# ADAPTIVE SPATIAL-TEMPORAL FILTER (ASTF) AND ENHANCED RECURRENT NEURAL NETWORK (ERNN) FOR VIDEO SUMMARIZATION

**S. China Ramu , Professor in CSE DEPT CBIT Email:chinaramu_cse@cbit.ac.in**

**N. Srinivas, Research Scholar Rayalaseema University Kurnool, Associate professor in CSE Dept ,**

**VBIT Ghatkesar. E-mail:srinivas.bhaskar3@gmail.com**

**K. JayaSankar, Principal &amp; Professor in ECE, MGIT , GandipetHyderabad-75. E-mail: kottareddyjs@gmail.com**

**M.V. RamanaMurthy, HOD M&amp;H DEPT, MGIT GANDIPET ,Hyderabad-75, Retired Prof Mathematics &amp; Computer Science, OUHYD.E-mail: mv.rm50@gmail.com**

Received: 14 Feb 2020 Revised and Accepted: 25 March 2020

**ABSTRACT:** Video summarization is one among the potential schemes designed for efficient interpretation of visual content by choosing the meaningful frames present in the video clip. The fast evolving video content, video summarization, highlighting the automatic selection of significant and meaningful segments from videos, has found critical importance. But, the issue is a huge challenge owing to the subjective behavior of users having their individual choices on the summaries. Using the temporal dependency observed among video frames or sub shots is quite essential for the video summarization phase. The earlier work (i) temporal noise occurring in scenes under variations in illumination or lighting, (ii) Convolutional Neural Network (CNN)'s capability deficit to exhibit spatiall invariance to the input data, shot edge detection is not carried out automatically. In order to get over these problems, proposed model presented a deep learning technique called as Enhanced Recurrent Neural Network (ERNN) for video summarization, noise cancellation is performed with the help of two filtering algorithms such as Mean Filter (MF) and Adaptive Spatial-Temporal Filter (ASTF). The discussed work includes three steps (i) sampling of video frames, (ii) preprocessing, (iii) frame feature extraction. In the first step, the input video are split into sampling the video frames, and are sampled with the intent of reducing the further computational overhead. In the second step, preprocessing of the sampling of video frames is carried out for noise elimination. Once the noise suppression is completed, this kind of frames leads to increased computational time, and this kind of frames can lead to the generation of high quality video summary, and two filtering algorithms known as MF and ASTF are utilized for noise elimination. At the end, Enhanced Recurrent Neural Network (ERNN) is proposed for the frame feature extraction. Once the preprocessed frames are chosen as the input video for this stage, it is sent through ERNN comprising of two layers, which includes the first layer of shot edge detection is used for the encoding of brief video sub shots cut from the actual video, and keyframe extraction carried out by the hidden state and every subshot becomes the input for the second layer in computing its confidence to act as a primary subshot, and then identical keyframe videos are eliminated in key shot, and at last, a video with best accuracy is summarized. In addition, experiments are conducted on two datasets (Thumb1K and TVSum50) and the excellent performance of ERNN is demonstrated in comparison with various benchmarked techniques for video summarization. The results of experiments are reported in terms of metrics such as Mean Average Precision (MAP) and F-measure (F1 score).

**KEYWORDS:** Convolutional Neural Network (CNN), Mean Filter (MF), Adaptive Spatial-Temporal Filter (ASTF), Mean Average Precision (MAP), Enhanced Recurrent Neural Network (ERNN), Video summarization, Keyframe.

## 1. INTRODUCTION

Recently, the ubiquity of camera devices has led to the mass capture of large number of videos and online sharing. On a daily basis, immense amount of video data inundates the social networking platforms on the Internet. It makes it convenient for the users to have access to video information. But, it also increases the time consumed in data browsing. Therefore, there is a urgent demand for an effective means of dealing with these massive video data [1].

The massive usage of video websites such as YouTube, Yahoo Video, and social networks such as Facebook, Google+ have been the influence behind the enormous rise of video contents all on the Web. Luckily, video summary can be of extreme help in this era of data explosion. It provides the video summary to the viewers by producing a short version of the video content. With the aim of managing the exponentially rising amount of videos being made available on the Internet and also for extracting useful information out of them, much focus have been shifted to video summarization, a technique whose objective is generating a brief summary of a video clip, so that the consumers are given an artificial and meaningful visual gist of video content [2].

In the recent times, approaches for automated video content summarization have gained prominent focus owing to its commercial viability particularly for home video applications. A short video summary, spontaneously, must focus on the video content and have very less redundancy when maintaining the balance coverage of the actual video. However, a video summary, must be diverse from video trailers where particular contents are deliberately concealed so that the intrigue of a video is increased manifold. Approaches in automated video summarization can be broadly classified into two important techniques, which include static storyboard summary and dynamic video skimming [3].

The former includes a set of fixed key frames made up of video shots, where as the latter is a much smaller version of video consisting of consecutive chosen video clips. Static storyboard permits browsing of video content nonlinearly by trading off the temporal growth of a video. On the contrary, dynamic video skimming, helps in preserving the time-evolving characteristic of a video through linear and continuous browsing of particular clips of video content according to a certain time length. In the case of both techniques, the suitable choice of video parts has a primary role to play in increasing the entropy information and the quality perceived of a video summary [4].

Video summarization approaches help in the development of concise versions of the actual input data by identifying the highly significant and relevant content existing within the video content. The video summaries obtained can be employed in different applications, such as movie (post-) production interactive browsing and search systems, providing the user with the capability of effective access to video streams[5].

Various approaches are different in terms of the content type utilized, the analysis carried out and the kind of video summary format. If the kind of the information used is considered, it may belong to common or domain based type (e.g., sports, news, movies and so on.) and also the information in addition to the video (extrinsic information, that a person provides). The extraction of the objects, events, perspectives and features are carried out by evaluating the modalities available for skimming out the spontaneous semantics from the video content[6].

The concised semantic content, which must to be added in the target summary is generally indicated to be a key for still key frames or a video extraction. This procedure may use temporal video segmentation, such that the derived key frames indicate a video section. The procedure of key frame extraction is also called as “key-framing”, or “story-boarding” or “static video summarization” [7].

A video skim is basically a video that is brief compared to the input video, called as “dynamic video summarization”. In the case of video summarization approaches having applications in post-production of movie, the benchmark technique uses key frame extraction and video skimming approaches. Generally, longer videos having more than one shot go through temporal segmentation, either by manual means or automatically by using shot detection algorithms. A saliency score for every frame is then computed and the very important frames are chosen to act as the key frames. Video clips present around every key frame are merged with the help of a fade-in fade-out approach for generating the summarized video skim [8].

Dynamic summary has a short segment of the video shots merged in a chronological series and similar to a brief version of the actual video content. Moreover, a video thumbnail, which is the first one seen by a person while browsing or searching for videos, could be considered to be a specific type of video summary at the greatest degree of abstraction, including just one single frame. Both static storyboard and dynamic video skimming can yield a user-friendly means for video browsing, and also be exploited in a broad array of applications, like activity recognition, event detection, saliency detection and video embedding [9].

Video summarization is a huge challenge due to its subjective nature - users possess their individual choices on the summaries. Different evaluators provide high inter-annotator agreement on the summaries of the same video content. Therefore, it is feasible to choose the significant and useful segments from the video clips, which can cater to a major segment of choices on an overall. In order to resolve this issue, unsupervised techniques [14] frequently selected frames or shots from videos using few manually chosen conditions like visual focus, representation and significance. But, manual criteria frequently do not succeed in making the different videos suitable on the web. Contrary to unsupervised ones, supervised techniques[15] made the system learn directly from man-made summaries on the ways of choosing the subsets, so that the evaluation metrics acquired from human perceived quality is satisfied. Even though at times efficient, they were dependent hugely on human annotations that is difficult to get.

Past few years, researchers have highlighted on the summarization of videos that are edited, whose preprocessing is done by editors, like news, TV program, video ads etc. Edited videos are generally done using short structures, and many of the shots are meaningful. For the summarization of this sort of videos, research expertshave focused on designing prototypes for exploiting the video's structural information (i.e., the association among the visual shots), and then choose the most characteristic components. A majority of these frameworks depend on low-level features such as appearance and motion [10]. Presently, the widespread usage of camera devices have led to the capture of excessive number of videos. Usually, they are created with pertinent shots and with no editing done. Therefore, these videos, represented as raw videos, are generally filled with repetitive and meaningless content. Many techniques for the summarization of raw video are object-oriented, highlighting much on explaining video content (i.e., what is exactly presentin the shot). Generally, the shots having significant objects are chosen into the summary [11].

However, the progress and wide spread usage of audio-visual capturing tools have made efficient approaches for dynamic video skimming to be the most sought after. Assume that a majority of individuals will get uninterested by an unedited and long-span home video lasting for hours together. A tool, which can automatically reduce the actual video when keeping most of the events preserved by focusing just on the significant content would be of immense help to many users. Approaches for dynamic video skimming are using expectation maximization (EM), singular value decomposition (SVD), motion model, utility framework, attention model, and semantic analysis. Several approaches are primarily based on visual information with the excluding the techniques such as where audio and linguistic information are also included with the aim of extracting the semantic meaning [12].

Practically, humans are considered best at information summarization and their cognition with words. When humans see a video, it is typical for them to provide a summarized form the video content and then share it with one another through words, and their statements can be considered to be the summary of the video in text format. Various individuals may talk about diverse contents, however the fundamental subject and semantic meaning still remains the same. Frames or shots having the highest relevance to the common expression taken from individuals are truly "summary worthy" as they depict the semantic information extracted from the video content and indicate what the individuals think about[13].

