Principles of Video Segmentation Scenarios

M. R. KHAMMAR[1], YUNUSA ALI SAI'D[1], M. H. MARHABAN[1], F. ZOLFAGHARI[2],
[1]Electrical and Electronic Department, Faculty of Engineering
University Putra Malaysia, 43400 UPM Serdang, Selangor
[2]Computer Science Faculty, University of Sistan and Baluchestan, Zahedan, Iran
Khammar_m@yahoo.com

Abstract :

Video segmentation is the first step toward automatic video processing such as browsing, retrieval, and indexing. Many algorithms and techniques have been proposed a few years ago. They can cover the topic of video segmentation from different angles and it is beneficial to review the most important properties of them in brief in order to clarify the subject and find out the latest challenges and drawbacks. In this paper, the important parameters which are involved in video segmentation are discussed and video shot detection systems are compared together.

Key words:

video segmentation, shot detection, video processing, feature extraction**.**

1 Introduction:

Today audio and video media are the most important impact of the media on human societies, and of course, it includes the very large volume of information With the arrival of digital systems for producing, recording, and playback of multimedia information and also providing communication infrastructure for transfer high-volume data growth speed the media is dramatically increasing. According to statistics presented in 2010 about half of the data traffic on the Internet is related to the video information, Meanwhile, based on forecasts, in 2014, over 90% of global Internet network capacity to transmit video information will be designated[1].

Therefore, encounter with video signals is a part of human life that cannot be ignored, secondly, it has been a major problem subsequently formed and is important on how to deal with high volume of information. On the other hand, knowledge of search is an important parameter. Optimal mechanism carries out, a traditional method of search field according to the name of each film done. In this case, there is no real recognition regards to exact film content, so, the name is based on overall interpretation and the details of film sequence are not considered.

With advances Emerging in technology in the field of image processing (and video) towards classification, video searching based on content to have an understanding. From the perspective of time, video can be seen as a sequence of constituted blocks, this approach create a hierarchical structure and a video at different levels to form a long sequence of components.. The lower level is divided into smaller building blocks. Building blocks in the upper levels will have more time. Hierarchical structure of video sequences from the perspective of time, components, can be seen in Figure 1,Building blocks in this structure, respectively, from bottom to top are:

1) **Frame:** The smallest element is non-degradable video that each frame alone is equivalent to an image
2) **Shot:** A shot is a sequence of frames as they have been joined by a camera so that the camera profile (location, rotation angle, zoom, etc.) were constant or slightly altered
3) **Scene**: A set of consecutive shots are shown one-place or area, and spaces. Picture elements in a scene are constant over time, the scene usually describes a particular event or concept stage
4) **Clip:** A consecutive series of scenes linked together to tell a short story.
5) **Video:** The sequence of clips that are linked in terms of meaning and a general story to tell.

