International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 8 Ver IV || Aug 2018 || PP 69-75

Convolutional Neural Network (CNN) and GMM Supervectors in Video Concept Detection.

Ritika D Sangale1,Nita S Patil2,Sudhir D Sawarkar3
1(Department of Computer Engineering, DattaMeghe College of Engineering, Navi Mumbai, India)
2(Department of Computer Engineering, DattaMeghe College of Engineering, Navi Mumbai, India)
3(Department of Computer Engineering, DattaMeghe College of Engineering, Navi Mumbai, India)
Corresponding Author: Ritika D Sangale

Abstract *: Video concept detectionplays an important role in digital image processing. Video concept detectionis defined as detecting the concepts that are present in the videoshots. Concept canbeanything of users' interest. Video concept detectionisassigning the labels to the input videoshots. For getting the accuratevideoshots, different types of techniques are used. The basic approach of concept detectionis to use classification algorithm. There are manydifferent CNN structures for video or image classification likeAlexNet, Overfeat, VGG, Network-in-network, GoogLeNet, ResNet. In thispaperwe propose to use CNN and SVM, to build Concept classifierswhichpredict the relevance between images or videoshots and a given concept. DeepConvolutional neural networks (CNN) are a special type of feed-forward networks. DCNN performverywell on visual recognition tasks. We are usingAlexnet architecture pre-trained on the ImageNETdataset on TRECVID dataset and using GMM supervectors corresponding to apply on different types of audio and visualfeatures.*

Keywords *-Shot Boundary Detection, Keyframe Extraction, Feature Extraction, Deep CNN, GMM Supervectors, SVM, Score Fusion.*

Date of Submission: 17-08-2018 Date of acceptance: 31-08-2018

I. INTRODUCTION

The main buildingmodules of state-of-the-art concept detection systemsinclude feature extraction and fusion, and classifiertraining. Thus, what kind of features and whatclassifier models are adopted has critical impacts onthe performance of concept detection. The goal of video concept detection, or high-level feature extraction, is to build mapping functions from the low-level features to the high-level concepts with machine learning techniques.

Thus, state-of-the-art approaches in the field of image andvideo retrieval focus on semantic concepts serving as an intermediate descriptionto bridge the semantic gap between the data representation and the humaninterpretation. Semantic concepts, also known as high-level features, can be,forexample, objects, sites, scenes, personalities, events or activities.

Detecting video into shots is the first processing step in analysis of video content for indexing, browsing and searching. A shot is defined as an unbroken sequence of frames taken from on camera. Shot is further divided into keyframe. Keyframes are representative frames from the entire shot. Various features are extracted from these keyframes and stored as feature vector. These features are given to classifier which predicts presence or absence of the particular concept in given shot based on the previously trained model. The basic approach of concept detection is to use classification algorithms, typically support vector machines (SVMs), to build concept classifiers which predict the relevance between images or video shots and a given concept.

We use deep Convolutional Neural Network (DCNN) trained on the ImageNet dataset to extract features from video shots. A 4096-dimensional feature vector is extracted from the key-frame of each video shot by using the CNN. Convolutional Neural Networks(CNNs), are discriminatively trained via back-propagation through layers of convolutional Filters and other operations such as rectification and pooling. And also Each video shot is modelledby a Gaussian-mixture model (GMM) supervector.

II. LITERATURE SURVEY

JingweiXu and LiSong, and RongXie2016 [1], proposed a novel SBD framework based on representative features extracted from CNN. This scheme is suitable for detection of both CT and GT boundaries. Authorsachieves excellent accuracy in shot Boundary Detection.

Ravi Mishra et al 2014 [2], presents a comparison between the two detection methods like block matching algorithm and dual tree complex wavelet transform for shot detection in videos in terms of various parameters like false rate, hit rate, miss rate tested on a set of different video sequence.