The earlier research work, which primarily depends on manually developed criteria or practically in economic human annotations, does not succeed in attaining good results often. It is to be noted that the side information corresponding to a video (e.g., enclosing text like topics, questions, descriptions, comments, etc) denotes a type of manually-selected semantics of video content. Even though this side information is precious for video summarization, it is not given due importance in the available techniques. The earlier work (i) temporal noise that happens in scenes under illumination or light variations, (ii) Convolutional Neural Network (CNN)'s inability to exhibit spatial in variance to the input data, shot edge detection are not performed automatically.

Making the best use the temporal dependency existing among the video frames or sub shots holds massive significance for the video summarization process. In order to get over these challenges, an Enhanced Recurrent Neural Network (ERNN) is introduced for video summarization, noise elimination is carried out with the help of two filtering algorithms, which include Mean Filter (MF) and Adaptive Spatial-Temporal Filter (ASTF). The proposed ERNN model comprises of three steps, which are (i) sampling of video frames, (ii) preprocessing, (iii) frame feature extraction. In the first step, input video are divided by sampling the video frames, and this is done for decreasing the computational overhead. In the second step, preprocessing is done on sampled video frames and carried out for noise suppression. Once the noise elimination is completed, this kind of frames leads to increased computational time, and they can produce superior quality video summary, and two filtering algorithms named MF and ASTF are utilized for noise elimination. In the last step, Enhanced Recurrent Neural Network (ERNN) is proposed for frame feature extraction. Once the preprocessing is completed, the frames are chosen to be input video for this step, and ERNN consists of two layers, where the first layer of shot edge detection is used for encoding short video sub shots sniped from the actual video, and keyframe extraction carried out by hidden state of every subshot becomes the input to the second layer for computing its confidence to become a prime subshot, and identical keyframe videos in key shot are eliminated, non-identical keyframe videos are provided as the result and thus an accurate video summarization is done.

The remaining section of the article is arranged as given: Section 2 provides the review on the relevant literature on video summarizationand the available techniques are briefly explained. The proposed technique on video summarizationemploying deep learning technique Enhanced Recurrent Neural Network (ERNN) in studied in section 3. Section 4 provides the experimental results and discussion on the proposed work. The conclusion and the intended future work is studied in Section 5.

## 2. LITERATURE REVIEW

Recently, video summarization is extremely popular in video websites such as YouTube, Yahoo Video, and social networks such as Facebook, Google+ have encouraged the tremendous increase in video contents all over the Web. A Video Summarization can be defined as a summary that characterizes a high-level perspective on the actual video clip and can be helpful in video browsing and search systems. Various techniques are utilized for choosing the key frames. A survey on the literature is carried out in this section on video summarization, which is then followed by analysing the their audio, video and textual contents. The advantages and drawbacks of different video summarization techniques are clearly explained in this review section.

Zhang et al [16] introduced a Context-Aware Video Summarization (CAVS) framework, which uses the sparse coding technique empowered with generic sparse group lasso for learning a dictionary on the video features and a dictionary composed of spatiotemporal feature correlation graphs. The sparsity guarantees that the a major portion of the meaningful features and associations are preserved. The feature correlations, indicated using a dictionary comprising of graphs, show the way in which motion regions associate with one another on a global scale. During the processing of a novel video segment using CAVS, both the dictionaries get updated when online. Especially, CAVS goes through through all the video segments to decide whether the learned dictionaries help in sparse representation of new features combined with the feature correlations. Otherwise, the dictionaries get updated, and then the respective video clips are included in the summarized video stream. The results acquired on four openly available data sets, mainly consisting of surveillance videos and a brief number of other videos present online, reveal the efficiency of the novel technique.

Li and Merialdo [17] suggested an Optimized Balanced Audio Video Maximal Marginal Relevance (OB-MMR) for the video summarization process. This algorithm is desirable for summarizing both individual and multiple videos. OB-MMR is accomplished through parameters' optimization in Balanced Audio Video Maximal Marginal Relevance (AV-MMR), called as the balancing aspect between audio and visual information existing in the video, but also the significance of facial and audio changes among the audio parts of diverse categories. Hence, OB-MMR attains relatively superior result compared to the earlier algorithms, such as Video-MMR and Balanced AV-MMR. Also, it is feasible to choose the optimized parameters for every type of videos, resulting in potential automated algorithms designed for video summarization during the big-scale experiments carried out in the future.

Al-Musawi and Hasson [18] presented a Histogram of Oriented Gradient and Correlation Coefficients approaches for artificial noisy video summarization. In applications involving real-time like surveillance applications, lighting variations or shadowing for motion objects may happen in surveillance video content. Several video summarization techniques attempt to build the video summary based on the supposition that noise or illumination values stay constant all through the video frames. This algorithm has been used on the newly introduced multi-model dataset developed by concatenating the actual information and the dynamic man-made information. This dynamic information is constructed with the help of Random Number Generator function. The experiments carried out on this dataset reveals the efficiency of the novel algorithm in comparison with the classical dataset.

Yuan et al [19] introduced a Deep Side Semantic Embedding (DSSE) for the generation of video summaries by exploiting the publicly accessible side information. The DSSE builds a latent subspace by connecting the hidden layers belonging to the two uni-modal autoencoders, which insert the video frames and side information, correspondingly. Particularly, through the interactive reduction of the loss in semantic relevance and the feature reconstruction loss incurred in the two uni-modal autoencoders, the relative common information between video frames and side information could be learned fully. Hence, their semantic relevance can be measured with more efficiency. At last, semantically informative portions are chosen from videos by reducing their distances with the side information present in the latent subspace built. In addition, experiments are carried out on two datasets which include Thumb1K and TVSum50 datasets and it is proven that the performance of DSSE is much superior compared to different benchmark techniques for video summarization.

Mygdalis et al [20] studied about a Subclass Support Vector Data Description (SSVDD) for video summarization, which functions on a video segment. Automated video segment selection depending on a learning step, uses primary video portion concepts. A hierarchical learning approach is designed, which comprises of two stages. In the first stage, an unsupervised mechanism is carried out to decide on the important types of video segments. The next step involves a supervised learning mechanism conducted for all the primary kinds of video segment in an independent manner. For the latter scenario, as the primary training examples are only present, the problem is treated to be a one-class classification problem. With the purpose of considering the subclass information, which may show up in the video segment varieties, a new format of the SSVDD is introduced, which uses the subclass information during its optimization step. The proposed technique is used in three Hollywood movies, in which the performance comparison of the novel Subclass SVDD (SSVDD) algorithm is done with those of relevant techniques. The results of experiments depict that the usage of both hierarchical learning and the novel SSVDD technique is a major contribution to the ultimate classification performance.

Liet al [21]studied about an ensemble summarization model both for edited video and unprocessed video summarization. On the whole, the research work can be classified into three types: 1) Four frameworks are developed to acquire the characteristics of video summaries, i.e., having significant individuals and objects (saliency), representative to the video content (representativeness), no identical key-shots (versatility) and evenness of the storyline (storyness). Particularly, these frameworks are suitable for both edited videos and unprocessed videos. 2) An extensive score function is devised using the weighted combination of the above stated four frameworks. It is to be observed that the weights of the four frameworks in the score function, represented as property-weight, are learned in a supervised fashion. In addition, the property-weights are learned for both edited videos and unprocessed videos, correspondingly. 3) The training set is built with both edited videos and unprocessed videos to balance for the the deficit of training information. Especially, all the training videos are encompassed with a pair of mixing-coefficients that can minimize the structural disorder in the training set that the uneven mixture has caused. In addition, proposed framework is tested on three datasets, inclusive of edited videos, brief unprocessed videos and lengthier raw videos. The results of experimentshave demonstrated the efficiency of the newly introduced model.

Cong et al [22] introduced a Dictionary Selection Model (DSM) for summarizing the end user videos. DSM that uses sparsity consistency, with a dictionary made up of key frames is chosen such that the reconstruction of the actual video can be efficiently done using this symbolic dictionary. An effective global optimization algorithm is presented to resolve the dictionary selection framework with the convergence rates as (where is the iteration counter), as opposed to classical sub-gradient descent techniques. Since a video sequence can be summarized using any number of key frames provided, the technique yields a powerful solution for video summarization. It renders a flexible solution for both key frame extraction and video skim generation, as one can choose any number of key frames for the actual videos' representation. Experiments carried out on a manually labeled standard dataset and comparison analysis with the benchmark techniques show the benefits of the algorithm. These results prove the benefits of DSM technique.

Ainasoja et al [23] studied about a Bag-of-Words (BoW) for Keyframe-based video summarization. Keyframes indicate salient and different contents present in an input video and a summary is created through the temporal expansion of the keyframes to key shots that are concatenated to a sequential dynamic video summary. In this technique, keyframes are chosen from videos, which depict semantically identical content. In the case of scene detection, an ordinary yet efficient dynamic improved version of a video BoW technique that yields over segmentation (high recall) for keyframe selection is proposed. In the case of keyframe selection, two efficient schemes including: local region descriptors (video content) and optical flow descriptors (motion content) are investigated. In addition, many intriguing observations are listed. 1) Whenvideo segments (visually identical content) can be efficiently identified by region descriptors, optical flow (motion variations) yields good keyframes. 2) But, the desirable parameters of the motion descriptor based keyframe selection differ between videos and average performances stay less. In order to prevent more complicated processing, a human-in-the-loop step is introduced where the user chooses keyframes generated by the three popular techniques. 3) Human aided and learning-independent technique attains much better accuracy compared to learning-based techniques and for several videos, it is quite equivalent with the mean human accuracy.

