CONTENT BASED VIDEO RETRIVAL USING DEEP NEURAL NETWORK

P. Karunakar Reddy[1] Ashwini Golla[2] Dr Venkataramana K[3] T Rathna Reddy[4]

[1]Assistant professor, Dept.of CSE, Siddhartha Institute of Engineering & Technology, Ibrahimpatnam, Hyderabad, Telangana, India.
[2]Assistant professor, Dept.of CSE, Siddhartha Institute of Engineering & Technology, Ibrahimpatnam, Hyderabad, Telangana, India.
[3]Associate professor, Dept.of CSE, Siddhartha Institute of Engineering & Technology, Ibrahimpatnam, Hyderabad, Telangana, India.
[4]Student, Dept.of CSE, Siddhartha Institute of Engineering & Technology, Ibrahimpatnam, Hyderabad, Telangana, India.

Abstract: — Retrieving videos from large repositories using image queries is important for many applications, such as brand monitoring or content linking. We introduce a new retrieval architecture, in which the image query can be compared directly with database videos—significantly improving retrieval scalability compared with a baseline system that searches the database on a video frame level. Matching an image to a video is an inherently asymmetric problem its challenging problems are large intra-class variations, and tremendous time and space complexity. In this paper, we develop a new deep convolutional neural network (deep CNN) to learn discriminative and compact binary representations of faces for face video retrieval. We will use some combined features and can retrieve the data more accurately by using neural network. Neural network will help us to get more accuracy for video retrieval. We can implement proposed work using MATLAB 2016 a software which is high level technical computing language. Both by subjective and objective analysis we can get better results for deep neural network.

Keywords: *Image Retrieval, video Retrieval, Neural Network, MATLAB, subjective and objective analysis;*

I. INTRODUCTION

Artificial intelligence is a computer to imitate human intelligence behavior, to achieve computer image recognition, natural language recognition and some specific intelligent tasks and other work. The research direction of artificial intelligence has been half a century history, is recognized as one of the modern high-tech core in the world, due to the discipline of artificial intelligence itself has characteristics of extensive, its application has been researched in-depth the various disciplines and fields and achieved significant results . At the same time, combining and utilizing other disciplines, the research and development of artificial intelligence will bring new ideas, so as to promote the development of artificial intelligence. Bionics interdisciplinary combination of biological science, engineering and technology, through learning, imitation, copy and biological system structure, function and working principle of the reengineering, to improve the existing or even create new mechanical and technological process, modern bionics has spawned many fields [1]. And Research on them from the mechanism of biological evolution inspired put forward many new methods to solve the problem of artificial intelligence, such as based on the fundamental mechanisms of biological immune system that mimics the body's immune system, for the first time using artificial immune algorithm to solve the optimization problem. By drawing on the evolution law of biology, namely "survival of the fittest, survival of the fittest" the genetic mechanism was proposed for the genetic algorithm to solve the problem of searching for the large-scale and complex system [2]. At the same time, there are methods for searching the path of food recruitment of ants and the bees forage for nectar, and the ant colony algorithm and bee colony algorithm are proposed.

In within a decade of the 21st century, mankind has been working to study how to imitate human brain work and expression of the information ability, and realize the so called artificial intelligence (AI). Human beings receive a lot of data in the perception of information, but it can accurately capture the key factors and save it for future use. Until the 1940s, with the breakthrough progress has been made in the study of neuroanatomical, neurophysiology and neuronal electrophysiological process, people for the human brain structure and the basic working unit have more and more fully understanding and Research on personnel began to try to imitate the structure and working principle of the human brain to construct to achieve something similar to the human brain with this algorithm of recognition and memory function. The brain learning system is composed of interconnected neurons with an unusually complex network system, using the simplified signal propagation mechanism to mimic some basic functions of human brain neurons, it lays the foundation for the development of early neural computing. Based on artificial neuron model, increasing the learning mechanism proposed perceptron model to solve some of the problems in the field of character recognition, first artificial neural network theory is applied to practical problems [3]. And it is proved that the network constitutes a two-layer perceptron can be input to the linear classification, and proposed a hidden layer perceptron unit is a very important research direction. But after that due to the rapid development of computer hardware technology, and other fields, and artificial neural network in dealing with the nonlinear classification problem is not a breakthrough, the field in quiet period for a long time. Generalizing the nonlinear data of artificial neural network of information storage and retrieval function, and through equation of dynamic equation and the learning algorithm is proposed, important formulas and parameters of network algorithm and the theoretical basis for future study and the structure of the artificial neural network are provided.

