dudekula, K.V., Syed, H., Basha, M.I.M., Swamykan, S.I., Kasaraneni, P.P., Kumar, Y.V.P. and Azar, A.T. (2023) 'Convolutional neural network-based personalized program recommendation system for smart television users', *Sustainability*, Vol. 15, No. 3, p.2206.

Elfeki, M., Wang, L. and Borji, A. (2022) 'Multi-stream dynamic video Summarization', *WACV 2022: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, pp.339–349.

Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D. and Remagnino, P. (2019) 'Summarizing videos with attention', *Computer Vision – ACCV 2018 Workshops: Proceedings of the 14th Asian Conference on Computer Vision*, Perth, Australia, pp.39–54.

Fati, S.M., Senan, E.M. and Azar, A.T. (2022) 'Hybrid and deep learning approach for early diagnosis of lower gastrointestinal diseases', *Sensors*, Vol. 22, No. 11, p.4079.

Feng, X., Wang, L. and Zhu, Y. (2020) 'Video summarization with self-attention based encoder-decoder framework', *ICCST 2020: proceedings of the IEEE International Conference on Culture-oriented Science and Technology*, Beijing, China, pp.208–214.

Gaayathri, R.S., Rajest, S.S., V.K. Nomula, and Regin, R. (2023) 'Bud-D: enabling bidirectional communication with ChatGPT by adding listening and speaking capabilities', *FMDB Transactions on Sustainable Computer Letters*, Vol. 1, No. 1, pp.49–63.

Ganesan, J., Azar, A.T., Alsenan, S., Kamal, N.A., Qureshi, B. and Hassanien, A.E. (2022) 'Deep learning reader for visually impaired', *Electronics*, Vol. 11, No. 20, p.3335.

Ghafoor, H.A., Javed, A., Irtaza, A., Dawood, H., Dawood, H. and Banjar, A. (2018) 'Egocentric video summarization based on people interaction using deep learning', *Mathematical Problems in Engineering*, Vol. 2018, No. 11, pp.1–12.

Ghauri, J.A., Hakimov, S. and Ewerth, R. (2021) 'Supervised video summarization via multiple feature sets with parallel attention', *ICME 202: Proceedings of the IEEE International Conference on Multimedia and Expo*, Shenzhen, China, pp.1–6.

Ghozali, M.T., Amalia Islamy, I.D. and Hidayaturrohim, B. (2022) 'Effectiveness of an educational mobile-app intervention in improving the knowledge of COVID-19 preventive measures', *Informatics in Medicine Unlocked*, Vol. 34, No. 10, p.101112.

Ghozali, M.T., Hidayaturrohim, B. and Dinah Amalia Islamy, I. (2023) 'Improving patient knowledge on rational use of antibiotics using educational videos', *International Journal of Public Health Science (IJPHS)*, Vol. 12, No. 1, p.41.

Govindharaj, A. and Mariappan, A. (2020) 'Design and analysis of novel Chebyshev neural adaptive backstepping controller for boost converter fed PMDC motor', *International Journal of Automation and Control*, Vol. 14, Nos. 5–6, pp.694–712.

Gygli, M., Grabner, H., Riemenschneider, H. and Van Gool, L. (2014) 'Creating summaries from user videos', *Computer Vision – ECCV 2014: Proceedings of the 13th European Conference*, Springer International Publishing, Zurich, Switzerland, pp.505–520.

Han, M.X., Hu, H.M., Liu, Y., Zhang, C., Tian, R.P. and Zheng, J. (2017) 'An auto-encoder-based summarization algorithm for unstructured videos', *Multimedia Tools and Applications*, Vol. 76, pp.25039–25056.

Hsu, T.C., Liao, Y.S. and Huang, C.R. (2023) 'Video summarization with spatiotemporal vision transformer', *IEEE Transactions on Image Processing*, Vol. 32, pp.3013–3026.

Huang, J.H., Yang, C.H.H., Chen, P.Y., Chen, M.H. and Worring, M. (2023) *Conditional Modeling Based Automatic Video Summarization*, arXiv preprint arXiv:2311.12159.

Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S.W. and de Albuquerque, V.H.C. (2021) 'A comprehensive survey of multi-view video summarization', *Pattern Recognition*, Vol. 109, No. 1, p.107567.

Jeba, J.A., Bose, S.R. and Boina, R. (2023) 'Exploring hybrid multi-view multimodal for natural language emotion recognition using multi-source information learning model', *FMDB Transactions on Sustainable Computer Letters*, Vol. 1, No. 1, pp.12–24.

Ji, Z., Jiao, F., Pang, Y. and Shao, L. (2020) 'Deep attentive and semantic preserving video summarization', *Neurocomputing*, Vol. 405, No. 9, pp.200–207.