Xu et al [24] presented a Scene Cluster Model (SCM) for scene video query and summarization. It aims at developing a novel model for distributed multiple-scene global comprehension that carries out the clustering of surveillance scenes using their capability of describing individual behaviors and also finds the subset of activities that are shared against the shot unique within every cluster. The research work works with variable and piecewise interscene association through the semantic clustering of scenes based on the correspondence between semantic activities and preferentially shares the activities across video shots within the clusters. In addition, the means of using this organized representation of several scenes to boost the typical surveillance jobs, including scene activity interpretation, cross-scene query-by-example, behavior classification and summarization to be generalized to several scenes with minimal supervised labeling demands is shown. Primarily, by finding the associated shots and shared activities, it is feasible to attain cross-scene query-by-example (as opposed to a common within-scene query) and for annotating the behavior in a new shot with no labels that is essential for deploying the surveillance system scale practically. At last, it can yield video summarization skills, which specifically use the redundancy both within and across video shots by exploiting the proposed multiscene framework. In every case, it also demonstrates the way in which multiscene model helps in the enhancement of an ensemble of standard single-scene framework and a simple model of all thevideo shots.

Sigari et al [25] suggested a Fuzzy Inference System (FIS) fast highlight detection and scoring technique for broadcast soccer video summarization. It employs an on demand feature extraction and a FIS. The discussed technique divides the video to highlight and analyze their content employing an on-demand feature extraction technique. After this, every highlight is given a score with the help of a FIS as per the content analyzed. The score assigned decides the significance of the events happened in the highlight. This technique is helpful for scalable video summarization. The proposed technique for on-demand feature extraction is typically a heuristic model of attention control, which decreases the computational complexity of the algorithm to a larger extent. Besides, FIS provides a simple and reliable solution for content analysis. The results of experiments show that the proposed technique is rapid and processes nearly 130 frames per second on a personal computer. Moreover, objective and subjective analysis validate that the proposed technique yields superior quality results for highlight detection, scoring and video summarization.

## 3. PROPOSED METHODOLOGY

In this technical work, an Enhanced Recurrent Neural Network (ERNN) algorithm is proposed for video summarization. The discussed video summarization comprises of three important steps. (i) sampling of video frames, (ii) preprocessing, (iii) frame feature extraction. In the first step, the input video are split into video frames by sampling, video frames are sampled with the aim of reducing the computational complexity. In the second step, preprocessing the sampling number of video frames is used for noise elimination. Once the noise suppression is done with this kind of frames increasing the computational complexity, they can also yield superior quality video summary, two filtering algorithms known as Mean Filter (MF) and Adaptive Spatial-Temporal Filter (ASTF) used for noise elimination. At last, Enhanced Recurrent Neural Network (ERNN) is presented for frame feature extraction. Once the preprocessing is completed, the frames are chosen in the form of input video for this step, ERNN consists of two layers, shot edge detection in which the first layer is used for encoding small video subshots snipped from the actual video, and keyframe extraction carried out by hidden state of every subshot becomes the input to the second layer for computing its confidence to become a key subshot, and after this, identical keyframe videos are eliminated in keyshot, non-identical keyframe videos are obtained and ultimately, the summarization of an accurate video is obtained. The functional flowchart of proposed ERNN technique for video summarization is depicted in figure 1.

### *A. Frame Sampling*

The sampling of video frames are done with the aim of reducing the computational complexity incurred further. In order to prevent frame comparisons made redundantly, frames must be sampled with suitable technique. The sampling depends on the inference that a visual redundancy exists among particular number of frames per second. The most popular mechanism is uniform sampling using a constant rate of sampling. The input video cliphas to be split into frames prior to the processing step. Rather than processing for the complete set of input frames, a subset of frames is produced with even sampling. Then the subset is selected on the basis of predefined sampling rate, a significant parameter, which has a direct impact on the quality of summary finally [26].

The sampling rate is based on the frame rate corresponding to the input video. A lesser sampling rate resulting in missing a key-frame, and a higher sampling rate can produce a higher number of repetitive frames to show up in the output with the key-frames retained. The sampling rate utilized in the available techniques is one frame per second. Its performance is quite good with high frame rated videos such as sports visuals and is not desirable for low frame ratedvisuals such as cartoon videos where the frames are different even between the frames, which spansfor a one second owing to a sudden variation in the video content. In order to deal with this problem, the sampling rate selected in this stage is about two frames per second. The video shots are made from input frames in which every video shots includes frames corresponding to one second in the input video clip. Then the frames are sampled by choosing the first and middle frame of every segment [27].

Uniform sampling is one of the most popular techniques used for frame sampling. The key is to choose every $k^{th}$ frame from the video where the videolength decides the value of k. A popular preference of length for a summarized video is 5% to 15% of the actual video, which implies that every 20th frame in the 5% or every 7th frame in case of 15% length of the summarized video is selected.

### *B. Preprocessing*

Noise indicates that the pixels in the videos depict diverse values of intensity values rather than the actual pixel values that the camera captures. Noise removal algorithm refers to the process of eliminating or decreasing the noise from the video frame. Many techniques for video summarization do not consider temporal noise that happens in the visuals under varyingluminance or light. Noise is a popular issue in digital cameras owing to which few errors may be observed in one of the two sensor cameras, or no clarity in some sensor information with noise exposure. The video summarization, which depends on motion detection mixed with noisy video generates incorrect motion object vectors.

For practical applications, the feature extraction process in computer vision and image processing has to exhibit resilience towards variations in brightness or illumination or to frame disturbances like noise or haziness. The luminance or brightness variations of few points between sequential frames in video segments frequently happen because of the changes in parameters of various video cameras, or moving of objects from one segment to another section of the visual can be varied with diverse lighting conditions. These problems will result in processing the video stream inaccurately. Nearly all the video summarization techniques for one single stationary camera or multi-camera video do not consider the variations in illumination or to the available noise signals that occur in few video frames. They are based on the supposition that noise or luminance values are constant all through the video frames [18].

In order to resolve these problems, two filtering algorithms are presented for noise elimination for video frames. The smoothening of videos is a necessity which helps in the noise removal and for this, the best filters or popular filters are used in many video summarization applications. It is resolved with the help of various algorithms. In accordance, noises are discerned with neighboring information and are eliminated with the help of the best filtering methods with no effect on the actual video; it is emphasized for the video quality improvement for analysis. In order to boost the video quality, filtering algorithms such as Mean Filter (MF) and Adaptive Spatial-Temporal Filter (ASTF) are used. The frames are then sampled by choosing the input frames/videos for eliminating the noise using both filtering algorithms.

### *1. Mean Filter (MF)*

It is an ordinary sliding-window spatial filter, which substitutes the middle value in the window with the mean value of all the pixels present in the window. The arithmetic mean filter is defined to be the average of all pixels existing within a local site of frames.

Processing these kind of frames will lead to an increased computational complexity, and also this kind of frames can result in low quality summary being generated. Therefore, these kinds of frames are neglected in pre-processing step i.e. prior to the construction of the ranked graph. It must have a higher normalized variance between histogram bins as it adopts homogeneous distribution. Therefore, the mean and standard deviation of variance vector are computed on all the frames present in a video. Therefore the threshold ($T$) is adaptively calculated as:

$$T = Mean + \beta * Standarad\ Deviation$$

Where $\beta$refers to the user specified positive constant. In case, the variance an image of a video is greater than that threshold ($T$).

In the case of mean filtering, the window size $L_k$having the highest sample variance is the filter duration utilized for the samplegiven; the window size giving the least variance, is chosen. Making use of the window having the highest sample variance for mean filtering guarantees best elimination of surging peaks and impacts of camera/object motion activity in regions of peak activity. The technique needs that the set of probable filter durations must be defined before processing satisfying the requirement of defining the variance-associated parameters or thresholds is therefore omitted.

Original environments are frequently subjected to unwarranted conditioins referred as noise. Gaussian noise is regarded the most common kind of noise that follows normal distribution. Noisy signal distorted by Gaussian noise can be defined with the expression:

$$g(x,y) = f(x,y) + \sigma * randn(size(f)) + \mu$$

Where $g(x,y)$refers to the signal with additive Gaussian noise; $f(x,y)$indicates the actual signal; $\sigma$stands for the standard deviation; $\sigma^2$refers to the variance; and $\mu$indicates the mean. $randn()$ for the generation of random numbers following a Gaussian normal distribution [18]. The probability density function for a Gaussian distribution having mean $\mu$and variance $\sigma^2$ can be expressed as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

The standard deviation or the variance gives the power of Gaussian noise signal. The conventional Gaussian noise generator yields the same approximated linear power for each of the video frames given. In dynamic scenarios, noise features generally vary with time, and it is important to employ a simulation method for making changes to the noise features. The resultant output of the algorithm is a noisy video having frames including non-linear or dynamic noise components. Mean Square Error (MSE) is applied as an error metric between the actual video signals and the noise-mixed video signals.

Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator (involves a process for the estimation of an unmonitored quantity) provides a measure of the average valueof the squares of the errors, implying the average squared difference between the estimated values and the original one. MSE is basically a risk function, associated with the assumed value of the squared error loss [28].

