

VIDEO SHOT BOUNDARY DETECTION BY STRUCTURAL ANALYSIS OF LOCAL IMAGE FEATURES

Murat Birinci, Serkan Kiranyaz, Moncef Gabbouj

Department of Signal Processing
Tampere University of Technology

ABSTRACT

In this paper a novel shot boundary detection (SBD) algorithm is proposed in order to detect both abrupt and gradual transitions. Visual content changes between different shots are detected via structural analysis of local image features. A top-down approach is utilized in order to find the location of the transition accurately, while keeping the computational load minimal avoiding unnecessary feature extraction and matching. Experimental results prove that the proposed method outperforms the state-of-art methods both in performance and efficiency.

1. INTRODUCTION

The amount of available video content is growing exponentially with the development in content creation technology. Breaking up the video into its basic components, i.e. shots, is the fundamental step for efficiently analyzing and managing such content. However, factors such as low video quality, various transition types between shots, intense camera and object motion turn shot-boundary-detection (SBD) into a more challenging problem. While there are numerous methods for detecting shot boundaries, most of them suffer from performance, computational cost or even both.

The idea of an SBD algorithm is based on the same assumption that there is a visual discontinuity between consecutive shots. Such discontinuities have been aimed to be detected via various visual descriptors such as color histograms [1], DCT coefficients [2] or motion activity [3]. While such global descriptors yield acceptable results for hard-cuts, they suffer from performance in the presence of gradual transitions and camera or object motion. In [4], authors compared frames based on focus of attention (FOA) of the viewer and obtained reasonable results for hard-cut and fade type gradual transitions. Park *et al.* in [5] used an object recognition algorithm, namely SIFT [6], in order to exploit the similarities between frames and hence detect shot boundaries. Their assumption was that if certain amount of matches are present between frames, than those frames are considered to belong to the same shot. They compared consecutive frames for detecting abrupt shot changes (hard-

cut) and non-adjacent frames with a fixed distance apart for detecting gradual transitions. However, their method considerably suffered from the heavy computational cost of the SIFT algorithm. Moreover, by relying on the difference between adjacent frames with a fixed threshold, their accuracy is relatively low under high motion and sudden illumination changes. A parallel approach is carried out by Huang *et al.* [7] where they used Contrast-Context Histograms (CCH) [8] as their descriptor. Instead of applying a fixed threshold to the number of matched points between consecutive frames, they processed all the frames in the video and compared every adjacent frame. Then, they check the local minima and maxima of the number of matches between consecutive frames and determine candidate boundaries. They further match non-adjacent frames at those local minima in order to increase the accuracy and also handle gradual transitions. Due to the light computation of CCH, their method is considerably faster than [5]. Yet despite their satisfactory performances, both approaches suffer from an excessive computational complexity due to the bottom-up approach, i.e. feature extraction and matching is performed for *every* single frame in the video. On the other hand, while skipping a predetermined number of frames as in [5] seems to be a viable solution for decreasing computational cost, it may, as well, result in a reduced accuracy.

In order to address the aforementioned problems and perform accurate shot boundary detection under practical computational costs, in this paper we propose a *top-down* approach where the whole process is modeled based on the well known “*Information Seeking Mantra*” described by Shneiderman in [9]: “*Overview first, then zoom and filter*”. That is to say, the algorithm first performs a global search and then localizes (or “*zooms in*”) where a shot boundary exists. Therefore, unnecessary feature extraction and matching operations can be avoided by filtering out the irrelevant frames. This is completely the opposite approach followed in [7], where authors compare every adjacent frame and “*zoom out*” whenever there is a candidate shot boundary. Moreover we perform a structural analysis over the keypoints and judge frame similarity based on the structural similarity instead of blindly performing a one-to-one irregular keypoint matching.

The rest of the paper is organized as follows. In Section 2 proposed method is introduced in detail. Experiments conducted in a benchmark video repository are presented in Section 3. Section 4 derives some important conclusive remarks and suggests possible future work.

