A Survey on Video Segmentation

Dalton Meitei Thounaojam, Amit Trivedi, Kh. Manglem Singh and Sudipta Roy

Abstract This paper presents a short survey on video segmentation. Due to the growth in multimedia information, an effective video indexing and video retrieval is necessary. This can be achieved when an effective video segmentation tools and algorithms are available. MPEG-compressed videos are mostly used by researchers for video segmentation. *Shot boundary detection*, *color histogram characteristics*, *DC-images*, *motion vector* and *motion compensation*, *threshold-based detector*, etc. are mostly used for video segmentation.

Keywords Shot · MPEG · Fade · Dissolve · Motons · Color histogram

1 Introduction

With the growth of the multimedia information, many researchers have been attracted toward *video indexing* and *video retrieval*. Better performance video indexing and retrieval system can be achieved when proper video segmentation tools and algorithm are applied to the system. Video-on-demand, digital video

D. M. Thounaojam (✉) · S. Roy
Assam University, Silchar, India
e-mail: dalton.meitei@gmail.com

S. Roy
e-mail: sudipta.it@gmail.com

Kh. Manglem Singh
National Institute of Technology Manipur, Manipur, India
e-mail: manglem@gmail.com

D. M. Thounaojam · A. Trivedi
National Institute of Technology Silchar, Silchar, India
e-mail: rivedi19@gmail.com

D. P. Mohapatra and S. Patnaik (eds.), *Intelligent Computing, Networking, and Informatics*, Advances in Intelligent Systems and Computing 243,
DOI: 10.1007/978-81-322-1665-0_91, © Springer India 2014

libraries, distance education, geographical information systems, etc. are some applications of video segmentation.

It is needed to divide the video data for effective indexing and retrieval into basic elements called shots. A shot [1] represents a sequence of frames captured when camera starts rolling and until it stops. The main problem of segmenting a video sequence into shots is the ability to distinguish between scene breaks and normal changes which may be due to the motion of large objects or the motion of the camera. When special effects are involved, two shots are merged using gradual transition. The types of gradual transitions used mostly are *dissolve, fade in*, and *fade out*. A fade is a gradual transition between a scene and a constant image (fade out) or between a constant image and a scene (fade in). A dissolve is a gradual transition from one scene to another, in which the first scene fades out and the second scene fades in.

In [2], video segmentation is performed based on the types of editing operation that an editor can perform when editing two shots. They are identity class (cuts), spatial class (page translates, page turns, shattering edits), chromatic class (fade in, fade out, and dissolve), and spatio-chromatic class (image morphing and wipes).

The Sects. 2, 3, and 4 of this paper contains a short survey of video segmentation based on text feature, audio feature and image and motion feature.

2 Text Feature Extraction

Hauptman and Smith [3] used scenes, audio signal, and word relevance in the video to segment the video for better indexing and retrieval which is to be implemented for the Informedia Digital Video Library Project at Carnegie Mellon University. Two main techniques are used in natural language analysis.

1. Natural structural text markers such as punctuation are used to identify segments of video.
2. Term frequency/inverse document frequency (TF/IDF) is used to identify critical keywords and their relative importance for the video document.

3 Audio Feature Extraction

Hauptman and Smith [3] used Sphinx-II speech recognizer to recognize spoken words, which uses semi-continuous hidden Markov models (HMM) to model context-dependent phones (triphones), including between-word context. To detect breaks between utterances, they used a modification of signal-to-noise ratio (SNR) techniques which compute signal power.

Boreczky and Wilcox [4] describes a technique for segmenting video by using an audio distance based on the acoustic difference in intervals just before and after the frames. An audio distance based on sliding two-second intervals is calculated for extracting audio features.

Huang et al. [5] computes a feature vector for each clip and then calculates an audio dissimilarity index for each clip to detect audio breaks for segmentation.

4 Image and Motion Feature Extraction

Fukunaga and Hostetler [6] proposed a general nonparametric density and mean shift estimators for application in pattern recognition procedure and its efficacy on low-level vision tasks such as tracking has been considered.

Black [7] presented a method for segmenting image sequences by using intensity and motion information which adopts *Markov random fields* approach.

