Automatic Shot Boundary Detection Using Adaptive Thresholds

A. Miene, Th. Hermes, G. T. Ioannidis, and O. Herzog
TZI - Center for Computing Technologies, University of Bremen
Universitätsallee 21-23, D-28359 Bremen, Germany
{andrea|hermes|gtis|herzog}@tzi.de

Abstract

This paper describes the contribution of the TZI to the shot detection task of the TREC 2003 video analysis track (TRECVID).

The approach comprises a feature extraction step and a shot detection step. In the feature extraction, three features are extracted: a frequency-domain approach based on FFT-features, a spatial-domain approach based on changes in the image luminance values, and another spatial domain approach based on gray level histogram differences. Shot boundary detection uses then adaptive thresholds based on all extracted features of the complete video. The final shot list is a combination of shots which result from an independent examination of all three features.

1 Introduction

The Center for Computing Technologies (TZI), University of Bremen, Germany, participated in the video analysis track in the shot detection task.

Many approaches of shot boundary detection are proposed in literature. An overview is given in [Lienhart, 1999, Yusoff et al., 1998]. The principal methodology of shot boundary detection is to extract one or more features from every nth frame of a video sequence, to compute the difference of features for consecutive frames, and to compare these differences to a given threshold. Each time the threshold is exceeded, a shot boundary is detected. Cut transitions are detected more accurately than gradual (e.g. dissolve, fade out/in etc.) ones [Smeaton and Over, 2003]. Gradual transitions especially when dealing with low quality video material need also frame to frame comparisons at greater temporal distances [Adams et al., 2003]. Special modules for detecting photographic flashes [Quénot et al., 2003] are also important in the accurate detection of shot boundaries. The various approaches differ concerning the used features.

Based on experimental results, we selected three shot boundary detection approaches which were combined to an improved shot boundary detection methodology: a frequency-domain approach based on FFT-features, a spatial-domain

approach based on changes in the image luminance values, and another spatial domain approach based on gray level histogram differences. The approach is divided in two steps - feature extraction and shot boundary detection. In the first step, the features of the three different methods for the measurement of shot boundaries within the video are extracted. The second step detects the shot boundaries based on the previously extracted features. The advantage of this methodology is the possibility to set adaptive thresholds for the shot boundary detection considering all extracted features of the complete video sequence. The adaptive threshold is set to a percentage of the maximum of all calculated difference values of the video. In the case of gradual changes, often multiple shot boundaries are detected. Therefore multiple detected shot boundaries that follow each other within a short temporal interval are grouped together and a gradual change is detected beginning with the first and ending with the last shot boundary in the interval. The shot boundaries detected by examining the three features independently are then combined to a complete list of shot boundaries.

Section 2 describes the approach in more detail and section 3 presents the results. An outlook and future work is given in section 4.

2 Shot detection

The shot boundary detection system we used for TRECVID 2003 is based on the approach presented in [Miene et al., 2001]. As mentioned before, the approach can be divided into two main steps. The first step is to extract all needed features from a video. The second step is to detect the shot boundaries based on the previously extracted features. In the following, the steps from the feature extraction up to the shot list generation will be described in detail.

2.1 Feature Extraction

In this step all needed features for the shot boundary detection are extracted. We analyze an image sequence concerning the following features in the frequency and in the spatial domain;

2.1.1 Fast Fourier Transform (FFT) Feature Extraction

First, each frame is converted to a gray-scale image and then it is converted with the FFT into the frequency space [Vellaikal and Kuo, 1996]. Each frame then consists of a real- and an imaginary-part. R_{Sum} is calculated by adding values from the lower frequencies of the real-part and I_{Sum} by adding the appropriate values of the imaginary-part. In our implementation, we take 25 values for each part. Finally, the sum of the absolute differences of the real- and the imaginary-part for each consecutive frames is calculated:

$$F_{Total}(n, n-1) = |R_{Sum}(n) - R_{Sum}(n-1)| + |I_{Sum}(n) - I_{Sum}(n-1)|. \quad (1)$$

2.1.2 YUV Feature Extraction

Each frame is converted to the YUV 1:1:1 format and then all Y-values of a frame of size $w \cdot h$ are summed up:

$$Y_{Sum} = \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} Y(x, y). \tag{2}$$

Then the absolute differences

$$Y_{Diff}(n, n-1) = |Y_{Sum}(n) - Y_{Sum}(n-1)|. \tag{3}$$

of each two consecutive frames are calculated.

2.1.3 Gray Histogram x^2 Feature Extraction

For the feature extraction part, each frame is converted into a grayscale image. Then a histogram H_G is created. Subsequently, the squared differences between each two consecutive frames

$$H_{G_{Diff}}(n, n-1) = \sum_{i=0}^{255} \frac{(H_G(n)(i) - H_G(n-1)(i))^2}{Max(H_G(n)(i), H_G(n-1)(i))} \tag{4}$$

are calculated. $H_G(n)(i)$ denotes a grayscale histogram value at index i of frame n. $Max(H_G(n)(i), H_G(n-1)(i))$ denotes the maximum of both grayscale histogram values $H_{Gray}(n)(i)$ and $H_{Gray}(n-1)(i)$, and is used as a normalization factor.

