Scene Segmentation in Artistic Archive Documentaries

Dalibor Mitrović, Stefan Hartlieb, Matthias Zeppelzauer, Maia Zaharieva

Interactive Media Systems Group
Vienna University of Technology
Favoritenstrasse 9-11, Vienna, Austria
{mitrovic, hartlieb, zeppelzauer, zaharieva}@ims.tuwien.ac.at

Abstract. Scene segmentation is a crucial task in the structural analysis of film. State-of-the-art scene segmentation algorithms usually target fiction films (e.g. Hollywood films). Documentaries (especially artistic archive documentaries) follow different montage rules than fiction films and consequently require specialized approaches for scene segmentation. We propose a scene segmentation algorithm targeted at artistic archive documentaries. We evaluate the performance of our technique with archive documentaries and contemporary movies and obtain satisfactory results in both domains.

Keywords: Archive film material, documentaries, scene boundary detection

1 Introduction

Films have a hierarchical structure that is the result of the editing process. On the lowest structural level there is the single *frame*. A number of continuously recorded frames form a *shot* and a sequence of shots belonging together a *scene*. The set of all scenes is the *film*. In modern fiction films, such as Hollywood films, scenes usually depict activities with the same dramatic incident or location [1]. This definition is not applicable to documentaries and especially not for artistic archive documentaries. In artistic archive documentaries, shots constituting a scene are related on a higher abstraction level. For example, in a fiction film, a scene may show two people driving in the car and talking to each other. All the shots depicting this conversation form the scene. In an artistic archive documentary, a scene consists of shots that e.g. show how electricity is brought to a village. These shots show someone installing a power line, peasants using an electrical thresher and several houses of the village with electrical lighting. All these shots are recorded at different locations and at differing times. Their cohesion is much lower than for the shots in the fiction film example. Due to the low cohesion of shots there is only little a priori knowledge (e.g. about composition rules) that can be incorporated into the segmentation process. The most important clue for scene segmentation in documentaries is the repeated appearance of visually similar shots and motives. We aim at exploiting this clue by applying

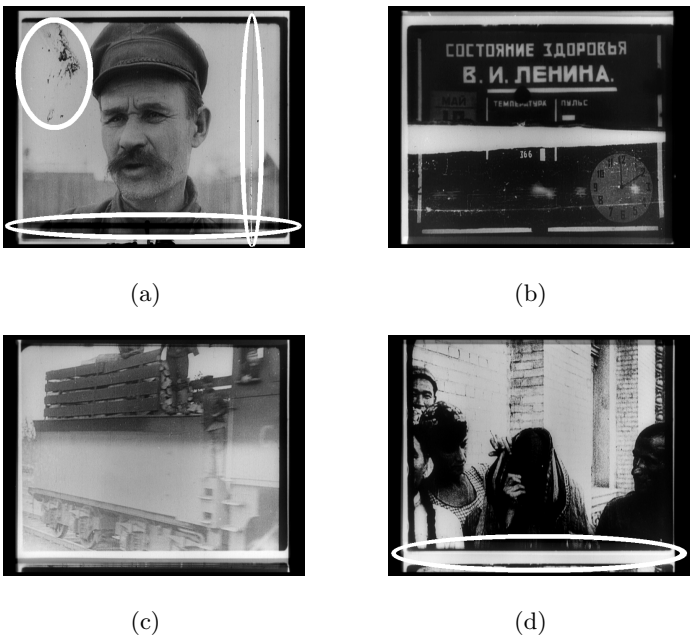
several *orthogonal* features for the detection of visual similarities in parallel and by merging their individual results. By this approach we attempt to compensate for the lack of a priori knowledge.

Existing techniques for scene segmentation are usually tailored to modern films. The techniques can be divided into three categories: graph-based approaches [5, 9], model-based approaches [2, 11], and merge-(and-split)-based approaches [4, 8]. Yeung et al. propose a graph representation for video content where nodes represent the shots and edges represent the temporal flow of the story [9]. The graph is split into sub-graphs representing scenes based on temporal (time windows) and visual constraints (color and luminance similarity). In general, graph-based approaches are not applicable to scene detection in artistic archive documentaries since these films are mostly shot in an open and highly dynamic environment. Zhai et al. present a statistical method for the detection of video scenes. The method is based on the Markov Chain Monte Carlo technique and relies on model priors, visual constraints (color similarity), and temporal consistency [11]. Such model-based approaches work well for scenes where scene content can be represented by a parametric probabilistic model. The experimental nature of artistic archive documentaries does not allow for the definition of such generic models. Rasheed et al. propose a two-pass approach for scene boundary detection in Hollywood films and TV shows [4]. In the first stage, the authors perform shot clustering based on color similarity. In the next stage, over-segmented scenes are merged again based on shot lengths and motion content analysis in the scenes. Recently, Wang et al. proposed a method based on overlapping links for the detection of video scenes [8]. The authors perform an iterative backward and forward search for similar shots in a video. Shot similarity is defined as the combination of visual (color) similarity and consistent motion characteristics. Since artistic archive documentaries frequently have no color information and exhibit a large amount of shaking and flicker that reduces the reliability of motion features, scene segmentation in this context requires more robust content-based features. In this paper, we adapt the overlapping links method (which is independent of content-based features) for the segmentation of scenes in artistic archive documentaries.