Ji, Z., Xiong, K., Pang, Y. and Li, X. (2019) 'Video summarization with attention-based encoder-decoder networks', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 6, pp.1709–1717.

Joshi, M., Shen, Z. and Kausar, S. (2023) 'Enhancing inclusive education on leveraging artificial intelligence technologies for personalized support and accessibility in special education for students with diverse learning needs', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 3, pp.125–142.

Kalaivani, P. and Roomi, S.M.M. (2017) 'Towards comprehensive understanding of event detection and video summarization approaches', *ICRTCCM 2017: Proceedings of Second International Conference on Recent Trends and Challenges in Computational Models*, IEEE, Tindivanam, India, pp.61–66.

Kanyimama, W. (2023) 'Design of a ground based surveillance network for Modibbo Adama University, Yola', *FMDB Transactions on Sustainable Computing Systems*, Vol. 1, No. 1, pp.32–43.

Khan, H., Hussain, T., Khan, S.U., Khan, Z.A. and Baik, S.W. (2024) 'Deep multi-scale pyramidal features network for supervised video summarization', *Expert Systems with Applications*, Vol. 237, No. 3, p.121288.

Kota, B.U., Stone, A., Davila, K., Setlur, S. and Govindaraju, V. (2021) 'Automated whiteboard lecture video summarization by content region detection and representation', *ICPR 2020: Proceedings of 25th International Conference on Pattern Recognition*, IEEE, Milan, Italy, pp.10704–10711.

Lai, P.K., Décombas, M., Moutet, K. and Laganiere, R. (2016) 'Video summarization of surveillance cameras', *AVSS 2016: Proceedings of 13th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Colorado Springs, CO, USA, pp.286–294.

Lal, S., Duggal, S. and Sreedevi, I. (2019) 'Online video summarization: predicting future to better summarize present', *WACV 2019 IEEE Winter Conference on applications of Computer Vision,* Waikoloa, HI, USA, pp.471–480.

Liang, B., Li, N., He, Z., Wang, Z., Fu, Y. and Lu, T. (2021) 'News video summarization combining surf and color histogram features', *Entropy*, Vol. 23, No. 8, p.982.

Liu, T., Meng, Q., Vlontzos, A., Tan, J., Rueckert, D. and Kainz, B. (2020a) 'Ultrasound video summarization using deep reinforcement learning', *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: Proceedings of 23rd International Conference*, Springer International Publishing, Lima, Peru, pp.483–492.

Liu, Y.T., Li, Y.J. and Wang, Y.C.F. (2020b) 'Transforming multi-concept attention into video summarization', *ACCV 2020: Proceedings of the Asian Conference on Computer Vision*, Springer Cham, Kyoto, Japan, pp.498–513.

Lodha, S., Malani, H. and Bhardwaj, A.K. (2023) 'Performance evaluation of vision transformers for diagnosis of pneumonia', *FMDB Transactions on Sustainable Computing Systems*, Vol. 1, No. 1, pp.21–31.

Mademlis, I., Tefas, A. and Pitas, I. (2018) 'A salient dictionary learning framework for activity video summarization via key-frame extraction', *Information Sciences*, Vol. 432, No. 3, pp.319–331.

Mahasree, M., Puviarasan, N. and Aruna, P. (2022) 'Interpolation-based reversible data hiding with blockchain for secure e-healthcare systems', *Blockchain Applications for Healthcare Informatics*, Academic Press, Elsevier, pp.373–400.

Manju, A., Arivukarasi, M. and Mahasree, M. (2022) 'AEDAMIDL: an enhanced and discriminant analysis of medical images using deep learning', *ICSTCEE 2022: Proceedings of third International Conference on Smart Technologies in Computing, Electrical and Electronics,* IEEE, Bengaluru, India, pp.1–8.

Mayya, V. and Nayak, A. (2019) 'Traffic surveillance video summarization for detecting traffic rules violators using R-CNN', *IC4S 2017: Proceedings of Advances in Computer Communication and Computational Sciences*, Springer, Singapore, Vol. 1, pp.117–126.

Medentzidou, P. and Kotropoulos, C. (2015) 'Video summarization based on shot boundary detection with penalized contrasts', *ISPA 2015: Proceedings of 9th International Symposium on Image and Signal Processing and Analysis*, IEEE, Zagreb, Croatia, pp.199–203.

Nagaraj, B.K. and Subhashni, R. (2023) 'Explore LLM architectures that produce more interpretable outputs on large language model interpretable architecture design', *FMDB Transactions on Sustainable Computer Letters*, Vol. 1, No. 2, pp.115–129.

Nair, M.S. and Mohan, J. (2022) 'VSMCNN-dynamic summarization of videos using salient features from multi-CNN model', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 14, No. 7, pp.14071–14080.

