International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Key Feature Extraction for Video Shot Boundary Detection using CNN

Neelam Labhade Kumar, Yogeshkumar Sharma, Parul S. Arora

Abstract: ***Now days as the progress of digital image technology, video files raise fast, there is a great demand for automatic video semantic study in many scenes, such as video semantic understanding, content-based analysis, video retrieval. Shot boundary detection is an elementary step for video analysis. However, recent methods are time consuming and perform badly in the gradual transition detection.***

In this paper we have projected a novel approach for video shot boundary detection using CNN which is based on feature extraction. We designed couple of steps to implement this method for automatic video shot boundary detection (VSBD). Primarily features are extracted using H, V&S parameters based on mean log difference along with implementation of histogram distribution function. This feature is given as an input to CNN algorithm which detects shots which is based on probability function. CNN is implemented using convolution and rectifier linear unit activation matrix which is followed after filter application and zero padding. After downsizing the matrix it is given as a input to fully connected layer which indicates shot boundaries comparing the proposed method with CNN method based on GPU the results are encouraging with substantially high values of precision Recall & F1 measures. CNN methods perform moderately better for animated videos while it excels for complex video which is observed in the results.

Keywords: Video Shot Boundary Detection (VSBD), Convolutional Neural Networks (CNN). Rectifier linear Unit (ReLU)

I. INTRODUCTION

With the tremendous development of video data, content based video analysis and organization tools such as indexing, browsing and retrieval have drawn much attention. Video Shot Boundary Detection (SBD) is usually the preliminary step for those technologies. Great efforts have been made to improve the accuracy of SBD algorithms. However, most works are based on signal rather than interpretable features of frames.

Manuscript published on January 30, 2020.
* Correspondence Author
Neelam Labhade-Kumar*, Research Scholar, J. J. T. University, Rajasthan. Assistance Prof. JSPM's ICOER Wagholi, Maharashtra, India. Email: neelam.labhade@gmail.com
Dr. Yogeshkumar Sharma, Associate Prof. Shri J. J. T. University, Churella, Jhunjhunu, India. Email: dr.sharmayogeshkumar@gmail.com
Dr. Parul S. Arora, Associate Prof. JSPM's ICOER Wagholi, Maharashtra, India. Email: parulsarora@gmail.com

Deep learning refers to the shining branch of machine learning that's supported learning levels of representations. Convolutional Neural Networks (CNN) is one reasonably deep neural network. During this work, we tend to give an in depth analysis of method of CNN algorithm each the forward process and back propagation. Then we applied the actual convolutional neural network to implement the standard shot boundary detection by MATLAB. [4] Convolutional Neural Networks (CNN) is one reasonably feed forward neural network. The Convolutional Neural Network was originally planned for the task of zip code recognition. Each convolutional neural network was excessively applied to character recognition. Training was initially based on error back propagation and radiant descent.

A. Features of CNN

- An standard Neural Networks usually takes features as inputs, for this problem CNN take image array as inputs, therefore it has vector, size of (image width*height) as a input. Conv Nets are used primarily to look for patterns in an image, no need to provide features, and therefore the CNN understands the proper features by itself.[5]
- Ordinary neural networks don't scale well for full sized images, let's say that input images size =100(width) * 100 (height) * 3 (rgb), Then network has to process 30,000 neurons that are extremely costly within the network.
- CNN is an efficient recognition algorithm that is widely utilized in pattern recognition and image processing.
- Structure of CNN is very easy.
- To train the CNN very less training parameters are required.
- CNN has highest flexibility.
- Its weight shared network structure makes it more similar as biological neural networks. It reduces the complexness of the network model and also the range of weights.
- CNN is especially used to establish displacement, zoom and alternative types of distorting invariance of two-dimensional graphics.
- The fully connectedness of these networks makes them prone to over fitting data.