Ganesh. I. Rathod, Dipali. A. Nikam 2013 [3], worked on a Square histogram based model using frame segmentation and automatic threshold calculation. In this scheme the keyframe is extracted by using a reference frame approach per shot. A total of around 40 videos of different types are tested on this model and the model is able to detect all shot boundaries and is storing the suitable frames as keyframes to represent the video summary. An efficiency of almost 95% to 98% is observed using this algorithm.

Samira Pouyanfar and Shu-Ching Chen 2017 [4], a novel ensemble deep classifier is proposed which fuses the results from several weak learners and different deep feature sets. In their proposed framework is designed to handle the imbalanced data problem in multimedia systems.

Alex Krizhevsky, IlyaSutskever, Geoffrey E. Hinton2012 [5],author trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet ILSVRC-2010 contest into the 1000 different classes.

Nakamasa Inoue and Koichi Shinoda and Zhang Xuefeng and Kazuya Ueki 2014 [9],authors proposed a high-performance semantic indexing system using a deep CNN and GMM supervectors with the six audio and visual features. Theresult was 28.1 % in terms of Mean InfAP, which was ranked third among participating teams in the semantic indexing task.

III. OVERVIEW OF OUR VIDEO CONCEPT DETECTION SYSTEM

The video is a one which plays a major role in the today's life which consists of number of frames. A set of frames will form the shots and the group of shots will produce the scenes, the combination of different scenes will form the video. A shot is defined as the consecutive frames from the start to the end of recording in a camera. It shows a continuous action in an image sequence. Video stream is taken as an input and converts into different scenes. This is called as scene segmentation [1].

In our system video is taken as input, video clip is composed into three steps: 1.shot boundaries detection, 2.shot selection, and 3.key frame extraction within the selected shot.Firstthevideo is divided into shots, then the best shot is determined,and finally a representative frame is extracted from theselected shot. The major methods for dividing video into shot are pixel differences, statistical differences, histogram comparisons, edge differences, and motion vectors.

We use edge difference method. In this method, first video stream is taken as an input and frames of video are read in memory. After this the frame is converted into grey scale. Here we use canny edge detection method on each frame. After that edge difference between two successive frames is calculated. After calculating the edge difference of all frames, the mean value of the video can be determined and then standard deviation and threshold value can be obtained. There are two thresholds for two different shot transitions i.e. cut and gradual. As per the threshold value cut or gradual transition is detected from the different scenes. After the detection of the shot boundary the valid Key frame is extracted from the different frames.

In our proposed system video shot is taken as input to deep CNN to train on the Alexnet for extracting features and GMM supervectors. In Deep Convolutional Neural Network (CNN), from each keyframe we extract a 4096 feature at sixth layer to train SVM for each Concept. In total seven visual and audio low- Level features are extracted from video data and pass to SVM for test the concept. After video shots are represented by GMMs, GMM Supervectors are extracted by Combiningnormalizedmean Vectors.SVM are used to train discriminative models for each Semantic Concept. The relationships between shots are useful for detecting semantic concept. After calculating probabilities score are fused and then we use re-ranking method to re-evaluate score of video shot by using shot-score distribution. Re-ranking method defines a video-clip score as the maximum value of shot scores among all the shots in a video clip.Highest Probability Match will shows the videoconceptdetection result.