2 Material

The material employed in this investigation are artistic documentaries produced in the Soviet Union in the 1920s and 1930s. In contrast to news- and sports broadcasts and fiction films, these documentaries do not contain any narrative structure. In fact the director Dziga Vertov strongly opposed any narration in the Hollywood sense. Consequently, the films share a collage-like style. Vertov relied on the viewer's mind to connect the depicted activities and locations. Scenes in the context of Vertov's films represent topics on an abstract level. For example one scene shows people and machines building a hydro-electric dam. The next scene shows the flooding of different villages to symbolize that villages are flooded with energy. Eventually, it is shown how the villages are electrified.

The original material is 35mm black-and-white silent film. We employ frame-by-frame digitized backup copies of the films, because the original films do not exist anymore. The films are digitized at PAL resolution. Unfortunately, the state of the material has degraded significantly, during storage, copying, and playback over the last decades. Numerous artifacts were introduced into the material (see Figure 1). The most common artifacts we have to deal with include scratches, introduced by dirt in the projectors during playback, dust, liquids, etc. copied into the images, visible framelines, and frame displacements introduced by shrinking[1] and copying under suboptimal conditions.



(a) (b)

(c) (d)

Fig. 1. Artifacts often found in archive film. Figure 1(a) shows a frame with several artifacts: dirt (top left), vertical scratch (right), frameline copied into the image (bottom). The frame in Figure 1(a) has a tear in the middle. Errors introduced during copying include unwanted changes in intensity and contrast as well as visible framelines depicted in Figures 1(c) and 1(d), respectively.

3 Method

We propose a merge-based approach which is optimized for artistic archive documentaries. Firstly, we employ local features to identify shots. Secondly, we ex-

[1] Shrinking refers to the process of physical (horizontal and vertical) contraction of the filmstrip over time.

tract different orthogonal features for the detection of visual similarities. Thirdly, we extend the *overlapping links method* of Wang et al. [8] (for the grouping of visually similar shots into scenes) to multiple features by adding a merging step. Finally, we introduce a postprocessing step that assigns shots that were not assigned to scenes in the previous steps.

3.1 Shot Boundary Detection

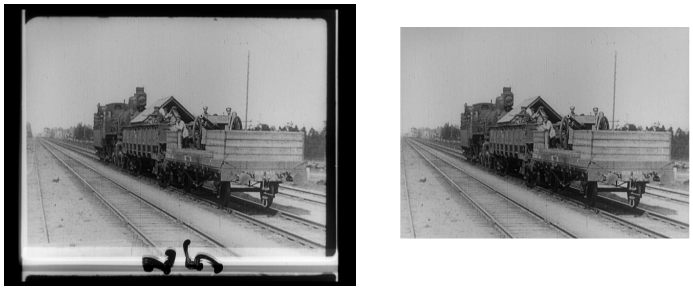
The employed shot boundary detection (SBD) technique consists of two stages. The first stage utilizes scale-invariant feature transform (SIFT) keypoints [3] to measure the similarity between two consecutive frames. We take the number of matched keypoints in two frames as an indicator for similarity. If the number of matching keypoints drops below a fixed threshold the technique recognizes a shot boundary. The results of this simple approach deteriorate in two cases, namely dissolves and sequences with fast motion. During dissolves (frames of the preceding shot are blended with frames of the following shot) the number of matching keypoints does not drop below the threshold and the shot boundary is missed. We observe that missed dissolves do not impede scene segmentation because dissolves usually do not coincide with scene boundaries in the material under consideration.