The recent discovery of neuroscience provides some clues to the basic rules of the brain's expression of the mammalian brain. A key finding is a lot of perception related cerebral neocortex neocortical and no clear preprocess sensory signals, but let these signals through a complex hierarchy module communicate, and with the passage of time, this module can learn based on observed signals showed some regular characteristics to describe the observed signals. And this feature in the primate visual system performance is more obvious. The process can be divided into some successive processing stages: edge detection, the shape of the then gradually increased to more complex visual shape [4]. The findings lead to the emergence of the depth in the field of machine learning, the commitment to research with the neocortex that can show some similar characteristics of the presentation of information capacity calculation model. 50 years ago, the pioneer Bellman in the field of dynamic programming theory and optimization control pointed out that the high dimensions in the data were the fundamental obstacle to many scientific and engineering applications. In pattern recognition, a lot of difficulties are learning, the complexity of the algorithm is relative to the dimension of the data to present the index level of growth, Richard bellman, this phenomenon is known as the curse of dimensionality. To avoid the curse of dimensionality of the mainstream approach is by preprocessing the data to achieve the data dimension reduction to can be some of the existing methods to effectively deal with the degree, this reduction process dimension is generally referred to as the feature extraction. Therefore, the intelligent process of many pattern recognition systems is transformed to the high difficulty of manual design and the feature extraction process based on the application of specific applications. However, if the feature extraction process is errors or imperfect, the performance of the classifier will inevitably be limited. In addition to the real life of the data in addition to the space dimension, the time dimension is also very important. A group of observed continuous data is capable of conveying a number of specific information, or the actual meaning of the event or observed data is generally inferred from the similar data of the time [5]. Thus, the modeling of temporal components in the observation data is especially important in the information expression. Therefore, the consistency of the observation data to capture the dependence of the data in the time and space, is considered to be the fundamental goal of the depth learning system. If can get a learning system with the depth of the robustness, can through this hierarchy system training on a large data set, then the extraction system of information as input to follow a relatively simple classification system, came to an equally robust pattern recognition system. Here the robustness of the said classification results for data in the transformed and distorted can keep invariant features, this kind of transformation and distortions include noise, scaling, rotation, illumination changes, displacement and so on. Deep neural network algorithm is widely concerned in academic field, which is based on fast data analysis and forecasting. The depth of the neural network is an automatic learning sample characteristics of the input method, and to study the characteristics of data is more essential characterizations and through layer by layer initialization "to overcome the difficulty in training. Therefore, it has been a wide range of academic and industrial circles, and became as research fields. Therefore, the depth of the neural network as a learning model of complex hierarchical probabilistic method in various fields has been widely used. At present, it has been applied to the field of speech recognition, recognition of handwriting font, traffic signs, face recognition and other image processing fields, showing superior performance of learning.

II. LITERATURE SURVEY

Deep learning was proposed by Schmidhuber and Jürgen in 2015 [6], which is a new area of machine learning. Deep learning has made great progress in speech recognition. It depends not only on the ability of the parallel processing of the large data, but also on the algorithm, and this algorithm is the depth of learning. June 2012, "New York Times" disclosed the Google brain project, the project by Stanford University Professor Andrew Ng and computer systems for the top expert Jeff Dean co dominant, 16000 CPU core parallel computing platform training a known for the depth of the neural network of the machine learning model. The human visual system is classified by the processing of information. From low-level edge features extracted to shape (or target), to the higher target, target behavior, namely low - level features into the high-level features, characterized by a low to high said more and more abstract. That depth study is how to learn from this process, the process is the process of modeling. For a pair of images, the pixel level feature has little value. The smaller graphics can be made of basic hook, so the complicated concept graphics need a higher level feature, that is, the high-level representation is composed by the bottom layer.

