Journal of Interdisciplinary Mathematics
ISSN: 0972-0502 (Print), ISSN: 2169-012X (Online)
Vol. 27 (2024), No. 2, pp. 369–382
DOI : 10.47974/JIM-1857

Mathematical modelling and statistical analysis in designing deep learning-based shot boundary detection and aesthetic assessment of videos

Madhura Phatak *
Department of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
MIT World Peace University (MIT-WPU)
Pune
Maharashtra
India

Manasi Patwardhan †
Tata Consultancy Services-Research
TCS
Pune
Maharashtra
India

Abstract

The main goal of this study is to create a strong deep learning-based system for finding shot boundaries and judging the quality of movies by combining mathematical models and statistical analysis. Aesthetic evaluation helps people understand visual material better, and shot border recognition is an important part of video analysis. We come up with a new approach that uses convolutional neural networks (CNNs) to improve the accuracy of shot border recognition by extracting features and using mathematical modeling. The model's performance is improved through statistical analysis, which makes sure that changes between shots are correctly identified. Proposed method also includes judging how something looks by combining deep learning methods with mathematical models that understand how things look. This two-in-one model gives a full picture of video material, which helps with both content selection and improving the user experience. We test our method on a variety of video datasets and show that it works better than other methods. The way our deep learning approach uses mathematical models and statistical analysis together not only raises the bar for finding shot boundaries, but it also helps the new field of judging the aesthetic quality of video material. This study creates a useful and flexible tool for analyzing videos, which

---

* *E-mail:* **madhuraphatak17@gmail.com (Corresponding Author)**
† *E-mail:* **manasi.kelkar@gmail.com**

opens the door to better uses in fun, content creation, and multimedia. The proposed method is implemented MATLAB and the performances evaluated based on performance metrics such as such as accuracy, precision, sensitivity, specificity and recall. And the accuracy of the proposed system is obtained as 93.12%. And to assure that the proposed system functions effectively in aesthetic assessment of video a contrast is made with the conventional technique such as Hand-Crafted Feature (HCF) technique, Neural Network (NN) and Support Vector Machine (SVM).

Subject Classification: *03D78.*

Keywords: *Shot boundary detection, Aesthetic classification, AlexNet CNN, Quality video, Video processing.*

1. Introduction

Because of improvements in communication technology, multimedia has grown very quickly. This has caused a huge amount of video data to flood in from many places, such as police cameras, long films, and sports events [1, 2]. Video processing, an important step before information gets to watchers, includes tasks like summary, abstraction, labeling, search, and user interface design in various fields [3]. Two important methods in video processing are shot border recognition and beauty rating [4]. These are used in the beginning stages of video analysis and have a big impact on the whole process. A "shot" is a series of pictures that are taken one after the other by a camera [5, 6]. It is important to find the borders between shots. These [7] changes can happen quickly (cut transition) or slowly, using fades, wipes, or splits to connect frames. Aesthetic evaluation, on the other hand, looks at how visually appealing films are to people, with better aesthetic quality being better [8, 9]. Image aesthetics have been tried to be estimated in the past using handmade features, optimization, and grouping [10], but there are still problems with detecting shot boundaries and judging video aesthetics. New methods for judging beauty have been studied, including Machine Learning (ML) and optimization techniques like Neural Networks (NN), Support Vector Machines (SVM), Genetic Algorithms (GA), and fuzzy systems [13, 14].

The proposed system has some major contribution in the process of analysis of video content which are given as follows.

- To develop deep learning-based technique for aesthetic assessment of video using statistical analysis.

- To attain effective classification among appealing and non-appealing video, the proposed system is designed.
- The proposed method is designed for efficient detection among cut transition and gradual transition in video shot to analysis aesthetic assessment of video.
- Deep learning based approach is used in this proposed system for performing automatic aesthetic classification.

Several approaches were developed by various authors to perform shot boundary detection boundary and aesthetic assessment of video. Among them fewer approaches were designed on the basis of ML algorithms. Some of the research work related to shot boundary detection and aesthetic assessment of video will be reviewed in this section. [11] had designed orthogonal polynomial (OP) for achieving efficient shot boundary detection. In this present work for identifying the transitions present within the video sequences, the features that are obtained based on orthogonal transform domain was utilized. The frames in the [12] video were represented with the help of moment in order to restrain the information content. For finding moment, the developed OP namely squared Krawtchouk-Tchebichef polynomial was used. The obtained [16] moments will be joined together to generate the feature vector. At last to classify the different transition Support Vector Machine (SVM) was used.

