Deep Learning Approach for Shot Boundary Detection

[1]Kevadkar Pragati Ashok, [2]Jadhav Dattatraya A
[1]Student,[2]Professor
[1]Department of Electronics & Telecommunication Engineering,
[1] JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India.

Abstract : Detecting shot boundary and gradual shot change happens to the prominent research problem in the field of video retrieval or indexing. Video shot boundary detection (SBD) is commonly an important and first step for indexing, and retrieval, video data, browsing, and content-based video analysis and many other such technologies. There have been great efforts put forward to enhance the precision of SBD algorithms. But many of these works are oriented towards signal and not towards interpretable features of frames. In our paper, we put forward a video shot boundary detection structure based on Convolutional Neural Networks (CNNs). Candidate segment selection is adopted which determines the locations of shot boundaries and removes most non-boundary frames. This kind of preprocessing method helps in improving both the speed and accuracy of the SBD algorithm. Later on, features of frames in a shot are synthesized and semantic labels for the shot are generated. This modular CNN network will be superior to the state-of-art methods on RAI Dataset with better than average real-time deduction speed even on just one mediocre GPU. Randomly generated transitions using selected shots from the TRECVID IACC.3 dataset will be employed under the training process. The proposed system achieved an average precision of 0.966, recall of 0.864 and F1 score of 0.912.

Index Terms **- CNN, RAI Dataset, Shot Boundary Detection (SBD), Video Frame**

I. INTRODUCTION

With the development of multimedia technology and the speedy growth of video content available online and offline, digital video has grown rapidly in terms of both quantity and quality. Subsequently, video browsing, indexing, retrieval, and other such video analyzing technologies have drawn tremendous attention. Being the initial step of all these mentioned technologies video shot boundary detection is also a fundamental step. As the transitions information does not exist in the video format, it is an important step in shot boundary detection systems to be automated for video management and retrieval. A common way to build-up a video is to make use of a shot composition, where shots get delimited with the aid of transitions.

A video shot can be defined as a series of images that are captured by an individual camera in a continuous run[1]. There exist two types of transitions between any two adjacent shots: Gradual Transition (GT) and Cut Transition (CT). CT comprises of the final frame of the former shot and the initial frame of the upcoming shot. While GT commonly comprises multiple frames and the transition among the subsequent shots is more softly. Thus we understand that video shot boundary detection is a process through which we identify the transition between every two neighboring connected shots [2]. Shot boundary detectors must be able to distinguish between shot transitions and any abrupt changes in a video which can be caused because of the partial blockage of the scene by an object passing closer to the camera. Movement of a faster-moving object or movement of the camera in the scene cannot is mistaken for a shot transition. This may show that some semantic representation of a scene is necessary to correctly segment a video.

By far most works have not paid attention to the frame content. But when we consider the definition of video shots, it becomes a natural thought of using the contents in the frame to detect shot. In our paper, we suggest using CNN in extracting interpretable features of the high-level present in the frames, hence encouraging the precision of SBD. Recently, CNN's [3] has been showcased as an effective class of models for understanding components of images and videos. In CNN, features such as angles and edges which are said to be low-level can be learned with the assistance of various initial layers. And as the networks go on deeper low-level features can be used in extracting high-level features[4]. Thus the proposed paper tries to achieve the shot boundary algorithm by temporal segmentation using CNN.

II. LITERATURE SURVEY

This part of the paper focuses on previous researches and works carried out in the domain of shot boundary detection. In 1996 Boreczky et al. [5] proposed a discussion that describes different techniques of video shot boundary detection and presented a study on how different shot boundary detection algorithms perform.

In 1999 Gauch et al. [6] introduced an innovative method for Shot Change Detection. His work stated, if two look alike shots were joined with a gradual crossfade, then the visual changes might become much smaller than expected. Shot detection in the VISION (video indexing for searching over networks) system is done by combining three-parameter:

- The mean brightness of each video frame,
- The change in pixel values among two consequent frames and
- The change in color distribution between two consecutive frames.

These 3 quantities are checked by comparing with dynamic thresholds which help in identifying the boundaries when shot changes.

In 2001, Heng et al. [7] suggested a shot detection technique that was object-based. His works proposed finding information with the aid of a timestamp which had a transferring mechanism from multiple frames. Gradual transitions were efficiently handled by this mechanism. This algorithm showed many advantages of traditional algorithms.