Arman et al. [8] proposed an algorithm for detecting scene change using the DCT coefficient of MPEG- or JPEG-encoded video sequences before decoding.

Alattar [9] developed a dissolve detector and a dissolve handling algorithm for the DVI PLV 2.0 which has been incorporated with PLV 2.0 and currently being used with Intel Corporation's DCF.

Zhang et al. [10] used pairwise comparison, likelihood ratio, and histogram comparison algorithms as video partitioning metrics. The twin-comparison technique is also used to detect dissolve.

Hampapur et al. [11] adopted a feature-based classification approach to the problem of edit detection. A chromatic and spatial edit is achieved by manipulating the color or intensity space and pixel space of the shots being edited, respectively. The segmentation is achieved by using a modified two-class discrete classifier.

Tonomura et al. [12] propose a method for shot analysis based on video X-ray images. A sliced image from spatiotemporal images, based on video intensity data, is extracted for edge detection. They used an interframe differencing operation for detecting cuts.

Meng et al. [13] proposed an algorithm for the detection of abrupt scene change and special editing effects such as dissolve in a compressed MPEG/MPEG-2 bit-stream with minimal decoding of the bit-stream. Scene changes are easily detected with DCT DC coefficients and motion vectors. The detection is performed with only a partial decoding of the compressed bit-stream: minimal decoding for the DCT DC coefficients for the *I*- and *P*-frames and motion vectors for the *P-frames* and *B*-frames.

Hampapur et al. [2] presented a video edit model which is based on a study of video production processes. The techniques proposed in this paper treat video segmentation as edit boundary detection. This model captures the essential aspects of video editing. Video features extractors for measuring image sequence properties are designed based on the video edit model. The work presented in this paper has taken the top down approach to video segmentation resulting in the use of effect models which have allowed the design of simple and effective techniques for segmenting digital video.

Hauptman and Smith [3] used *histogram difference analysis* for image analysis. TF/IDF technique is used to identify critical keywords and their relative importance for the video document.

Wang [14] described a technique for unsupervised video segmentation based on watersheds and temporal tracking. The motion estimation algorithm for arbitrarily shaped objects is based on hierarchical block-matching and least-square approximation. It is computationally efficient and generally provides motion parameters of reasonable accuracy. The algorithm for motion projection allows the proposed technique to track fast-moving objects. The algorithm for marker extraction and the modified watershed algorithm can detect the appearance and disappearance of objects and can segment newly appearing objects.

Boreczky and Wilcox [4] described a technique for segmenting video using *HMM*. Features for segmentation include an image-based distance between adjacent video frames and an estimate of motion between the two frames. The histogram feature measures the distance between adjacent video frames based on the distribution of luminance levels. It is simple, easy to compute, and works well for most types of video. Motion vectors are computed using an exhaustive-search block-matching algorithm in a 24×24 window for nine evenly distributed 40×40 pixel blocks.

Huang et al. [5] presented a scheme for video segmentation by integrating image, motion, and audio information which is used to differentiate between shot break and scene break. Color histogram of each video frame is calculated to detect color break. The difference in motion histogram between two successive frames is calculated to detect motion break.

Salembier and Marques [15] discussed region-based representations of image and video that are useful for multimedia services such as those supported by the MPEG-4 and MPEG-7 standards. Classical tools related to the generation of the region-based representations are discussed. It mentions that object tracking enables video object (in the sense of MPEG-4) creation and it opens the door to content-based functionalities.

Meier and Ngan [16] presented the *video object plane* (VOP) segmentation algorithm which separates foreground objects from the background based on motion information. The model points consist of edge pixels detected by the Canny operator which is later used for VOP extraction. MPEG-4 video is used.

Allunbasak [17] introduced a framework for selecting statistically optimal thresholds in temporal video segmentation algorithms. Statistics for various temporal video events such as shot boundary, zoom, and pan events are collected. Then, optimal thresholds are estimated so as to minimize a statistical cost function defined in terms of the complied statistics.