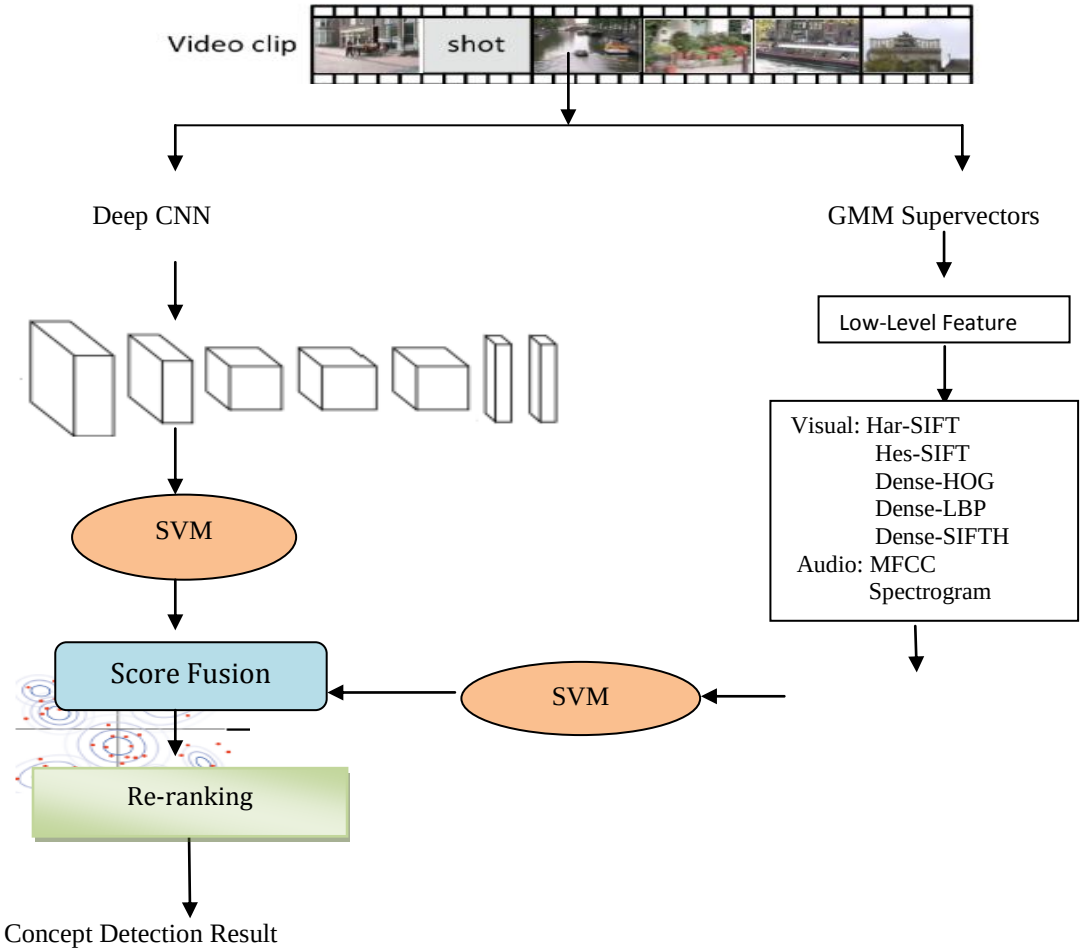

Fig 1: Proposed Architecture

1. Deep Convolutional Neural Networks

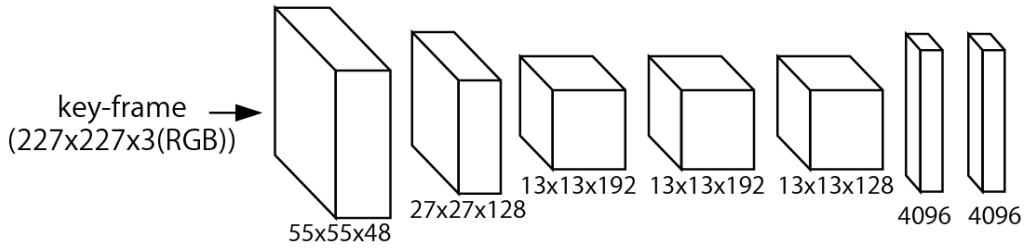

Fig 2: Architecture of Deep CNN [5]

Layer Description in Architecture of Deep CNN
Layer 0: Input Keyframe: Size: 227 x 227 x 3

1. Layer 1: Max-Pooling with 48 filters, size 11×11, stride 4, padding 0: Size: 55 x 55 x 48, (227-11)/4 + 1 = 55 is the size of the outcome and 48 depth because 1 set denotes 1 filter and there are 48 filters.
2. Layer 2: Max-Pooling with 128 filters, stride 2: Size: 27 x 27 x 128, (55 – 3)/2 + 1 = 27 is size of outcome and depth is 128 because pooling is done independently on each layer.