### 2. *Adaptive Spatial-Temporal Filter (ASTF)*

Noise reduction algorithm depends on an ensemble of a spatial Wiener and temporal filters and then improving the resultant output of the ensemble. Spatial Wiener filter is a popular filter as far as spatial noise filtering is concerned. This filter is capable of eliminating the noise with efficiency. Regretfully, it can result in blurring particularly in the region of lesser levels of noise. To prevent this blurring, the filtering operation in the regions having highlevels of noise are increased and reduced in the regions having lesser degrees of noise. The filtering operation is controlled by using the mask size of 5×5 in the regions containinggreater degrees of noise and less number of image features. On the other side, the mask size of 3 × 3 is applied in the regions containing less degree of noise and several image features. In the regions containinggreater degrees of noise, the mask size of 5 × 5 will perform better than the mask size of 3 × 3 in terms of noise elimination.

But, using the mask size of 5 × 5 frequently results in blur effectparticularly in the regions containing less degree of noise. In order to merge the benefits of both masks, the threshold ($T1$) is used to decide the mask that must be used as per the noise level. More exactly, the mask size of 5 × 5 is applied in the regions having highdegree of noise with the aim of reducing the highest degrees of noise; but, using the mask size of 3 × 3 in the regions having low levels of noise has a predominant role to play in the preservation of the image edges.

$$f_n(i,j) = \begin{cases} \mu > T1 & Mask\ size\ of\ 5 \times 5 \\ otherwise & Mask\ size\ of\ 3 \times 3 \end{cases}$$

where $\mu$stands for the amount of noise and $T1$indicates atest threshold (optional).

Now assume $f_n$represent the $n^{\text{th}}$ current frame that is filtered using spatial Wiener filter. The temporally filtered $n^{\text{th}}$ frame $T_n$can be formulated as

$$T_n(i,j) = \rho . f_{n-1}(i,j) + (1-\rho). f_n(i,j)$$

where

$$\rho(i,j) = \frac{\sigma_n^2(i,j)}{\sigma_r^2(i,j) + \sigma_n^2(i,j) + \sigma_r(i,j).\sigma_n(i,j)}$$

and $f_{n-1}$indicates the earlier frame that is filtered with spatial Wiener filter.

In order to improve the temporally filtered frame $T_n(i,j)$, the motion field below is selected to regulate the procedure of noise elimination as per the areas information (stationary or movable):

$$M(i,j) = \begin{cases} 0 & |f_n(i,j) - T_n(i,j)| < T, \\ 1 & otherwise, \end{cases}$$

The above motion field can find the following aspects:

(i) If$M(i,j) = 0$, the changes over $f_n(i,j)$ and $T_n(i,j)$ at the spatial position $(i,j)$ will be close to zero; otherwise said, $f_n \approx T_n$

(ii) If $M(i,j) = 1$, the changes over $f_n(i,j)$ and $T_n(i,j)$at the spatial position $(i,j)$will be substantial. Therefore, the noiseelimination must be terminated in this scenario; else, the important information of the image in this region will be eliminated that in turn, will result in a blurry image.

### *C. Feature Extraction*

Feature Extraction can detect patterns, and these are redundant media segments, which can either be detected through the comparison of chunks over the media dimensions or with templates using cross-correlation or auto-correlation, and media content summarization. The features of the obtained key frames could be color, edge or motion background, and specifies the boundary existing between the objects overlapping. This implies that in case the edges in an image can be detected with accuracy, all the objects can be found and the measurement of fundamental characteristics like area, perimeter, and shape can be done[30].

Edges specify the boundaries between the areas in an image, and is useful in the segmentation and object identification. The image edges exhibit no variations in the gray value distribution. The edge matching rate is utilized for matching the edges of neighboring frames to eliminate the recurring frames.

The formula for computing the edge matching rate is as below:

$$P(f_i, f_{i+1}) = s/n$$

where, $n = Max(f_i, f_{i+1})$

$$S = \sum_i^m \sum_j^n h(i,j)$$

Where $m$and $n$specify the height and the width of the image, $f_i$and $f_{i+1}$indicates the sequential frame and $h(i,j)$indicates the difference between the $f_i$ and $f_{i+1}$.

The technique for getting the values (features) meaningful and non-repetitive is associated with reduction in dimensionality. If the input data to an algorithm is very big for processing and it is doubted to have redundancy (redundant behavior of images shown as pixels), then it can be modified into a minimized set of features also referred as a Feature Vector. The features chosen are found to have the relevant information taken from the input data, such that the necessary task can be carried out with the help of this minimized representation rather than the entire initial data of video.

### *1. Frame Feature Extraction*

Frame feature extraction has a critical role in a keyframe extraction algorithm that has a direct impact on the algorithm performances.

The visual feature that is optimum for video processing applications, must meet many important conditions[31]:

- Reliability: A frame feature must exhibit large in variance for the same frame content under different kinds of changes, like format conversion or content editing.
- Diverseness: Features obtained for diverse video frames must be definitely unique.
- Compactness: A frame feature must have a inconsequential size, in comparison with the data size of the actual video frame.
- Lesser complexity: The algorithm for the feature extraction must incur the least computational complexity.

## 2. *Shot Edge Detection and Keyframes Extraction*

**Shot Edge Detection:**One significant step for preprocessing for several video analysis tasks is the extraction of shot boundaries detection. The objective of this step is to partition the input video into segments called shots, where a shot is specified as a set of video frames captured contiguously using a single camera.

**Keyframes Extraction:** A keyframe is a compact representation of an identified video shot using a single frame. Generally, a keyframe is selected to be the first or the final frame of a shot, even though few techniques make use of any single frame within a shot. To begin the extraction process, the first frame is called as a key frame. After this, the frame difference is computed between the present frame and the last key frame extracted. In case the frame difference meets a particular threshold criteria, then the present frame is considered as key frame. This procedure is iterated for every frame in the video.

In this technical work, Enhanced Recurrent Neural Network (ERNN) for Shot Edge Detection and Keyframes Extraction is proposed.

### *Recurrent Neural Network (RNN)*

Recurrent Neural Network (RNN) belongs to the group of artificial neural networks where the correlations between nodes establish a directed graph along a temporal series. Obtained from feed forward neural networks, RNNs can make use of their internal state (memory) for processing variably lengthened series of inputs. This renders them suitable for tasks like unsegmented, connected handwriting recognition or speech identification.

The “recurrent neural network” term is employed with no difference to specify two broad categories of networks having an identical generic structure, which include finite impulse and infinite impulse. A finite impulse recurrent network is a directed acyclic graph, which can be unwound and substituted with a tight feed forward neural network, while an infinite impulse recurrent network is basically a directed cyclic graph, whose unrolling cannot be done [32].

A standard RNN is built by expanding a feed forward network having an additional feedback connection, such that it can help in sequence modelling. In reality, it can comprehend the input sequence $(x_1, x_2, \dots, x_n)$ into another sequence $(y_1, y_2, \dots, y_n)$repetitively using the equations below:

$$h_t = \phi(W_h x_t + U_h h_{t-1} + b_h)$$

$$y_t = \phi(U_y h_t + b_y)$$

where $h_t$refers to the hidden state, t indicates the t-th time step, $\phi$refers to the activation function, and W, U and b stand for the training weights and biases. By principle, the standard RNN must function effectively for sequence modeling. But, it is pretty difficult to get the gradient vanishing problem trained. After this, LSTM is developed to deal with this problem, the most common version of standard RNN. Particularly, it is expanded from standard RNN having an additional memory cell, used for selective memorizing of the earlier inputs [33].

Practically, there are different versions of LSTM, and they are identical to one another. Comprehensively, the computation of hidden state $h_t$ and memory cell $c_t$ is expressed as:

$$i_t = \sigma(W_{ix} x_t + U_{ih} h_{t-1} + b_i)$$

$$f_t = \sigma(W_{fx} x_t + U_{fh} h_{t-1} + b_f)$$

$$o_t = \sigma(W_{ox} x_t + U_{oh} h_{t-1} + b_o)$$

$$g_t = \phi(W_{gx}x_t + U_{gh}h_{t-1} + b_g)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \phi(c_t)$$

where $\sigma$represents the sigmoid function, and all the $W_s, U_s, b_s$indicate the training weights and bias. Also, $i_t$, $f_t$ and $o_t$refer to three gates, which hold the highest importance to LSTM. Strongly, the input gate has the control if the current input $x_t$, has to be recorded or forgotten, gate ft determines whether to omit the earlier memory cell $c_{t-1}$, and the output gate $o_t$decides the information in the present memory cell $c_t$moved to the hidden state $h_t$.

***Enhanced Recurrent Neural Network (ERNN)***

Using the temporal dependency among video frames or subshots is quite essential for the video summarization job. In practice, RNN offers good performance at temporal dependency modeling, and has attained remarkable performance in several video-centric jobs, like video captioning and classification. But, RNN does not the adquate capability to deal with the video summarization task, as classical RNNs, LSTM included, can only work with small videos, whereas the videos in the summarization process are generally much longer. In order to deal with this issue, an Enhanced Recurrent Neural Network (ERNN) is proposed for video summarization. The inspiration behind the design of an improved RNN is to increase its potential to use the long-limit temporal dependency of the videos. In fact, its real inspiration is the operation of one-dimensional convolution. Figure 2 illustrates the forms of improved convolution and RNN. Factually, it is a comparison analysis between improved RNN and single long RNN, the filters in (a) (i.e., $w$ and $w'$) are replaced with low length RNNs. In comparison with (b), improved RNN offers better efficiency in exploiting the longer temporal dependency, at the same time, the computation operations is considerably minimized. As shown in the first layer of Figure 2(a), a one-dimensional filter $w$ is used for exploiting the sequential information by carrying out convolutional operations on the input sequence $x$:

$$y = w * x$$

where $y$represents the output sequence, and $*$ indicates the convolutional operation. It can be found that even though the filter $w$ is much shorter compared to $x$, at every time step, it functions suitably on a subsequence of $x$ and produces a much short length sequence $y$as the output. Especially, in case the convolution stride is fixed as $n, |y|$ is onlu $1/n$ of $|x|$, where$|.|$represents the length of the sequence. In addition, in the second layer, one more filter $w'$ is used on $y$, and produces short length sequence $y'$as the output. Generally, many filters can also be used for higher layers, till the ultimate output is produced. After this, the improved structure is created, and the lengthier sequence $x$ is processed with multiple improved small filters.