B. Construction of CNN

A convolutional neural network contains input layer and output layer, in addition to multiple hidden layers as shown in Fig.1. The hidden layers of a CNN generally include a

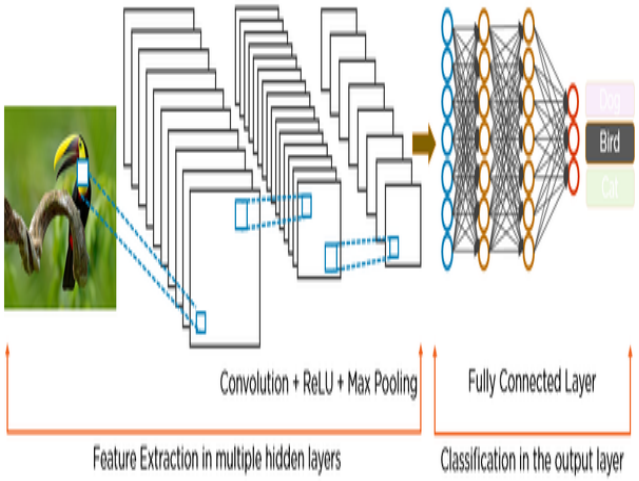


Fig.1 Structure of Convolutional Neural Network

series of convolutional layers that convolve with a multiplication or different dot product. The activation function is usually a RELU layer, and is later followed by pooling layers, totally connected layers and normalization layers, named as hidden layers. As a result of their inputs and outputs are masked by the activation function and final convolution.[2] The ultimate convolution usually involves back propagation so as to more accurately weight the end product. Although the layers are informally stated as convolutions, this can be only by convention. Arithmetically, it is strictly a sliding dot product or cross-correlation. This has import for the indices within the matrix, in this it affects how weight is decided at a specific index point.

II. LITERATURE REVIEW

The first convolutional neural network is predicated on weight sharing that was planned by D. E. Rumelhart, and G. E. Hinton in 1986. [3]

Wenjing Tong et.al. (2015), suggest a new video shot boundary recognition framework based on interpretable TAGs cultured by Convolutional Neural Networks (CNNs).

Tui Liang et.al. (2017), in "A Video Shot Boundary Detection Approach based on CNN Feature" projected a new approach which used CNN model to extract features of video sequence parallelly based on GPU, so it can simplify the expression of video and decrease the calculation time for shot detection, and took local frame similarity and dual-threshold sliding window similarity into consideration to increase recall and precise of shot detection.[1]

Saad Albawi et.al (2017) explain and define all the elements and important issues related to CNN, and how these elements work. In addition, they also state the parameters that effect CNN efficiency.[9]

Rikiya Yamashita et.al (2018) offers a perspective on the basic concepts of CNN and its application to various radiological tasks, and discusses its challenges and future directions in the field of radiology.[8]

III. PROPOSED ALGORITHM

Algorithm for projected system using CNN is shown in fig 2 which is explain as below

- Desired number frames from a video are captured. Each frame is converted from RGB space to HVS colour space.
- Plot histogram of H, V, S frames.
- Extract features using Mean Log Difference Method.
- Provide features of input image into convolution layer.
- Convolutions are done on the image and apply ReLU activation to the matrix.
- Perform pooling to decrease dimensionality size.
- Flatten the output and feed into a fully connected layer (FC Layer).

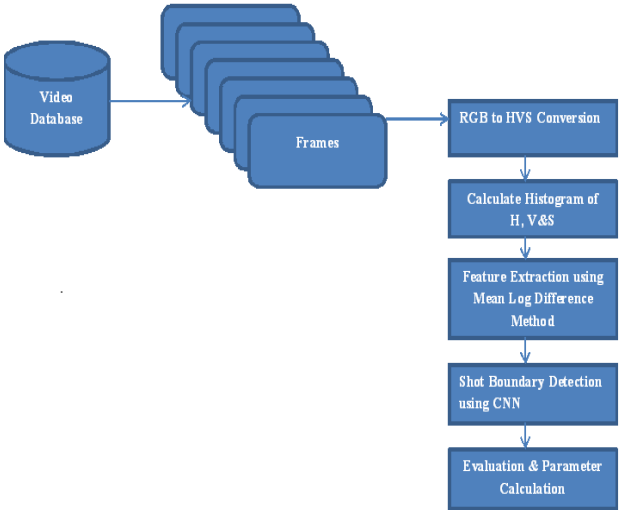


Fig.2-Proposed Algorithm of CNN

IV. RESULTS & DISCUSSION

In order to evaluate our method and compare with other methods, we employed 5 video clips like sport, movie, cartoon, wildlife and news and extracted frame sequence of each video. Then we manually marked the shot boundary indexes of transitions of each video and compared same with detected shots by CNN.