Fast motion causes difficulties for the SBD technique because fast motion introduces significant differences in consecutive frames. The number of matching SIFT keypoints drops to zero and a shot boundary is detected. The false differences lead to over-segmentation and motivate a second stage of our SBD technique. The second stage heuristically corrects over-segmentation by combining shots that are below a reasonable length. Note that over-segmentation is not critical, because similar but falsely split shots are later merged during scene segmentation.

3.2 Scene Segmentation

Our scene segmentation technique relies on the results of the SBD to extract keyframes which are used for the similarity measurements. Each shot is represented by the first frame of the shot. After these keyframes have been selected, additional preprocessing is necessary. This preprocessing consists of image cropping in order to remove the frame borders that do not carry any important information, see Figure 2. For the selected film material the crop masks were determined manually for each film.

The next computational step is the extraction of visual content-based features. We employ three content-based features: block-based intensity histograms (BBH), the edge change ratio (ECR) and SIFT keypoints. For the BBH we divide the image into blocks and compute the intensity histogram for each block. The edge change ratio (ECR) is an edge-based measure for the dissimilarity of images [10]. ECR is the ratio of appearing and disappearing edges in two (consecutive) frames. SIFT provides descriptions of salient points in the image. These three features represent orthogonal information, namely intensity, edges,



(a) original (b) pre-processed

Fig. 2. A frame prior to (Figure 2(a)) and after (Figure 2(b)) preprocessing. Entirely black areas are cropped to facilitate similarity matching. Cropping further removes potentially misleading information like the horizontally handwritten numerics *2* and *5* at the bottom of Figure 2(a).

and salient keypoints. Consequently, by combining the features we are able to capture a larger spectrum of visual similarities. Furthermore, the features have the potential to mutually compensate for weaknesses. For example in situations where keypoints can hardly be detected (e.g. in a shot that mainly shows homogeneous areas like sky), the intensity histograms may provide a more accurate description).

We compute the similarities between the shots using appropriate distance measures for the features. We can limit the similarity computation to a time window of several preceding and following shots without losing too much information. The similarity for the SIFT feature is expressed by the number of matching keypoints. If the number of matching keypoints of two frames exceeds the threshold T_{SIFT} we consider the frames as similar. We compare BBHs using the absolute sum of bin-wise differences and the number of matching blocks. We employ two thresholds $T_{BBH_{Bin}}$ for the bin-wise differences and $T_{BBH_{Block}}$ for the number of matching blocks. $T_{BBH_{Bin}}$ is used to decide whether two image blocks are similar and $T_{BBH_{Block}}$ is used to decide whether two frames are similar. The ECR is a feature which is always computed for two frames and therefore it is used directly as a measure for dissimilarity. If the ECR is lower than the threshold T_{ECR} the frames are considered to be similar.

These similarity computations yield three different similarity scores for each pair of shots. We use these scores separately for scene segmentation. The scene segmentation algorithm groups all shots that are similar and inside a specified time window into scenes. Additionally, shots that are between two similar shots are assigned to the group of the surrounding similar shots. Figure 3 illustrates this process. At the end of the shot grouping we obtain three different scene segmentations (one for each content-based feature).

We merge these three segmentations using the set operation *union*. For an illustration see Figure 4. For example consider Shot 2-4 in the first and second row of Figure 4. Shot 3 and Shot 4 are part of one scene according to BBH (first

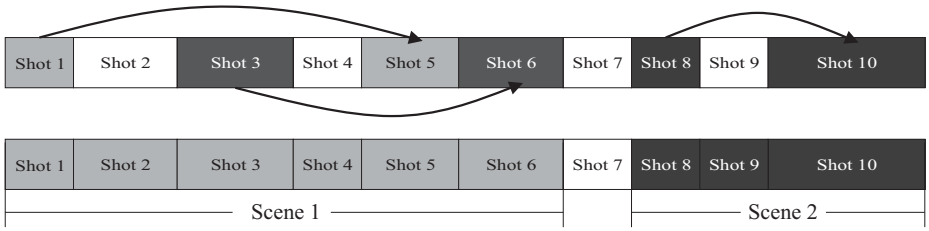


Fig. 3. The similarity computations show which shots belong together (indicated by the arrows and shading). Shots that have no similarities (Shot 9) but exist between matching shots are assigned to the scene defined by the surrounding matching shots (Shot 8 and Shot 10).

row in Figure 4). Shots 2 and 3 are part of one scene according to SIFT (second row in Figure 4). The union operation combines these overlapping sets of shots into one scene with the shots 2, 3, and 4. The output of this procedure are the so called *core scenes* (last row in Figure 4).

