A Comparison of Scene Change Localization Methods over the Open Video Scene Detection Dataset

Taras Panchenko[1†] and Igor Bieda[1†],

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Abstract
Scene change detection is an important topic because of the wide and growing range of its applications. Streaming services from many providers are increasing their capacity which causes the industry growth. The method for the scene change detection is described here and compared with the State-of-the-Art methods over the Open Video Scene Detection (OVSD) – an open dataset of Creative Commons licensed videos freely available for download and use to evaluate video scene detection algorithms. The proposed method is based on scene analysis using threshold values and smooth scene changes. A comparison of the presented method was conducted in this research. The obtained results demonstrated the high efficiency of the scene cut localization method proposed by authors, because its efficiency measured in terms of precision, recall, accuracy, and F-metrics score exceeds the best previously known results.

Keywords:
scene change detection; OVSD dataset; scene cut; scene break; scene localization.

1. Introduction

In this work, we continue the research of the problem of scene changes localization [1] in video using the method proposed by authors in [2] being applied to new data. The method for scene changes detection [3] gives the possibility to find the limits and duration of a scene or track moving objects [1]. But the primary task is to determine scene change fact (time tick) in the video stream. The method from [2,3] described in Section 4 here is a new combination of techniques to achieve good performance for this task. By performance here we mean precision, recall, accuracy, and *F*-metric.

Scene cut detection is useful in many cases:

- to divide the video stream into more or less independent and useful parts (for different purposes: advertising, effective pausing),
- to analyze the presence of characters of interest in a certain area of observation and their behavior (the activities of actors and analysis of the frequency of their appearance),
- for security purposes – to find episodes in recordings with some specific activities; for games such as football – to find points of interest in time, tracking players and their actions,
- to track human activity or the movement of objects in a general context and find important points of high interest in video streams or recordings,
- for faster navigation through the video,
- to more convenient navigation through movies, TV shows, or gameplay recordings thereafter,
- to identify the moments of the greatest interest; it will improve the browsing experience for end-users,
- and many others.

To clarify what we mean by scene we would like to introduce the internal hierarchical structure of the video. It is as follows:

- lower level – consists of frames, which are a series of static images,
- middle level – consists of a sequence of frames taken by the same camera at a certain point in time, also called a snapshot.
- accordingly, scenes are at the top hierarchical level – this means they are located above the pictures, frames, and snapshots (Fig. 1).

The main objective of the research is to develop a precise method for scene edge localization on the video stream.

There exist methods based on computer vision (CV) and machine learning (ML) techniques [1,2] for annotation, simplified navigation, in case the user only needs a certain part of the video to further process the splitting of scenes, but they do not show good results. The computational complexity of such methods is one of the issues for real-life applications.

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.1

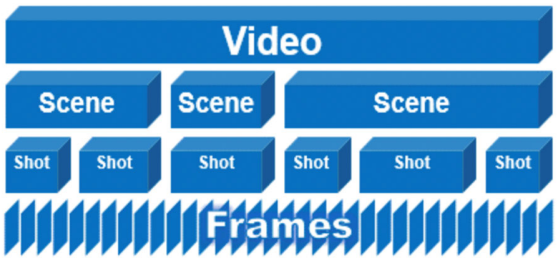

Fig. 1. The hierarchical structure of the video

Open Video Scene Detection [4,5] – is an open dataset of movies and films created from Creative Commons licensed videos freely available for download and use to evaluate video scene detection algorithms. These videos contain open-source data, making them ideal for use by researchers in both academy and industry.

As the input for this research, we use an open dataset from [4,5], which is used for both public and scientific purposes. As the proposed method for scene change localization for movies, films and videos has shown good results [2], in this paper we decided to compare our method with the existing method described in [4,5] by Daniel Rothman, Dror Porat, Gal Ashur from IBM Research. It introduced the dataset which consists of 21 open-source videos: Elephants Dream, Big Buck Bunny, Sintel, Tears of Steel, Cosmos Laundromat – First Cycle, Valkaama, 1000 Days, Boy Who Never Sleep, CH7, Fires under Water, Honey, Jathia's Wager, La Chute d'une Plume, Lord Meia, Meridian, Oceania, Pentagon, Route 66, Seven Dead Men, Sita Sings the Blues and Star Wreck.