2. PROPOSED SBD METHOD

The proposed method utilizes local image features and detects shot boundaries by revealing frame (dis-)similarities. The main contributions of the method can be discussed in two parts: First, is the structural analysis performed on the keypoints for verifying the frame (dis-)similarity. This basically utilizes (solid) object (or object parts) tracking between (non-adjacent) frames whilst camera/object view may change due to zoom, scale or rotations. Second is the top-down nature of the algorithm, which provides significant improvements in terms of computational complexity and SBD accuracy as compared to recent techniques such as [5] and [7].

2.1. Similarity Matching via Structural Analysis

Employing local image features over salient key-points for the purpose of SBD is favorable since they are robust against camera/object variations in the video such as background motion, camera zoom-in and zoom-out effects, object movement, scaling, rotation within the scene. Particularly, scale and affine invariance capability of such features strengthens the robustness of SBD. On the other hand, relying solely on the number of matches or the ratio of the matched keypoints may yield erroneous results. This is mainly due the imperfect discrimination power of the underlying feature descriptor. Since their notion of similarity typically relies on applying a predetermined threshold to individual feature distances, their performance on matching the same object(s) is far from reliable. In an earlier work [10], it has been demonstrated that the matching process should be more than a mere sum of the independently matched keypoints; instead an organization should be imposed in order to converge to human visual perception. A structural analysis over local features has been proposed in [10] by applying *Gestalt's laws of grouping* as the organization measure, namely *spatial proximity*, which significantly emphasizes both similarities and dissimilarities among frames (see Fig.1).

Let us consider the match between keypoints p_1 and p_2 from $F(n)$ and $F(m)$ respectively, where $F(n)$ and $F(m)$ are the n^{th} and m^{th} frame of a video. While it is rather intuitive that keypoints in close proximity of p_1 should match to the keypoints in close proximity to p_2 , without the structural constraint in [10] (spatial proximity), they can match anywhere in $F(m)$ (see Fig.1.c) causing erroneous similarity measures. However, such structural analysis assures that "objects" (or object parts) are matched as a whole to each

other instead of individual points in $F(n)$ matching to arbitrary and possibly irrelevant points in $F(m)$. While this fact may deteriorate generic object matching algorithms due to intense object and camera transformation, in case of video – particularly in the proposed algorithm – possible content change within the same shot is limited. While there are more elaborate structural analysis methods available, such as RANSAC [11], they are computationally complex for our purpose and hence not preferred. The details on how the structures are formed and matched between two frames are omitted due to space limitations and the reader is referred to [10] for further details.

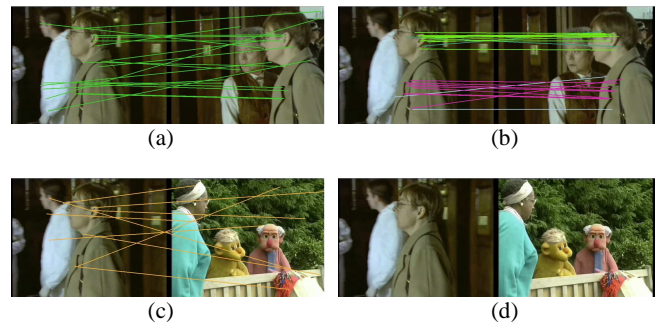


Fig.1. Frame similarities with (b and d), and without (a and c) structural analysis.

The video content has significant effect on the extracted number of keypoints from each frame. Therefore, it will be misleading to simply rely on the number of matched keypoints in order to judge the similarity between two frames. Consider a simple example where $F(n)$, $F(m)$ and $F(k)$ have 320, 350 and 1500 keypoints respectively. If there are 200 matches between both $F(n)$ and $F(m)$, and between $F(m)$ and $F(k)$, it would lead to incorrect conclusions to regard the similarity as the same. So, in order to avoid such inaccurate decisions, after the keypoints are matched based on the structural analysis, the similarity judgment between two frames is based on the *ratio* of the matched keypoints to the total number of keypoints, in order to avoid the bias due to the variations in the number of extracted keypoints. Therefore the difference in the similarity will be signified, as in the above example (200/350 between $F(n)$ and $F(m)$, and 200/1500 between $F(m)$ and $F(k)$).