Nayak, S.C. (2021) 'Artificial chemical reaction optimisation of recurrent functional link neural networks for efficient modelling and forecasting of financial time series', *International Journal of Automation and Control*, Vol. 15, No. 6, pp.669–691.

Obaid, A.J., Bhushan, B., Muthmainnah, S. and Rajest, S.S. (Eds.) (2023) 'Advanced applications of generative AI and natural language processing models', *Advances in Computational Intelligence and Robotics*, IGI Global, USA, DOI: 10.4018/979-8-3693-0502-7.

Padmanabhan, J., Rajest, S.S. and Veronica, J.J. (2023) 'A study on the orthography and grammatical errors of tertiary-level students', *Handbook of Research on Learning in Language Classrooms through ICT-Based Digital Technology*, pp.41–53, IGI Global, USA.

Potapov, D., Douze, M., Harchaoui, Z. and Schmid, C. (2014) 'Category-specific video summarization', *Computer Vision–ECCV 2014: Proceedings of 13th European Conference*, Springer International Publishing, Zurich, Switzerland, pp.540–555.

Psallidas, T. and Spyrou, E. (2023) 'Video summarization based on feature fusion and data augmentation', *Computers*, Vol. 12, No. 9, p.186.

Rafiq, M., Rafiq, G., Agyeman, R., Choi, G.S. and Jin, S.I. (2020) 'Scene classification for sports video summarization using transfer learning', *Sensors*, Vol. 20, No. 6, p.1702.

Rajest, S.S., Singh, B., J. Obaid, A., Regin, R. and Chinnusamy, K. (2023) 'Recent developments in machine and human intelligence', *Advances in Computational Intelligence and Robotics*, IGI Global, USA, DOI: 10.4018/978-1-6684-9189-8.

Ray, K., Saharia, S. and Sarma, N. (2021) 'Detection and identification of *Ascaris lumbricoides* and *Necator americanus* eggs in microscopic images of faecal samples of pigs', *International Journal of Automation and Control*, Vol. 15, No. 3, pp.378–402.

Regin, R., Khanna, A.A., Krishnan, V., Gupta, M., Rubin Bose, S. and Rajest, S.S. (2023) 'Information design and unifying approach for secured data sharing using attribute-based access control mechanisms', *Recent Developments in Machine and Human Intelligence*, pp.256–276, IGI Global, USA.

Reshan, M.S.A., Amin, S., Zeb, M.A., Sulaiman, A., Alshahrani, H., Azar, A.T. and Shaikh, A. (2023) 'Enhancing breast cancer detection and classification using advanced multi-model features and ensemble machine learning techniques', *Life*, Vol. 13, No. 10, p.2093.

Rochan, M., Ye, L. and Wang, Y. (2018) 'Video summarization using fully convolutional sequence networks', *ECCV 2018: Proceedings of the European Conference on Computer Vision*, Springer Cham, Munich, Germany, pp.358–374.

Singh, K., Regin, R., Sharma, P.K., Narendra, Y.V., Bose, S.R. and Rajest, S.S. (2023) 'Fine-grained deep feature expansion framework for fashion apparel classification using transfer learning', *Advanced Applications of Generative AI and Natural Language Processing Models*, pp.389–404, IGI Global, USA.

Sirajudheen, M.A.S., Bose, S.R., Kirupanandan, G., Arunagiri, S., Regin, R. and Rajest, S.S. (2023) 'Fine-grained independent approach for workout classification using integrated metric transfer learning', *Advanced Applications of Generative AI and Natural Language Processing Models*, pp.358–372, IGI Global, USA.

Sivapriya, G.B.V., Ganesh, U.G., Pradeeshwar, V., Dharshini, M. and Al-Amin, M., (2023) 'Crime prediction and analysis using data mining and machine learning: a simple approach that helps predictive policing', *FMDB Transactions on Sustainable Computer Letters*, Vol. 1, No. 2, pp.64–75.

Song, Y., Vallmitjana, J., Stent, A. and Jaimes, A. (2015) 'TVSum: summarizing web videos using titles', *CVPR 2015*: *Proceedings of the IEEE conference on computer vision and pattern recognition,* Boston, MA, USA, pp.5179–5187.

Sreeja, M.U. and Kovoor, B.C. (2021) 'A unified model for egocentric video summarization: an instance-based approach', *Computers and Electrical Engineering*, Vol. 92, No. 6, p.107161.

Sreeja, M.U. and Kovoor, B.C. (2023) 'A multi-stage deep adversarial network for video summarization with knowledge distillation', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 14, No. 8, pp.9823–9838.