There are plenty of methods for this task developed. We will mention some of them in our work. In this paper, we compare the best-known methods and results for this task solution [4,5] with the method proposed by the authors and depicted in Section 4.

The best-known result for the scene cut detection is presented in [4] and [5]. Here we concentrate on these results and their comparison with one proposed by the authors. The method, proposed by authors in [2,3] was applied to the dataset of movies with different characteristics, like lighting and dynamics.

The aim and the main contribution of this research are the following:

- to confirm or improve the known results, and to compare the proposed method's quality with other previously known;
- to re-check the effectiveness of the method on different datasets.

2. Open Video Scene Detection Dataset

It may seem not so obvious at first glance, but the problem of scene change detection seriously lacks large datasets [1]. The main reason for this is that long, heterogeneous videos that consist of several scenes are in most cases copyrighted. These are films, talk shows, and programs that are assets to their creators or copyright holders. Copyright holders are unlikely to ever make them publicly available.

We have deliberately not used the Blip10000 dataset [6], which contains videos from blip.tv. That content is created by users who have gone beyond simple video capture techniques on platforms like YouTube and Flickr, so it is unusual for a regular movie and is out of our scope. Such content is commonly referred to as User Generated Semi-Professional Content (SPUG). Blip.tv users post video content in series on one specific topic, post at regular intervals, and target a wide audience. All this makes real research on such videos impossible since they often do not contain the expected markers of scenes changes.

To demonstrate the effectiveness of our method, we decided to use the Open Video Scene Detection Dataset (OVSD) [4,5]. This dataset is of interest for scientific purposes, as mentioned above. This dataset was gathered from Creative Commons licensed videos that are freely available to download, use and reuse. Dataset consists of short or full-length films of various genres, including animation, documentary, drama, crime, comedy, and sci-fi movies. They all contain enough number of scenes. To the best of our knowledge, this dataset is the only video scene detection dataset that contains entire films and is available free of charge with minimal legal restrictions.

In our previous works [2,3] we used the dataset of 11 videos, and it showed good results with 94% of scene cut detected.

So, we will use 21 videos of the OVSD dataset with a total duration of more than 18 hours. These videos are longer and cover a much wider range of genres. Taken together, this provides us with high quality and broad assessment tool for detecting alteration of video scenes (see Table 1).

Table 1. Details of the OVSD Dataset

Video Name	Duration, min	Number of Scenes	Genre
1000 Days	43	23	Drama
Boy Who Never Slept	69	37	Comedy, Romance
CH7	86	45	Crime
Fires Beneath Water	76	63	Documentary
Honey	86	21	Drama
Jathia's Wager	21	16	Drama, Sci-Fi
La Chute D'une Plume	10	11	Animation
Lord Meia	37	28	Crime, Comedy
Meridian	12	10	Mystery, Sci-Fi
Oceania	54	32	Drama, Mystery
Pentagon	50	32	Comedy, Drama
Route 66	103	56	Documentary
Seven Dead Men	57	35	Crime
Sita Sings the Blues	81	53	Animation, Comedy
Star Wreck	103	56	Comedy, Sci-Fi

3. Overview of the existing methods

In this section, we analyze the existing solutions for frame edge detection and video scene detection. We provide a short description and links for more detailed information.

The State-of-the-Art in dividing the video into its components – frames, was considered basically a solved problem [7].

This is due to the relative homogeneity of information in one frame, which makes it possible to accurately determine. Measuring the difference in pixels and histograms between frames as one increases the intervals allows you to detect sharp and gradual transitions between frames. Some of the modern methods for detecting the border of a shot are described in [8,9] and can be used out of the box with impressive results.

There are also methods that are aimed at creating a complete video splitting, and not searching for a single scene. For a more general overview, see [2,3,10]. So, for example, Baraldi [11] extracts a normalized histogram for each image and uses a distance metric consisting of Bhattacharya distance and time distance. They use a similarity matrix to perform spectral clustering with automatic adjustment of the number of clusters.

Also, Baraldi emphasizes the inconsistency of clustering in time, and the scene boundaries are denoted between each pair of adjacent images belonging to different clusters.

Sakarya and Telatar [12] plot the video frames by weighing the edges using the temporal and spatial similarity function. The dominant set of images is detected and marked as a scene. Enforcing the temporal consistency constraint allows scene edge selection using the mean and standard deviation of the positions of the shots. The process continues until the entire video is distributed across scenes.