Truong et al. [18] improved cut detection method by using color histogram differences by utilizing an adaptive threshold computed from a local window on the luminance histogram difference curve. Based on the mathematical models for producing ideal fades and dissolves, different clues (e.g., monochrome frames) for discovering the existence of these effects are proposed, and constraints on the characteristics of frame luminance mean and variance curves are derived analytically in this approach to eliminate false positives caused by camera and object motion during gradual transitions.

Jadon et al. [19] proposed a scheme using the Rayleigh distribution for fuzzification of frame-to-frame property. They used histogram intersection for detecting abrupt change, combination of pixel difference and histogram intersection in detecting a gradual change and a combination of pixel difference, histogram intersection and edge-pixel count is used to detect a fade. Three-dimensional euclidean distance is used for computing pixel difference and Sobel edge detection method is used for edge-pixel count.

Huang and Liao [1] used *DC image difference* and *histogram-based difference* to measure the differences between frames. A simple gradient operator (i.e., *Sobel masks*) is used to compute the gradient image for edge detection.

Lo and Wang [20] proposed a video segmentation method using a histogram-based fuzzy c-means (HBFCM) clustering algorithm. The HBFCM clustering algorithm is a hybrid of the shot change detection algorithm (pixel-based, histogram-based, and block-based algorithms) and the clustering algorithm (K-means clustering and fuzzy c-means clustering).

Xu et al. [21] used *DC-images* which are extracted by parsing the MPEG-1 video stream without full decoding; then, grass pixels are detected according to the grass detector; and view labels are obtained by quantizing the grass-ratios into 3 levels according to appropriate threshold values adaptively set in the initialization phase.

Khan and Shah [22] proposed a *maximum posteriori probability* (MAP) framework that uses multiple cues, such as spatial location, color, and motion, for segmentation. Weights are assigned to color and motion terms, which are adjusted at every pixel, based on a confidence measure of each feature. The appropriate modeling of probability density functions (PDFs) of each feature of a region is also discussed. The correct modeling of the spatial PDF imposes temporal consistency among segments in consecutive frames.

Patras et al. [23] used *watershed* segmentation algorithm. A filtering with morphological operators with a small (3 × 3) structuring element is used for a nonlinear smoothing of the current frame.

DeMenthon [24] adopted a hierarchical clustering method which operates by repeatedly applying mean shift analysis. Each pixel of a 3D space–time video stack is mapped to a 7D feature point, and the clustering of these feature points provides color segmentation and motion segmentation.

Lu et al. [25] proposed a novel scheme for automatic video segmentation which fully exploits both temporal information from background and spatial information from foreground. In this technique, the background regions of one scene are first warped into a large sprite image, which includes all visible parts of background throughout the sequence. Temporal segmentation produces *independent foreground component* (IFC) image with the rough locations of the foreground objects.

Spatial segmentation produces homogeneous regions with precise edges reserved. The final segmentation is complete based on temporal and spatial segmentations by using boundary extraction strategy.

Sifakis et al. [26] applied empirical probability distribution to the test frame for determining the frame differences which later predict the camera motion. A multi-label fast-marching-level set algorithm which is an extension of the well-known fast-marching algorithm is applied to find the interframe difference. To localize the boundary of the moving object, region-growing algorithm is used.

Chien et al. [27] proposed a new fast watershed algorithm, named *P-Watershed*, for image sequence segmentation. By utilizing the temporal coherence property of the video signal, this algorithm updates watersheds instead of searching watersheds in every frame, which can avoid a lot of redundant computation. An intra–interwatershed scheme (*IP-Watershed*) is also proposed to further improve the results.

Porter et al. [28] proposed a motion-based algorithm to identify shot cuts which deals inherently with object and camera motion. It uses block-matching motion compensation to generate an interframe difference metric which is achieved by calculating the normalized correlation between blocks and locating the maximum correlation coefficient.

Guimaraes et al. [29] transformed the video segmentation problem into problem of 2D image pattern detection. They used discrete line analysis, max tree analysis and topological and morphological tools to detect fade transitions, flashes, and cuts, respectively.