Influenced by this, a identical improved structure is constructed for RNN. In fact, the filters in various layers of Figure 2(a) are substituted with small RNNs, and the convolutional operation is similar toconsecutive processing of multiple shorter subsequences that are sniped from the longer input sequence with or with no overlap (the subsequence length equals to the filter RNN's length). Particularly, the RNN present in the first layer makes use of short-range temporal dependency, and higher layer RNNs capture the longer ones. By intuition, the long-range temporal dependency is acquired by the improved format of various small RNNs. Also, in comparison with single RNN operating directly on the long sequence as shown in Figure 2(b), improved RNN can not just decrease the information loss found in long sequence modeling, which is elaborately studied for the video summarization process.

y'
W'
y
W
Input
x
y
RNN
Input
x

**(a) Enhanced One-dimensional Convolution** **(b) A Single Long Recurrent Neural Network (RNN)**

**Figure 1: The blocks of Enhanced Convolution and RNN**

***Video Summarization with ERNN***

The videos for summarization generally have long timeframes, i.e., nearly many thousand frames. In addition, the video composition is apparently made up of layerssuch that the frames create the subshots and those subshots create the video. Hence, the improved RNN is very suitable for the video summarization process. In this section, ERNN is designed for the video summarization process.

Figure 3 illustrates the framework of the newly introduced technique ERNN for the video summarization process, it has two layers, where the first layer is actually a LSTM and a bi-directional LSTM (forward and backward) forms the second layer. The two layers uses the intra-subshot and inter-subshot temporal dependency, correspondingly, and the resultant output of the second layer is used for the prediction of the confidence of every subshot to be chosen into the summary. The details on summarization of the video with ERNN is given as below.

First, the frame sequence $(f_1, f_2, \dots, f_T)$ is divided into multiple subsequences, represented as subshots $(f_1, f_2, \dots, f_s), (f_{s+1}, f_{s+2}, \dots, f_{2s}), \dots., (f_{m.s+1}, f_{m.s+2}, \dots, f_T)$, where $f_i$indicates the feature of frame $i, T$ stands for the overall number of frames in the video, $m$refers to the number of subshots, and $s$indicates the length of every subshot. In practice, in case the final subshot is smaller in length compared to$s$, then it is concatenated with zeros. After this, the subshots are given as input to the first layer LSTM, expressed as below:

$$\tau_i = LSTM(f_{i.s+1}, f_{i.s+2}, \dots, f_{(i+1).s})$$

where $LSTM\ (.)$ is the short name for RNN Equations, $\tau_i$represents the final hidden state of the i-th subshot. In fact, the short temporal dependency in the subshot is acquired by $\tau_i$. Therefore, it is considered to represent the i-th subshot.

Subsequently, the sequence $(\tau_1, \tau_2, \dots, \tau_m)$is taken as the input to the second layer. Like it is stated above, a bi-directional LSTM acts as the second layer. In fact, bi-directional LSTM comprises of a forward LSTM and a backward LSTM. The primary difference between them is that the backward LSTM works in an inverse manner. Hence, the computation in the second layer is expressed as:

$$h_t^f = LSTM\left(\tau_t, h_{t-1}^f\right)$$

$$h_t^b = LSTM(\tau_t, h_{t+1}^b)$$

where $h_t^f$ and $h_t^b$refer to the t-th output hidden state of forward LSTM and backward LSTM, correspondingly.

At the end, the output of the second layer is used for predicting the confidence of a particular subshot to be chosen into the video summary. It is derived as:

$$p_t = softmax\left(\tanh\left(W_p\left[h_t^f, h_t^b, \tau_t\right] + b_p\right)\right)$$

where $W_p$ and $b_p$refer to the parameters that have to be learned. The softmax function is used for limiting the sum of the elements in $p_t$ to have the value 1. In fact, $p_t$refers to a two-dimensional vector, and all elements specifies the probability of the t-th subshot being a key or non-key. It can be noticed from the Equation above that $p_t$is unitedly decided by the hidden state of forward and backward LSTM, i.e., $h_t^f$ and $h_t^b$, along with the t-th subshot $\tau_t$representation. It is due to the fact that for subshot t, $h_t^f$and $h_t^b$acquires the front and back temporal dependency, correspondingly, and $\tau_t$has the intra-subshot dependency. Each of this information is extremely necessary for deciding whether to choose the subshot t into the summary.

ERNN is trained on an end-to-end manner. With the manual generation of the reference summaries, the parameters in ERNN are learned using:

$$\ominus = \arg\min_{\ominus} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{m^{(i)}} L\left(p_t^{(i)}, g_t^{(i)}\right)$$

where $N$refers to the number of videos present in the training set. $m^{(i)}$indicates the number of shots present in video i. $L(.)$refers to the loss function, which provides a measure of the cross-entropy between the generated probability distribution $p_t^{(i)}$ and the ground truth $g_t^{(i)}$. In fact, $g_t^{(i)}$refers to a binary vector (specifies if the subshot is a key or not) or decimal vector (depicts the confidence of the subshot to act as a key subshot).

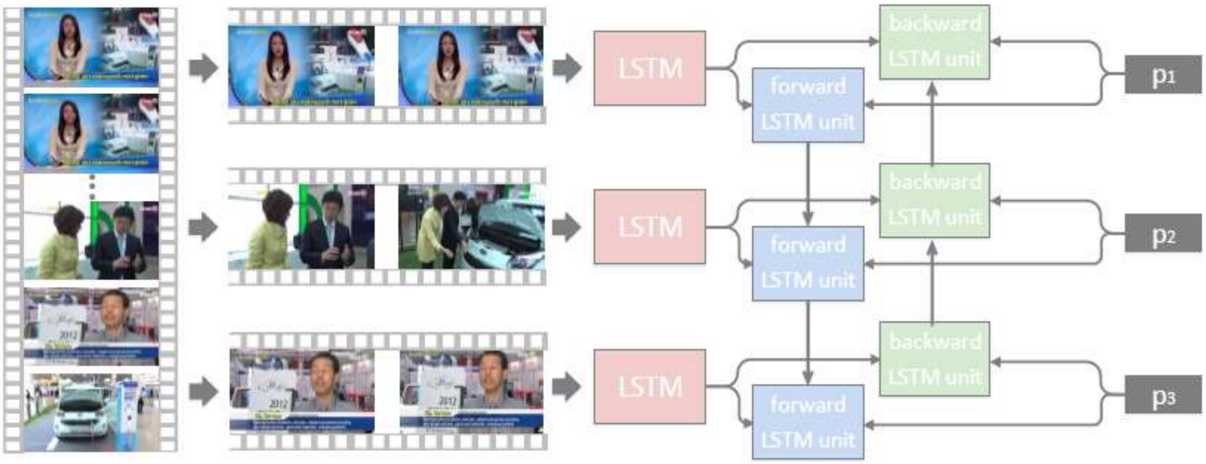


**Figure 2: Architecture of the ERNN Proposed technique for Video Summarization**

**Video Thumbnail Selection:** Video thumbnail can be considered to be the most compact stable video summary as it has to explain the video information present in one image. Therefore, few experiments are carried out for video thumbnail selection. Once the video frame level semantic relevance scores have been acquired, the frames can be ranked by their score numbers and the frame having the greatest semantic relevance score can be taken to be the video thumbnail.

**Dynamic Video Summarization:**For the generation of a dynamic video summary having length $l$, in the first step, a video segmentation algorithm is used for getting the video shots, and in the next step, the shot-level semantic relevance scores is computed by calculating the mean of the frame level semantic relevance scores within every shot.

### *3. Similar Keyframes Elimination*

It has been commonly noticed that a video generally includes few useless frames like entirely dark frames, completely white frames, faded frames. As these frames usually depict considerable difference from a typical frame, it is very much possible that they are chosen to be key-frames. Once a frame is identified to be a key frame, the computation of standard deviation of pixels in the frame is done. In case the standard deviation is quite less (approaching zero) then that frame is taken to be an useful frame and is omitted. When a key frame is chosen, its comparison is done with the next candidate frame. The procedureis performed again and again for every candidate frame. Once the group of key frames are generated, the redundancy is decreasedfurtherby eliminating those key frames that exhibit much similarity with one another. Once the extraction of keyframes are done, there can be identical keyframes, which show up at various temporal positions in the video. Identical keyframes are removed, and just the most useful ones are retained.

## 4. RESULTS AND DISCUSSION

The proposed ERNN framework for video summarization is tested on two datasets, which include Thumb1K and TVSum50 individually. The results of experiments are computed in terms of metrics such as Mean Average Precision (MAP) and F-measure (F1 score). Each video has been tested with its original length. MATLAB software is utilized for testing in this research. In this results and discussion section, the comparison analysis between three approaches including proposed ERNN system, available CNN and DNN results are carried out.