As to the method suggested by Daniel Rotman, Dror Porat, Gal Ashour [4,5]. The sense of this is to group consecutive frames into scenes, and the resulting optimization problem can be efficiently solved using a new dynamic programming scheme. It directly performs time-consistent video scene detection and has the advantage of being parameterless, making it robust and applicable to multiple genres of video.

There are also exist plenty of methods for this task [13-17] ranging from scene classification and features extraction (for example, text from natural scene) to behavior extraction and scene understanding based on different approaches from classical CV to LSTM and other NN-models.

4. The proposed method application and the discussion

First, we determined possible indicators of a scene change. We do not clearly define such events, but in the proposed approach we consider the following:

- fade-in / fade-out (gradual transitions): the video gradually disappears, the next part of which appears, but with (most probably) a changed background (scene);
- instant or prompt (sharp) scene change: the background in the video changes quickly between two consecutive frames;
- analysis of the shift of color components and changes in the values of objects (contrast, intensity, saturation, brightness) – in the context of different color spaces.

As for the first attempt, we analyze the following indicators:

- intensity (average value of channels R, G, B) and contrast (the difference between the maximum and minimum saturation of pixels in the area divided by the sum of the minimum and maximum saturation);
- analysis of the contours of image objects helps to determine their structure, movement, and other changes;
- frame difference analysis helps to track how much the content of the frame has changed (not just the contours);
- facial recognition to track movement – identification of characters.

For Open Video Scene Detection Dataset, when finding differences between two images, we suggest using the Canny operator [18] to search for image contours. The Canny operator makes it possible to obtain a matrix with boundaries. However, the matrix is a large set of values that are slowly processed.

Therefore, we propose to expand the boundaries on these matrices by intersecting the contours of one frame with the contours of another frame, and from the resulting overlays to find the contours of differences as arrays of coordinates, and then to find out how much they have changed between two consecutive frames, describing it with one value.

Thus, we will calculate the difference not between the locations, but between the chains themselves from the points that form the contours. This path is much more representative, as it allows us to find how much the content of the frame has changed, not just the contours:

- convert two consecutive frames to black and white, convert the view to HSV color scheme and remove saturation (S);
- reduce the noise exposure when finding boundaries by Gaussian blur;
- application of the Canny operator and obtaining a boundary matrix for two frames;
- expansion of the received borders;
- superimposing extended boundaries of one frame on another and vice versa, which allows us to obtain two values, that represent a measure of the boundaries difference;
- search for contours as arrays of coordinates of points that form boundaries on the mismatch matrix,
- find out the difference in contours by taking a maximum of two.

To do this, we introduce a numerical measure from 0 to 1, which describes the difference between the contours of objects between two frames. This indicator is a reliable signal of a scene change because it characterizes how much the scene could change in terms of the contour of the object or change the boundaries.

More details about the method that is used can be found in [2] and [3]. Intensity and saturation analysis gives us key points where scenes potentially change. Then, we localize these peaks. There must be a certain threshold to detect them.

This threshold is dynamic and adaptive, as it depends on the quality of the video and can detect unimportant peaks or miss a scene change in the video if its value is less than required. Therefore, we use a recursive adaptive threshold adjustment, which allows the video to be divided into pieces according to the first basic threshold based on the change in intensity (Fig. 2).

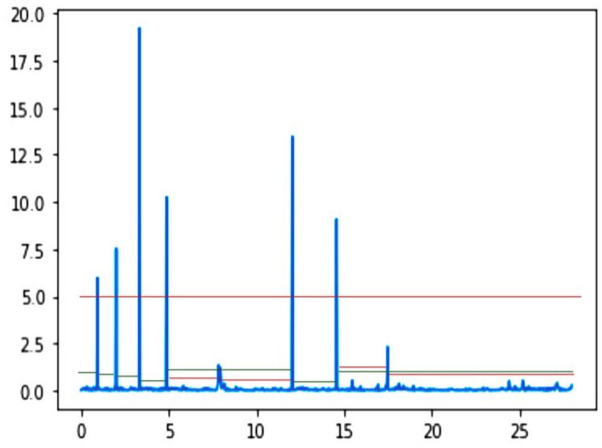

Fig. 2. Adaptive saturation thresholding for scene change detection

After that, the process of dividing the video continues recursively, as long as it remains possible. Thus, we calculate a new threshold level and find the peaks that exceed this level. Details on this can be found in [2].