Qi et al. [30] used supervised classification methods for video shot segmentation: a whole-frame *color histogram difference* and an 8×8 block-wise histogram difference between frames, both in the YUV color space. Combining both global- and block-wise histogram differences makes the system insensitive to object motion but very sensitive to cut boundaries, and it is followed by k-nearest neighbor classification.

Wang et al. [31] proposed an *anisotropic kernel mean shift*. The application of mean shift to an image or video consists of two stages. The first stage is to define a kernel of influence for each pixel. This kernel defines a measure of intuitive distance between pixels, where distance encompasses both spatial (and temporal in the case of video) and color distances. The second iterative stage of the mean shift procedure assigns to each pixel a mean shift point, M(xi), initialized to coincide with the pixel. These mean shift points are then iteratively moved upwards along the gradient of the density function defined by the sum of all the kernels until they reach a stationary point (a mode or hilltop on the virtual terrain defined by the kernels). For the color domain, they use an *Epanechnikov kernel* with a profile $kr(z) = 1 - |z|$ if $|z| < 1$ and 0 otherwise.

Fang et al. [32] proposed a hybrid scheme for temporal segmentation of videos, where a fuzzy logical approach is pursued to combine a number of features in performing shot-cut detection. These features include color histogram intersection, motion compensation, and texture change for abrupt shot boundary detection, and edge variance for gradual shot-cut detection.

Boccignone et al. [33] defined a novel approach to partitioning of a video into shots based on a *foveated* representation of the video. The proposed scheme aims at detecting both abrupt and gradual transitions between shots using a single technique, rather than a set of dedicated methods.

Apostoloff and Fitzgibbon [34] presented an automatic video segmentation algorithm that uses spatiotemporal features to regularize the segmentation. Detecting *spatiotemporal T-junctions* that indicate occlusion edges, they learn an occlusion edge model that is used within a color contrast sensitive MRF to segment individual frames of a video sequence. T-junctions are learn and classified using a support vector machine, and a Gaussian mixture model is fitted to the (foreground, background) pixel pairs sampled from the detected T-junctions. Graph cut is then used to segment each frame of the video, showing that sparse occlusion edge information can automatically initialize the video segmentation problem.

Joyce and Liu [35] presented two algorithms for the detection of gradual transitions in video sequences. The first is a dissolve detection algorithm utilizing certain properties of a dissolves trajectory in image space. It is implemented both as a simple threshold-based detector and as a parametric detector by modeling the error properties of the extracted statistics. The second is an algorithm to detect a wide variety of wipes based on image histogram characteristics during such transitions. Both algorithms operate in the compressed domain, requiring only partial decoding of the compressed video stream.

Wang et al. [36] proposed a method which can detect the scene change boundaries accurately through the *macroblock* information of B-frame and P-frame. The algorithm can extract the macroblock information from the encoded bit-stream directly without total decompression for MPEG bit-stream. The macroblock coding of a B-frame has four types, intro-type, forward-type, backward-type, and bidirectional-type. Motion compensation is used.

Tan et al. [37] presented a metric based on *blocked color histogram* (BCH) for interframe difference. They use SVM classifier for pattern recognition.

Pritch et al. [38] proposed video synopsis which is a temporally compact representation of video that enables video browsing and retrieval. The video synopsis uses energy minimization which represents input video sequence in 3D space–time volume. The cost function used in this paper corresponds to a 3D Markov random field (MRF) where each node corresponds to a pixel in the 3D volume of the output movie and can be assigned any time value corresponding to an input frame.

Tziakos et al. [39] proposed an approach with a graph based on the similarity of frames and enriched with the temporal information from the sequence processed by Laplacian eigenmaps. This makes it possible to visualize the manifold of motion in the scene and to detect unusual events in a low-dimensional space. The proposed framework enables the separation of actions without the need of an object detector or object tracker.

Yin et al. [40] presented an algorithm for the efficient segmentation of foreground and background in monocular video sequences. *Depth-based labeling* is used in monocular system to learn to imitate stereo motion features, called *motons*. The base weak classifier used in this paper is the widely used decision stump. A useful framework called tree-cube taxonomy is used for constructing strong classifiers by combining weak classifiers in different ways.