A.Video Segmentation – The first step in any video data management is segmentation of the video track into smaller units, enabling the succeeding processing operations on video shots, such as video indexing, semantic depiction or tracking of the selected video data and identifying the frames where a changeover takes place from one shot to another.

B.Manual Shot Boundary Detection-

manually the video is selected and closly observerd the shot change and it is recorded. Similarly the whole videois visualised manually to determine video shots and recorderd as shown in Fig.3 . In the further stage we have implemented a ground truth generator (GUG) to which we give total no of frams and the frame no where a shot is detected visually. The GUG shows the time line and the instant where shot is identified. In the ater stage the output of GUG is givento the computation of evaluation of parametrs.

Fig 3: Manual Shot Detection of Video.

C.RGB-HVS Conversion and its Histogram-

RGB-HVS Conversion and its Histogram is as shown in fig.4 (a)-4(d). In RGB feature extraction method R, G, B are co-related to the colour luminance which is similar to intensity therefore to separate colour information from luminance is difficult. RGB defines colour in terms of a combination of primary colours. In circumstances where shading portrayal assumes an essential job, the HVS shading model is frequently favored over the RGB model. To beat this restriction here we use HVS for highlight extraction which is only Hue, Value and Saturation. The HVS model depicts hues also to how the human eye will in general observe shading. In circumstances where shading portrayal assumes a vital job, the HVS shading model is regularly favored over the RGB model. Mathematical equation for RGB-HVS conversion is given by eq. no1-eq.no-6b.

$$Max = \max(r, g, b) \quad (1)$$

$$Min = \min(r, g, b) \quad (2)$$

$$\Delta = Max - Min \quad (3)$$

$$h = \frac{g - b}{\Delta} \quad ------Max = r \quad (4a)$$

$$h = 2 + \frac{b - r}{\Delta} \quad ------Max = g \quad (4b)$$

$$h = 4 + \frac{r - g}{\Delta} \quad ------Max = b \quad (4c)$$

$$V = Max \quad (5)$$

$$S = \frac{\Delta}{Max} \quad ------Max > 0 \quad (6a)$$

$$S = 0 \quad ------Max = 0 \quad (6b)$$

Fig.4(a) Original Frame

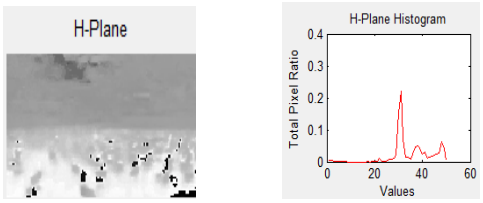


Fig.4(b) H plane and its Histogram.

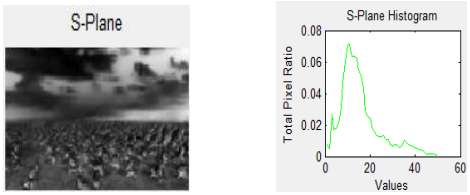


Fig.4(c) S plane and its Histogram.

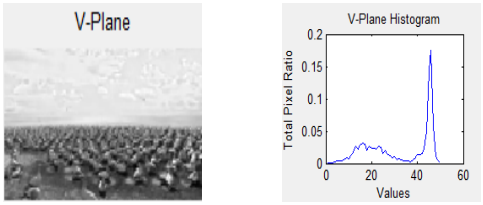


Fig.4(d) V plane and its Histogram

Fig.4 RGB-HVS Conversion & its Histogram.