2. Proposed Method

The tremendous growth in the field of multimedia had led to the development of digital information system in terms of both quantity and

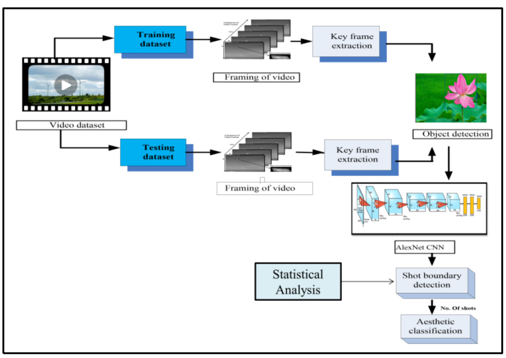


Figure 1

Architecture of Proposed System

quality. So analysis of video content automatically [17] is considered as significant and it is essential for further processing of video. For achieving effective processing of video, generally the approach such as aesthetic assessment of video was considered as vital. Above Figure 1, shows the architecture of the proposed system.

2.1 *Collection of Video dataset*

The video dataset shown in Figure 2 take an in input to proposed system to created, through conducting the different survey. Generally, many number of shots will be joined together to form a video. The collection of videos is the foremost step in video analysis and it is considered as significant.

2.2 *Training phase*

The training video dataset that was created in previous step is considered in order to perform further processing.

2.2.1 Framing of video

After the video trimming the next step employed in the proposed system is the framing of video. The individual frames will appear as an image. The processing can be done effectively with the help of framed video.

2.2.2 Extract key Frame from Image

Usually, many numbers of frame will be present in video [19]. The sequence of frame which is included to be representative and which possess the summary related to the entire content of video is referred to as key frame.

Key frame extraction algorithm

One way to get a key frame from an algorithm [18] is to look at the absolute difference between the histograms of two or more frames.

Step 1: Extraction of frames from video

- The initial step carried out in this algorithm is extraction of the frames from the trimmed video.

Step 2: Calculate histogram

- Once the frames have been extracted, the next step is to find the difference in the histograms of two frames that come after each other.

$$d(H1,H2)=\sum_{I}\frac{(H1(I)-H2(I))^2}{H1(I)} \tag{1}$$

Step 3: Estimate mean and standard deviation of absolute difference

- The mean and standard deviation of the absolute difference of the histogram.

$$\mu=\frac{\Sigma_1^N diff(i)}{N} \tag{2}$$

$$\sigma=\sqrt{\sum\frac{(diff(i)-\mu)^2}{N}} \tag{3}$$

Step 4: Compute the threshold value

- The value for threshold related to absolute difference of histogram will calculated based on mean and standard deviation value obtained in previous step.

$$T=\mu+\sigma \tag{4}$$

Step 5: Proceed till the end of video

- The above process for key frame extraction will be kept going until the end of the video is finished so that all of the key frames in the chosen video can be gathered.

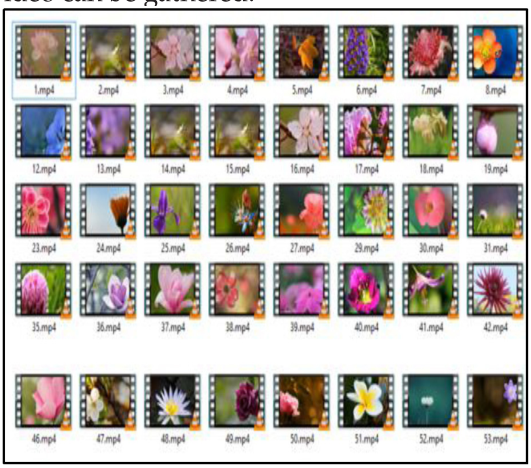


Figure 2

Sample Dataset

3. GrabCut algorithm

The GrabCut algorithm [20],[21] is a way to neatly divide a picture into focus and background areas based on what the user says. An repeated improvement process in the program makes the division better.