### *Datasets*

**Thumb1K [34]:** Thumb1K comprises of 1037 query-video pairs gathered from Bing. The dataset yields nearly 20 key frames in the form of representative thumbnails for every video, and these candidate thumbnails are obtained with the help of a representative attributes based technique. Each of the candidate thumbnails are marked using five diverse scores, which include Very Good (VG), Good (G), Fair (F), Bad (B), and Very Bad (VB).

**TVSum50 [35]:** TVSum50 includes 50 videos that are taken from YouTube in 10 groups specified in the TRECVid Multimedia Event Detection (MED). The dataset yields the title of the video and a significant score of 1 (not important) to 5 (very important) to each one of shots that are of even-length (2s) for the entire video. The frame level important scores get the same label like their relevant shots and 20 diverse important scores are labeled with the help of 20 diverse individuals for every video shot.

### *Experimental Settings*

**Shot Segmentation:**For the generation of video summary on TVSum50 dataset, at first, a video is temporally segmented into disconnected intervals employing KTS, which is a kernel-based change point detection algorithm extensively applied in video summarization tasks.

**Baseline Methods:**Even though the technique can be applied on both video thumbnail selection and dynamic video summarization, Thumb1K just yields nearly 20 visual characteristic and elaborate candidate thumbnails with none of the actual videos, few video summarization technique are unsuitable for this dataset as video summarization is developed on the entire video stream.

**Evaluation Metrics:**The video thumbnail selection is evaluated interms of two conditions, including HIT@1 computing the hit ratio for the first chosen thumbnail and Mean Average Precision (MAP) computing the mean precision for each of the candidate thumbnails. The MAP value is computed using the expression

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Preceision(R_{jk})$$

Where query set is specified by Q, for the j$^{th}$ query-video pair, $m_j$ positive thumbnails exists, $Preceision(R_{jk})$indicates the mean precision at the position of retrieved k$^{th}$ positive thumbnails. HIT@1 and MAP in two diverse scenarios, which include fixed thumbnails having VG score in the form of positive samples and fixed thumbnails having VG or G score in the form of the positive samples.

The quality of generated video summary by several human annotations [34]. Particularly, with a newly introduced summary $S$ and a user summary $B_i$given by the i$^{th}$ annotator, the precision $p_i$ and recall $r_i$ is computed, as per the temporal overlap between the two. So the pair wise F1 metric is formulated as below

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2p_i r_i}{p_i + r_i}$$

Where N refers to the number of user summaries for each video, and N is fixed at 20 in TVSum50 dataset. The results of video summarization is evaluated by the mean F1 score of all of the video streams.

### *Video Thumbnail Selection*

The performance of ERNN technique is first evaluated in video thumbnail selection step, Table 1 provides the summary of the MAP scores, the HIT@1 results are provided in Figure 4.

**Table 1: The MAP of various techniques for Video Thumbnail Selection**

| Technique | MAP(VG) | MAP(VG&G) |
|---|---|---|
| DNN | 56.1 | 76.1 |
| CNN | 57.6 | 77.4 |
| ERNN | 60.2 | 79.5 |

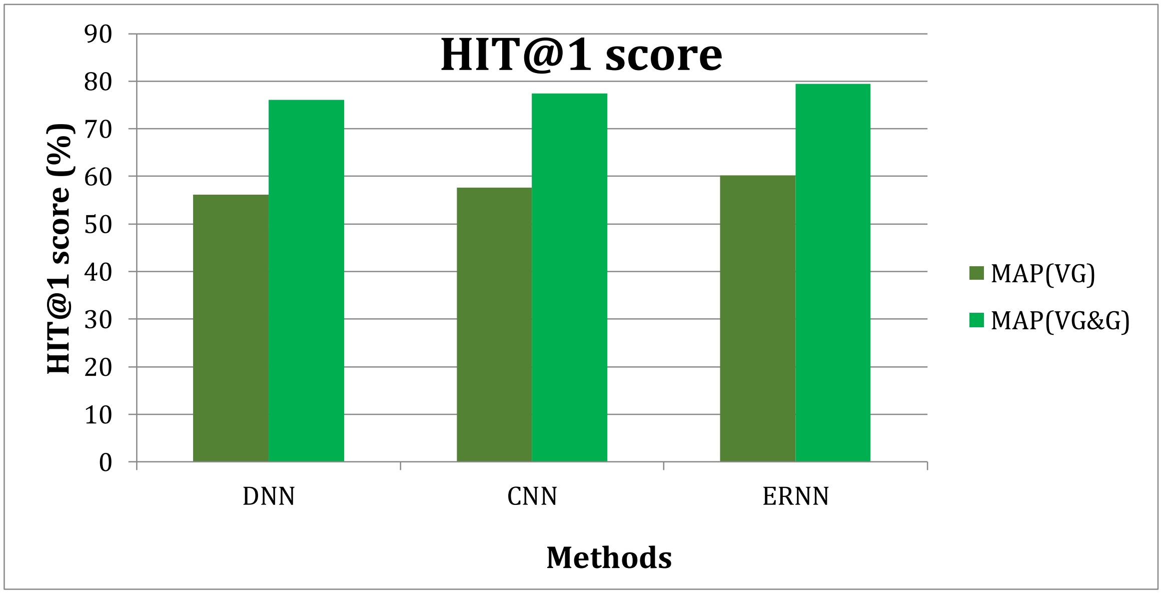


**FIGURE 3: HIT@1 SCORE OF VARIOUS TECHNIQUES FOR VIDEO THUMBNAIL SELECTION (MAP(VG), MAP(VG&G))**

Table 1 illustrates the MAP of various techniques for the process of video thumbnail selection. Figure 4 illustrates HIT@1 score of various techniques for video thumbnail selection. Figure 4 provides the comparison of HIT@1 score values for both MAP(VG), MAP(VG&G). The novel ERNN attained greater HIT@1 score in comparison with other available techniques such as DNN and CNN. The novel ERNN attained a greater value of 60.2%, while the available techniques DNN and CNN attained lower values as 56.1%, 57.6% for MAP(VG). The novel ERNN attained greater value of 79.5%, while other existing techniques such as DNN and CNN attained much less values such as 76.1%, 77.4 % for MAP(VG&G).

### *Dynamic Video Summarization*

Table 2 illustrates the HIT1@1 performance curves when the positive score equals to VG and G, various tradeoff parameters $\alpha$ and various subspace dimensions (hidden layer unit numbers) for the novel ERNN. Table 3 illustrates MAP performance curves when the positive score is equal to VG and G, various tradeoff parameters $\alpha$ and diverse subspace dimensions (unit numbers of the hidden layer) for the discussed ERNN.

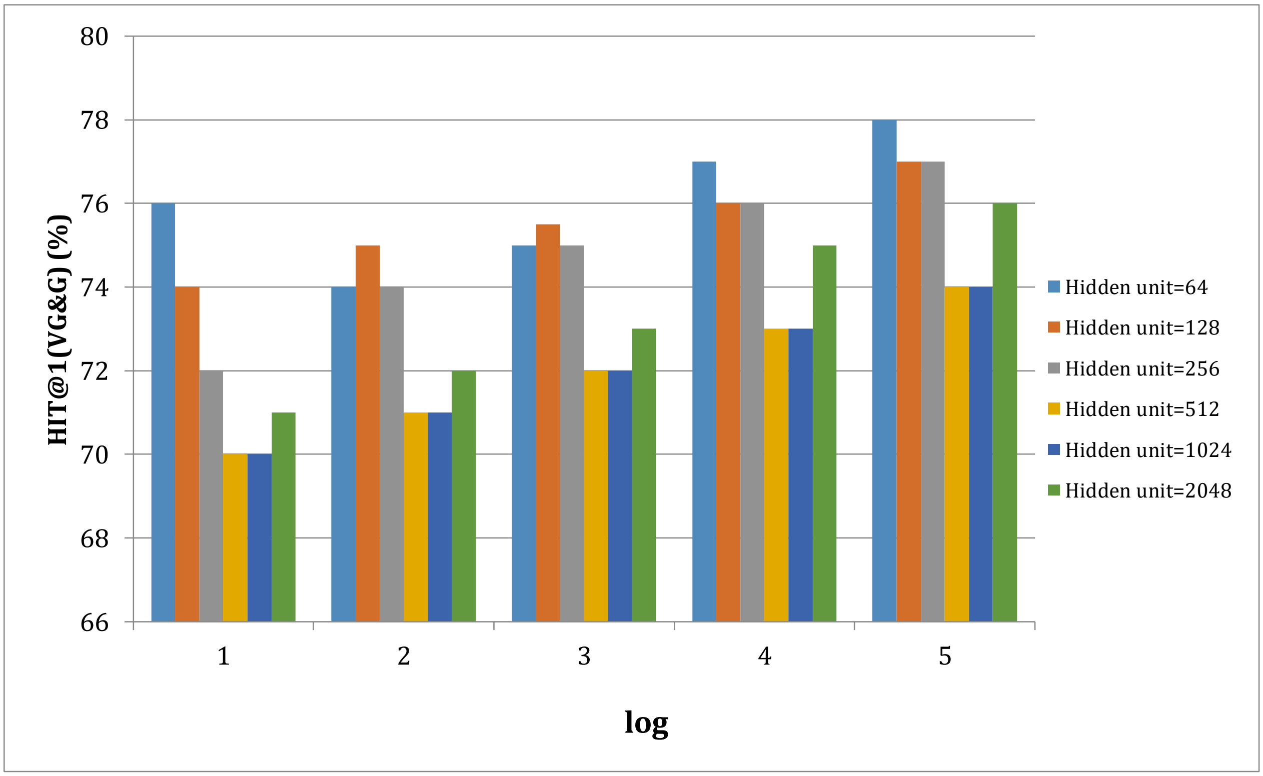
**Table 2: HIT1@1 Performance Curves when Positive Score is equal to VG and G (ERNN)**

| Method/Metric<br>log $\alpha$ | Hidden unit=64 | Hidden unit=128 | Hidden unit=256 | Hidden unit=512 | Hidden unit=1024 | Hidden unit=2048 |
|---|---|---|---|---|---|---|
| 1 | 76 | 74 | 72 | 70 | 70 | 71 |