A picture histogram is graphical portrayal of the color dispersion in a computerized picture. It plots the quantity of pixels for each color worth. This histogram is a representation appearing number of pixels in a picture at each extraordinary strength worth found in that picture. A histogram is a precise portrayal of the dissemination of numerical information. It is a gauge of the likelihood circulation of a constant variable. Mathematical equation for histogram calculation is given by eq. no7

Histogram Calculation

$$H(i) = \sum_{i=1}^{255} sum(I_{Pixel\ value=i}) \quad (7)$$

D.Means Log Difference Method (MLD) -

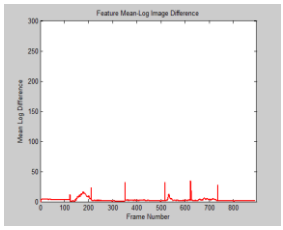
We have combined differentiation between current frame (I1) and the previous frame (I2) with the function for finding mean. Mathematical expressions for calculating mean log difference is given by equation no 8a-8d.In the next step we have converted to double precision floating point of the value corresponding to current and previous frame. Further this value is divided with the mean value and assigned to the respective variables. In final step difference between these values and its absolute value in the integer form is calculated .This is depicted in corresponding Fig 5.(a)-5.(e). Mathematical equation for mean log difference is given by eq. no8(a)-8(d)

$$Mean(\mu_1) = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} I_1\ (i,j)}{R * C} \quad (8a)$$

$$Mean(\mu_2) = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} I_2\ (i,j)}{R * C} \quad (8b)$$

$$Mean\ Difference\ Image = abs\left(\frac{I_1}{\mu_1} - \frac{I_2}{\mu_2}\right) \quad (8c)$$

$$Log\ Difference\ Image = |Log(I_1) - Log(I_2)| \quad (8d)$$



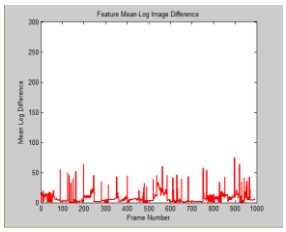
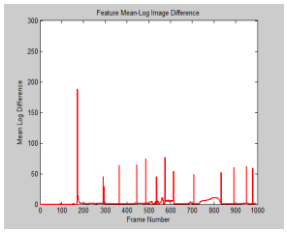


Fig.5(a) MLD of "Wildlife" **Fig.5(b) MLD of "Movie"**



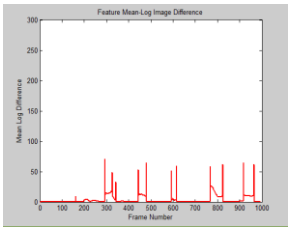


Fig.5(c) MLD of "Cartoon" **Fig.5(d) MLD of "Sports"**

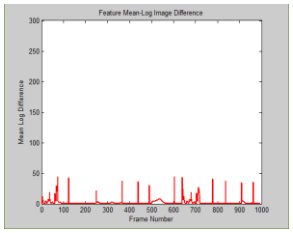


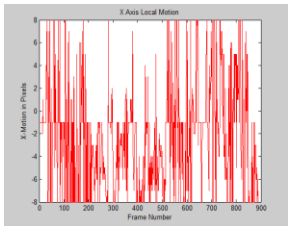
Fig.5(e) MLD of "News"

E. Local Motion Vector

Fig 6(a)-6(e) shows Local Motion Vector of experimental video. The local motion vectors of previous and subsequent frame re calculated in such a way that they indicate motion relative to previously transmitted frames. Motion vectors can be assigned to all blocks of frames of the video taken in to consideration. Block filtering within frame or block within previous frames is performed in which motion vectors are calculated. Feature based block matching method is incorporated to calculate motion vector. The X and Y components of the motion vector are required separately to calculate motion vectors in horizontal and vertical direction therefore those areas of the pixels/ frames which optimally represents global motion are selected separately according to X component and Y component of the motion vector. A block row and column are identified how's X components of motion vectors have little spread. While considering block row and block column only X components of motion vectors are taken in to account while their Y component are not considered. The row whose X component of motion vector have smallest extend is chosen as it is assumed for this o that this block doesn't substantially shows local motion in X direction as X vector is comparatively large spread in local motion. In similar way the column whose X components have the smallest value is selected. As a result block R and block C have been observed where motion vectors have smallest values and therefore it can be further used for shot boundary detection calculation. After calculating respective X & Y Components of local motion in terms of blocks some of absolute difference between consecutive block is calculated. Sum of absolute difference is found by resizing the input frames then finding the absolute value between the two frames finally the sum of this difference gives sum of absolute difference. Mathematical equation for linear local motion is given by eq. no9 &10