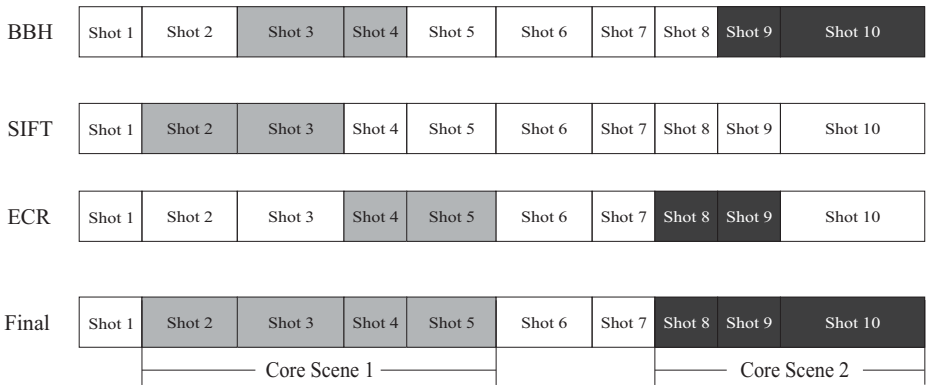


Fig. 4. The three scene segmentations obtained by the content-based features, BBH, SIFT, and ECR are combined. The result of this combination are the core scenes.

3.3 Post-processing

Usually, there are still unassigned shots between the core scenes. These shots are assigned to scenes by iteratively repeating the scene segmentation for the unassigned shots. With each iteration we decrease the similarity requirements (manipulating the similarity thresholds accordingly) until the shots are assigned to adjacent scenes (see Figure 5).

The *iterative* reduction of the similarity thresholds avoids the unwanted merging of scenes. If we set the thresholds too low from the beginning, scene

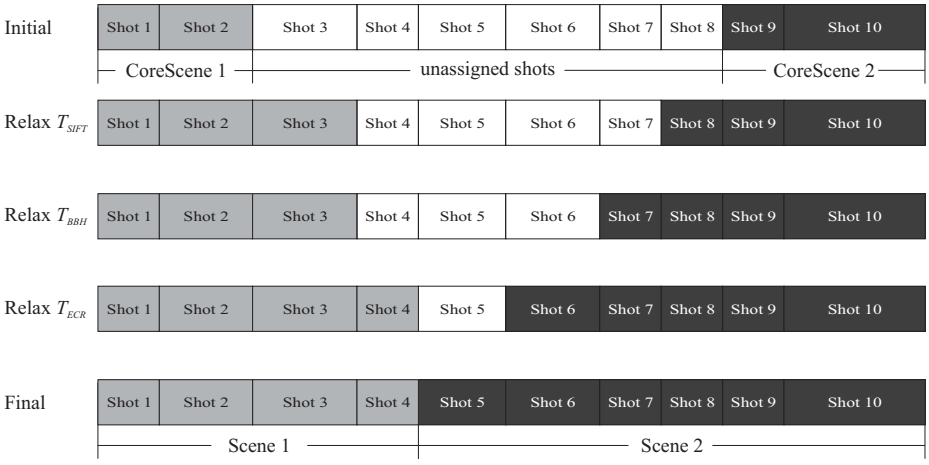


Fig. 5. Iterative repetition of the scene segmentation in order to assign the unassigned shots. The similarity requirements are decreased until all shots are assigned.

segmentation would be too tolerant yielding an under-segmentation of the film. An example is shown in Figure 6.

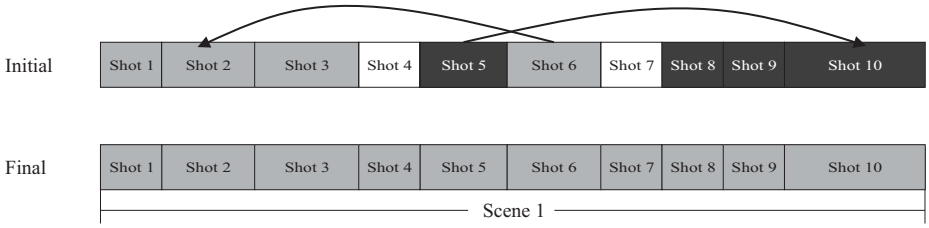


Fig. 6. Unwanted scene merging occurs, when the initial similarity requirements are too low. Arrows indicate similar shots, shading indicates assignment to scenes. The iterative reduction of similarity thresholds prevents this situation.

