Scene Boundary Detection Approaches: A Review

Mayur Akewar [1]

[1]Shri Ramdeobaba College of Engineering and Management

March 28, 2024

Abstract

Scene boundary detection is a crucial task in video analysis, facilitating tasks such as video summarization, indexing, and content-based retrieval. In this paper, we present a comprehensive review of various approaches and methodologies employed for scene boundary detection. We analyze the strengths and limitations of different techniques and highlight emerging trends and challenges in the field.

Scene Boundary Detection Approaches: A Review

Mayur Akewar
Alumni, Department of Computer Science & Engineering
Shri Ramdeobaba College of Engineering & Management
Email: mayurakewar87@gmail.com

***Abstract*—Scene boundary detection is a crucial task in video analysis, facilitating tasks such as video summarization, indexing, and content-based retrieval. In this paper, we present a comprehensive review of various approaches and methodologies employed for scene boundary detection. We analyze the strengths and limitations of different techniques and highlight emerging trends and challenges in the field.**

***Index Terms*—Video**

I. Introduction

Scene boundary detection plays a vital role in video processing and analysis, enabling the segmentation of videos into meaningful units for further analysis and retrieval. Over the years, numerous approaches and algorithms have been proposed to address this task, ranging from traditional methods based on visual cues to more recent techniques leveraging deep learning and machine learning algorithms. In this paper, we provide an overview of scene boundary detection approaches, examining their evolution, underlying principles, and applications. By synthesizing existing literature and analyzing recent advancements, we aim to provide insights into the state-of-the-art techniques and future directions in scene boundary detection research.

II. Literature Review

Various methodologies have been put forth to address scene boundary detection, employing advanced techniques in deep learning and machine learning that primarily leverage image and text features. These approaches involve intricate algorithms and models designed to automatically identify transitions between scenes within video content. By harnessing the capabilities of deep learning and machine learning, these methods aim to enhance the accuracy and efficiency of scene boundary detection processes, contributing to the broader field of video processing and analysis. In [1], the authors address the underexplored issue of semantic video indexing in their paper, emphasizing its potential to enhance video search, monitoring, and surveillance experiences. They propose a comprehensive pipeline for video structure mining using deep features, consisting of feature extraction and filtering, shot clustering, and labeling stages. Deep convolutional networks are employed for feature extraction. The evaluation of the pipeline demonstrates high-quality scene detection and annotation, assessed through various metrics. The paper also includes an analysis of current segmentation and annotation metrics. The presented work has practical implications for real-time semantic video annotation. In [2], the authors explore the realm of video scene analysis, acknowledging its increasing significance in various applications like real-time vehicle tracking, pedestrian detection, surveillance, and robotics. They highlight the persisting challenges in achieving accuracy and the need for more precise algorithms. The paper delves into recent advancements in deep learning algorithms tailored for real-time video scene analysis. The authors conduct a comprehensive review of recent developments, emphasizing datasets and their limitations. Special attention is given to challenges in real-time video scene analysis, encompassing activity recognition, scene interpretation, and video description. In [3], the authors address the growing challenge of managing multimedia data on the internet, particularly focusing on video organization, summarization, and retrieval where scene detection is crucial. Existing shot clustering algorithms for scene detection often treat temporal shot sequences as unconstrained data. While graph-based methods consider temporal relations, they typically rely on low-level features for determining shot similarities. The authors propose a novel temporal clustering method leveraging graph convolution networks and shot node link transitivity. In [4], the authors propose a novel framework for shot boundary detection in videos, leveraging dynamic mode decomposition (DMD). This method addresses the challenges posed by weak boundaries and sudden changes in brightness or foreground objects. By extracting temporal foreground and background modes from video data using DMD, shot boundaries can be detected based on sharp changes in amplitude. The algorithm reduces error detection in scenarios with rapid illumination changes or quick foreground object or camera movements. It achieves high detection accuracy even in cases where color changes are subtle, illumination changes slowly, or foreground objects overlap. The effectiveness of the proposed method is demonstrated through shot boundary detection across various video content types. In [5], the authors present a novel framework for saliency detection in videos, emphasizing the limitations of existing research primarily focused on still images. Their approach involves a deep learning-based hybrid spatiotemporal saliency feature extraction framework, integrating high-level features obtained through deep learning with other features. This deep learning model proves more effective than traditional handcrafted methods for extracting hidden features. The proposed model considers both spatial and temporal characteristics by taking several consecutive frames as input for computing saliency maps. Evaluation on five databases with complex scenes involving human gaze

demonstrates the superior performance of the proposed hybrid feature framework compared to five other state-of-the-art video saliency detection approaches.