$$X(i) = \sum_{i=1}^{Frames} dist(currentframe_{ROI^x}, oldframe_{ROI^x}) \quad (9)$$

$$Y(i) = \sum_{i=1}^{Frames} dist(currentframe_{ROI^y}, oldframe_{ROI^y}) \quad (10)$$



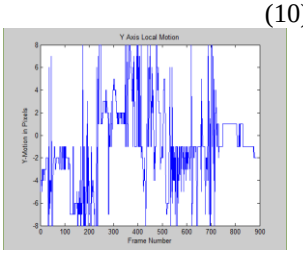
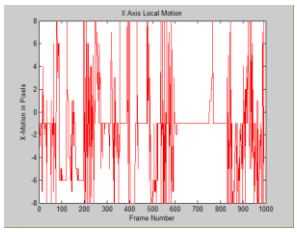


Fig.6(a) Local Motion Vector of "Wild Life"



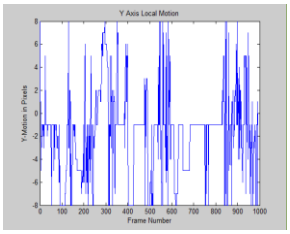
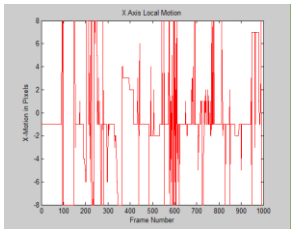


Fig.6 (b) X-Y Local Motion Vector of "Movie"



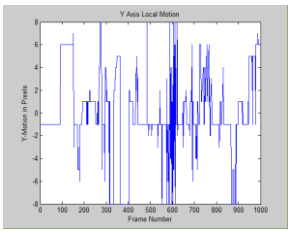
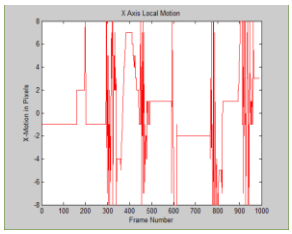


Fig.6(c) X-Y Local Motion Vector of "Cartoon"



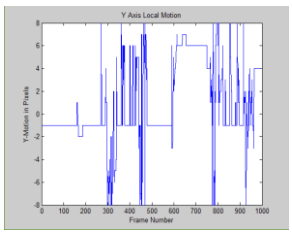


Fig.6 (d) X-Y Local Motion Vector of "Sports"

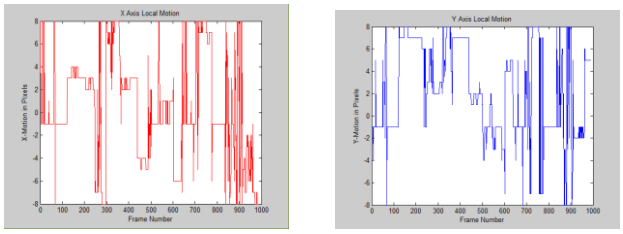


Fig.6 (e) X-Y Local Motion Vector of "News"

F. Shot Boundary Detection using CNN

The mean log difference values are given as a input for calculating CNN decision considering probability function. Incorporating handle function and displaying legend access a graph between frame number on X axis and CNN decision with probability on Y axis is represented in fig 7(a)-7(e). The peak represents not only CNN shot decision value but also shot probability function values which are calculated during feature extraction process. In GUI represented function also calculate count of total shot boundaries detected by algorithm. While this graph represents the same values in the form of shots. That means the area between two peaks represents one shot. In this way the total count calculated and the shots detected by CNN shot detection shot probability function are analogous to each other.

