Reducing False Positives in Video Shot Detection

Nithya Manickam
Computer Science & Engineering Department
Indian Institute of Technology, Bombay
Powai, India - 400076
mnitya@cse.iitb.ac.in

Sharat Chandran
Computer Science & Engineering Department
Indian Institute of Technology, Bombay
Powai, India - 400076
sharat@cse.iitb.ac.in

Abstract

Video has become an interactive medium of daily use today. However, the sheer volume of video makes it extremely difficult to browse and find required information. Organizing the video and locating required information effectively and efficiently presents a great challenge to the researchers. This demands a tool which would break down the video into smaller and manageable units called shots.

Traditional shot detection methods use histograms, or temporal slice analysis to detect hard-cuts and gradual transition for video. However, to our knowledge there is no system which is robust to sequences that contain illumination changes, camera effects, and other effects such as fire, explosion, and synthetic screen split manipulations. Traditional systems produce false positives for these cases; i.e., they claim a shot break when there is none.

We propose a shot detection system which reduces errors even if all the above effects are cumulatively *present in one sequence. The similarity between successive frames are computed by finding the correlation. Correlation sequence is analyzed using a wavelet transformation, which is used to locate the location of shot breaks. We achieve better accuracy in detecting hard-cuts when compared with other techniques.*

1. Introduction

In recent times, the demand for a tool for searching and browsing videos is growing noticeably. This has led to computer systems internally reorganizing the video into a hierarchical structure of frames, shots, scenes and story. A frame at the lowest level in the hierarchy, is the basic unit in a video, representing a still image. *Shot detection techniques* are used to group frames into shots. Thus, a shot designates a *contiguous sequence of video frames recorded by an uninterrupted camera operation*. A scene is a collection of shots which present different views of the same event and contain the same object of interest. A story is a collection of scenes that defines an unbroken event. Figure 1 illustrates this paradigm.

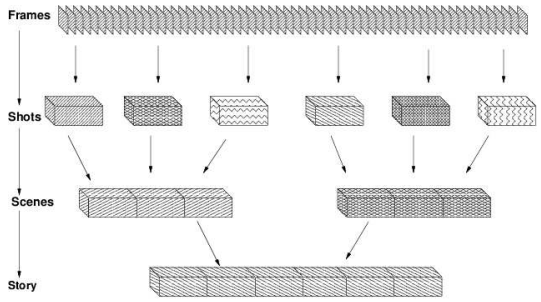

Figure 1. Hierarchical structure of video.

Video shot detection forms the first step in organizing video into a hierarchical structure. Intuitively, a shot captures the notion of a single semantic entity. A *shot break* signifies a transition from one shot to the subsequent one, and may be of many types (for example, fade, dissolve, wipe and hard (or immediate)). Our interest lies in improving shot break detection by *reducing the number of places erroneously declared* as shot breaks (false positives).

A wide range of approaches have been investigated for shot detection but the accuracies have remained low. The simplest method for shot detection is *pair-wise pixel similarity* [11], where the intensity or color values of corresponding pixels in successive frames are compared to detect shot-breaks. This method is very sensitive to object and camera movements and noise. A *block-based approach* [5, 6] divides each frame into a number of blocks that are compared against their counterparts in the next frame. Block based comparison is often more robust to small movements falsely declared as shot-break. Sensitivity to camera and object motion, is further reduced by *histogram comparison* [7, 2]. However, all these methods per-

Figure 2. A movie excerpt featuring Aishwarya Rai. Lightning creates unpredictable lighting changes.

Figure 3. Fast camera motion makes individual frames undecipharable.