In that very same year, a simple method of histogram comparison was proposed by Lee et al. [8]. This method did not look upon the spatial information present in the frame and was unsuitable during highly sudden luminance changes. Later, Liao et al. [9] proposed a smarter way to dissolve detection in a video by using a model based on the binomial distribution. This method helped in determining the threshold required for discriminating the dissolve caused because of motions. Liu et al. [10] suggested an algorithm that had a constant false alarm ratio (CFAR) for video segmentation. For video cut detection, a theoretical strategy for determining threshold using the non-parametric CFAR was developed. It also had the capability of finding a controllable precision.

Fang et al. [11] used fuzzy logic to integrate hybrid features that can detect shot boundaries correctly. The shot boundary with gradual shot cuts and sudden shot cuts were detected by a different process. Different features were used for the fuzzy logic approach based temporal segmentation of videos. Cao et al. [12] proposed a classifier based approach to find the wipes and digital video effects. Six parameters that are the causes of the formation of feature vector were evaluated for each frame in a temporal window. A supervised SVM classifier based on feature vectors classifies the frames as no shot change gradually shot change and sudden shot change categories. An automatic shot detection technique proposed by Huang, et al. [13] performed very poorly due to high wrong detection rates caused by the camera or object motion. The problems in the shot detection mechanism were tried to be overcome by this work, which uses local key points to match with video frames that detected both sudden and gradual change efficiently. Changes with a long time are very hard to locate using low-level features.

The shot can be easily identified by comparing objects between two back to back frames [14]. On one hand effect of camera motion and object motion can be easily avoided by detecting objects inside the frames. Simultaneously it can detect efficiently both gradual transitions and abrupt transitions. Thus a single approach can be used for various shot boundary detection. ASCD can be taken as an automatic operation in a real-world application. The adaptive threshold can be calculated from the following frame of the last shot change to the previous one of a current frame. It also uses histogram variation of successive frames to set an automatic weighting factor. An automatic SCD algorithm using mean or average and variance-based have a detection rate higher as compared with the pixel-based procedure.

K-means clustering was proposed by Xu et al. [15] in the shot detection systems. It first extracted the color feature and then obtained the dissimilarity of video frames using the features. Grouping of video frames was performed by using the graph-theoretical algorithm as suggested in Xu et al.'s [16]. Detecting the transition in the shot was done using block-based motion as proposed by Park et al. [17]. The block-wise similarity in motion and modified displaced frame difference (DFD) is together used for the detection of shot changes. Mishra et al. [18] described an algorithm that at first mined structure features coming from each video frame [19] making use of dual-tree complex wavelet transform was computed between consecutive frames. Lu et al. [19] presented a Video Shot Boundary Detection technique that comprised of singular value decomposition (SVD) and segment selection with Pattern Matching. In this work shot, boundaries' position and gradual transitions' calculation of lengths was done using adaptive thresholds and most of the boundaryless frames were removed at the same time. Later a study comparing dual-tree complex wavelet transform based SBD algorithm [20] and block matching SBD algorithm were performed, for multiple parameters like a miss, the rate hit rate, false rate, verified on a set of the different video sequence.

To summarize in brief, we have seen the technology used for shot boundary detection emanate from studying how different shot boundary algorithms perform to using the VISION system for shot detection. Also, we have come across various practices for shot boundary detection such as studying SBD based in object detection, binomial distribution, fuzzy logic to integrate hybrid features, SVM based classifier approach, dual tree complex wavelet transform, singular value decomposition (SVD), Block matching SBD algorithm. Thus this literature survey gives us an insight into the initial works and developments in the domain of shot boundary detection.

III. PROPOSED SYSTEM

3.1. Block Diagram

Retrieval of a video and its analysis are challenging tasks due to the various video types of and special effects and several transitions that can be added. Along with this, a variety of other factors can be proven as a big challenge to SBD applications. In shot boundary detection, the fast object or camera motions, the special effects and the various illumination changes that may occur in a scene may lead to error detections. A robust shot boundary detection method should perform good detections for all types of transitions for any video sequence with reduced manual predefined parameterization. Fig. 1. Shows a block diagram of the claimed shot boundary detection system using a convolutional neural network.