Mr. Pradeep Chivadshetti (2015) [7] proposed a Traditional video retrieval methods fail to meet technical challenges due to large and rapid growth of multimedia data, demanding effective retrieval systems. In the last decade Content Based Video Retrieval (CBVR) has become more and more popular. The amount of lecture video data on the World Wide Web (WWW) is growing rapidly. Therefore, a more efficient method for video retrieval in WWW or within large video archives is urgently needed. This paper presents an implementation of automated video indexing and video search in large video database and also present personalized results. Proposed system works in three different phases, in the first phase video segmentation and key frame detection is performed to extract meaningful key frames. Secondly, OCR, HOG and ASR algorithms are applied over the keyframe to extract textual keyword. In the third phase, Color, Texture and Edge features are also extracted. Finally, search similarity measure is performed on the extracted features that are saved in SQL database and the output with personalised re-ranking results as per interest is presented to the users.

Mohd.Aasif Ansari, HemlataVasishtha [8] proposed a paper of Content Based Video Retrieval Systems performance is analysed and compared for three different types of feature vectors. These types of features are generated using three different algorithms; Block Truncation Coding (BTC) extended for colors, Kekre’s Fast Codebook Generation (KFCG) algorithm and Gabor filters. The feature vectors are extracted from multiple frames instead of using only key frames or all frames from the videos. The performance of each type of feature is analysed by comparing the results obtained by two different techniques; Euclidean Distance and Support Vector Machine (SVM). Although a significant number of researchers have expressed dissatisfaction to use image as a query for video retrieval systems, the techniques and features used here provide enhanced and higher retrieval results while using images from the videos. Apart from higher efficiency, complexity has also been reduced as it is not required to find key frames for all the shots. The system is evaluated using a database of 1000 videos consisting of 20 different categories. Performance achieved using BTC features calculated from color components is compared with that achieved using Gabor features and with KFCG features. These performances are compared again with the performances obtained from systems using SVM and the systems without using SVM.

Markus Muhling, (2017) [9] proposed a paper, that present deep learning approaches to support professional media production. In particular, novel algorithms for visual concept detection, similarity search, face detection, face recognition and face clustering are combined in a multimedia tool for effective video inspection and retrieval. The analysis algorithms for concept detection and similarity search are combined in a multi-task learning approach to share network weights, saving almost half of the computation time. Furthermore, a new visual concept lexicon tailored to fast video retrieval for media production and novel visualization components are introduced. Experimental results show the quality of the proposed approaches. For example, concept detection achieves a mean average precision of approximately 90% on the top-100 video shots, and face recognition clearly outperforms the baseline on the public Movie Trailers Face Dataset.

Khokher et al., (2012) [10] proposed a Content based video retrieval has a wide spectrum of promising applications, motivating the interests of the researchers worldwide. This paper represents an overview of the general strategies used in visual content-based video retrieval. It focuses on the different methods for video structure analysis, including shot segmentation, key frame extraction, scene segmentation, feature extraction, video annotation, and video retrieval method. This work helps the upcoming researchers in the field of video retrieval to get the idea about different techniques and methods available for the video retrieval.

Angadi, Shanmukhappa, and Vilas Naik (2016) [11] proposed a new approach for hot event detection and summarization of news videos. The approach is mainly based on two graph algorithms: optimal matching (OM) and normalized cut (NC). Initially, OM is employed to measure the visual similarity between all pairs of events under the one-to-one mapping constraint among video shots. Then, news events are represented as a complete weighted graph and NC is carried out to globally and optimally partition the graph into event clusters. Finally, based on the cluster size and globality of events, hot events can be automatically detected and selected as the summaries of news videos across TV stations of various channels and languages. Our proposed approach has been tested on news videos of 10 hours and has been found to be effective.

III. PROPOSED METHOD

The challenges behind the design and implementation of the content based video browsing; indexing and retrieval systems have attracted researchers from much compliance. It is widely accepted that successful solution to the problem of understanding and indexing the videos requires combination of information from different sources such as images, audio, text, speech etc. Videos have the following characteristics:

1) Much richer content than individual images;

2) Huge amount of raw data; and

3) Very little prior structure.

These characteristics make the indexing and retrieval of videos quite difficult. In the past, video databases have been relatively small, and indexing and retrieval have been based on keywords annotated manually. More recently, these databases have become much larger and content based video indexing and retrieval is required, based on the automatic analysis of videos with the minimum of human participation. Content based video retrieval has a wide range of applications such as quick video browsing, analysis of visual electronics commerce, remote instructions, digital museums, news video analysis [1], intelligent management of the web videos and video surveillance. A video may have an auditory channel as well as a visual channel. The available information from videos includes the following:

1) Video metadata, which are tagged texts embedded in videos, usually including title, summary, date, actors, producer, broadcast duration, file size, video format, copy-right, etc.