The comparison of the best-known results is presented in Table 2. Here are the results of the original method proposed by the authors of [4] dataset, [19] for comparison, and the method presented by the authors [2]. The metrics widely accepted are Coverage

(*C*) and Overflow (*O*) metrics [20], with a single *F* score as an overall quality measure for scene detection computed as the harmonic mean of *C* and 1 – *O*:

$$F = \frac{1}{\frac{1}{C} + \frac{1}{1-O}}$$

Table 2. Results on the OVSD Dataset. The best *F* score in each row is highlighted in bold

Video (short name)	[2]	[4]	[19]
1000	**0,52**	0,38	0,5
BBB	0,78	**0,83**	0,49
BWNS	**0,67**	0,63	0,61
CH7	**0,66**	0,63	0,52
CL	**0,70**	0,53	0,45
ED	0,59	**0,6**	0,56
FBW	0,57	0,57	**0,61**
Honey	**0,58**	**0,58**	0,38
JW	0,71	**0,75**	0,28
LCDP	**0,61**	0,53	0,18
LM	0,70	0,69	**0,71**
Meridian	**0,66**	0,63	0,63
Oceania	0,65	**0,67**	0,51
Pentagon	0,71	**0,73**	0,48
Route 66	**0,61**	0,54	0,36
SDM	0,80	0,68	**0,81**
Sintel	0,58	0,46	**0,59**
SStB	**0,46**	**0,46**	0,43
SW	0,56	0,55	0,4
ToS	0,6	0,5	**0,75**
Valkaama	0,69	0,63	**0,73**
Avg. F	**0,64**	0,6	0,52

To compare the method with others, we could say that the proposed one is a combination of known techniques, while others try to introduce the new feature engineering. Our method [2,3] is based on frame light intensity analysis while others (including [4,19]) are more oriented on complex analysis.

The main idea and intuition behind our method is a dynamic threshold, which shows us the signals of prompt scene cut, and the additional analysis of the smooth fade-in / fade-out transitions through the contours changing analysis using Canny operator, blurring, and the difference analysis, as described above.

One of the main conclusions is that our proposed technique shows better results by the score, but still has room for improvements because the audio analysis could be added for a more comprehensive picture.

The quality of our approach [2] was confirmed by this research conducted over the OVSD dataset [4], which showed better results of the proposed method [3] in comparison with existing [5].

5. Summary and the conclusions

The task of scene change detection is actual in many areas and important for many reasons. The proper attribution of separate scenes could help to navigate through video effectively, localize important pieces, and analyze different aspects like specific character presence or its behavior.

In this research, we apply the approach proposed by authors [2] to detect scene changes in video over the OVSD dataset [4], which is:

- open for general access (the use of videos which does not violate copyright restriction and are freely available);
- the dataset contains a vast amount of video materials of different types and genres: from amateur video filming to cartoons.

This dataset allows a more objective understanding of the effectiveness of the proposed in [2] method over different data: video records and movies. The relevance, as well as use cases and applications, were also described in the introduction.

The method described in Section 4 is a new combination of different previously known techniques, which proved its effectiveness due to this research results.

The proposed method turned out to be more accurate than [4] and [19], the best-known results. Results in Table 2 demonstrate that the proposed approach provides minor gains compared to other known methods. The proposed method shows high results on the selected dataset, which demonstrates its high efficiency.

The results obtained for this dataset once again ensure us in the quality of the developed method. As one can see from the table, the scene change detection did not always correctly mark the start and stop points of the scene, therefore, in the future, we plan to add audio analysis to detect scene changes. Also, we hope

that threshold levels could be adapted in a better way, to become more selective and to get higher results.

References

[1] M. Del Fabro, L. Boszormenyi. State-of-the-Art and Future Challenges in Video Scene Detection: a Survey[J]. Multimedia systems, 2013, 19(5): 427–454.

[2] Igor Bieda, Anton Kisil, Taras Panchenko. An Approach to Scene Change Detection[C]. The 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2021), 2021: 489-493.

[3] Igor Bieda. Scene Change Localization in Video[J]. Taras Shevchenko National University of Kyiv Visnyk, Physical and Mathematical Sciences Series, 2021, 1: 57–62.