4 Experiments & Results

The investigated artistic archive documentaries have an experimental style and a complex temporal structure. We perform experiments with two films from the soviet filmmaker Dziga Vertov: “The Eleventh” and “Man with a Movie Camera”. The films were selected because there is a common agreement among film scientists about the scene segmentation of these films. This is not the case for all films originating in this time where the segmentation into scenes usually

is highly interpretative. The characteristics of the two films are summarized in Table 1.

Title	Length	Frames	Shots	Scene boundaries
The Eleventh	58'	63123	660	21
Man with a Movie Camera	1h28'	95768	1782	38

Table 1. Characteristics of the investigated films.

We perform a detailed analysis of the approach (as a whole) and the different components (features, processing steps, etc.). The experiments are structured as follows:

1. Performance of single features: which of the employed features are most beneficial for scene segmentation.
2. Evaluation of the overall system' performance (based on the selected features from the first step).
3. Analysis of postprocessing: what is the benefit of the proposed postprocessing step (assignment of unassigned shots to core scenes)?
4. Performance on contemporary material: we perform scene segmentation on contemporary movies enabling a comparison with the performance of other state-of-the-art methods.

In a preliminary study, we analyze the performance of single features. Therefore, we select a subset of the video material (the first 200 shots of "The Eleventh") and analyze how many similar shots are detected by the individual features. A manual annotation reveals that 98 shots show significant similarities. For each feature we empirically optimize the decision thresholds. Similarity comparison by single features focuses on the optimization of precision while the next step (scene scegmentation) targets the improvement of recall. The results for all three features are summarized in Table 2.

Feature	Correct	Missed	False Positives	Recall	Precision
Block-based histogram (BBH)	50	48	0	0.51	1.00
Edge Change Ratio (ECR)	46	52	0	0.47	1.00
SIFT	48	50	0	0.49	1.00

Table 2. Performance of the single features.

The three features show an equally good performance (recall of approximately 0.50 at a precision of 1.00). This does not mean that the features represent the same information. An evaluation of the results reveals that each feature finds a different subset of similar shots. That means that they capture different types of similarities and complement each other. Consequently, we select all three features for the final system to retain as much information about similarities as possible.

For the SIFT feature the optimal decision threshold T_{SIFT} is 40, which means that two frames are considered similar if more than 40 feature points are similar. The decision threshold for ECR, T_{ECR} is 23% which means that two keyframes are similar if less than 23% of their edges are different. The two thresholds for BBH are: $T_{BBH_{Bin}} = 0.5$ and $T_{BBH_{Block}} = 55\%$. That means that two keyframes are considered similar if the difference between their histogram is lower than 0.5 in more than 55% of the blocks. Experiments with block-size (5x5, 10x10, 20x20) showed that this parameter has only little influence on the results. A block-size of 10x10 is chosen for the final system. The thresholds are set rather strict to avoid false positives at this stage of processing.

The performance measures for the entire system for the investigated films are summarized in Table 3. Recall is the ratio of correctly detected shot boundaries (SBs) and the total number of SBs according to the ground truth. Precision is the ratio of correctly detected SBs and the total number of retrieved SBs. For both films recall is above 90%. In "The Eleventh" only one scene boundary is not detected. For "Man with am Movie Camera" three scene boundaries cannot be detected. The (compared to recall) lower precicion values indicate that the approach performs an over-segmentation of the films.

Title	Correct	Missed	False Positives	Recall	Precision
The Eleventh	20	1	11	0.95	0.65
Man with a Movie Camera	35	3	50	0.92	0.41

Table 3. The performance of scene segmentation for the historic documentaries.

Next, we investigate the performance of the proposed postprocessing step (iterative assignment of unassigned shots to core scenes). An analysis of the core scenes obtained by the scene segmentation (without postprocessing) reveals that a significant amount of shots are not assigned to any core scene, see Table 4. These results show that the postprocessing is crucial for robust scene segmentation.

Title	# Shots	# Unass. Shots	Percentage unassig. shots
The Eleventh	660	203	31%
Man with a Movie Camera	1787	460	26%

Table 4. Assignment of shots to core scenes results in a significant portion of shots that are not assigned to any core scene.

The postprocessing iteratively makes the decision thresholds more tolerant to assign the unassigned shots to core scenes. Our experiments show that the largest benefit in this postprocessing is provided by the SIFT feature followed by BBH and ECR. From this observation we conclude that the SIFT feature is the most expressive feature in our study. Additionally, we observe that only a

few iterations are necessary to assign all unassigned shots (3 iterations for "Man with a Movie Camera" and 5 iterations for "The Eleventh").