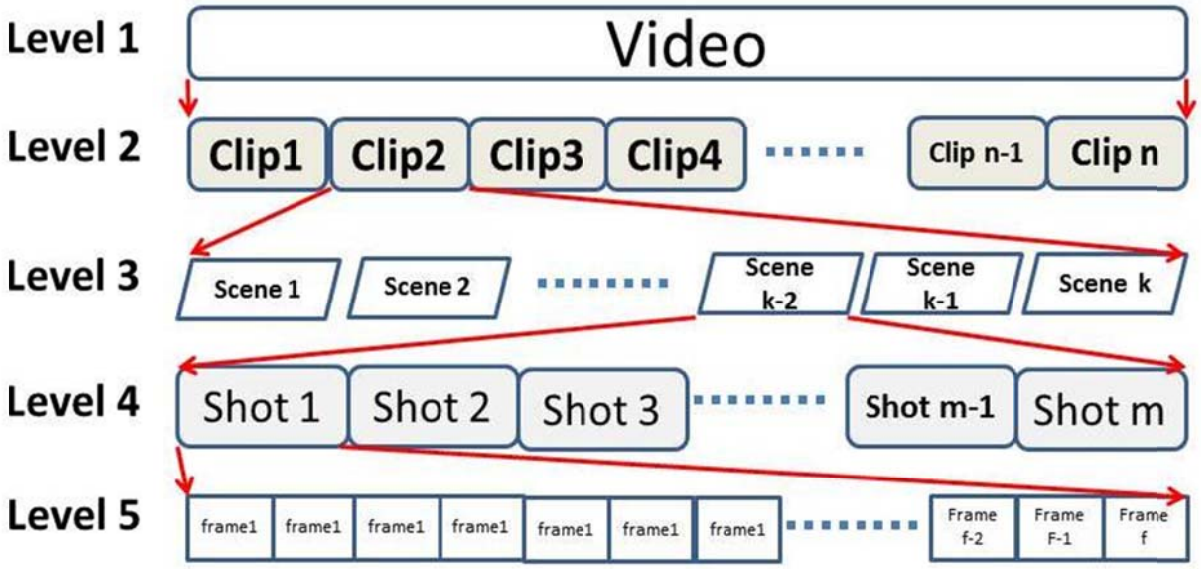


Figure 1: The video structure from the perspective of long sequence of building blocks

Search, tagging and content analysis of individual frames of a video dealing with very difficult and in many cases are unnecessary. Therefore, a preliminary processing step in video processing to understand and make decisions based on content video (video segmentation) is called a shot. Therefore, to determine the mechanism for detection of successive frames in a similar format or group. with proximity features that belong to two consecutive images is a shot [2]. In reviewing a shot detection system, several key questions can be posed as follows:

A. What characteristics of each frame should be considered from detection process parameters such as environment photo effect or physical contents in the scene is independent, that means selecting the properties of a frame should be done with sufficient wisdom.
B. In a comparison of two quantities adjacent frames to determine the similarity, minimal changes should be accepted or not. The threshold should be determined fixed or variable based on a function of statistical parameters or frames? How is the system response in each of these scenarios?
C. Shot changes are done by elimination or gradual manner. In order to go from one shot to another shot, transition exists or not? And in what conditions are they did gradually and in any case, how is the system?

Comprehensive study materials and techniques mentioned above require scanning algorithms presented in this paper, so the structure of shot detection system is presented and discussed in Section 2. The conclusion are shown in section 3.

2 Structure of a shot detection system:

The general flowchart of a shot detection system is presented in Figure 2. The first step in a given video file and all its frames are available in a finite chain.On each frame, we attempted to extract features. The question here is: What characteristics can be obtained from an image based on a review of published literature in this area? What is the fundamental point that a large number of features extracted for each frame as well as computational complexity and execution time can be greatly affected? so this should help in identifying the features that are distinctive [2].

a) **Luminance:** For a black & white frame pixel values associated with the allocation of bits for each pixel in a defined range, for example, if the allocation is 8 bits per pixel, pixel values between zero to two hundred and fifty-five will vary. The averaging of all pixel values for two adjacent frames in a shot that will bring the numbers closer together and the two adjacent frames of a large number will offer two different shots. The same procedure can extend to a color frame with the description that averaging for each frame must separate each component colors for R, G, B. And then based on three results of the survey, calculate the Euclidean distance to find out weather the current frame is belong to cureent shot or not.

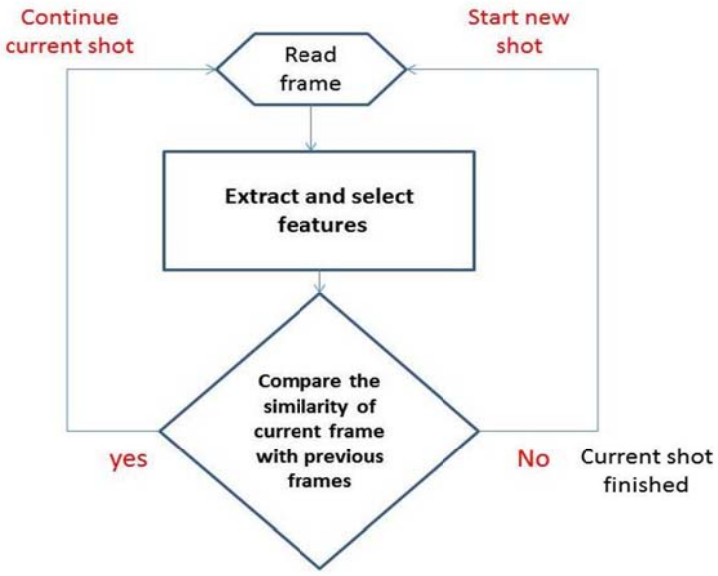
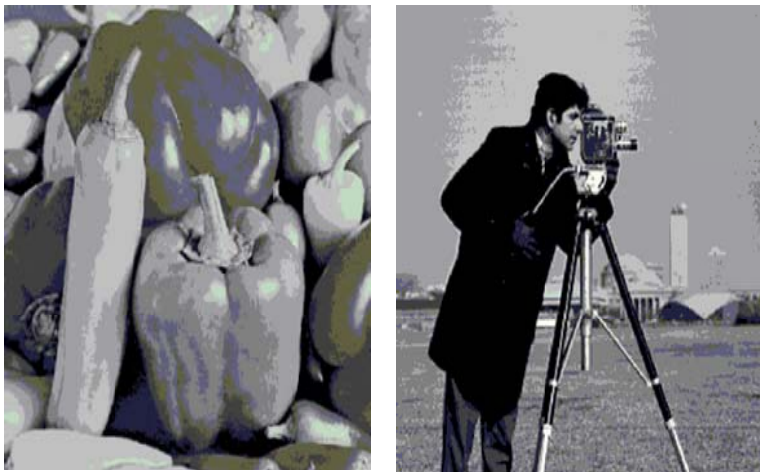