Hong et al. [41] proposed the dual threshold method to segment the video into segments and extract key frames from each segment. SIFT features are extracted from the key frames of the segments. Then, SVD-based method is proposed to match two video frames with SIFT point set descriptors.

5 Conclusion

This paper presents a short survey on video segmentation. Video segmentation is applied to video indexing and video retrieval system. MPEG-compressed videos are used by the researchers for video segmentation as the compressed feature of macroblock coding of B-frames intro-type, forward-type, backward-type, and bidirectional-type is used directly without total decompression for MPEG bit-stream [36]. Motion compensation and motion vector are effectively used in MPEG bit-streams.

Video segmentation using *pattern recognition*, *SVM,* and *foveated shot detection* is rarely used in video segmentation. This can be used for effective object-based or content-based video segmentation. *Foveation mechanisms* [33] have never been taken into account for video segmentation, while there are some recent applications to video compression.

References

1. Huang, C.L., Liao, B.Y.: A robust scene-change detection method for video segmentation. IEEE Trans. Circ. Syst. Video Technol. **11**(12), 1281–1288 (2001)
2. Hampapur, A., Jain, R., Weymout, T.E.: Production model based digital video segmentation. Multimedia Tools Appl. **1**(1), 9–46 (1995)
3. Hauptman, A.G., Smith, M.A.: Text, speech and vision for video segmentation: The *Informedia TM* Project. In: Proceeding of AAAI Fall Symposium Computational Models for Integrating Language and Vision, Boston (1995)
4. Boreczky, J.S., Wilcox, L.D.: A Hidden Markov model framework for video segmentation using audio and image features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (1998)
5. Huang, J., Liu, Z., Wang, Y.: Integration of audio and visual information for content-based video segmentation. In: International Conference on Image Processing, ICPI, vol. 3, pp. 526–529 (1998)
6. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory **21**(1), 32–40 (1975)
7. Black, M.J.: Combining intensity and motion for incremental segmentation and tracking over long image sequences. ECCV (1992)