In [6], the authors introduce ShotCoL, a self-supervised shot contrastive learning approach designed for scene boundary detection in movies and TV episodes. ShotCoL learns a shot representation that maximizes the similarity between nearby shots compared to randomly selected shots, achieving state-of-the-art performance on the MovieNet dataset with only 25% of the training labels, 9x fewer model parameters, and 7x faster runtime. Additionally, ShotCoL proves effective in novel applications, specifically finding timestamps for minimally disruptive ad cue-points in a new dataset called AdCuepoints, comprising 3,975 movies and TV episodes. In [7], the authors propose a straightforward yet effective method for video shot boundary detection (SBD), focusing on expediting the process and simplifying it without compromising recall and accuracy. The model incorporates a top-down zoom rule, image color features, and local descriptors, along with a motion area extraction algorithm. Candidate transition segments are selected using color histogram and speeded-up robust features, followed by cut transition detection and gradual transition detection. Evaluation on the TRECVid2001 and TRECVid2007 video datasets demonstrates that the proposed method enhances recall, accuracy, and detection speed compared to other SBD methods. In [8], the authors address the growing number of videos online, emphasizing the inefficiency of content-based video browsing and retrieval due to the storage method in databases. They highlight the importance of automated video structure analysis, focusing on Shot Boundary Detection (SBD) as a crucial process for video indexing and retrieval. The paper reviews a wide range of SBD approaches, exploring their advantages and disadvantages, discussing developed algorithms, and presenting challenges and recommendations for future research. In [9], the authors introduce a novel key frame selection method for object detection in intelligent systems, such as pedestrian tracking and video surveillance. Traditional methods using SIFT features suffer from high key frame selection error rates. In this paper, the proposed approach utilizes object detection and image quality for key frame selection. Object detection, like identifying pedestrians or vehicles, is employed to assign a quality score to each training frame. Frames containing objects receive high-quality scores. The algorithm uses a CNN based on the AlexNet architecture for deep feature representation extraction. Key frames are extracted through a combination of mutual information entropy and SURF image local features. Through comprehensive experiments, the feasibility of the key frame extractor based on a convolutional neural network is verified, involving model training and a study on quality assessment models. In [10], the authors address the need for effective and efficient video summarization in the face of the explosive growth of video data. They propose a novel approach that considers the block-sparsity of candidate keyframes, formulating the video summarization problem as a block sparse dictionary selection model. The paper introduces a simultaneous block version of Orthogonal Matching Pursuit (SBOMP) algorithm for model optimization and explores two keyframe selection strategies for each block. Experimental results on benchmark datasets, VSumm and TVSum, demonstrate that the SBOMP-based video summarization method outperforms several state-of-the-art sparse representation-based methods in terms of F-score, redundancy among keyframes, and robustness to outlier frames. In [11], the authors propose a Detect-to-Summarize network (DSNet) framework for supervised video summarization, comprising both anchor-based and anchor-free counterparts. The anchor-based method generates temporal interest proposals to identify and locate representative content in video sequences, while the anchor-free method predicts importance scores and segment locations directly without pre-defined temporal proposals. Unlike existing methods, the DSNet leverages temporal consistency through the temporal interest detection formulation. The anchor-based approach involves dense sampling of temporal interest proposals with multi-scale intervals and extraction of long-range temporal features. The anchor-free approach predicts importance scores and segment locations directly. In [12], the authors present a solution to the Natural Language Video Localization (NLVL) problem, aiming to localize the video segment corresponding to a natural language description in a long and untrimmed video. Current NLVL methods, categorized into anchor-based and anchor-free approaches, have inherent drawbacks, such as susceptibility to heuristic rules or failure to exploit segment-level interaction. In response, the authors propose a novel Boundary Proposal Network (BPNet), a two-stage framework. In the first stage, an anchor-free model generates high-quality candidate video segments with boundaries. The second stage incorporates a visual-language fusion layer to model multi-modal interaction and a matching score rating layer for alignment. BPNet is evaluated on three NLVL benchmarks, demonstrating superior performance compared to state-of-the-art methods.