Algorithm:

1. Initialization:
   - Define an initial bounding box around the object of interest.
   - Assign each pixel to either the foreground or background based on its location inside or outside the bounding box.
2. Gaussian Mixture Model (GMM):
   - Represent the image as a mixture of Gaussian distributions for both foreground and background.

$$P(x \mid G_i) = \left(\frac{1}{(2\pi)^{\frac{3}{2}} |\Sigma_i|^{\frac{1}{2}}} \right) * \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right) \tag{5}$$

3. Graph Cuts:
   - Construct a graph where pixels are nodes, and edges represent the relationships between pixels.
   - The optimal segmentation by minimizing the energy function.

$$E(U,V) = \Sigma_{\{p \in U\}} Dp + \Sigma_{\{(p,q) \in V\}} V\_\{pq\} \tag{6}$$

4. Iteration:
   - Refine the segmentation by iteratively updating the GMM parameters and optimizing the graph cuts.

---

4. Design of CNN model: AlexNet

There are eight layers in AlexNet CNN as shown in Figure 3, including three Fully-Connected Layers [22], [23], and two Convolutional and Pooling layers. It lets you train on more than one GPU by spreading neurons across GPUs [24]. Using a model that has already been taught for Image Classification tasks, AlexNet is very good at finding objects in pictures. It works with pictures that are 224 x 224 in size.

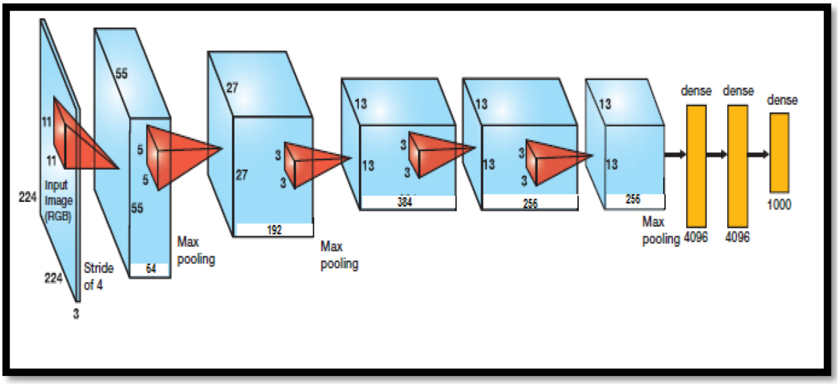


Figure 3

Representation of AlexNet

Algorithm:

1. Input Layer:
   - Accepts a color image with dimensions (224 x 224).
2. Convolutional Layer 1 to 5
   - Applies the first convolution operation with a filter size of 11x11, a stride of 4, and 96 filters.
   - ReLU Activation function:

$$a^{(1)(l,m,n)} = ReLU\left(\sum_{x=0}^{10}\sum_{y=0}^{10}\sum_{z=0}^{3} w^{(1)(d,p,q,r)}, * x(1)(i+l, j+m, n) + b^{(1)(p)}\right)$$

3. Max Pooling Layer 1 to 3
   - Performs max pooling with a filter size of 3x3 and a stride of 2.

$$a^{(2)(l,m,n)} = \max(l = 0 \text{ to } 2)\max(m = 0 \text{ to } 2) a^{(1)(2i+l, 2j+m, k)}$$

4. Fully Connected Layer 1:
   - Connects all neurons from the previous layer to each neuron in this layer.

$$a^{(9)(k)} = ReLU\left(\sum_{l=1}^{6}\sum_{r=1}^{6}\sum_{s=1}^{2566} w^{(6)(i,j,l,k)} * a^{(8)(i,j,l)} + b^{(6)(k)}\right)$$

5. Fully Connected Layer 2:
   - Connects all neurons from the previous layer to each neuron in this layer.

$$a^{(10)(k)} = ReLU(\Sigma(l = 1 \text{ to } 4096)w^{(7)(l,k)} * a^{(9)(l)} + b^{(7)(k)})$$

6. Fully Connected Layer 3 (Output Layer):
   - Multi-class classification.

$$a^{(11)(k)} = \frac{e^{z^k}}{\Sigma(l=1 \text{ to } 1000)e^{z^l}}$$

$$z^k = \Sigma(l = 1 \text{ to } 4096)w^{(8)(l,k)} * a^{(10)(l)} + b^{(8)(k)}$$

---

5. Result and Discussion

Table 1 shows the difference between the real number of shots in a movie and the number of shots found by CNN and a handmade method. The results show that the proposed system is more accurate at finding shots than the old way that was done by hand.