2.2. The Top-Down Approach

In order to avoid the aforementioned drawbacks of the methods proposed in [5] and [7], we followed a top-down approach adopted from the well-known "Information-Seeking Mantra" [9]: *Overview first, then zoom and filter*. Therefore, a global search is first carried out by performing similarity matching every N^{th} frame in the video. Here, instead of applying a threshold to the match ratio mentioned in Section 2.1, we observe the change in the match ratio and decide on the shot boundary based on the significant

changes in the frame similarity. Note that such comparison only indicates whether or not a shot boundary exists, yet the actual shot boundary can be anywhere within that N -frame interval. So we further “zoom in” to that part of the video and gradually decrease the interval between frames.

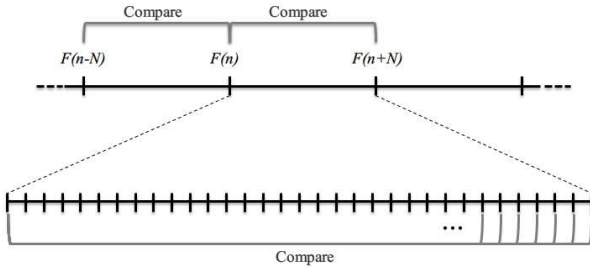


Fig. 2. The proposed top-down approach

Let us denote the n^{th} frame of a video as $F(n)$. Then the pseudo-code of the proposed method, which is also depicted in Fig. 2, can be given as in Table 1.

Table 1 The pseudo-code of the proposed SBD method.

1	for $n = 0, N, 2N, 3N, \dots$
2	Compare $F(n)$ to $F(n+N)$
3	If a shot boundary is detected then do:
4	for $m = n+N-1, n+N-2, \dots$
5	Compare $F(n)$ to $F(m)$
6	end
7	end
8	end

The change in similarity as m approaches n reveals the location of the shot boundary together with its nature, i.e. whether it is a hard-cut or a gradual transition (see Fig. 3). Especially for the latter case, the amount of visual change between adjacent frames is imperceptible and thus the advancement of such change as we “zoom in” provides significant information.

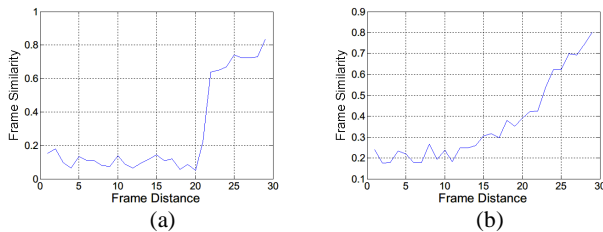


Fig. 3. The change in frame similarity as the interval grows.
(a) Hard-cut, (b) Gradual

3. EXPERIMENTAL RESULTS

In order to demonstrate the performance of the proposed method and reveal its advantages over the state-of-art methods, we performed SBD over 5 different video

sequences taken from the *Open Video Project* [12] with various qualities, dimensions and containing several types of transitions, as tabulated in Table 2.

Table 2 Test Sequences

Name	Number of frames	Number of transitions	Size
1955 Chevrolet Screen Ads	15802	78	480x368
Open Video2	1945	11	320x240
Open Video3	1870	20	320x240
History Of Flight	2801	22	720x480
Volcano Eruptions	3332	27	720x480

We used SURF [13] for both feature detection and description. Our choice is not only based on its speed, but also its high performance, which is also denoted in [14] that it has the highest coincidence with visual saliency models among other detectors such as SIFT [15] and MSER [16]. The parameter N (discussed in Section 2.1) is chosen to be 1 *sec*. In other words, shots in the videos are assumed to be at least 1 *sec*.

The overall performance results of the proposed method are shown as precision and recall values in Table 3. In order to validate its contribution, results are given both with and without structural analysis. The average precision and recall values are respectively 97.2% and 85.6% for the proposed method and 99% and 48.6% without structural analysis. It can clearly be seen that SBD without structural analysis, misses significant amount of shot boundaries. While the precision values are close to perfect without structural analysis (almost all retrieved boundaries are correct), recall values are quite low. This indicates that several frames belonging to different shots were regarded as similar, and hence several shot boundaries are missed. This is a clear revelation of the fact depicted in Fig.1.