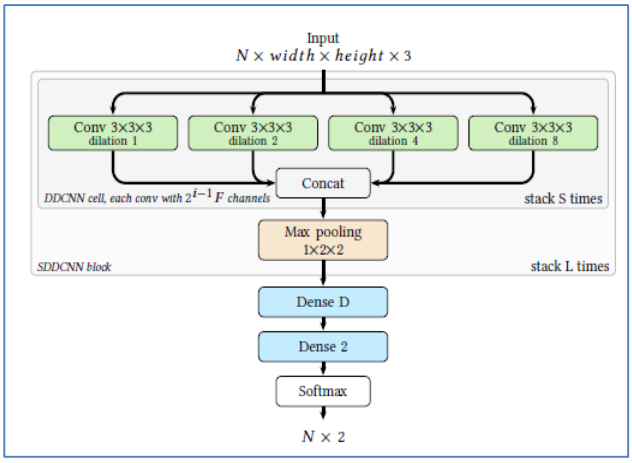

Fig. 1. Block diagram of the proposed CNN based shot boundary system

3.2. Dataset

The TRECVID IACC.3 dataset [22] was used as it has a set of predefined temporal segments. Thus, pairs of predefined segments can be non sequentially selected from the collection for automatic generation of transitions for the purpose of training. More specifically, we take segments of 3000 IACC.3 videos that are selected. Furthermore, segments that have more than 5 frames are only included and from the leftover set, every other part was chosen, which gives us 54884 segments as selected.

The examples of training will be generated only when the demand arises during training by randomly sampling 2 shots and connecting them with the assistance of the random type of transition. Training included only dissolves and hard cuts. The transition position was generated randomly. In dissolves, also the length was randomly generated from the interval. Each training sequence's length N was nominated to be 100 frames. Input frames size was set to 48 × 27 pixels. For model validation, an additional 100 IACC.3 videos were labeled manually, resulting in 3800 number of shots. The RAI dataset [30] was considered for testing purposes.

The database distribution for this approach is as shown in Table 3.1.

Table 3.1: Quantitative analysis

Database	Videos	Cut Transitions	Gradual Transition
Training	3539	122760	35698
Testing	500	5876	2422

3.3. Network Architecture

In our work, a scalable CNN architecture with multiple dilated 3D convolutional operations per layer is used (normally one is used) which results in the enhanced field view with fewer trainable parameters. the architecture is being trained on2 common types of transitions. The proposed system consists of two parts:(i) Training and (ii) Testing. In the training phase, the model is created by feeding training images with labels to the modified CNN architecture. To make validation accuracy of the system maximum,3 stage CNN network is used by hyper tuning the parameter.

The input frame will be tested in the testing phase with the created CNN model. The model classifies the frame according to its behavior. The proposed TransNet architecture (as shown in Fig. 1.) follows the work of Gygli [21], as well as other architectures for convolutional, which are standard. A sequence of N back to back video frames is taken as an input to the network and to which a series of 3D convolutions is applied giving back an estimate for every input frame. Each prediction shows how likely any given frame happens to be a shot boundary. The main building block of the model (Dilated DCNN cell) is curated as four 3D, (3×3×3) convolutional operations. The convolutions employ dilation at different rates for the time dimension and their then outputs are combined in the channel dimension. In this way, the amount of trainable parameters gets significantly reduced as when considered by comparing with standard 3D convolutions which have the same field of view. Stacked DDCNN block is formed by spatial max pooling which is preceded by several DDCNN cells on top of each. Features pulled out by the convolutional layers are refined by two fully connected layers that determine the possibility of shot boundary for every frame representation independently (layers' weights are shared). ReLU activation function is used in all layers but there is just one exception that is of the final fully connected layer. Stride 1 and the 'same' padding is employed in all of the convolutional layers.

3.4. Training

The proposed architecture provides the following meta-parameters that grid search investigates are:

- S stands for the count of DDCNN cells in an SDDCNN layer,
- L stands for the count of SDDCNN layers,
- F stands for the count of filters in the first set of DDCNN layers
- D stands for the count of dense layer neurons.