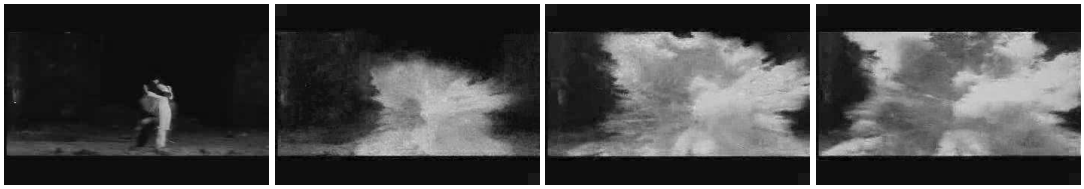

Figure 4. Explosion in a dimly lit scene causes considerable change in color and intensity.

Figure 5. Two different scenes are displayed simultaneously using split-screen methods. However, a shot break may be observed in only one of them.

form poorly when there are deliberate or inadvertent lighting variations.

At the cost of more processing, the *edge change ratio method* [10] handles slow transitions by looking for similar edges in the adjacent frames and their ratios. Three-dimensional *temporal-space methods* [3, 9] are better, but still sensitive to sudden changes in illumination. *Cue Video* [1] is a graph based approach, which uses a sampled three-dimensional RGB color histogram to measure the distance between pairs of contiguous frames. This method can handle special issues such as false positives from flash photography.

2. Problem Statement

As mentioned earlier, our main interest is in reducing false positives in challenging situations enumerated below.

1. *Illumination changes*: An example of this situation (inter-reflections, user-driven light changes, flash photography) is illustrated in Figure 2. In the movie excerpt, lighting causes the actress Aishwarya Rai to appear different. It is natural to the human, but confuses shot detection algorithms and even the camera as seen in the third frame!

2. *Camera effects*: By this we include effects such as

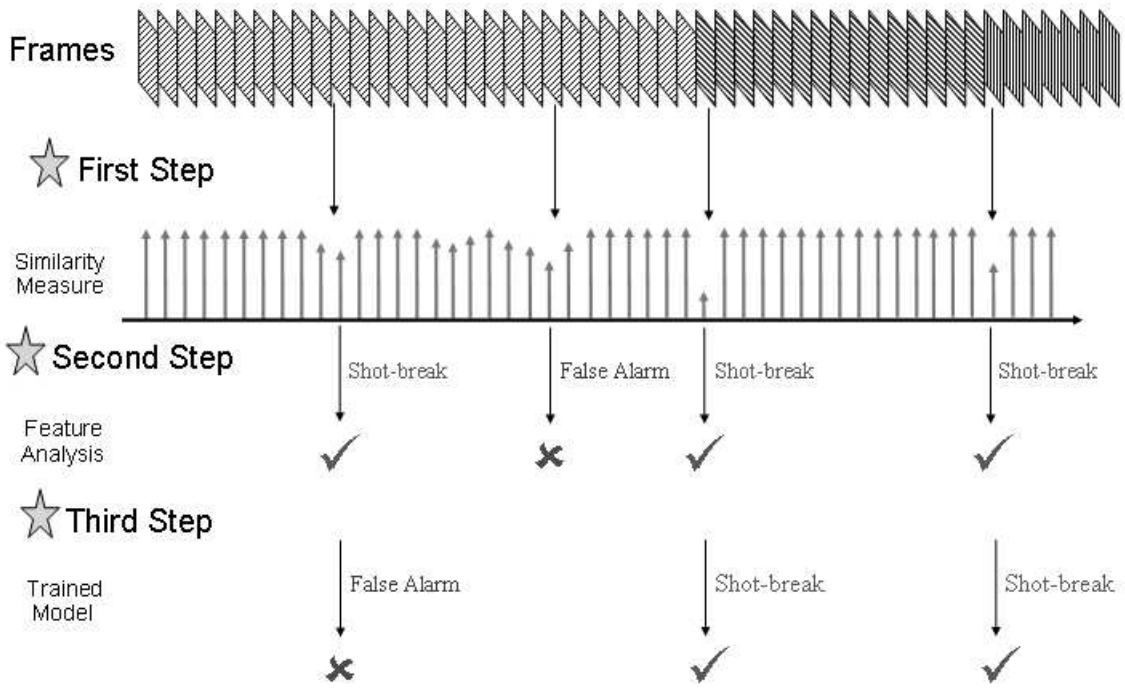

Figure 6. Our Apporach.

zooming and tilting of objects of interest, shaky handling of amateur video, fast object motion, and fast camera motion. An example is illustrated in Figure 3.