Su, M., Ma, R., Zhang, B. and Li, K. (2023) 'Recurrent unit augmented memory network for video summarisation', *IET Computer Vision*, Vol. 17, No. 6, pp.710–721, DOI: 10.1049/cvi2.12194.

Sulaiman, A., Anand, V., Gupta, S., Alshahrani, H., Reshan, M.S.A., Rajab, A. and Azar, A.T. (2023) 'Sustainable apple disease management using an intelligent fine-tuned transfer learning-based model', *Sustainability*, Vol. 15, No. 17, p.13228.

Suman, R.S., Moccia, S., Chinnusamy, K., Singh, B. and Regin, R. (Eds.) (2023) 'Handbook of research on learning in language classrooms through ICT-based digital technology', *Advances in Educational Technologies and Instructional Design*, USA, DOI: 10.4018/978-1-6684-6682-7.

Tahir, M., Qiao, Y., Kanwal, N., Lee, B. and Asghar, M.N. (2023) 'Privacy preserved video summarization of road traffic events for IoT smart cities', *Cryptography*, Vol. 7, No. 1, p.7.

Teng, X., Gui, X., Xu, P., Tong, J., An, J., Liu, Y. and Jiang, H. (2022) 'A hierarchical spatial-temporal cross-attention scheme for video summarization using contrastive learning', *Sensors*, Vol. 22, No. 21, p.8275.

Terbouche, H., Morel, M., Rodriguez, M. and Othmani, A. (2023) 'Multi-annotation attention model for video summarization', *CVPRW 2023: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp.3142–3151.

Waleed, J., Azar, A.T., Albawi, S., Al-Azzawi, W.K., Ibraheem, I.K., Alkhayyat, A. and Kamal, N.A. (2022) 'An effective deep learning model to discriminate coronavirus disease from typical pneumonia', *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, Vol. 13, No. 1, pp.1–16.

Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X. and Yao, C. (2018) 'Video summarization via semantic attended networks', *AAAI 2018: Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, Vol. 32, No. 1, pp.216–223.

Xiao, S., Zhao, Z., Zhang, Z., Guan, Z. and Cai, D. (2020) 'Query-biased self-attentive network for query-focused video summarization', *IEEE Transactions on Image Processing*, Vol. 29, No. 4, pp.5889–5899.

Xu, Y., Li, X., Pan, L., Sang, W., Wei, P. and Zhu, L. (2023) 'Self-supervised adversarial video summarizer with context latent sequence learning', *IEEE Transactions on Circuits and Systems for Video Technology*. Vol. 33, No. 8, pp.4122–4136.

Yuan, Y., Li, H. and Wang, Q. (2019) 'Spatiotemporal modeling for video summarization using convolutional recurrent neural network', *IEEE Access*, Vol. 7, No. 5, pp.64676–64685.

Yuan, Y., Mei, T., Cui, P. and Zhu, W. (2017) 'Video summarization by learning deep side semantic embedding', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 29, No. 1, pp.226–237.

Zhang, K., Chao, W.L., Sha, F. and Grauman, K. (2016) 'Video summarization with long short-term memory', *Computer Vision–ECCV 2016: Proceedings of 14th European Conference*, Springer International Publishing, Amsterdam, The Netherlands, pp.766–782.

Zhang, K., Grauman, K. and Sha, F. (2018) 'Retrospective encoders for video summarization', *ECCV 2018: Proceedings of the European Conference on Computer Vision*, Munich, Germany, pp.391–408.

Zhang, R., Qin, B., Zhao, J., Zhu, Y., Lv, Y. and Ding, S. (2023a) 'Locating X-ray coronary angiogram keyframes via long short-term spatiotemporal attention with image-to-patch contrastive learning', *IEEE Transactions on Medical Imaging*, Vol. 43, No. 1, pp.51–63.

Zhang, Y., Liu, Y., Kang, W. and Zheng, Y. (2023b) 'MAR-Net: motion-assisted reconstruction network for unsupervised video summarization', *IEEE Signal Processing Letters*, Vol. 30, No. 9, pp.1282–1286.

Zhao, B., Li, X. and Lu, X. (2017) 'Hierarchical recurrent neural network for video summarization', *MM 2017: Proceedings of the 25th ACM international conference on Multimedia*, Association for Computing Machinery, New York, NY, USA, pp.863–871.

Zhao, X., Tian, X., Li, Z., Tan, X., Zhang, Q., Chen, H. and Liu, S. (2021) 'Binary particle swarm optimisation and the extreme learning machine for diagnosing paraquat-poisoned patients', *International Journal of Automation and Control*, Vol. 15, Nos. 4–5, pp.427–443.

Zhou, K., Qiao, Y. and Xiang, T. (2018) 'Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward', *AAAI 2018: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, Louisiana, USA, Vol. 32, No. 1, pp.7582–7589.