This shows the efficiency of the proposed system is shot boundary detection. Following the process of shot boundary detection, the second process performed in CNN is aesthetic classification. The output obtained as result of shot boundary detection is sent as an input into classifier to achieve aesthetic classification of video.

Figure 4 shows a comparison between the picture that was given and the image that was split from it. The input image is the original picture that was sent in, and the split image shows how the image segmentation process worked by showing different areas or items as shown in Figure 4.

Table 1

Number of Shots Detected In Shot Boundary Detection

Sr. No	Contrast value	Shots present	Number of shots detected (handcrafted)	Number of shots detected (CNN)
1	High	5	3	5
2	Low	4	2	3
3	Low	2	2	2
4	High	2	1	2
5	low	5	2	5

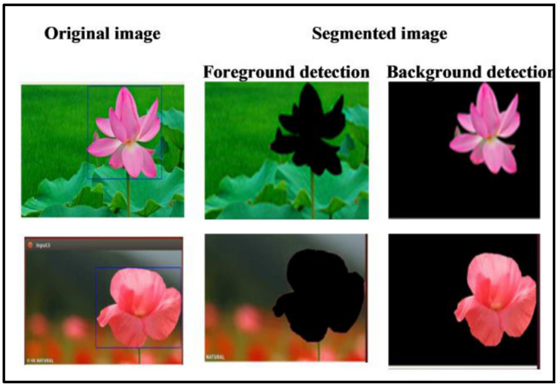


Figure 4

Representation of Input Image and Segmented Image

This picture helps us figure out how well the division process is working. The statistical parameters values that were found for the proposed system are shown in Table 2. A scientific method is used to figure out the numbers for these metrics. The accurate number will be the main factor that determines how well any suggested system works.

When the accuracy number is found to be 1, the suggested method is said to be 100% effective. The number for accuracy for this suggested system is 0.9312. The suggested system's memory, accuracy, sensitivity, and specificity were then found to be 0.9023, 0.9223, 0.9456, 0.9607, and 0.9123, respectively. These measures can be used to get a rough idea of how well the suggested system will work. Also, to show that the suggested system works well for classifying videos based on their aesthetics compared to other methods, a comparison study needs to be carried out.

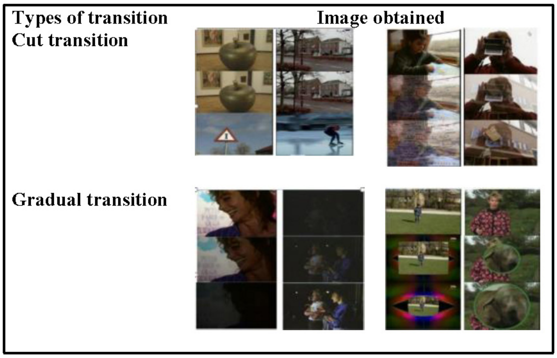


Figure 5

Shot boundary Detection

Table 2

Metrics for measuring how well the proposed system works

Performance metric	CNN AlexNet
(in %) Accuracy	93.12
(in %) Recall	90.23
(in %) Precision	92.23
(in %) F1 score	94.56
(in %) Sensitivity	96.07
(in %) Specificity	91.23

The purpose of the comparison study was to show that the planned system works better than other common methods. The only thing that can be used to compare is how well the created method works.