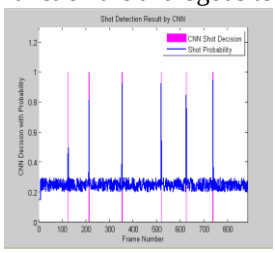


Fig.7(a) SBD of "Wildlife"

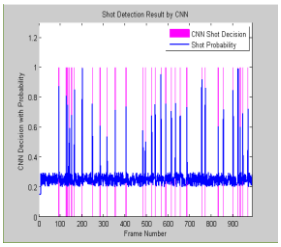


Fig.7(b) SBD of "Movie"

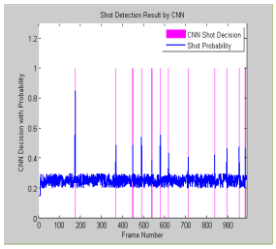


Fig.7(c) SBD of "Cartoon"

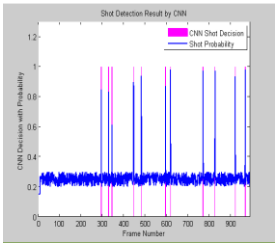


Fig.7(d) SBD of "Sports"

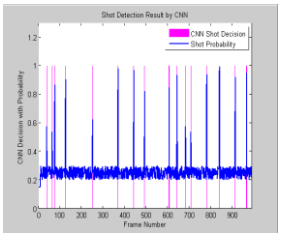


Fig.7(e) SBD of "News"

V. PARAMETER EVALUATION

Following parameters are used to evaluate the shot boundary detection method.

- **Precision:** Among the transitions (cut or gradual) detected, how many were true transitions. Mathematical expression for precision is given by eq.no10(a)

$$Precision = \frac{N_c}{N_c + N_f} \quad (10a)$$

Where
Nc is number of shot boundaries detected correctly
Nf is number of shot boundaries detected falsely

- **Recall**: For all possible transitions (cut or gradual) we marked manually, how many were detected. Mathematical expression for Recall is given by eq.no10(b)

$$Recall = \frac{N_c}{N_c + N_m} \quad (10b)$$

Where
Nc is number of shot boundaries detected correctly.
Nm is number of shot boundaries missed.

- F1: Defined as the comprehensive evaluation index of precise and recall. Mathematical expression for Recall is given by eq.no10(c)

$$F1 = \frac{Recall * Precision}{Recall + Precision} \quad (10c)$$

A good method should have high precision, high recall.
Table no 1 projects the experimental results of input videos taken in to consideration like cartoon, movies, wild life, nature, sports etc. the result shows consistency in terms of Recall, Precision and F1 measure. Closely when we observed the evaluation parameters for the given videos wild life frames depicts moderately less values of Recall, Precision and F1 measure due to abrupt change in semantic properties of video. This means CNN performs well for rest of the video except abrupt change in video feature as depicted in above table.

Table no.1 Experimental result of proposed CNN algorithm.

Sr.No	Video	R (%)	P (%)	F1 (%)
1.	Cartoon_1	92.30	90.46	91.37
2.	Cartoon_2	92.30	75.38	82.99
3.	Movie_1 (Chennai Express)	96.87	85.75	90.97
4.	Movie_2 (Bable)	94.11	98.00	96.01
5.	Movie_3 (Transformer)	72.72	82.66	77.37
6.	Wildlife	85.71	84.00	84.84
7.	Nature	70.00	86.22	77.26
8.	News_1	94.11	80.47	86.76
9.	Sports_1 (Tennis)	91.66	89.83	90.74
10.	Advertise (Bornvita)	95.45	93.54	94.49

Table 2 shows the comparison of evaluation parameters for different videos of CNN based o feature extraction method which is our proposed algorithm with CNN based on GPU[1].