[4] D. Rotman, D. Porat, G. Ashour. Robust Video Scene Detection Using Multimodal Fusion of Optimally Grouped Features[C]. IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), 2017: 44–47.

[5] D. Rotman, D. Porat, G. Ashour. Robust and Efficient Video Scene Detection Using Optimal Sequential Grouping[C]. IEEE International Symposium on Multimedia (ISM), 2016: 275-280.

[6] S. Schmiedeke, P. Xu, I. Ferrane, M. Eskevich, C. Kofler, M. A. Larson, Y. Este`ve, L. Lamel, G. J. Jones, T. Sikora. Blip10000: A social video dataset containing spug content for tagging and retrieval[C]. Proceedings of the 4th ACM Multimedia Systems Conference, 2013: 96–101.

[7] A. F. Smeaton, P. Over, A. R. Doherty. Video shot boundary detection: Seven years of trecvid activity[J]. Computer Vision and Image Understanding, 2010, 114(4): 411–418.

[8] QL. Baraldi, C. Grana, R. Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video[C]. International Conference on Computer Analysis of Images and Patterns, 2015: 801–811.

[9] E. Apostolidis, V. Mezaris. Fast shot segmentation combining global and local visual descriptors[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014: 6583–6587.

[10] I. Bieda, A. Kysil, V. Shevchenko. An approach for the scene change localization[C]. Proc. Problems of Decision Making under Uncertainties (PDMU-2021), 2021: 22.

[11] L. Baraldi, C. Grana, R. Cucchiara. Analysis and re-use of videos in educational digital libraries with automatic scene detection[C]. 11th Italian Research Conference on Digital Libraries, 2015: 155–164.

[12] U. Sakarya, Z. Telatar. Video scene detection using dominant sets[C]. 15th IEEE International Conference on Image Processing, 2008: 73–76.

[13] Vaibhav Goel, Vaibhav Kumar, Amandeep Singh Jaggi, Preeti Nagrath. Text Extraction from Natural Scene Images using OpenCV and CNN[J]. International Journal of Information Technology and Computer Science(IJITCS), 2019, 11 (9): 48-54.

[14] Ameni Sassi, Wael Ouarda, Chokri Ben Amar, Serge Miguet. Sky-CNN: A CNN-based Learning Approach for Skyline Scene Understanding[J]. International Journal of Intelligent Systems and Applications(IJISA), 2019, 11 (4): 14-25.

[15] Anupam Dey, Fahad Mohammad, Saleque Ahmed, Raiyan Sharif, A.F.M. Saifuddin Saif. Anomaly Detection in Crowded Scene by Pedestrians Behaviour Extraction using Long Short Term Method: A Comprehensive Study[J]. International Journal of Education and Management Engineering(IJEME), 2019, 9 (1): 51-63.

[16] Md. Arafat Hussain, Emon Kumar Dey. Remote Sensing Image Scene Classification[J]. International Journal of Engineering and Manufacturing(IJEM), 2018, 8 (4): 13-20.

[17] Lolith Gopan, E.Venkateswarlu, Thara Nair, G.P.Swamy, B.Gopala Krishna. Scene based Non-uniformity Correction for Optical Remote Sensing Imagery[J]. International Journal of Image, Graphics and Signal Processing(IJIGSP), 2017, 9 (12): 50-57.

[18] J. Canny. A Computational Approach To Edge Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8 (6): 679–698.

[19] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, I. Trancoso. Temporal video segmentation to scenes using high-level audiovisual features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21 (8): 1163–1177.

[20] J. Vendrig, M. Worring. Systematic evaluation of logical story unit segmentation[J]. IEEE Transactions on Multimedia, 2002, 4 (4): 492–499.

Taras Panchenko received the B.S. and M.S. degrees, from Taras Shevchenko National University of Kyiv (Ukraine). He received the PhD from the same University. He has been working as an Associate Professor in the Theory and Technology of Programming Department of the Faculty of Computer Science and Cybernetics of Taras Shevchenko National University of Kyiv. His research interest includes data science and data processing technologies, data engineering, and other aspects of computer science. He is a member of the Association for Computing Machinery.

Igor Bieda received the B.S. and M.S. degrees, from National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" (Ukraine). He is obtaining his PhD from Taras Shevchenko National University of Kyiv (Ukraine).