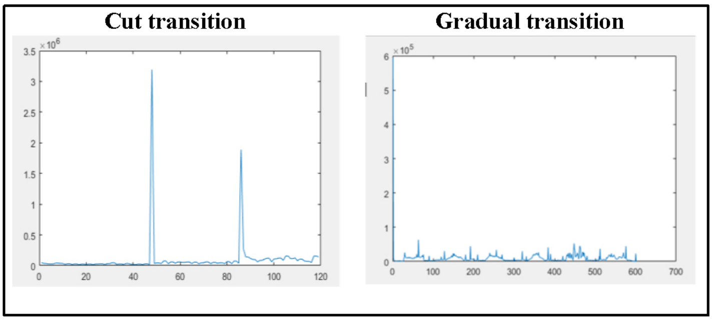
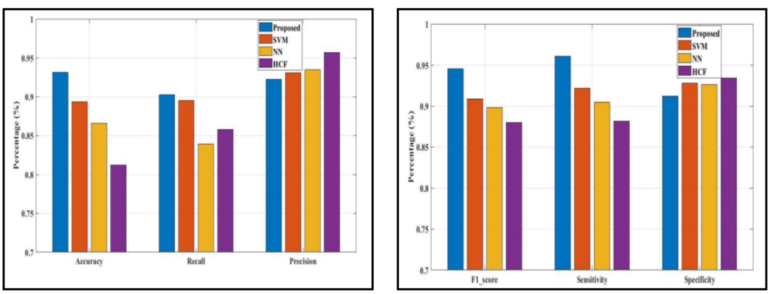


Figure 6

Graphical Representation of Shot Boundary Detection



(a) (b)

Figure 7

Comparison of (a) Accuracy, Recall and Precision (b) F1 Score, Sensitivity and Specificity

Table 3

Comparison Analysis for Proposed and Existing Method

Performance metric	HCF	NN	SVM	CNN Alexnet
Accuracy	0.8123	0.8656	0.8933	0.9312
Recall	0.8578	0.8390	0.8950	0.9023
Precision	0.9567	0.9345	0.9309	0.9223
F1 score	0.8797	0.8978	0.9089	0.9456
Sensitivity	0.8817	0.9047	0.9215	0.9607
Specificity	0.9340	0.9264	0.9279	0.9123

The success measures are used to compare how well different methods work. In order to do a comparison study, the same metrics that were used to judge the suggested method were also used to judge other methods that were already in use.

HCF, NN, and SVM are the standard ways to compare two situations. Table 3 and Figure 7 shows the success metrics that were found for both the traditional method and the suggested method. The above-mentioned Figure 6 shows a graph for shot border identification, similar to the one in Figure 5. There are two main types of changes that can be found in a movie. One is a blunt change, and the other is a smooth shift.

6. Conclusion

Modern video processing needs more complex methods, and this work fills that need for more complex video analysis. Our suggested system uses Convolutional Neural Networks (CNN) to do a great job of finding the edges of shots and classifying movies based on how they look. The important thing about this study is that the CNN-based method is automatic and works well, which helps improve the quality of video material. The first step in video analysis is to count the number of shots in a video. Next, the video's aesthetics can be categorized based on this information. In tests, the system works much better at judging how something looks than other methods, as shown by the large improvement in accuracy. The comparison study with other methods confirms that our suggested system works better. Combining mathematical models and statistical analysis, this study not only creates a new way of thinking about deep learning, but it also provides a strong and dependable way to process videos. Because of this, this work lays the groundwork for future improvements in multimedia apps. It shows how deep learning can be used to accurately find shot boundaries and judge the quality of a video, making the watching experience better for users.

References

[1] Bendraou, Youssef, E. Fedwa, A. Driss, and S. Ahmed, "Shot boundary detection via adaptive low rank and svd-updating, *Computer Vision and Image Understanding*, 161, 20-28 (2017).

[2] L Jingwei, and R. Xie, Shot boundary detection using convolutional neural networks, In 2016 Visual Communications and Image Processing (VCIP) 1-4 (2016).

[3] Sawitchaya, Tippaya, Suchada Sitjongsataporn, Tele Tan, Masood Mehmood Khan, and Kosin Chamnongthai, "Multi-modal visual features-based video shot boundary detection, IEEE Access 5, 12563-12575 (2017).

[4] Bogusław, Cyganek and M.Woźniak, Tensor-based shot boundary detection in video streams, New Generation Computing, 35(4), 311-340 (2017).

[5] Qi, Kuang, X. Jin, Q. Zhao, and B.Zhou, Deep Multimodality Learning for UAV Video Aesthetic Quality Assessment, Transactions on Multimedia, (2019).

[6] Fang-Lue, Zhang, X. Wu, R-L. Li, J. Wang, Z-H. Zheng and S-M. Hu, Detecting and removing visual distractors for video aesthetic enhancement, *IEEE Transactions on Multimedia*, 20(8), 1987-1999 (2018).