3. Layer 3: Convolution with 192 filters, stride 2: Size: 13 x 13 x 192, (27 – 3)/2 + 1 = 13 is size of outcome and Depth is 192 because of 192 filters.
4. Layer 4: Convolution with 192 filters, size 3×3, stride 1, and padding 1: Size: 13 x 13 x 192, Because of padding of (3-1)/2=1, the original size is restored and 192 depth because of 192 filters.
5. Layer 5: Max-Pooling with 128 filters, stride 2: Size: 13 x 13x 128, and Depth is128 because pooling is done independently on each layer.
6. Layer 6: Fully Connected with 4096 neuron: In this layer, each of the 13 x 13x 128, pixels are fed into each of the 4096 neurons and weights determined by back-propagation.
7. Layer 7: Fully Connected with 4096 neuron: Similar to layer6.finally Fully Connected with 1000 neurons this is the last layer and has 1000 neurons because ImageNET data has 1000 classes to be predicted[5].

CNNs are hierarchical neural networks, which reduce learning complexity by sharing the weights in different layers. CNN is proposed with only minimal data preprocessing requirements and only a small portion of the original data are considered as the input of small neuron collections in the lowest layer.

The obtained salient features will be tiled with an overlap to the upper layer in order to get a better representation of the observations. The realization of CNN may vary in the layers. However, basically they always consist of three types of layers: convolutional layers, pooling layers (or sub-sampling layers), and fully connected layers. The network designed was used for classification with 1000 possible categories[5].

The output of every convolutional layer and fully connected layer is put through RELU non-linearity. The output of the last fully connected layer sent to the 1000-way softmax layer, which produces 1000 probability values for 1000 class labels, where higher value corresponds to higher probability. Under probability distribution this neural network maximizes the average across the training cases of the log-probability of the correct label[5].

We use deep Convolutional Neural Network (CNN) trained on the ImageNET (Alexnet) dataset to extract features from video shots. A 4096-dimensional feature vector is extracted from the key-frame of each video shot by using the CNN. The first to fifth layers are convolutional layers, in which the first, second, and fifth layers have max-pooling procedure. The sixth and seventh layers are fully connected. The parameters of the CNN is trained on the ImageNET LSVRC dataset with 1,000 object categories. Finally, from each keyframe, we extract a 4096-dimensional feature at the sixth layer to train an SVM for each concept in the Video Concept Detection.

2 GMM Supervectors

In GMM Supervector used Local Feature Extraction, There are Seven types of visual and audio features are extracted from video data. Input Video is divided into Shot, this Video Shot is used to extract low-level feature.

2.1 Low-Level Feature:

1. SIFT features with Harris-Affine detector (SIFT-Har)

Scale Invariant Feature Transform (SIFT) proposed by Lowe is a local feature extraction method that is widely used for object categorization since it is invariant to image scaling and changing illumination.The SIFT feature is partially or totally invariant with respect to many common image deformations, including position, scale, illumination, rotation, and affine transformation.The Harris-Affine detector, which is an extension of the Harris corner detector, improves the robustness against affine transform of local regions.The Hessian affine detector is typically used as a preprocessing step to algorithms that rely on identifiable, characteristic interest points[9].