For the cartoon video the proposed method gives better results in terms of precision & F1 measures but recall parameter is moderately low. In sports video the proposed method performs substantially better for all the three parameters while movies video continues with the same trade. This indicates the proposed method shows better performance for complex videos with sudden change in video features while there is a tradeoff for the recall parameters in animated video

Table no.2 Comparison of results of proposed algorithm with CNN based on GPU

Sr. No	Video	CNN			CNN Based on GPU		
		R%	P%	F1%	R%	P%	F1%
1.	Cartoon	92.3	94.4	93.3	95.6	93.7	94.7
2.	Sports	91.6	94.4	93.2	89.9	85.7	87.6
3.	Movie	96.8	93.0	94.9	93.9	90.5	92.2
4.	Mean	93.6	93.9	93.8	91.7	91.1	91.4

The performance of proposed algorithm and CNN algorithm based on GPU is depicted in fig8(a)-8(c) . The relation between recall, Precision for all three videos i.e. Cartoon, sports & movie has been shown clearly[6]. The comparison of performance calculation indicates the proposed algorithm performs well for precision well for precision & F1 measures substantially for all three videos while there is a little tradeoff for the recall parameter in animation video

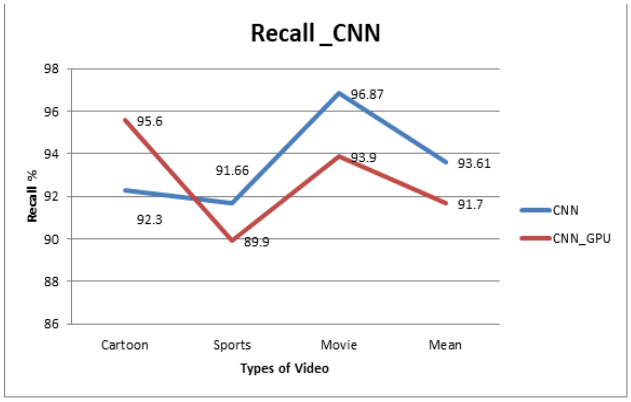


Fig.8a Parameter Evaluation- Recall

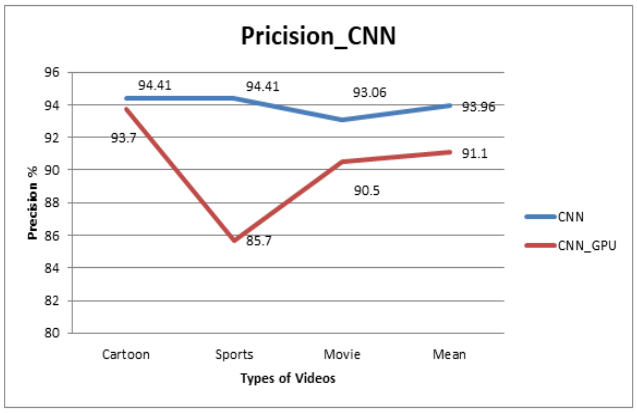


Fig.8a Parameter Evaluation- Recall

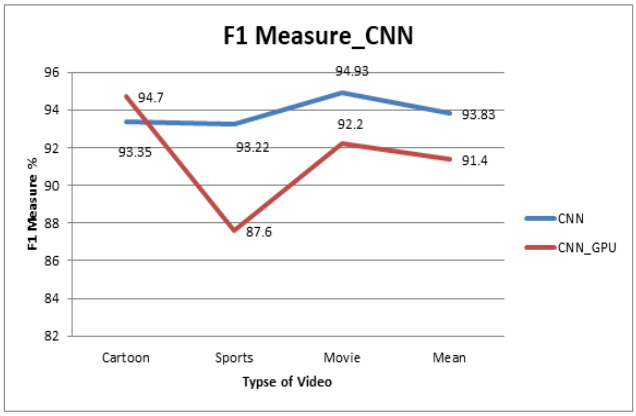


Fig.8a Parameter Evaluation- F1 measure

VI. CONCLUSION

In this paper we proposed a novel approach for video shot boundary detection method based on feature extraction using CNN. Initially features like H, V and S are extracted using mean log difference method & successively histogram distribution spectrum is defined.