2.2 Detection and Classification

The feature extraction step leads to three feature difference lists, one for each feature. First, a shot boundary detection based on each feature is performed, i.e. each value of the feature difference list is compared to a threshold which can be indicated either explicitly or adaptive. To determine the adaptive threshold, the maximum of all calculated difference values of the actual video is calculated. The adaptive threshold for the actual video is specified as a percental value of the maximum:

$$Th = \frac{Max\{H_{G_{Diff}}(1, 0), \ldots, H_{G_{Diff}}(n, n-1)\} \cdot Th_{percentage}}{100} \tag{5}$$

2.2.1 Merging of Shot Boundaries

For gradual changes like dissolves or wipes, the shot boundary detection often detects more than one boundary per shot. Therefore, all shot boundaries which belong to the same shot have to be merged into one boundary. This step is illustrated in Figure 1. Shot boundaries are merged together if the temporal distance between their occurrences is less than a threshold. The minimal frame

number of the merged shot boundaries determines the start, and the maximum frame number determines the end of the gradual change. The merging of shot boundaries is executed for each of the three shot boundary lists.

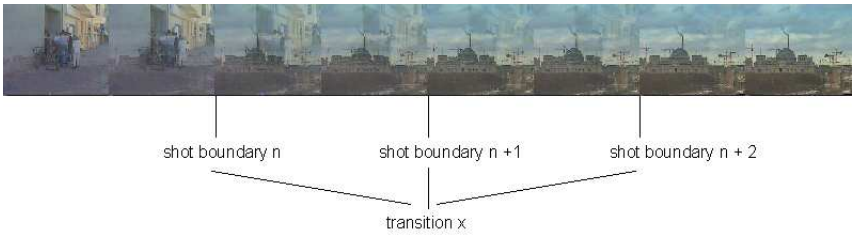

Figure 1: Merging of multiple detected shot boundaries [Miene et al., 2001].

2.2.2 Combination of Single Shot Boundary Lists

In this step, the three existing shot boundary lists are combined to one final shot list. Boundaries which overlap each other because more than one detection process has detected the same shot are joined to one boundary as shown in Figure 2.

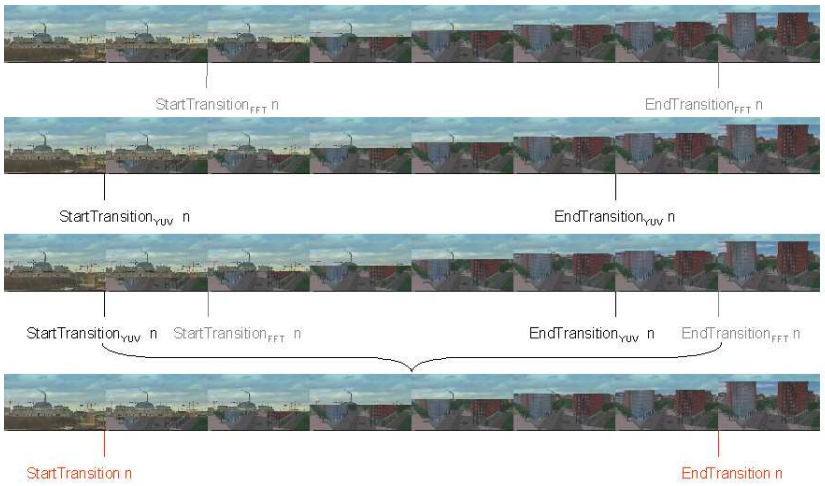

Figure 2: Combination of multiple detected shot boundaries [Miene et al., 2001].

All boundaries which are detected by more than one approach within a temporal threshold of 10 frames are joined to one boundary, whereas the minimum frame number determines the start and the maximum frame number the end of the shot.

2.2.3 Classification of Shot Boundaries

A boundary is classified as a hard cut, if

$$H_{G_{Diff}} > Th_{H_g} \vee F_{Total} > BreakThreshold_{FFT} \vee Y_{Diff} > BreakThreshold_{YUV} \quad (6)$$

All boundaries which are detected by the FFT and the YUV shot boundary detection and which are not classified as a hard cut are classified as wipe boundaries.

All remaining boundaries are classified as "unknown" and may be removed from the final shot list to increase the accuracy of the results.

3 Results

In order to evaluate our approach, we submitted 6 runs to TRECVID 2003. Table 1 lists the parameter settings for each run. For each method, a threshold for the feature difference is set adaptive, i.e. as a percental value relating to the maximal feature difference measure within the actual video. For the FFT and the YUV approach an additional threshold ($BreakThreshold_{FFT}$ and $BreakThreshold_{YUV}$) regarding the classification of detected cuts as hard cuts has to be specified (see sec. 2.2.3).