2) Audio information from the auditory channel.

3) Transcripts: Speech transcripts can be obtained by speech recognition and caption texts can be read using optical character recognition techniques.

4) Visual information contained in the images themselves from the visual channel. In this paper, we focus on the visual contents of the videos and give a survey on visual content-based visual retrieval and indexing.

The process of building indexes for videos normally involves the following three main steps:

1. *Video Parsing:*

It consists of temporal segmentation of the video contents into smaller units. Video parsing methods extract structural information from the video by detecting temporal boundaries and identifying significant segments, called shots.

2. Abstraction:

It consists of extracting the representative set of video data from the video. The most widely used video abstractions are: the “highlight” sequence (A shorter frame sequence extracted from the shot) and the key frame (images extracted from the video shot). The result of video abstraction forms the basis for the video indexing and browsing.

3. *Content Analysis:*

It consists of extracting visual features from key frames. Several techniques used for image feature extraction can be used but, they are usually extended to extraction of features that are specific to video sequences, corresponding to the notion of object motion, events & actions.

4. *VIDEO PARSING:*

Similarly to organizing a long text into smaller units, such as paragraph, sentences, words and letters, a long video sequence must be organized into smaller and more manageable components, upon which indexes can be built. The process of breaking a video into smaller units is known as video parsing. These components are usually organized in a hierarchical way with 5 levels, in decreasing degree of granularity: video, scene, group, shot and key frame. The basic unit is called as a shot. It is defined as a sequence of frames recorded contiguously and representing a continuous action in time or space. The most representative frame of a shot is called a key frame. A scene or sequence is formally defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story. A video group is an intermediate entity between the physical shots and semantic scenes and serves as a bridge between the two. Examples of groups are temporally adjacent shots and visually similar shots [4]. In the following sections we present few algorithms and techniques for video parsing at a shot level and boundary detection at a scene level.

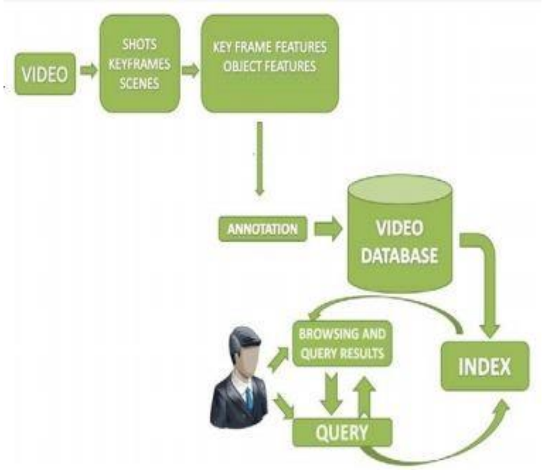

Fig3 System Architecture

5. *Shot Boundary Detection:*

Shot detection is the process of detecting boundaries between two consecutive shots, so that a sequence of frames belonging to a shot will be grouped together. There are different types of boundaries between shots. The simplest one is the cut, an abrupt change between the last frame of a shot and the first frame of a subsequent shot. Gradual boundaries are harder to detect. Examples include: dissolves, wipes, fadeins, and fade-outs. A fade is a "gradual means of closing or starting a shot, often used as a transitional device when one scene closes with the image disappearing (a fade-out) and the next scene comes into view as the image grows stronger and stronger (a fade-in) ." A dissolve is "a transition between two shots whereby the first gradually fades out as the second gradually fades in with some overlap between the two." A wipe is a transition "in which the new shot gradually appears while pushing or 'wiping' off the old." An additional level of difficulty is imposed by camera operations such as panning (the process of moving a camera horizontally around a fixed axis) and zooming (the apparent movement either toward or away from a subject). A robust shot boundary detection algorithm should be able to detect all these different types of boundaries with accuracy. The basis for detecting shot boundaries is the detection of significant changes in contents on consecutive frames lying on either side of a boundary. Automatic shot boundary detection techniques can be classified into seven main groups: Pixel-based: The easiest way to detect a shot boundary is to count the number of pixels that change in value more than determine if a shot boundary has been found.