8. Arman, F., Hsu, A., Chiu, M-Y.: Image processing on compressed data for large video databases. In: Proceedings of ACM Multimedia, pp. 267–272 (1993)
9. Alattar, A.M.: Detecting and compressing dissolve regions in video sequences with DVI multimedia image compression algorithm. ISCAS, pp. 13–16 (1993)
10. Zhang, H.J., Kankanhalli, A., Smoliar, S.W.: Automatic partitioning of full-motion video. Multimedia Syst. **1**(1), 10–28 (1993)
11. Hampapur, A., Jain, R., Weymouth, T.: Digital video segmentation. In: Processing of Multimedia, ACM (1994)
12. Tonomura, Y., Akutsu, A., Taniguchi, Y., Suzuki, G.: Structured video computing. IEEE Multimedia **1**(3), 34–43 (1994)
13. Meng, J., Juan, Y., Chang, S-F.: Scene change detection in a MPEG compressed video sequence. IS&T/SPIE Symposium Proceedings, vol. 2419, California (1995)
14. Wang, D.: Unsupervised video segmentation based on watersheds and temporal tracking. IEEE Trans. Circ. Syst. Video Technol. **8**(5), 539–546 (1998)
15. Salembier, P., Marques, F.: Region-based representations of image and video: segmentation tools for multimedia services. IEEE Trans. Circ. Syst. Video Technol. **9**(8) 1147–1169(1999)
16. Meier, T., Ngan, K.N.: Video segmentation for content-based coding. IEEE Trans. Circ. Syst. Video Technol. 9(8) 1190–1203(1999)
17. Altunbasak, Y.: A statistical approach to threshold selection in temporal video segmentation algorithms. In: IEEE International Conference on ASSP (2000)
18. Truong, B.T., Dorai,C., Venkatesh, S.: New enhancements to cut, fade, and dissolve detection processes in video segmentation. In: Proceedings of the eighth ACM international conference on Multimedia (2000)
19. Jadon, R.S., Chaudhury, S.K., Biswas, KK.: A fuzzy theoretic approach for video segmentation using syntactic features. Pattern Recogn. Lett. **22**(13), 1359–1369 (2001)
20. Lo, C.C., Wang, S.J.: Video segmentation using a histogram based fuzzy c-means clustering algorithm. Comput. Stand. Interfaces **23**(5), 429–438 (2001)
21. Xu, P., Xie, L., Chang, S.F., Divakaran, A., Vetro, A., Sun H.: Algorithms and system for segmentation and structure analysis in soccer video. In: IEEE International Conference on Multimedia and Expo (ICME) (2001)
22. Khan, S., Shah, M.: Object based segmentation of video using color, motion and spatial information. In: IEEE Proceeding on CVPR, vol. 2 (2001)
23. Patras, I., Hendriks, E.A., Lagendijk, R.L.: Video segmentation by MAP labeling of watershed segments. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 326–332 (2001)
24. DeMenthon, D.: Spatio-temporal segmentation of video by hierarchical mean shift analysis. Statistical Methods in Video Processing Workshop (2002)
25. Lu, Y., Gao, W., Wu, F.: Automatic video segmentation using a novel background model. In: IEEE International Symposium on Circuits and Systems (2002)
26. Sifakis, E., Grinias, I., Tziritas, G.: Video segmentation using fast marching and region growing algorithms. J. Appl. Sig. Process. **4**, 379–388 (2002)
27. Chien, S.Y., Huang, Y.W., Chen, L.G.: Predictive watershed: a fast watershed algorithm for video segmentation. IEEE Trans. Circ. Syst. Video Technol. **13**(5), 453–461 (2003)
28. Porter, S., Mirmehdi, M., Thosmas, B.: Temporal video segmentation and classification of edit effects. Image Vis. Comput. **21**(13), 1097–1106 (2003)
29. Guimaraes, S.J.F., Couprie, M., Araujo, A.D.A., Leite, N.J.: Video segmentation based on 2D image analysis. Pattern Recogn. Lett. **24**(7), 947–957 (2003)
30. Qi, Y., Hauptmann, A., Liu, T.: Supervised classification for video shot segmentation. In: Proceeding IEEE Conference Multimedia Expo, vol. 2, pp. 689–692 (2003)
31. Wang, J., Thiesson, B., Xu, Y., Cohen, M.: Image and Video Segmentation by Anisotropic Kernel Mean Shift. In: Computer Vision-ECCV, pp. 238–249. Springer, Berlin (2004)
32. Fang, H., Yin, Y., Norhashimah, P., Jiang, J.: A hybrid scheme for temporal video segmentation. In: Proceedings of the Third IEEE International Workshop on Electronic Design, Test and Applications, (2005)

33. Boccignone, G., Chianese, A., Moscato, V., Picariello, A.: Foveated shot detection for video segmentation. IEEE Trans. Circ. Syst. Video Technol. **15**(3), 365–377 (2005)
34. Apostoloff, N., Fitzgibbon, A.W.: Automatic video segmentation using spatiotemporal T-junctions. British Machine Vision Conference (2006)
35. Joyce, R.A., Liu, B.: Temporal segmentation of video using frame and histogram space. IEEE Trans. Multimedia **8**(1) 130–140 (2006)
36. Wang, X., Wang, S., Chen, H.: A fast algorithm for MPEG video segmentation based on macroblock. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (2007)
37. Tan, W., Teng, S., Zhang, W.: Research on video segmentation via active learning. Fourth International Conference on Image and Graphics (2007)
38. Pritch, Y., Rav-Acha, A., Peleg, S.: Non-chronological video synopsis and indexing. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1971–1984 (2008)
39. Tziakos, I., Cavallaro, A., Xu, L.Q.: Video event segmentation and visualisation in non-linear subspace. Pattern Recogn. Lett. **30**(2), 123–131 (2008)
40. Yin, P., Criminisi, A., Winn, J., Essa, I.: Bilayer segmentation of webcam videos using tree-based classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 30–42 (2011)
41. Hong, L., Lu, H., Xue, X.: A segmentation and graph-based video sequence matching method for video copy detection. IEEE Trans. Knowl. Data Eng. 1–1 (2012)