A batch size having 20 frames was used for training, for all investigated networks. To avoid overfitting, only 30 epochs were considered, each of which had 300 batches. According to our primary evaluations, results were not improved by dropout. Still, we plot to investigate training data augmentation and advanced forms of regularization in the future. Depending on the architecture, the entire training takes some hours to complete on GPU. Also for the case of dissolves, when the change is over several frames, the network has been trained to suggest only the middle frame as a shot boundary. This creates a divergence between the number of frames that do not comprise of a transition (99 in our case)and 'transition' frames (where each sequence contains only one). Increased weight of the transitions in the loss function was not able to produce improved results than reducing the acceptance threshold θ under commonly used 0.5; therefore, the approach explained latter is used.

3.5. Testing

During testing and validation, the list of shots is constructed as follows: The show begins at the first frame when the prediction falls below a threshold θ and ends at the first frame when the prediction surpasses θ. When processing a video, the input window is shifted by 50 frames between individual forward passes through the network. The F1 score is used as an evaluation metric which is the same metric as in [30]. The reported F1 score is computed as an average of individual F1 scores for every video. On analyzing we can see, when detected shots are considered to be a false negative, false positive, true positive. If the recognized shot transition coincides with the ground truth transition then a true positive is detected. And if the suggested transition does not coincide with the ground truth or the transition is detected for the second time then a false positive is detected. Lastly, if there is no transition which overlaps with the ground truth and the ground truth transition gets missed then a false negative is detected.

IV. IMPLEMENTATION

The flow diagram of the system is as shown in Fig. 2. Flowchart explains the system implementation. It starts with the loading input frame from the training data set or from the real-time video input. A series of 3D convolution network is applied to a sequence of N video frames which returns a prediction of every frame from the given input. Thus each prediction helps in expressing how likely is the given frame a shot boundary. If it detects a shot boundary it marks it as a frame output and saves it into a folder.

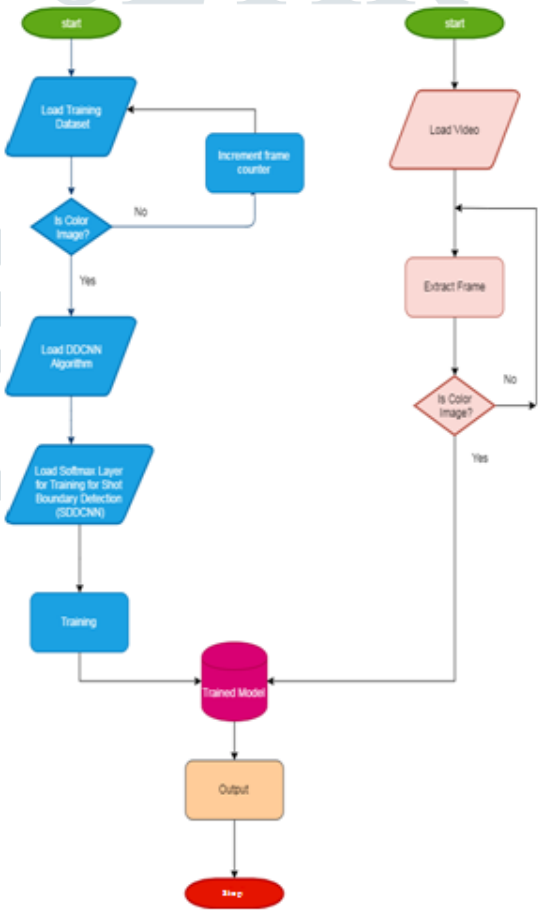

Fig.2. Flow chart of Proposed System

V. RESULTS

The proposed shot boundary detection system is implemented using the OpenCV library with the python language. The Keras and TensorFlow libraries are used to implement deep learning algorithms. The proposed algorithm is tested on the TRECVID IACC.3 dataset. The results are presented in qualitative and quantitative ways.

5.1. Qualitative Analysis

Qualitative analysis is the pictorial and non-statistical representation of the research. The results of the proposed system implementations have been shown below in Fig. 3 and Fig. 4.