Figure 2: General flowchart of a shot detection system

In this method, local variation of pixel values in the two-dimensional spatial is neglected, and it can bring the same situation that despite having similar mean but two frames are different. Figure 3 presents an example of that. Despite the small difference between the mean values, but the contents are quite different from the picture, so this feature cannot be considered as an efficient model for a shot detection.



r left

Sometimes it is better for averaging color frames not within pixel values in R, G, B, matrix but within matrix values such as H, S, I have done , because its sensitivity to light changes of environment is more less than and the results accuracy is correct. Anyway, luminance and color of each single pixel can consider as a feature for shot detection algorithms. This feature has been used in [3], [4].

b) **Luminance/color-histogram:** Two images presented in Figure 3, despite having an overall average of almost equal, but they have different histogram , Figure 4 shows histograms of two images

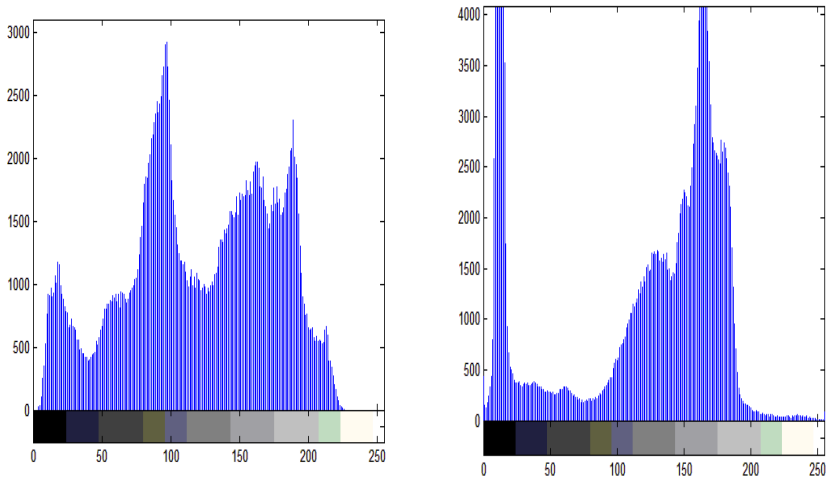


Figure 4: Histogram of two black and white images with same dimensions (fruit at the left and cameraman at the right)

In this method the histogram of two adjacent frames for intensity levels in a black & white image or color image are compared, and their similarity or differences is assessed. The method has the advantage that separate better shot implementation is provided and the simplicity of the method, with the sensitivity to the (translation) and (rotation) as well as zooming camera cuts from this method. This approach was applied by [6] and [7].

c) Edge detection

Edging is a suitable method for shot detection. The advantage of this method is that it has high independence than environment light changes or different of motion camera , and more importantly is that it is closer to the human visual system. References [5] and [8] have benefited from this approach in their works. The drawback of this method is high volume computing and its sensitivity to noise.

d) Transform coefficients

Implement DCT transform or wavelet and Fourier of all or a part of a frame of interest ROI (region of interest) can give the series of coefficients,. These coefficients can have good filters to measure the difference or similarity between frames to be used. The DCT in MPEG and wavelet in JPEG2000is are applied to image compression.

e) Motion

The movable parts of two adjacent frames are measured , the more motionless parts in two frames means its more similarity.