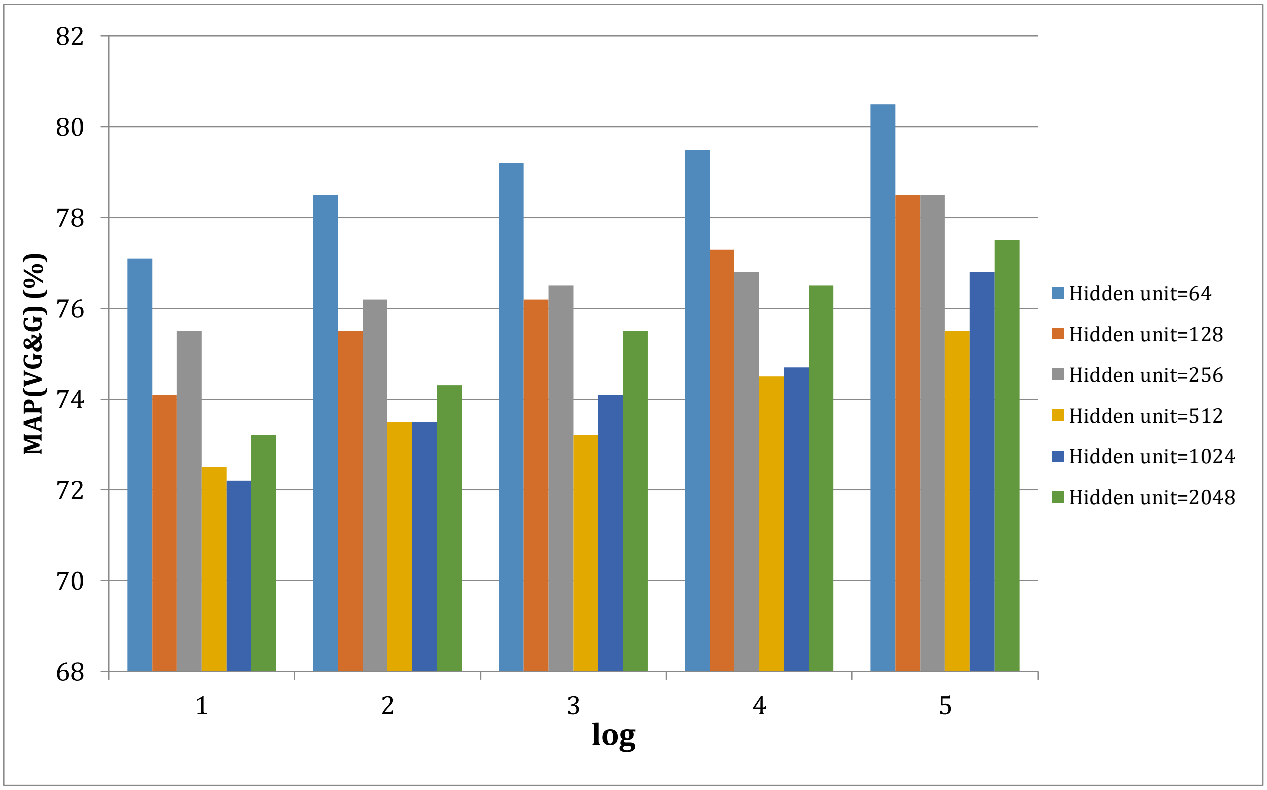
| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 74 | 75 | 74 | 71 | 71 | 72 |
| 3 | 75 | 75.5 | 75 | 72 | 72 | 73 |
| 4 | 77 | 76 | 76 | 73 | 73 | 75 |
| 5 | 78 | 77 | 77 | 74 | 74 | 76 |

**Table 3: MAP Performance Curves when Positive Score is equal to VG and G (ERNN)**

| Method/Metric<br>log $\alpha$ | Hidden unit=64 | Hidden unit=128 | Hidden unit=256 | Hidden unit=512 | Hidden unit=1024 | Hidden unit=2048 |
|---|---|---|---|---|---|---|
| 1 | 77.1 | 74.1 | 75.5 | 72.5 | 72.2 | 73.2 |
| 2 | 78.5 | 75.5 | 76.2 | 73.5 | 73.5 | 74.3 |
| 3 | 79.2 | 76.2 | 76.5 | 73.2 | 74.1 | 75.5 |
| 4 | 79.5 | 77.3 | 76.8 | 74.5 | 74.7 | 76.5 |
| 5 | 80.5 | 78.5 | 78.5 | 75.5 | 76.8 | 77.5 |



**(a) HIT1@1 Performance Curves when Positive Score is equal to VG and G**



**(b) MAP Performance Curves when Positive Score is equal to VG and G**

**Figure 4**

Figure 5 illustrates the HIT@1 and MAP performance curves with diverse tradeoff parameters $\alpha$ and diverse subspace dimensions (unit numbers of hidden layer). (a) HIT1@1 performance curves when positive score is equivalent to VG and G; (b) MAP performance curves when positive score is equivalent to VG and G.

Table 3 illustrates the pair wise F1 scores of various techniques on TVSum50 dataset. Ten categories are available in TVSum50 dataset: changing Vehicle Tire (VT), getting Vehicle Unstuck (VU), Grooming an Animal (GA), Making Sandwich (MS), Parkour (PK), Parade (PR), Flash Mob Gathering (FM), Beekeeping (BK), Attempting Bike Tricks (BT), Dog Show (DS). It is revealed that while seeing these videos, users are focused more towards seeing some particular content, such that the titles of the video can provide a meaningful directive to get hold of semantically informative frames or shots.

**Table 3: The F1 Score of Various techniques for Video Summarization**

| Category/Technique | DNN | CNN | ERNN |
|---|---|---|---|
| VT | 54 | 63 | 65 |
| VU | 55 | 56 | 60 |
| GA | 53 | 59 | 58 |
| MS | 50 | 55 | 64 |
| PK | 40 | 48 | 54 |
| PR | 46 | 50 | 58 |
| FM | 47 | 52 | 54 |
| BK | 45 | 44 | 51 |

| | | | |
|---|---|---|---|
| BT | 56 | 56 | 63 |
| DS | 47 | 50 | 55 |

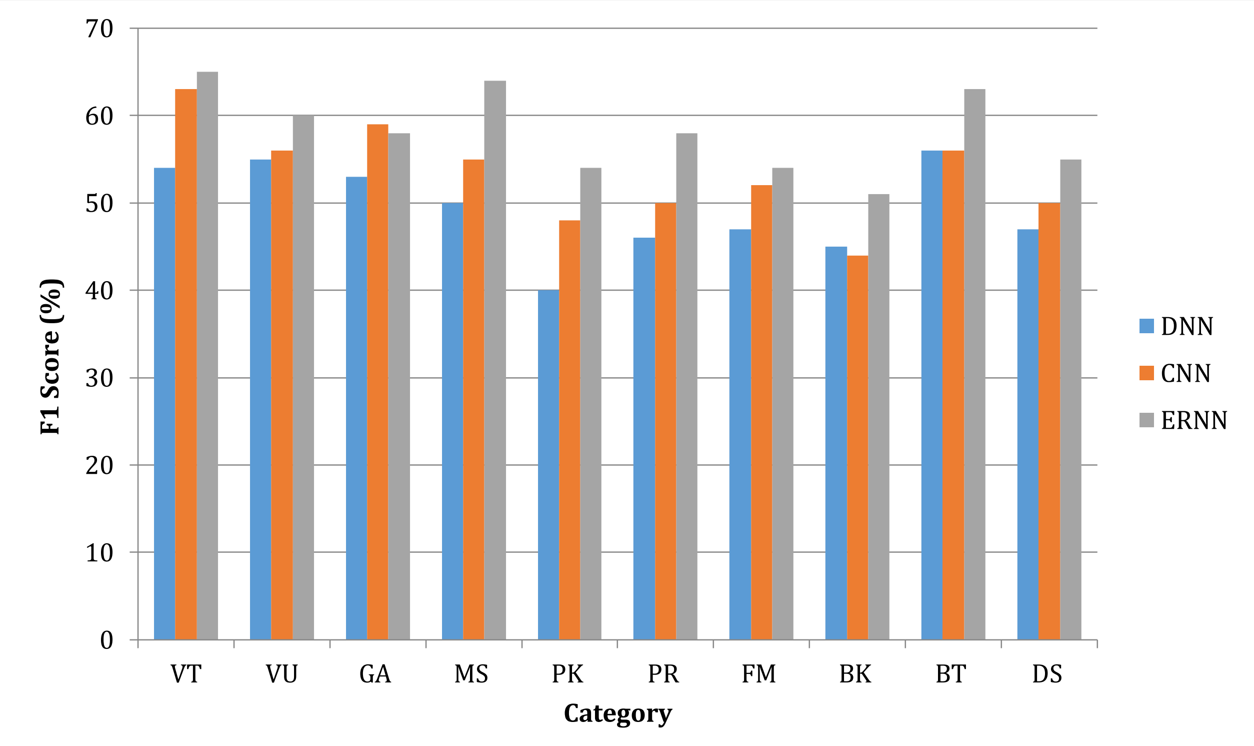


**Figure 5: The Pairwise F1 Scores of various techniques**

Figure 6 illustrates the pairwise F1 scores of various techniques on TVSum50 dataset. The f1 score value of discussed ERNN is compared with other available techniques including DNN and CNN. The discussed ERNN attained a bigger value, which is 65% while other available techniques such as DNN and CNN attained much lesser such as 54% and 63% for groups VT. There are few other 9 groups which have also attained much greater values of proposed ERNN technique, in comparison with other existing techniques.