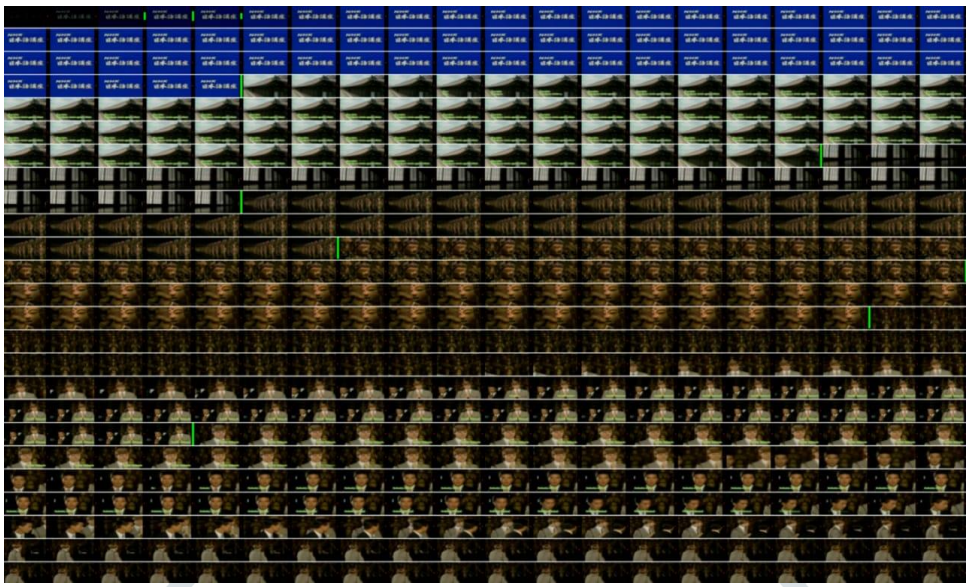

Fig.3. Results of shot boundary detected

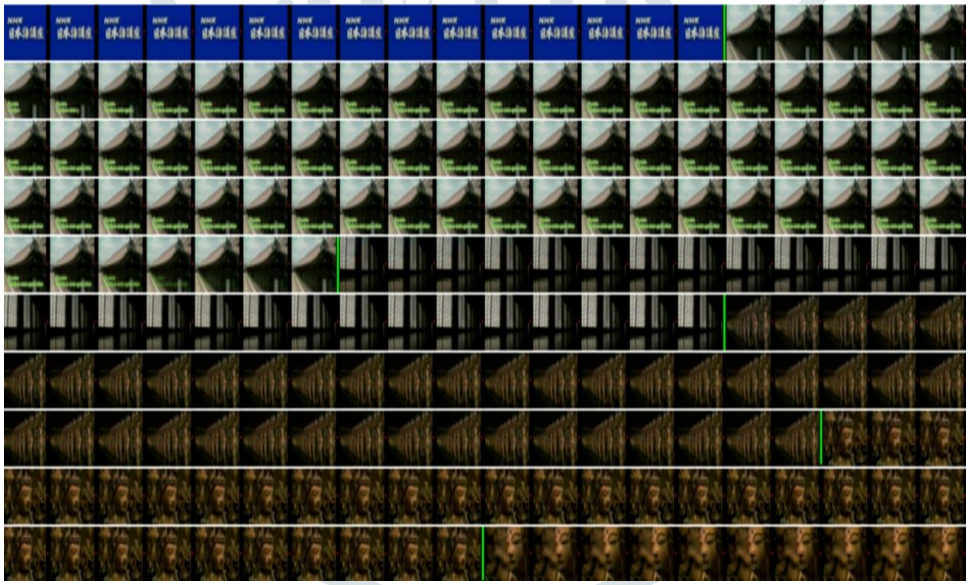

Fig.4. Results of shot boundary detected

In the implementation of the system videos of various categories such as sports, news, nature, infographic, movie, nature, and the cartoon is loaded into the system for extracting the frames. The light green color bar in results shows figures show the detection of shot boundary when the frame changes.

The GroundTruth (GT) frame transition of the testing video and the results of the proposed system is shown below.

GT -

{"transitions": [[64, 65], [136, 137], [164, 165], [206, 207], [239, 240], [277, 278], [363, 364], [531, 532], [572, 573], [632, 633], [704, 705], [1101, 1102], [1136, 1137], [1218, 1219], [1270, 1271], [1349, 1365], [1372, 1403], [1422, 1431], [1517, 1554], [1712, 1713], [2094, 2095], [2134, 2135], [2240, 2241], [2420, 2421], [2456, 2457], [2509, 2510]], "frame_num": 2775.0}

Pred-

[[0, 2], [5, 64], [65, 136], [137, 164], [165, 206], [207, 239], [240, 277], [278, 363], [364, 531], [532, 572], [573, 632], [633, 704], [705, 1101], [1102, 1136], [1137, 1218], [1219, 1270], [1271, 1362], [1363, 1712], [1713, 2094], [2095, 2134], [2135, 2240], [2241, 2420], [2421, 2456], [2457, 2509], [2510, 2774]]

From the above transitions, it is observed that the proposed system able to detect a maximum number of transitions effectively and correctly.