Finally, we evaluate the performance of our method for contemporary film material which enables the comparison with other state-of-the-art shot segmentation approaches. We apply the proposed approach to two films that are often used for scene segmentation in literature: "Forest Gump" and "Blade Runner". The results for "Blade Runner" are summarized in Table 5. This film is also analyzed in [6] where Sundaram and Chang apply two versions of their algorithm ("naive case" and "using refinement"), see Table 5.

Approach	SBs	Found	Correct	Missed	False Positives	Recall	Precision
"naive case" [6]	24	n/a	20	4	n/a	0.83	n/a
"using refinement" [6]	24	n/a	18	6	n/a	0.75	n/a
proposed approach	24	40	18	6	22	0.75	0.45

Table 5. The performance of scene segmentation for the film "Blade Runner" compared to the approach of [6].

From Table 5 we observe that our method performs comparably well to the approach in [6] although our method is not optimized for (high-quality) color video (for example we do not make use of color information). Furthermore, we do not use audio information as in [6]. Note that the comparison is limited in expressiveness since the authors of [6] do not provide information about the precision of their approach.

We further analyze the movie "Forest Gump". Since no ground truth segmentation is available, we manually analyze the film and identify 52 scene boundaries. The proposed method finds 68 scene boundaries, where 44 are correct and 8 are missed. The number of false positives is 24 which yields a recall of 0.85 and a precision of 0.65. This is a satisfactory result for state-of-the-art scene segmentation algorithms. We compare our method with that proposed by Vendrig and Worring [7]. The authors segment the movie into 152 "'logical story units". If we apply the same rules and preconditions for segmentation as in [7] we obtain 143 segments. A more precise comparison is not possible since the authors in [7] employ different evaluation criteria.

5 Summary

The specific characteristics of artistic archive documentaries require adapted analysis techniques. We have presented a novel scene segmentation technique that has been designed specifically for this type of film. The technique employs orthogonal features to group similar shots into scenes in a two-step process. In the first step, we identify core-scenes based on visual similarity. In the second step, we assign the remaining unassigned shots to the core scenes, by iterative adaptation of thresholds. We evaluate the technique using archive documentaries

and contemporary fiction films. The results for fiction films are satisfactory with an average recall of 0.80 and an average precision of 0.55. We achieve significantly higher recall for archive documentaries. An average recall of 0.94 and a precision of 0.53 prove the suitability of our technique to archive documentaries. We will direct future work toward reducing over-segmentation and thus increase the precision of the presented scene segmentation technique.

Acknowledgments

This work has received financial support from the Vienna Science and Technology Fund (WWTF) under grant no. CI06 024.

References

1. Beaver, F.E.: Dictionary of film terms: the aesthetic companion to film art. Peter Lang Publishing (2009)
2. Gu, Z., Mei, T., Hua, Z.S., Wu, Z., Li, S.: EMS: Energy minimization based video scene segmentation. In: IEEE International Conference on Multimedia and Expo. pp. 520–523 (2007)
3. Lowe, D.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
4. Rasheed, Z., Shah, M.: Scene detection in Hollywood movies and TV shows. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03). vol. 2, pp. II– 343–8 vol.2 (2003)
5. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE Transactions on Multimedia 7(6), 1097–1105 (2005)
6. Sundaram, H., C.S.: Video scene segmentation using video and audio features. In: IEEE International Conference on Multimedia and Expo, 2000. pp. 1145–1148. IEEE, Piscataway, NY, USA (2000)
7. Vendrig, J., Worring, M.: Systematic evaluation of logical story unit segmentation. IEEE Transactions on Multimedia 4(4), 492–499 (2002)
8. Wang, X., Wang, S., Chen, H., Gabbouj, M.: A Shot Clustering Based Algorithm for Scene Segmentation. In: Proceedings of the 2007 International Conference on Computational Intelligence and Security Workshops. pp. 252–259. IEEE Computer Society, Washington, DC, USA (2007)
9. Yeung, M., Yeo, B.L., Liu, B.: Segmentation of video by clustering and graph analysis. Computer Vision and Image Understanding 71(1), 94–109 (1998)
10. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying scene breaks. In: Proceedings of the 3rd ACM International Conference on Multimedia. pp. 189–200. ACM, New York, NY, USA (1995)
11. Zhai, Y., Shah, M.: Video scene segmentation using markov chain monte carlo. IEEE Transactions on Multimedia 8(4), 686–697 (2006)