In general, this method gives the highest value for the situation of high movement and secondly, we can combine this method with each above mentioned and achieve better accuracy.

Considering extraction we are to know which section of frame for the study to be selected. So, for a small unit area of study will decrease the detection correct rate and likely two adjacent shots lie in the form of the one-shot and, on the other hand, Big unit area of study or region of interest (ROI) will increase calculation and spending processing time, although better accuracy well-behaved. We both found it to be significant, so many will be considered in this chapter.

1) Single pixel: In many algorithms shot detection will select each pixel as a feature such as photo-intensity or edge direction. Thus for both frames will Correspond to pixel value between two frames, the high value difference is not acceptable as a one shot. Therefore, in cases where this method combined with (motion estimation) can provide a good result.

2) Rectangular block: in this method each frame divided into non-overlap blocks and then from each block characteristic such as average of intensity values or color for comparison are considered, the method has advantages of independent of camera changes and suitability for detection..

3) Arbitrarity shaped region:

Extraction can be made in the arbitrarily shape region on each frame. In some cases it can cause it to have a characteristic that is distinct. The method also can have high computation load induce and, final response is intensive to depend on the region and selected shape.

4) Whole frame:

In this method, the whole frame will analyze. For example in histogram method, all parts of each frame of the project has been considered previously.

Comparable Assessment: In order to determine the similarity and dissimilarity of two given frames, after feature extraction , we need some standard metric to do the assessment. Such as MSE (mean square error), correlation , PSNR (peak signal to noise ration), and so on.

$$MSE = \frac{\sum_{xy}(f_{xy} - g_{xy})^2}{MN}$$

$$PSNR = 10 \log 10 \frac{255^2}{MSE}$$

$$NC = \frac{\sum_x \sum_y (f_{xy} - \mu_f)(g_{xy} - \mu_g)}{\sqrt{\sum_x \sum_y (f_{xy} - \mu_f)^2 \ \sum_x \sum_y (g_{xy} - \mu_g)^2}}$$

Threshold selection: As earlier discussed, transition from one shot to another shot occur at different procedure, therefore it is important that process of comparison of characteristics among which frames are done. A detection method of shot can be done to only compare between two adjacent frames, if the result will determine which group of shot it belong to. This method is effective for conditions that shots are completely with different content, and shot transition is abrupt but in the case of other types, will not produce good results.

To overcome this problem of comparing multiple frames together, we compared first frame of the shot with the current frame with a threshold t1 and the current frame and the previous frame with the threshold t2. Naturally threshold t2 is smaller than t1.

In fact , after comparison assessment, we need to establish a threshold that will be used for a frame, adjacent frame , array of frames or previous frames, this threshold can be selected using any of the establish relationships or selected fixed value.

1) S**tatic thresholdl:** In this case a fixed threshold is selected to compare frames and it is clear that the threshold for each video should be selected manually regards to its differnet contents.This method will bring better result if video content shows similar properties over the time [9].

Figure 5: shot transition a- abrupt or cut mode b- fade mode

2) **Adaptive threshold**: using a fixed threshold regardless of the frame's contentt is not a logical method and the results is not impressive, so in this case, we need a statistical act to select a threshold that will producee better efficiency. This method demonstrated in [10].
3) **Probabilistic detection**: This method is based on the existing pattern of work, preferably on the shot and the extracted based on the assumed probability distribution model to estimate and evaluate features in the frames. This approach was applied by [3], [4].
4) **Trained classifier**: this method works based on the definition of a clustering approach. In this manner for each frame, it is necessary to determine just two difference cluster, shot change (it means a new shot start), or no shot change (the continuation of the previous shot). This method can be achieved through the implementation of an artificial neural network [11].