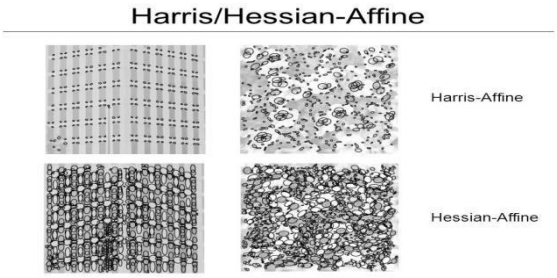

Fig 3:Harris and Hessian-Affine Detectors

SIFT features are extracted from every other frame, and principal component analysis (PCA) is applied to reduce their dimensions.

2. SIFT features with Hessian-Affine detector (SIFT-Hes)

SIFT features are extracted with the Hessian-Affine detector, which is complementary to the Harris-Affine detector. The combination of several different detectors can improve the robustness against noise. The Hessian affine uses a multiple scale iterative algorithm to spatially localize and select scale and affine invariant points.SIFT features are extracted from every other frame, and PCA is applied to reduce their dimensions[9].

3. SIFT and hue histogram with dense sampling (SIFTH-Dense)

SIFT features and 36-dimensional hue histograms are combined to capture color information. SIFT+Hue features are extracted from key-frames by using dense sampling (100x100 grid with 3 scales). Hue is an attribute closely related to the dominant wavelength of color signal.PCA is applied to reduce dimensions[9].

4. HOG with dense sampling (HOG-Dense)

The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image.32-dimensional histogram of oriented gradients (HOG) are extracted from up to 100 frames per shot by using dense sampling with 2x2 blocks. PCA is applied on dimensions of the HOG features[9].

5. LBP with dense sampling (LBP-Dense)

Local Binary Pattern (LBP) is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications. Perhaps the most important property of the LBP operator in real-world applications is its robustness to monotonic gray-scale changes caused, for example, by illumination variations. Another important property is its computational simplicity, which makes it possible to analyze images in challenging real-time settings[9].

Local Binary Patterns (LBPs) are extracted from up to 100 frames per shot by using dense sampling with 2x2 blocks to capture texture information. We follow the procedure in to extract LBP features. PCA is applied to reduce dimensions.

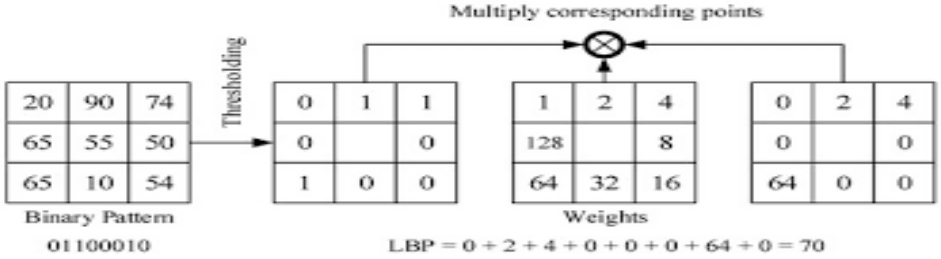

Fig 4:Example of LBP

6. **MFCC audio features (MFCC)**

Mel-frequency cepstral coefficients (MFCCs), which describe the short-time spectral shape of audio frames, are extracted to capture audio information. MFCCs are widely used not only for speech recognition but also for generic audio classification.

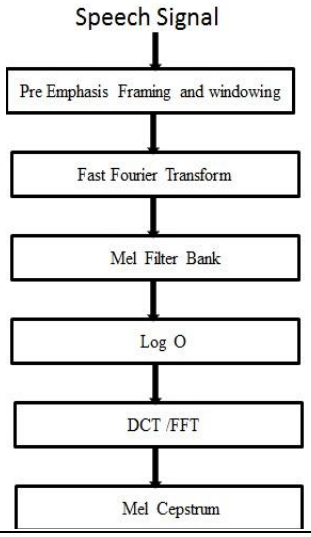

Fig 5: MFCC Derivation

Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, deemphasizes all other information. The speech signal is first divided into time frames consisting of an arbitrary number of samples [10]. ΔMFCCs, ΔΔMFCCs, Δ log-power and ΔΔ log-power are extracted in addition to the MFCCs. Here, "Δ" means the derivation of the feature.