3. *Special effects*: An example of this situation (fire, explosion, screen split) is illustrated in Figure 4. Split screen is another possibility shown in the last figure.

3. Our Approach

We propose a shot detection system which reduces errors even if all the above effects are cumulatively present in one sequence. The similarity between successive frames are computed by finding intensity-compensated correlation using ideas similar to the ones in [8]. We depart, by further analyzing these similarities using wavelet methods to locate the shot breaks and reduce false positives by analyzing the frames around the predicted shot-breaks. The method is summarized in the figure 6 and can be broken into three steps.

1. Extracting features representing the similarity between the successive frames helps to determine candidate points for shot breaks. Candidate points for shot breaks are where similarity is low; four frames are indicated in the portion marked "First Step" in Figure 6. This is further elaborated in Section A.

2. Analyzing features to detect plausible shot breaks. As shown in Figure 6 (Second Step) the second predicted shot break is dropped because it is a false alarm. This is further elaborated in Section B.

3. We refining the detected shot breaks using more involved techniques further reducing false positive. In Figure 6 (Third Step) the first candiate is now dropped. This technique is elaborated in Section C.

3.1. Similarity Computation

The similarity between two consecutive frames is computed using a normalized mean centered correlation.

The correlation between two frames f and g is computed

as

$$\frac{\sum_i (f(i) - m_f)(g(i) - m_g)}{\sqrt{\sum_i (f(i) - m_f)^2} \sqrt{\sum_i (g(i) - m_g)^2}}$$

where m_f and m_g are the mean intensity values of frame f and g respectively. A high correlation signifies similar frames, probably belonging to the same shot; a low value is an indication of an ensuing shot break.

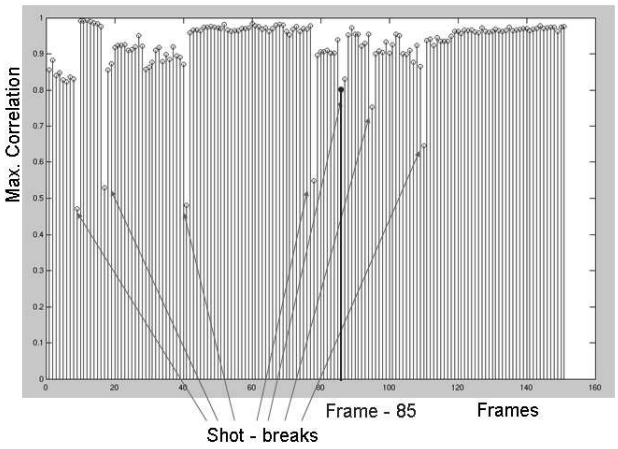

Figure 7. A sample correlation sequence. Low values might indicate shot breaks.

The maximum correlation values between successive frames are plotted as in Figure 7. The locations of shot breaks as identified by a human annotator are also indicated. From this diagram, it is also clear that placing an adhoc value as threshold to detect shot breaks will not work. A delicate shot break, like the one at frame 85 could be missed if a hard threshold is placed.

3.2. Shot Prediction

To overcome this difficulty, we consider the continuity of correlation values rather than the correlation values themselves, as an indicator of a shot. This is achieved using wavelet analysis. We have experimented with different wavelet transforms to detect this continuity and have observed that the Morlet wavelet results in a good discrimination between actual shot breaks and false positives.

The Morlet wavelet equation used in our computation is,

$$\psi(t) = Ce^{(\frac{-t^2}{2})} \cos(5t)$$

Morlet wavelet is a complex sine wave, localized with a Gaussian (bell shaped) envelope as shown in Figure 8.