6. *Statistics-based:*

Statistical methods expand on the idea of pixel differences by breaking the images into regions and comparing statistical measures of the pixels in those regions [8]. For example, Kasturi and Jain [12] use intensity statistics (mean and standard deviation) as shot boundary detection measures. This method is reasonably robust to noise, but slow and prone to generate many false positives (i.e., changes not caused by a shot boundary)

7. Histogram-based:

The most popular metric for sharp transition detection is the difference between histograms of two consecutive frames. In its simplest form, the gray level or color histograms of two consecutive frames are computed: if the bin-wise difference between the two histograms is above a threshold, a shot boundary is said to be found. Several variants of the basic idea have been proposed in the literature. Nagasaka and Tanaka [13] proposed breaking the images into 16 regions, using a x2- test on color histograms of those regions, and discarding the eight largest differences to reduce the effects of object motion and noise. Swanberg, Shu, and Jain [14] used gray level histogram differences in regions; weighted by how likely the region was to change in the video sequence. Their results were good because their test video (CNN Headline News) had a very regular spatial structure. Zhang, Kankanhalli, and Smoliar [15] compared pixel differences, statistical differences and several different histogram methods and concluded that the histogram methods were a good trade-off between accuracy and speed. They also noted, however, that the basic algorithm did not perform too well for gradual transitions as it did for abrupt cuts. In order to overcome these limitations, they proposed the twin-comparison algorithm, which uses two comparisons: one looks at the difference between consecutive frames to detect sharp cuts, and the other looks at accumulated difference over a sequence of frames to detection gradual transitions. This algorithm also applies a global motion analysis to filter out sequences of frames involving global or large moving objects, which may confuse the gradual transition detection. Additional examples of histogram-based shot detection techniques includes.

8. *Key frame based approach:*

This approach represents each video shot by a set of key frames from which features are extracted. Temporally close shots are grouped into a scene. An author in compute similarities between the shots using block matching of key frames, then similar shots are linked together and scenes are identified by connecting the overlapping links. Ngo et al. [16] extract and analyze the motion trajectories encoded in the temporal slices of image volumes Scene changes can be identified by measuring the similarities of the key frames in the neighboring shots. A limitation of this approach is that key frames cannot effectively represent the dynamic content of the shots.

9 FEATURE EXTRACTION:

The extraction of content primitives (referred to as "metadata" in the scope of the emerging MPEG-7 standard) from video programs is a required step that allows video shots to be classified, indexed, and subsequently retrieved. Since shots are usually considered the smallest indexing unit in a video database, content representation of video is also usually based on shot features. There are two types of features: those associated with key-frames only which are static by nature and those associated with the frame sequence that compose a shot which may include the representation of temporal variation of any given feature and motion information associated with the shot or some of its constituent objects. Representing shot contents at an object level through the detection and encoding of motion information of dominant objects in the shot is a new and attractive technique, because much of the object information is available in MPEG-4 video streams.

10.VIDEO ANNOTATION:

Video annotation is the allocation of video shots or video segments to different predefined semantic concepts, such as person, car, sky and people walking. Video annotation and video classification share similar methodologies: First, low- level features are extracted, and then certain classifiers are trained and employed to map the features to the concept/category labels. Corresponding to the fact that a video may be annotated with multiple concepts, the methods for video annotation can be classified as isolated concept-based annotation, context-based annotation, and integrated-based annotation

11. *Isolated concept based annotation:*

The annotation method trains a statistical detector for each of the concepts in a visual lexicon, and the isolated binary classifiers are used individually and independently to detect multiple semantic concepts correlations between the concepts are not considered. The limitation of isolated concept based annotation is that the associations between the different concepts are not modeled.

12. Context-based annotation:

The task of context-based annotation is to refine the detection results of the individual binary classifiers or infer higher level concepts from detected lower level concepts using a context-based concept fusion strategy. The limitation of context-based annotation is that the improvement of contextual correlations to individual detections is not always stable because the detection errors of the individual classifiers can propagate to the fusion step, and partitioning of the training samples into two parts for individual detections and conceptual fusion, respectively, causes that there are no sufficient samples for the conceptual fusion because of usual complexity of the correlations between the concepts.