7. Spectrogram Audio Features:

Spectrogram is used to extract the Audio features. A spectrogram is a visual representation of the spectrum of frequencies of sound or other signal as they vary with time or some other variable. There are many variations of format: sometimes the vertical and horizontal axes are switched, so time runs up and down; sometimes the amplitude is represented as the height of a 3D surface instead of color or intensity. The frequency and amplitude axes can be either linear or logarithmic, depending on what the graph is being used for.

Creating a spectrogram using the FFT is a digital process. Digitally sampled data, in the time domain, is broken up into chunks, which usually overlap, and Fourier transformed to calculate the magnitude of the frequency spectrum for each chunk. Each chunk then corresponds to a vertical line in the image; a measurement of magnitude versus frequency for a specific moment in time (the midpoint of the chunk). These spectrums or time plots are then "laid side by side" to form the image or a three-dimensional surface, or slightly overlapped in various ways, i.e. windowing.

2.2 Principal components analysis PCA:

Principal components analysis (PCA) is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information.Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.Traditionally, principal component analysis is performed on a square symmetric matrix. It can be a SSCPmatrix (pure sums of squares and cross products), Covariancematrix (scaled sums of squares and cross products), or Correlationmatrix (sums of squares and cross products from standardized data).We extract the low-level Feature from the Video Shot, low-level Feature extracted from every other frame, and principal component analysis (PCA) is applied to reduce their dimensions.

2.3 Gaussian Mixture Model (GMM):

A Gaussian mixture model is parameterized by two types of values, the mixture component weights and the component means and variances/covariances. A Gaussian mixture model (GMM) is a category of probabilistic model which states that all generated data points are derived from a mixture of a finite Gaussian distributions that has no known parameters. The parameters for Gaussian mixture models are derived either from maximum a posteriori estimation or an iterative expectation-maximization algorithm from a prior model which is well trained.

A Gaussian mixture model (GMM) is useful for modeling data that comes from one of several groups: the groups might be different from each other, but data points within the same group can be well-modeled by a Gaussian distribution.A GMM supervector consists of the parameters of a GMM for the distribution of low-level features extracted from a video clip. A GMM is regarded as an extension of the bag-of-words framework to a probabilistic framework, and thus, it can be expected to be robust against the data insufficiency problem. Probability density functions are used to model video shots[9].

The GMM parameters are estimated for each shot under the maximum a posteriori (MAP) criterion.The MAP solution for GMM means, namely MAP adaptation.The fast MAP adaptation technique, reduces computational costs for calculating posterior probabilities. After video shots are represented by GMMs, GMM supervectors are extracted by combining normalized mean vectors[9].

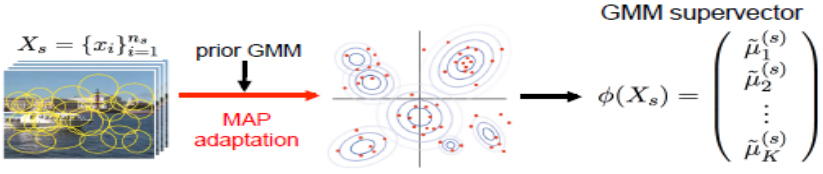

Fig 6: GMM Supervector [9]

We first make a GMM for a set of (a type of) feature vectors. Then, we construct a GMM supervector from each GMM by MAP adaptation. In the detection phase, we use it as an input for a SVM classifier. SVM are used to train descriminative models for each Semantic Concept.