As shown in the figure 8, there are equal number of positive and negative values in the mother wavelet and it sums

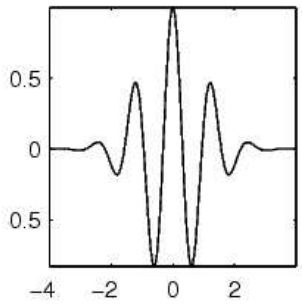

Figure 8. Morlet mother wavelet

to zero. Whenever there is no or little change in the correlation sequence, the wavelet transfrom returns zero value. If there is a hard cut, there is a discontinuity in the correlation value, which results in a distinctive PPNN pattern (two positive values followed by two negative values) in the lowest scale. At high scales the coefficient values are quite large. Hence hard cuts can be obtained by observing this pattern.

We graphically illustrate the power of the wavelet in Figure 9. The diagram shows a fluctuation in the correlation values from frames 215 up to 420. Out of these, frames 215 and 387 look like possible candidates for shot breaks. However, only frame 215 is an actual cut and frame 387 is a false positive (if reported as a cut).

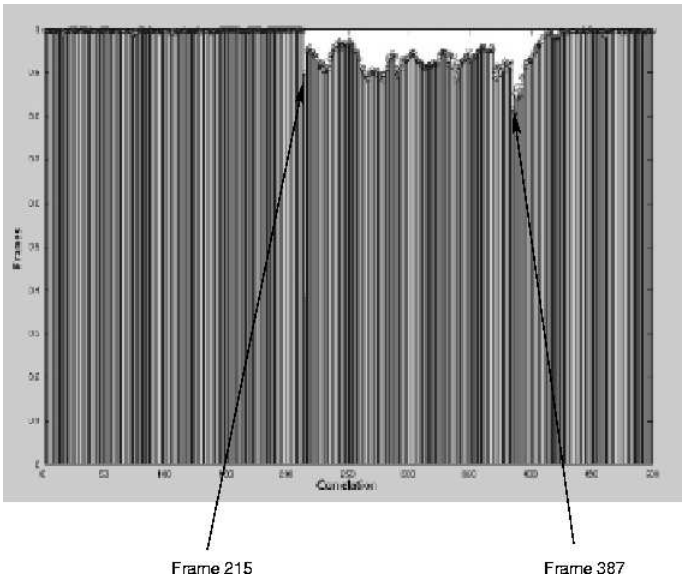

Figure 9. Sample correlation sequence

The corresponding Morlet wavelet transform in Figure **??**. The wavelet coefficients are high in all the scales

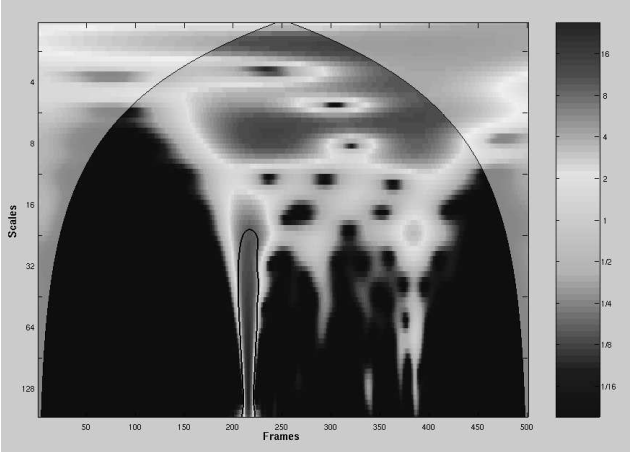

Figure 10. Morlet transform of the sequence shown in Figure 9.

around the frame 215, whereas the wavelet coefficients value around the frame 387 is not high at all the scales. Thus frame 215 is detected correctly as shot-break and frame 387 is dropped.