## 5. CONCLUSION AND FUTURE WORK

The explosive production in multimedia content owing to exponentially rising reach achieved by the internet, content in visual format has become extremely popular. The onset of Social media and improving video sharing in websites, such as YouTube, Yahoo Video, and social networks like Facebook, Google+ have inspired the widespread rise of video contents all over the Internet, and has gained considerable importance to video graphic content. Currently, Video Summarization, has emerged to be a distressing issue in deep learning field, which aims at the automatic evaluation of video content, and generating summary with video content having relevance. This technical work introduced an Enhanced Recurrent Neural Network (ERNN) for video summarization. ERNNcomprises of three important steps. (i) sampling of video frames, (ii) preprocessing, (iii) frame feature extraction. n the first step, the input video are divided into sampling the video frames, and are sampled with the intent of reducing the further computational overhead. In the second step, preprocessing of the sampling of video frames is carried out for noise elimination. Once the noise suppression is completed, this kind of frames leads to increased computational time, and this kind of frames can result in the generation of superior quality video summary, and two filtering algorithms known as MF and ASTF are utilized for noise elimination. At the end, Enhanced Recurrent Neural Network (ERNN) is proposed for the frame feature extraction. Once the preprocessed frames are chosen as the input video for this stage, it is sent through ERNN comprising of two layers, which includes the first layer of shot edge detection is used for the encoding of brief video sub shots cut from the actual video, and keyframe extraction carried out by the hidden state and every subshot becomes the input for the second layer in computing its confidence to act as a key subshot, and then identical keyframe videos are eliminated in keyshot, and at last, a video with best accuracy is summarized. In addition, experiments are conductedon two datasets (Thumb1K and TVSum50) and the best performance of ERNN is demonstrated in comparison with many benchmarked techniques for video summarization. The results of experiments are computed in terms of metrics such as Mean Average Precision (MAP) and F-measure (F1 score).

As a futuristic approach, ERNN technique can be improved by taking the the temporal correlation between video frames, the motion features and few other particular characteristics of videos into consideration when developing the video summarization framework. Other types of data such video tags, captions, comments etc will also be examined in the future works.

## 6. REFERENCES

[1] J. Tang, K. Wang and L. Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Transactions on Image Processing, Vol.25, No.7, Pp.3157-3166, 2016.

[2] X. Li, L. Mou and X. Lu, Video parsing via spatiotemporally analysis with images, Multimedia Tools and Applications, Vol.75, No.19, Pp.11961-11976, 2016.

[3] X. Lu, X. Zheng and X. Li, Latent semantic minimal hashing for image retrieval, IEEE Transactions on Image Processing, Vol.26, No.1, Pp.355-368, 2016.

[4] J. Han, K. Li, L. Shao, X. Hu, S. He, L. Guo, J. Han and T. Liu, Video abstraction based on fMRI-driven visual attention model, Information sciences, Vol.281, Pp.781-796, 2014.

[5] C.M. Tsai, L.W. Kang, C.W. Lin and W. Lin, Scene-based movie summarization via role-community networks, IEEE Transactions on Circuits and Systems for Video Technology, Vol.23, No.11, Pp.1927–1940, 2013.

[6] D. Tao, L. Jin, Y. Wang and X. Li, Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks, IEEE Transactions on Industrial Informatics, Vol.10, No.1, Pp.813-823, 2013.

[7] Y. Zhang, X. Chen, L. Lin, C. Xia and D. Zou, High-level representation sketch for video event retrieval, Science China Information Sciences, Vol.59, No.7, Pp.1-15, 2016.

[8] Y. Yuan, J. Wan, and Q. Wang, Congested scene classification via efficient unsupervised feature learning and density estimation, Pattern Recognition, Vol.56, Pp.159–169, 2016.

[9] D. Zhang, J. Han, J. Han and L. Shao, Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining, IEEE transactions on neural networks and learning systems, Vol.27, No.6), Pp.1163-1176, 2015.

[10] W. Wang, J. Shen, X. Li and F. Porikli, Robust video object cosegmentation, IEEE Transactions on Image Processing, Vol.24, No.10, Pp.3137–3148, 2015.

[11] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas and Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention, IEEE Transactions on Multimedia, Vol.15, No.7, Pp.1553–1568, 2013.

[12] W. Fu, J. Wang, L. Gui, H. Lu and S. Ma, Online video synopsis of structured motion, Neuro computing, Vol. 135, Pp.155–162, 2014.

[13] K. Streib and J. Davis, Summarizing high-level scene behavior, Machine Vision and Applications, Vol.25, No. 1, Pp.229–244, 2014.

[14] Y.J. Lee and K. Grauman, Predicting important objects for egocentric video summarization, International Journal of Computer Vision, Vol.114, No.1, Pp.38–55, 2015.

[15] N. Ejaz, I. Mehmood and S.W. Baik, Efficient visual attention based framework for extracting key frames from videos, Signal Processing: Image Communication, Vol.28, No.1, Pp.34–44, 2013.

[16] S. Zhang, Y. Zhu and A.K. Roy-Chowdhury, Context-aware surveillance video summarization, IEEE Transactions on Image Processing, Vol.25, No.11, Pp.5469-5478, 2016.

[17] Y. Li and B. Merialdo, Multi-video summarization based on OB-MMR, 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Pp.163-168, 2011.

[18] N.J. Al-Musawi and S.T. Hasson, Improving Video Streams Summarization Using Synthetic Noisy Video Data, International Journal of Advanced Computer Science and Applications, Vol.6, No.12, Pp.243-249, 2015.

[19] Y. Yuan, T. Mei, P. Cui and W. Zhu, Video summarization by learning deep side semantic embedding, IEEE Transactions on Circuits and Systems for Video Technology, Vol.29, No.1, Pp.226-237, 2017.

[20] V. Mygdalis, A. Iosifidis, A. Tefas and I. Pitas, Video summarization based on subclass support vector data description, IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES), Pp.183-187, 2014.

[21] X. Li, B. Zhao and X. Lu, A general framework for edited video and raw video summarization, IEEE Transactions on Image Processing, Vol.26, No.8, Pp.3652-3664, 2017.

[22] Y. Cong, J. Yuan and J. Luo, Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection, IEEE Transactions on Multimedia, Vol.14, No.1, Pp.66–75, 2012.

[23] A.E. Ainasoja, A. Hietanen, J. Lankinen and J.K. Kämäräinen, Keyframe-based Video Summarization with Human in the Loop, VISIGRAPP, Pp.287-296, 2018.

[24] X. Xu, T.M. Hospedales and S. Gong, Discovery of shared semantic spaces for multiscene video query and summarization, IEEE Transactions on Circuits and Systems for Video Technology, Vol.27, No.6, Pp.1353-1367, 2016.

[25] M.H. Sigari, H. Soltanian-Zadeh and H.R. Pourreza, Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference, International Journal of Computer Graphics, Vol.6, No.1, Pp.13-36, 2015.

[26] M. Fei, W. Jiang and W. Mao, A novel compact yet rich key frame creation method for compressed video summarization, Multimedia Tools and Applications, Vol.77, No.10, Pp.11957-11977, 2018.

[27] J. Li, T. Yao, Q. Ling and T. Mei, Detecting shot boundary with sparse coding for video summarization, Neurocomputing, Vol.266, Pp.66-78, 2017.

[28] X. Zhu, C.C. Loy and S. Gong, Learning from multiple sources for video summarization, International Journal of Computer Vision, Vol.117, No.3, Pp.247-268, 2016.

[29] A.A. Yahya, J. Tan and L. Li, Video noise reduction method using adaptive spatial-temporal filtering, Discrete Dynamics in Nature and Society, Pp.1-11, 2015.

[30] M.A.T. Jimenez, Summarization of video from Feature Extraction Method using Image Processing & Artificial Intelligence, ResearchGate, Pp.1-9, 2018.

[31] S. Cvetkovic, M. Jelenkovic and S.V. Nikolic, Video summarization using color features and efficient adaptive threshold technique, Przegląd Elektrotechniczny, Vol.89, No.2, Pp.247-250, 2013.

[32] Z. Ji, K. Xiong, Y. Pang and X. Li, Video summarization with attention-based encoder-decoder networks, IEEE Transactions on Circuits and Systems for Video Technology, Pp.1-9, 2019.

[33] Y. Yuan, H. Li and Q. Wang, Spatiotemporal modeling for video summarization using convolutional recurrent neural network, IEEE Access, Vol.7, Pp.64676-64685, 2019.

[34] Y. Song, J. Vallmitjana, A. Stent and A. Jaimes, Tvsum: Summarizing web videos using titles, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Pp.5179–5187, 2015.

[35] W. Liu, T. Mei, Y. Zhang, C. Che and J. Luo, Multi-task deep visualsemantic embedding for video thumbnail selection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Pp.3707–3715, 2015.