Table 3 Performance Analysis

Name	With Structural Analysis		Without Structural Analysis	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
1955 Chevrolet Screen Ads	96	90	95	69
Open Video2	100	91	100	27
Open Video3	100	95	100	60
History Of Flight	95	82	100	50
Volcano Eruption	95	70	100	37

Table 4 shows the computational times for the proposed and the competing methods, [5] and [7] which require processing of *all* frames in the video sequence. In order to perform a fair comparative evaluation in terms of computational complexity between the proposed and the competing methods, and demonstrate the improvement

gained with the top-down approach, Table 4 presents computation times obtained by processing all frames and comparing adjacent frames as in [5] and [7]. Note that in [7], an additional interval search is also performed for every candidate boundary (i.e. for every local minima), which is not included in the computational times reported in Table 4. The results clearly show that the proposed method increases computation speed *at least* 5 times while achieving a delicate SBD accuracy.

Table 4 Computation Time Analysis

Name	Time in sec. (<i>proposed</i>)	Time in sec. (<i>[5]and[7]</i>)
1955 Chevrolet Screen Ads	671,8	3821,65
Open Video2	43,81	252,175
Open Video3	101,71	289,909
History Of Flight	286,63	1399,56
Volcano Eruptions	260,13	1603,75

4. CONCLUSION

A new shot boundary detection algorithm is proposed where a top-down approach is used in order to minimize the computational cost without sacrificing the accuracy. Local image features are utilized to assess frame (dis)similarities. In order to further increase the accuracy and adapt the video content more competently, a structural analysis is performed on the keypoints. Experimental results prove the superiority of the proposed method not only in terms of performance, but also computational complexity. Planned future work includes performance improvements and utilization of a more in depth structural analysis that will not only handle intense camera and object motion, but also perform accurate classifications of the transition types between shots.

5. REFERENCES

[1] U. Gargi, R. Kasturi, and S.H. Strayer, "Performance characterization of video-shot-change detection methods," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, 2002, p. 1–13.

[2] K. Shen and E.J. Delp, "A fast algorithm for video parsing using MPEG compressed sequences", in Proc. ICIP, 1995, pp.2252-2255.

[3] A. Amel, B. Abdessalem, and M. Abdellatif, "Video shot boundary detection using motion activity descriptor," *Journal of Telecommunications*, vol. 2, 2010, pp. 54-59.

[4] G. Boccignone, A. Chianese, V. Moscato, and A. Picariello, "Foveated shot detection for video segmentation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, 2005, p. 365–377.

[5] M.-H. Park, R.-H. Park, and S. W. Lee, "Shot boundary detection using scale invariant feature matching," in Proc. SPIE Visual Communications and Image Processing, 2006, vol. 6077, pp. 569–577.

[6] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, 2004, p. 91–110.

[7] C.-R. Huang, H.-P. Lee, and C.-S. Chen, "Shot Change Detection via Local Keypoint Matching," *IEEE Transactions on Multimedia*, vol. 10, Oct. 2008, pp. 1097-1108.

[8] C.-R. Huang, C.-S. Chen, and P.-C. Chung, "Contrast context histogram—An efficient discriminating local descriptor for object recognition and image matching," *Pattern Recognition*, vol. 41, Oct. 2008, pp. 3071-3077.

[9] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," Proc. 1996 IEEE Symposium on Visual Languages, pp. 336-343.

[10] M. Birinci, F. Diaz-de-Maria, G. Abdollahian, E.J. Delp, and M. Gabbouj, "Neighborhood Matching for Object Recognition Algorithms Based on Local Image Features", in Proc. IEEE DSP/SPE Workshop, 2011.

[11] M.A.Fischler and R.C.Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, 1981, pp. 381–395.

[12] The Open Video Project [Online]. Viewed 2011 January 01. Available: <http://www.open-video.org>

[13] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, 2008, pp. 346-359.

[14] P. Harding and N.M. Robertson, "A Comparison of Feature Detectors with Passive and Task-Based Visual Saliency," Proc. of the 16th Scandinavian Conference on Image Analysis (SCIA), 2009, p. 725.

[15] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, 2004, pp. 91–110.

[16] J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. of British Machine Vision Conference, 2002, pp. 384-396.