[7] Xingmin, Shi, Z. Shan, and N. Zhao, Learning for an aesthetic model for estimating the traffic state in the traffic video, Neurocomputing, 181, 29-37 (2016).

[8] Ajani, S. N., Khobragade, P., Dhone, M., Ganguly, B., Shelke, N., & Parati, N., Advancements in Computing: Emerging Trends in Computational Science with Next-Generation Computing. *International Journal of Intelligent Systems and Applications in Engineering*, 12(7s), 546–559 (2023).

[9] A.Naci, Erdem and U.Halici, Applying Computational Aesthetics to a Video Game Application Using Machine Learning, *IEEE Computer Graphics and Applications*, 36(4), 23-33, (2016).

[10] M. Bende, M. Khandelwal, D. Borgaonkar and P. Khobragade, "VISMA: A Machine Learning Approach to Image Manipulation," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, pp. 1-5 (2023), doi: 10.1109/ISCON57294.2023.10112168.

[11] Yueying, Kao, K. Huang and S. Maybank, Hierarchical aesthetic quality assessment using deep convolutional neural networks, Signal Processing: Image Communication, 47, 500-510 (2016).

[12] K. Agnihotri, P. Chilbule, S. Prashant, P. Jain and P. Khobragade, "Generating Image Description Using Machine Learning Algorithms," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, pp. 1-6 (2023), doi: 10.1109/ICETET-SIP58143.2023.10151472.

[13] Wei-Tse, Sun, T-H. Chao, Y-H. Kuo and W.H. Hsu, Photo filter recommendation by category-aware aesthetic learning, *IEEE Transactions on Multimedia,* vol:19, no. 8, pp: 1870-1880 (2017).

[14] Alejandro, Hernández-García, F. Fernández-Martínez, and F.Díaz-de-María, "Comparing visual descriptors and automatic rating strategies for video aesthetics prediction, *Signal Processing: Image Communication*, 47, pp: 280-288.

[15] Shete, Dhanashri, and Prashant Khobragade. "An empirical analysis of different data visualization techniques from statistical perspective." AIP Conference Proceedings. Vol. 2839. No. 1. AIP Publishing (2023).

[16] Sadiq H., Abdulhussain, Basheera M. Mahmmod, M. I. Saripan, S.A.R. Al-Haddad, and Wissam A. Jassim, Shot boundary detection based on orthogonal polynomial, *Multimedia Tools and Applications,* 78(14), 20361-20382 (2019).

[17] Jaydeb, Mondal, M.K. Kundu, S. Das and M. Chowdhury, Video shot boundary detection using multiscale geometric analysis of nsct and least squares support vector machine, *Multimedia Tools and Applications*, 77(7), 8139-8161 (2018).

[18] Sawitchaya, Tippaya, S. Sitjongsataporn, T.Tan, M.M. Khan and K. Chamnongthai, Multi-modal visual features-based video shot boundary detection, *IEEE Access*, 5, 12563-12575 (2017).

[19] Luming, Zhang, R. Hong, L. Nie and C. Hong, A biologically inspired automatic system for media quality assessment, *IEEE Transactions on Automation Science and Engineering*, 13(2), 894-902 (2015).

[20] Xin, Jin, L.Wu, X. Zhou, G. Zhao, X. Zhang, X. Li and S. Ge, Predicting aesthetic radar map using a hierarchical multi-task network, Pattern Recognition and Computer Vision, 41-50, (2018).

[21] R. Hannane, A. Elboushaki, K. Afdel, P. Naghabhushan and M. Javed, An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram, *International Journal of Multimedia Information Retrieval*, 5(2), 89-104 (2016).

[22] Klaus, Schoeffmann, M.D. Fabro, T. Szkaliczki, L. Böszörmenyi and J. Keckstein, Keyframe extraction in endoscopic video, *Multimedia Tools and Applications*, 74(24), 11187-11206 (2015).

[23] Zhang, Yong, Y. Jiazheng, L. Hongzhe and L. Qing, GrabCut image segmentation algorithm based on structure tensor, *The Journal of China Universities of Posts and Telecommunications*, 24(2), 38-47 (2017).

[24] Jingwei, Xu, L. Song and R.Xie, Shot boundary detection using convolutional neural networks, In Visual Communications and Image Processing (VCIP), 1-4 (2016).