3.3. Reduction of False Positives

After detecting possible locations of shot breaks, we improving the prediction by analyzing the frames around predicted shot breaks in greater detail. Following measures are used for the same.

1. For the predicted frames, cross-correlation is computed by moving one frame over the other. It results in good correlation even in the case of fast motion frames (either due to camera or the object of interest). If cross-correlation is not done, we miss true positives.

2. Due to random lighting variations, the gray-scale value of successive frames in a shot might differ considerably. The false positives resulting from this are reduced by passing the frames through median filters and taking correlations.

3. We handle the low correlations resulting from sub shots by dividing the frame into four overlapping sub-frames and then taking the correlation of corresponding sub-frames. In case of sub-shots or in the case where text or object appears suddenly in a screen, one of these four correlation values might reflect the actual relation between the frames excluding the new object thereby such false positives are eliminated.

4. Results & Conclusion

Our system can process more than 30 frames per second with the accuracy required for the normal usage. We have tested our system on the data comprising of

- News videos each having around 500 hard cuts, containing different types of events. These are in multiple languages (notably Chinese and English).
- Short videos taken from motion pictures and from NASA. These involve some of the challenging problems mentioned in Section 2.
- Low-quality home video with varying lighting conditions and fast, shaky motion.

Table 1 shows the experimental results for various news channel videos containing problems like flash light, fast camera motion, shaky handling of camera, low quality of video. The ground truth for these experiments was generated manually with the help of about 20 research groups around the world [4]. As the results reflect, our system is successful in reducing the false positives considerably.

Methods	True Positive Ratio	False Positive Ratio
Pixel Comparison	0.8824	0.0040
Block Comparison	0.7059	0.0055
Histogram Comparison	0.8235	0.0088
Temporal Slice	1.0000	0.0031
Our Method	0.9706	0.000091

Table 2, shows the comparison between our system and existing shot-detection systems for a test video where we deliberately introduce a combination of all the challenging problems mentioned in Section 2.

Methods	True Positive Ratio	False Positive Ratio
Pixel Comparison	0.4667	0.0690
Block Comparison	0.7333	0.0460
Histogram Comparison	0.4667	0.1240
Temporal Slice	0.8000	0.0120
Our Method	1.0000	0.0020

In summary, our method considerably reduces false positives.

References

[1] B. A. et al. IBM research trec-2002 video retrieval system. *TREC Proc*, Nov. 2002.

[2] B. Funt and G. Finlayson. Color constant color indexing. *Pattern Analysis and Machine Intelligence, IEEE*, pages 522 – 529, May 1995.

[3] C. W. Ngo, T. C. Pong, and R. T. Chin. Detection of gradual transitions through temporal analysis. *Computer Vision and Pattern Recognition, IEEE Conference*, pages 36 – 41, June 1999.

[4] NIST. *TREC Video Retrieval Evaluation*. http://www-nlpir.nist.gov/projects/trecvid, 2005.

[5] S. Shahraray. Scene change detection and content-based sampling of video sequence. *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, pages 2 – 13, Feb. 1995.

[6] D. Swanberg, C. Shu, and R. Jain. Knowledge guided parsing in video database. *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, pages 13 – 24, May 1993.

[7] D. Swanberg, C. Shu, and R. Jain. Knowledge guided parsing in video database. *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, pages 13 – 24, 1993.

[8] T. Vlachos. Cut detection in video sequences using phase correlation. *SPLetters*, pages 173–175, July 2000.

[9] C. Yeo, Y.-W. Zhu, Q. Sun, and S.-F. Chang. A framework for sub-window shot detection. *Multimedia Modelling Conference, Proceedings of the 11th International*, pages 84–91, 2005.

[10] R. Zabih, J. Miller, and K. Mai. Feature-based algorithms for detecting and classifying scene breaks. *Third ACM Conference on Multimedia*, pages 189 – 200, Nov. 1995.

[11] H. Zhang, A. Kankanhalli, and S. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, pages 10 – 28, 1993.