5.2. Quantitative Analysis

The quantitative analysis of the proposed system is calculated with precision, recall, and F1 score. Eq.1- Eq.3 represents them mathematically respectively.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1\ score = \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Here, TP symbolizes true positive value that defines the occurred movement as shot boundary detected. FP symbolizes false positive value. FN is a false negative value.

In Table II. given below category of input video, video name, the precision of system, recall and F1 score are listed. As mentioned previously in testing, the F1 score is used as an evaluation metric.

Table 3.1: Quantitative analysis

Video Name	Precision	Recall	F1 Score
0EXCdXUN_fk.mp4	0.98	0.82	0.89
2i9mB1EQV7k.mp4	1	0.88	0.94
5L5used_AY0.mp4	0.98	0.83	0.9
7z0Bs6SHRx4.mp4	0.96	0.9	0.93
agu2oLm-QKA.mp4	0.91	0.89	0.9
Overall	**0.966**	**0.864**	**0.912**

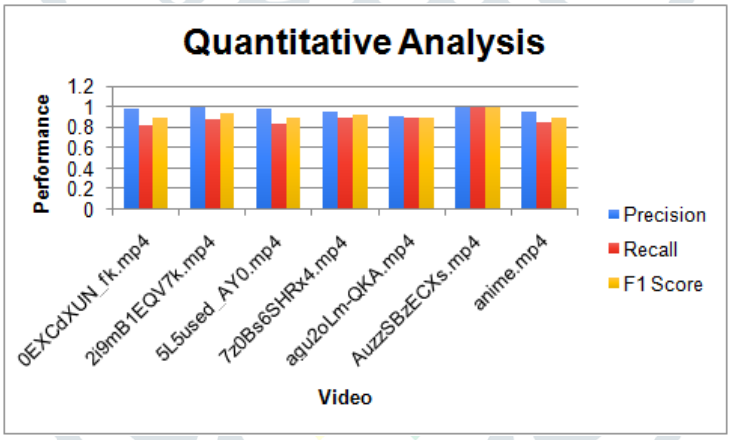

Fig.5. Graph of quantitative analysis of the proposed system

Fig. 5. demonstrates the performance of the system as a comparison between various categories of input videos. It tries to represent the comparison for precision, recall and F1 score. Thus we get the idea of our system will react to the above-mentioned categories of input.

VI. Conclusion

Shot change detection proves to be a very stimulating task. In our paper, we present an unusual approach to shot boundary detection based on Convolution Neural Network (CNN). When consecutive shot boundaries change with smaller variations and their backgrounds are very similar many states of art methods fail to detect boundaries with high accuracy. The method proposed in this paper uses RAI Dataset with the CNN network to superior results than traditional practices. We have tested the proposed system on the TRECVID IACC.3 dataset and made use of Keras and TensorFlow algorithms. Such use of shot boundary detection finds huge applications in software in post-production videos, content-based video retrieval, automated indexing, and summarization application. The proposed algorithm achieved an average precision of 0.966, recall of 0.864 and F1 score of 0.912.

The proposed system gives shot boundary detection implementation for the cut transition. So the important direction for future work is to establish a system to implement shot boundary detection for a gradual transition and fade transition. Another suggestion for future work is implementing the system with more improved algorithms and enhanced training data sets to achieve even better accuracy.

References

[1] C. Cotsaces, N. Nikiolaidis, and I. Pitas, Video shot detection and condensed representation. A review, IEEE Signal Process. Mag, Vol. 23, no. 2, pp. 28-37, Mar. 2006.

[2] J. H. Yuan, H. Y. Wang, L. Xiao, W. J. Zheng, J. M. Li, F. Z. Lin, and B. Zhang, A formal study of shot boundary detection, IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 2, pp. 168-186, Feb. 2002.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, 1998.