Shot transition: Another important section is the shots structural point of view, which shows how different situations may occur and appear as described below:

1) cut mode: in this case two-shot in this situation is quite different and represent different scenes , so last frame of the current shot and the first frame of next shot is completely different. (Fig. 5 –a)
2) dissolve mode: In this approach, next shot frames appear in the final frames shot so with time and approaching the end of the shot, step by step now reduced the amount of pixel frames shot (fade out) and the amount of pixels of frames the new shot is increased (fade in), this continues in similar pattern. (Fig. 6-a)

Figure 6: shot transition a- dissolve mode b- wipe mode

3) **fade mode**: In this mode at the end of a shot, we have an empty frame and the next shot will load after that. (Fig 5 –b)
4) **wipe mode**: In this case, shot changes start by the next shot frames with a majority present ,but just in small special region, so, the pixel values of new shot is replaced in other frames and it increase gradually until cover the whole frame. In this case the pixel values of last shot will disappeare and they replace by new pixels of new frames. (Fig 6 –b)

Shot detection assessment: The last point of the process to determine how the shot was fractioned assessment. This section is applicable by two basic parameters. *Recall, and precision.* We're looking at a video scan system to let users specify a video shot. Ideally, the algorithm can be used to count the number of all shots correctly but two problems in this regards may be occur and while working on a video may be emerging in practice. The first one is that the algorithm is not able to detect the whole shots and therefore there are some absent shots in the final result.

$$Recall = \frac{D}{D + D_M}$$

D_M: Total shots that has not been identified by the algorithm, D: number of diagnosed shots. *Recall* value is ideal for a variety of numerical algorithms is close to one that provides the possibility to compare algorithms. The second important issue is that algorithms can be identify some shots that is a failure.

$$Precision = \frac{D}{D + D_F}$$

D_F: virtual numbers are detected. If the value of Precision be one it means that there is no failure in shot detection and all detection shots are real shots.

Brief comparison: A brief comparison between some common shot detection methods are shown in table 1.

Table 1: Compare Some Popular Algorithms

Method	Advantages	Disadvantages
Pixel-comparison	Simple, easy to implement	Computationally heavy , very sensitive to moving object or camera motion
Block based	Performs better than pixel	Cannot identify dissolve, fade, fast moving objects
Histogram comparison	Performance is better, detect cut, fade, wipe and dissolve	Fails if the two successive shots have same histogram. Cannot distinguish fast object or camera motion
Edge change ratios	Detect cut, fade , wipe and dissolve	Computationally heavy, fails when there is a large amount of motion

3 Conclusion:

Video shot detection is the first step toward semantic analysis. In this paper, the most important parameters related to shot detection algorithms and techniques are reviewed and clarify in order to present a good insight on the subject. In case of fast object motion or camera motion and also fast illumination changes still remain the challenges.

REFERENCES

[1] Index, C.V.N., Forecast and Methodology, 2009-2014. White paper, CISCO, June. 2.
[2] Cotsaces, C., N. Nikolaidis, and I. Pitas, Video shot detection and condensed representation. A review. Signal Processing Magazine, IEEE, 2006. 23 (2): p. 28-37.
[3] Lelescu, D. And D. Schonfeld, Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. Multimedia, IEEE Transactions on, 2003. 5 (1): p. 106-117.
[4] Hanjalic, A., Shot-boundary detection: unraveled and resolved? Circuits and Systems for Video Technology, IEEE Transactions on, 2002. 12 (2): p. 90-105.
[5] Nam, J. And A.H. Tewfik, Detection of gradual transitions in video sequences using B-spline interpolation. Multimedia, IEEE Transactions on, 2005. 7 (4): p. 667-679.
[6] Zhang, H.J., A. Kankanhalli, and S.W. Smoliar, Automatic partitioning of full-motion video. Multimedia systems, 1993. 1 (1): p. 10-28.
[7] Z. Cernekova, C. Kotropoulos, and I. Pitas, "Video shot segmentation using singular value decomposition," in Proc. 2003 IEEE Int. Conf. Multimedia and Expo, Baltimore, Maryland, July 2003, vol. 2, pp. 301–302.
[8] Zabih, R., J. Miller, and K. Mai, A feature-based algorithm for detecting and classifying production effects. Multimedia systems, 1999. 7 (2): p. 119-128.
[9] Cernekova, Z., C. Kotropoulos, and I. Pitas. Video shot segmentation using singular value decomposition. In Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. 2003: IEEE
[10] Yu, J. And M. Srinath, An efficient method for scene cut detection. Pattern Recognition Letters, 2001. 22 (13): p. 1379-1391.
[11] Lienhart, R. Reliable dissolves detection. InProc. SPIE. 2001.