The relationship between shots are useful for detecting semantic concept.After Calculating Probabilities in score fusion use re-ranking method to re-evaluate scores of video shots by using shot-score distribution.We used re-ranking video shots accordingto a query or a concept. Our hypothesis is that videos have rather homogeneous contentsand that the presence of a given concept in a video depends a lot on the nature of thevideo itself. Scores (here homogeneous to probabilities) are computed independentlyfor all video shots as their likeliness to contain a target concept using classifiers (ornetworks of classifiers) that were trained on the development set. The re-ranking isactually done by a re-scoring which is done in two steps. First, we compute a globalscore for each video for containing the target concept; this score is computed from thescores of all the shots within the video. Then, we re-evaluate the score of each shotaccording to the global score of the video it belongs to.Re-ranking method defines a video-clip score as the maximum value of shot scores among all the shots in a Video clip.Highest Probability Match will shows the Video Concept Detection Result[11].

IV. CONCLUSION

In this paper we discussed and studied a few audio and visual Local Feature Extraction techniques and described the working of those techniques in relation to Video Concept Detection. We proposed videoconceptdetection using a deep CNN and GMM super vectors with the audio and visual features. Alexnet will be used in our implementation. From each key frames we extract a 4096 features in deep convolutional neural network, used the sixth layer feature and then fed to train SVM for each concept to obtain probability score. The GMM parameters are estimated for each shot under the maximum a posteriori adaptation by constructing tree structure GMM approximated to a single Gaussian. After video shots are represented by GMMs, GMM supervectors are extracted by combining normalized mean vectors.Output of DCNN and GMM supervector are linearly combined to give the Scores. This score is again re-evaluated using re-ranking method to predict the final score and detected concept labels.

Journal Papers:

[1]. JingweiXu, Li Song, RongXie "Shot Boundary Detection Using Convolutional Neural Networks" in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Ghent, 2016.

[2]. Ravi Mishra ,S.K.Singhai,M. Sharma "Comparative study of block matching algorithm and dual tree complex wavelet transform for shot detection in videos‖ Electronic system, signal processing and computing technologies(ICESC)", 2014 International Conference, Jan 2014 .

[3]. Ganesh. I. Rathod, Dipali. A. Nikam , " An Algorithm for Shot Boundary Detection and Key FrameExtraction Using Histogram Difference". International Journal of Emerging Technology and Advanced Engineering ,ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013 .

[4]. Samira Pouyanfar and Shu-Ching Chen "Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning" International Journal of Semantic Computing World Scientific Publishing Company, 2017.

[5]. Alex Krizhevsky, IlyaSutskever, Geoffrey E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks" University of Toronto,2012.

[6]. Weiming Hu, Senior Member, IEEE, NianhuaXie, Li Li, XianglinZeng, and Stephen Maybank "A Survey on Visual Content-Based Video Indexing and Retrieval" IEEE TRANSACTIONS ON SYSTEMS,VOL. 41, NO. 6, NOVEMBER 2011.

[7]. Hsin-Yu Ha, Yimin Yang, Samira Pouyanfar,HaimanTian, and Shu-Ching Chen "Correlation-based Deep Learning for Multimedia Semantic Concept Detection " Florida International University,2016.

[8]. AshwinBhandare, Maithili Bhide, PranavGokhale, RohanChandavarkar"Applications of Convolutional Neural Networks" AshwinBhandare et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (5), 2016.

[9]. Nakamasa Inoue and Koichi Shinoda and Zhang Xuefeng and Kazuya Ueki "Semantic Indexing Using Deep CNN and GMM Supervectors", Tokyo Institute of Technology, Waseda University,2014.

[10]. Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition",International Journal For Advance Research In Engineering And Technology,Volume 1, Issue VI, July 2013.

[11]. B. Safadi and G.Qunot. Re-ranking by Local Re-scoring for Video Indexing and Reterival.InOroc.of CIKM,2011.

Ritika D Sangale "Convolutional Neural Network (CNN) and GMM Supervectors in Video Concept Detection. "International Journal of Engineering Science Invention(IJESI), vol. 7, no. 8, 2018, pp. 69-75