In the later part CNN is incorporated for shot boundary detection. CNN implementation incorporates application of filters and padding of zeros column wise & row wise is done if required. Then convolution of image along with rectifier linear unit operation is performed to give matrix pooling method is to down scale the matrix which is then flattened and given as input to fully connected layers. This implies the output classified function in terms of peak values. The experimental result shows the outstanding performance of proposed algorithm in terms of precision, Recall and F1 measure. The values are consistent with number of videos which shows rigidity of the proposed algorithm. In future some motion features of videos, objects and camera will be considered for gradual transition feature, to get better and accurate shot detection for gradual transition.

REFERENCES

1. Rui Liang, Qingxin Zhu, Honglei Wei, Shujiao Liao, "A Video Shot Boundary Detection Approach based on CNN Feature" 2017 IEEE International Symposium on Multimedia ISBN: 978-1-5386-2937-6/17 © 2017 IEEE DOI 10.1109/ISM.2017.97
2. Ritika Dilip Sangale, "Overview of Video Concept Detection Using (CNN) Convolutional Neural Network",International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 12 | Dec-2017, e-ISSN: 2395-0056, p-ISSN: 2395-0072,PP 1265-1267
3. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representation by error propagation. In: Rumelhart, D.E., McClelland, J.L. (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 318–362. MIT Press, Cambridge (1986)
4. Neelam S. Labhade,Dr P.S.Arora,Dr.Yogeshkumar Sharma, "Study of neural networks in video Processng",Journal of Emerging Technologies and Innovative Research(JETIR),Val-6,Issue-3 PP-330-335,ISSN-2349-5162,March 2019.
5. Ganesh. I. Rathod, Dipali. A. Nikam, "An Algorithm for Shot BoundaryDetection and Key Frame Extraction Using Histogram Difference" International Journal of Emerging Technology and Advanced Engineering,Website: www.ijetae.com (ISSN 2250-2459,ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013)

6. Rui Liang, Qingxin Zhu, Honglei Wei, Shujiao Liao, "A Video Shot Boundary Detection Approach based on CNN Feature", 2017 IEEE International Symposiumon Multimedia, 978-1-5386-2937-6/17 IEEEDOI 10.1109/ISM.2017.97,Pp489-49
7. W. Tong, L. Song, X. Yang, H. Qu and R. Xie, "CNN-based shotboundary detection and video annotation," in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Ghent, 2015, pp. 1-5.
8. Rikiya Yamashita,Mizuho Nishio,Richard Kinh Gian Do, Kaori Togash,"Convolutional neural networks: an overview and application in radiology",August 2018, Volume 9, Issue 4, pp 611–629 Insights into Imaging
9. Saad Albawi ; Tareq Abed Mohammed ; Saad Al-Zawi, "Understanding of a convolutional neural network", 2017 International Conference on Engineering and Technology (ICET), 21-23 Aug. 2017, **INSPEC Accession Number:** 17615756, **DOI:** 10.1109/ICEngTechnol.2017.8308186 ,IEEE, Antalya, Turkey

AUTHORS PROFILE

Ms.Neelam S.Labhade has obtained ME in Signal Processing in 2014 and now persuing PhD in electronic engineering fromJ.J.T.University,Jhunjhunu,Rajasthan. Area of research is video processing.She has published more than 15 papers in several reputed journals.

Dr.Yogesh kumar Sharma has obtained his Ph.D in computer science. His area of research is wireless communication and image processing. He has published more than 50 papers in several reputed journals like Springer, Elsevier, and IEEE etc. He is working as Associate Professor in J.J.T.University,Jhunjhunu,Rajasthan.

Dr. Parul S. Arora has obtained her Ph.D in Electronics Engineering. Her area of research is Video shot boundary detection, Image enhancement and retrieval, Face recognition. She has published papers in several reputed journals like Springer, Elsevier, IEEE etc. She is working as Associate Professor in JSPM's ICOER Wagholi Pune.