The parameter "maximal concentration delay" specifies the maximal temporal difference in frames up to which two shot are merged together.

Removal of transitions classified as "unknown" may be switched on or off.

The optional restriction "at least 2" claims that a shot boundary is integrated into the final shot list only if it appears in at least two of three single shot boundary lists. This restriction may be switched on or off.

The optional restriction "GH x^2 or FFT+YUV" claims that a shot boundary is integrated into the final shot list if it detected by the GH x^2 shot analysis or by both the FFT and the YUV shot analysis. This restriction may be switched on or off.

Table 2 lists the results measured by precision and recall for each of the 6 runs.

4 Future Work

In this paper, we presented the shot detection approach used at TRECVID 2003. This approach is a combination of three different methods using features in the frequency and in the spatial domain. The results show that our approach detects the so-called hard cuts with a good accuracy.

The ongoing work will now concentrate on the improvement of the detection of gradual changes. Furthermore, also our recall concerning the detection of hard cuts could be improved.

Another challenging problem is the overall improvement of the recall of detected shot boundaries and the decrease of false detected ones.

Parameters	run 1	run 2	run 3	run 4	run 5	run 6
FFT						
threshold %	9	9	9	3	9	9
max. conc. del.	25	25	25	25	35	45
break th. %	25	25	25	5	25	10
YUV						
threshold %	1.94	1.94	1.94	1	1.94	1
max. conc. del.	25	25	25	25	35	45
break th. %	11.25	11.25	11.25	10	11.25	15
GH x^2						
threshold %	13	13	13	13	13	8
max. conc. del.	25	25	25	25	35	15
Other						
Remove unknown	off	off	off	on	on	off
At least 2	off	on	off	on	off	on
GH x^2 or FFT+YUV	off	off	on	off	off	off

Table 1: Parameter settings.

References

[Adams et al., 2003] Adams, B., Iyengar, G., Neti, C., Nock, H., Amir, A., Permuter, H., Srinivasan, S., Dorai, C., Jaimes, A., Lang, C., Lin, C.-Y., Natsev, A., Naphade, M., Smith, J., Tseng, B., Ghosal, S., Singh, R., Ashwin, T., and Zhang, D. (2003). Ibm research trec 2002 video retrieval system. In Voorhees, E. and Buckland, L., editors, *Information Technology: The Eleventh Text Retrieval Conference, TREC 2002*, NIST Special Publication 500-251, pages 289–298.

[Lienhart, 1999] Lienhart, R. (1999). Comparison of Automatic Shot Boundary Detection Algorithms. In *Proc. SPIE Vol. 3656 Storage and Retrieval for Image and Video Databases VII*, pages 290–301, San Jose, CA, USA.

[Miene et al., 2001] Miene, A., Dammeyer, A., Hermes, T., and Herzog, O. (2001). Advanced and adapted shot boundary detection. In Fellner, D. W., Fuhr, N., and Witten, I., editors, *Proc. of ECDL WS Generalized Documents*, pages 39–43.

[Quénot et al., 2003] Quénot, G., Moraru, D., Besacier, L., and Mulhern, P. (2003). Clips at trec 11: Experiments in video retrieval. In Voorhees, E. and Buckland, L., editors, *Information Technology: The Eleventh Text Retrieval Conference, TREC 2002*, NIST Special Publication 500-251, pages 181–187.

[Smeaton and Over, 2003] Smeaton, A. and Over, P. (2003). Video track report. In Voorhees, E. and Buckland, L., editors, *Information Technology:*

	All		Cuts		Gradual			
Run	Recall	Prec.	Recall	Prec.	Recall	Prec.	Frame-R.	Frame-P.
1	0.581	0.744	0.700	0.878	0.292	0.395	0.394	0.462
2	0.483	0.792	0.590	0.947	0.223	0.386	0.410	0.468
3	0.522	0.777	0.631	0.921	0.258	0.403	0.389	0.475
4	0.399	0.662	0.449	0.970	0.279	0.295	0.546	0.305
5	0.472	0.731	0.557	0.899	0.266	0.375	0.436	0.335
6	0.322	0.707	0.354	0.942	0.242	0.375	0.576	0.264

Table 2: Evaluation results.

The Eleventh Text Retrieval Conference, TREC 2002, NIST Special Publication 500-251, pages 69–85.

[Vellaikal and Kuo, 1996] Vellaikal, A. and Kuo, C.-C. J. (1996). Joint spatial-spectral indexing for image retrieval. In *Proc. IEEE International Conference on Image Processing*, pages 867–870.

[Yusoff et al., 1998] Yusoff, Y., Christmas, W., and Kittler, J. (1998). A study on automatic shot change detection. *Lecture Notes in Computer Science*, 1425.