13. Integration-based annotation:

This annotation method simultaneously models both the individual concepts and their correlations: The learning and optimization are done simultaneously. The entire set of samples is used simultaneously to model the individual concepts and their correlations. The limitation of the integration-based annotation.

14. QUERY AND VIDEO RETRIEVAL:

Once video indices are obtained, content-based video retrieval can be performed. On receiving a query, a similarity measure method is used, based on the indices, to search for the candidate videos in accordance with the query. The retrieval results are optimized by relevance feedback, etc. In the following, we review query types, similarity matching, and relevance feedback.

IV. CONCLUSION & FUTURE WORK

We have presented a review on recent developments in visual content-based video indexing and retrieval. The state of the art of existing approaches in each major issue has been described with the focus on the following tasks: video structure analysis including shot boundary detection, key frame extraction and scene segmentation, features extraction of static key frames, objects and motions, video annotation, query type and video retrieval methods, video search including interface, similarity measure and relevance feedback.

V. REFERENCE

[1] Y. X. Peng and C.W. Ngo, “Hot event detection and summarization by graph modeling and matching,” in Proc. Int. Conf. Image Video Retrieval, Singapore, pp. 257–266.

[2] A. F. Smeaton, “Techniques used and open challenges to the analysis, indexing and retrieval of digital video,” Inform. Syst., vol. 32, no. 4, pp. 545–559.

[3] Y. Y. Chung, W. K. J. Chin, X. Chen, D. Y. Shi, E. Choi, and F. Chen, “Content-based video retrieval system using wavelet transform,” World Sci. Eng. Acad. Soc. Trans. Circuits Syst., vol. 6, no. 2, pp. 259–265.

[4] Y. Rui and T.S. Huang. Unified framework for video browsing and retrieval. In A. Bovik, editor Handbook of Image and Video Processing chapter 9.2.Academic Press, San Diego.

[5] H.-J Zhang. Content-based video browsing and retrieval. In B. Furht, editor, Handbook of Internet and Multimedia Systems and Applications, chapter 712. CRC Press, Boca Raton.

[6] Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural networks*, *61*, pp.85-117.

[7] Chivadshetti, P., Sadafale, K. and Thakare, K., 2015, December. Content based video retrieval using integrated feature extraction and personalization of results. In *2015 International Conference on Information Processing (ICIP)* (pp. 170-175). IEEE.

[8] Ansari, M.A. and Vasishtha, H., Enhanced Video Retrieval and Classification of Video Database Using Multiple Frames Based on Texture Information.

[9] Mühling, M., Korfhage, N., Müller, E., Otto, C., Springstein, M., Langelage, T., Veith, U., Ewerth, R. and Freisleben, B., 2017. Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications*, *76*(21), pp.22169-22194.

[10] Khokher, A. and Talwar, R., 2012, April. Content-based image retrieval: Feature extraction techniques and applications. In *International conference on recent advances and future trends in information technology (iRAFIT2012)* (pp. 9-14).

[11] Angadi, S. and Naik, V., 2016. Dynamic Summarization of Video Using Minimum Edge Weight Matching in Bipartite Graphs. *International Journal of Image, Graphics and Signal Processing*, *8*(3), p.9.

[12] Antani, S., Kasturi, R. and Jain, R., 1998, August. Pattern recognition methods in image and video databases: past, present and future. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (pp. 31-53). Springer, Berlin, Heidelberg.

[13] Tanaka, M., Kohno, Y., Nakagawa, R., Ida, Y., Takeda, S., Nagasaki, N. and Noda, Y., 1983. Regional characteristics of stress-induced increases in brain noradrenaline release in rats. *Pharmacology Biochemistry and Behavior*, *19*(3), pp.543-547.

[14] Swanberg, D., Shu, C.F. and Jain, R.C., 1993, April. Knowledge-guided parsing in video databases. In *Storage and retrieval for Image and Video Databases* (Vol. 1908, pp. 13-24). Spie.

[15] Zhang, H., Kankanhalli, A. and Smoliar, S.W., 1993. Automatic partitioning of full-motion video. *Multimedia systems*, *1*(1), pp.10-28.

[16] Ngo, C.W., Pong, T.C., Zhang, H.J. and Chin, R.T., 2000, September. Motion-based video representation for scene change detection. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* (Vol. 1, pp. 827-830). IEEE.