III. Conclusion

In our paper, we explored various methods for detecting scene boundaries in videos. We addressed common issues encountered in methods and reviewed the methods.

References

[1] Stanislav Protasov, Adil Mehmood Khan, Konstantin Sozykin, Muhammad Ahmad *Using deep features for video scene detection and annotation*. Signal, Image and Video Processing 2018, [Online]. Available: https://link.springer.com/article/10.1007/s11760-018-1244-6

[2] Qaisar Abbas, Mostafa E. A. Ibrahim, M. Arfan Jaffar *Video scene analysis: an overview and challenges on deep learning algorithms* . Multimedia Tools and Applications 2018, [Online]. Available: https://link.springer.com/article/10.1007/s11042-017-5438-7

[3] Yingjiao Pei, Zhongyuan Wang, Heling Chen, Baojin Huang, Weiping Tu *Video scene detection based on link prediction using graph convolution network*. Proceedings of the 2nd ACM International Conference on Multimedia in Asia 2021, [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3444685.3446293

[4] Chongke Bi, Ye Yuan, JiaWan Zhang, Yun Shi, Yiqing Xiang, Yuehuan Wang, RongHui Zhang *Dynamic Mode Decomposition Based Video Shot Detection* . IEEE Access 2018, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8334241

[5] Zheng Wang, Jinchang Ren, Dong Zhang, Meijun Sun, Jianmin Jiang *A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos* . Neurocomputing 2018, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0925231218301097

[6] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, Raffay Hamid *Shot Contrastive Self-Supervised Learning for Scene Boundary Detection* . Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021.

[7] Shangbo Zhou, Xia Wu, Ying Qi, Shuyue Luo, Xianzhong Xie *Video shot boundary detection based on multi-level features collaboration* . Signal, Image and Video Processing volume 2020, [Online]. Available: https://link.springer.com/article/10.1007/s11760-020-01785-2

[8] Sadiq H. Abdulhussain, Abd Rahman Ramli, M. Iqbal Saripan, Basheera M. Mahmmod, Syed Abdul Rahman Al-Haddad, Wissam A. Jassim *Methods and Challenges in Shot Boundary Detection: A Review* . Entropy 2018, [Online]. Available: https://www.mdpi.com/1099-4300/20/4/214

[9] Mingju Chen, Xiaofeng Han, Hua Zhang, Guojun Lin, M.M. Kamruzzaman *Quality-guided key frames selection from video stream based on object detection* . Journal of Visual Communication and Image Representati 2019, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1047320319302998

[10] Mingyang Ma, Shaohui Mei, Shuai Wan, Junhui Hou, Zhiyong Wang, David Dagan Feng *Video summarization via block sparse dictionary selection* . Neurocomputing 2020, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0925231219314699

[11] Wencheng Zhu, Jiwen Lu, Jiahao Li, Jie Zhou *DSNet: A Flexible Detect-to-Summarize Network for Video Summarization* . IEEE Transactions on Image Processing 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9275314

[12] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, Jun Xiao *Boundary Proposal Network for Two-stage Natural Language Video Localization* . Proceedings of the AAAI Conference on Artificial Intelligence 2021, [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16406

[13] Partha Pratim Ray *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope* . Internet of Things and Cyber-Physical Systems 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266734522300024X

[14] M. Plakal and D. Ellis *Yamnet* . nan 2020, [Online]. Available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

[15] *Yamnet* , [Online]. Available: https://www.tensorflow.org/hub/tutorials/yamnet