[4] M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901, 2013.C. Cotsaces, N. Nikiolaidis, and I. Pitas, Video shot detection and condensed representation. A review, IEEE Signal Process. Mag, Vol. 23, no. 2, pp. 28-37, Mar. 2006.

[5] Boreczky, J.S., Rowe, L.A.: Comparison of video shot boundary detection techniques. Journal of Electronic Imaging 5(2), 122–128 (1996)

[6] Gauch, J.M., Gauch, S., Bouix, S., Zhu, X.: Real-time video scene detection and classification. Information Processing and Management 35, 381–400 (1999)

[7] Heng, W.J., Ngan, K.N.: An Object-Based Shot Boundary Detection Using Edge Tracing and Tracking. Journal of Visual Communication and Image Representation 12, 217–239 (2001)

[8] Lee, M.-S., Yang, Y.-M., Lee, S.-W.: Automatic video parsing using shot boundary detection and camera operation analysis. Pattern Recognition 34, 711–719 (2001)

[9] Liao, H.Y.M., Su, C.W., Tyan, H.R., Chen, L.H.: A motion-tolerant dissolve detection algorithm. In: IEEE 2nd Pacific-Rim Conference on Multimedia, vol. 2195, pp. 1106–1112 (2002)

[10] Liu, T.-Y., Lo, K.-T., Zhang, X.-D., Fengc, J.: A new cut detection algorithm with constant false-alarm ratio for video segmentation. J. Vis. Commun. Image R. 15, 132–144 (2004)

[11] Fang, H., Jiang, J., Feng, Y.: A fuzzy logic approach for detection of video shot boundaries. Pattern Recognition 39, 2092–2100 (2006)

[12] Cao, J., Cai, A.: A robust shot transition detection method based on support vector machine in compressed domain. Pattern Recognition Letters 28, 1534–1540 (2007) 10.

[13] Huang, C.-R., Lee, H.-P., Chen, C.-S.: Shot Change Detection via Local Keypoint Matching. IEEE Transactions on Multimedia 10(6), 1097–1108 (2008)

[14] Kim, W.-H., Moon, K.-S., Kim, J.-N.: An Automatic Shot Change Detection Algorithm Using Weighting Variance and Histogram Variation. In: ICACT, pp. 1282–1285 (2009

[15] Xu, L., Xu, W.: A Novel Shot Detection Algorithm Based on Clustering. In: 2010 2nd International Conference on Education Technology and Computer (ICETC), pp. 1570–1572 (2010)

[16] Xu, W., Xu, L.: A Novel Shot Detection Algorithm Based on Graph Theory. In: 2010 2nd International Conference on Computer Engineering and Technology (ICCET), pp. 3628–3630 (2010)

[17] Park, M.-H., Park, R.-H., Lee, S.-W.: Efficient Shot Boundary Detection for Action Movies Using Blockwise Motion-Based Features. In: Bebis, G., Boyle, R., Koracin, D., Parvin, B. (eds.) ISVC 2005. LNCS, vol. 3804, pp. 478–485. Springer, Heidelberg (2005)

[18] Mishra, R., Singhai, S.K., Sharma, M.: Video shot boundary detection using dual-tree complex wavelet transform. In: 2013 IEEE 3rd International Advance Computing Conference (IACC), pp. 1201–1206 (2013)

[19] Lu, Z.M., Shi, Y.: Fast Video Shot Boundary Detection Based on SVD and Pattern Matching. IEEE Transactions on Image Processing 22(12), 5136–5145 (2013)

[20] Mishra, R., Singhai, S.K., Sharma, M.: Comparative study of block matching algorithm and dual tree complex wavelet transform for shot detection in videos. In: 2014 International Conference on Electronic System, Signal Processing and Computing Technologies (ICESC), pp. 450–455 (2014)

[21] Gygli, Michael. "Ridiculously Fast Shot Boundary Detection with Fully Convolutional Neural Networks." 2018 International Conference on Content-Based Multimedia Indexing (CBMI) (2017): 1-4.

[22] George Awad, Asad Butt, Jonathan Fiscus, Martial Michel, David Joy, WesselKraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. 2017. TRECVID 2017: Evaluating Ad-hoc and instance Video Search, Events Detection, Video Captioning, and Hyperlinking. In Proceedings of TRECVID 2017. NIST, USA.