COMPARISON OF SHOT BOUNDARY DETECTORS

Jan Nesvadba [1], Fabian Ernst [1], Jernej Perhavc [1], Jenny Benois-Pineau [2], Laurent Primaux [2]
1 Philips Research, Eindhoven, The Netherlands, jan.nesvadba@philips.com, fabian.ernst@philips.com
2 Labri, Univ. of Bordeaux, France, jenny.benois@labri.fr, primaux@labri.fr

ABSTRACT

A video Cut Detector (CD), a member of the Shot Boundary Detector (SBD) group, is an essential element for spatio-temporal audiovisual (AV) segmentation and various video-processing technologies. Platform, processing and performance constraints forced the development of various dedicated CDs. Future platforms allow the usage of advanced CD algorithms with higher reliability. In order to enable an appropriate trade-off decision to be made between reliability and the required processing power, benchmarking of four CD algorithms has taken place on bases of a generic, culture-diverse multi-genre AV corpus. In terms of complexity / performance trade-off, a field-difference-based CD proved to be optimal.

1. INTRODUCTION

Nowadays, terabytes of storage capacity on Consumer Electronic (CE) In-Home networks no longer belong to the realm of fiction. Consequently, users of such networks are confronted with a content management problem e.g. to retrieve desired AV content within the network. This problem can partially be solved by means of content descriptors, so called metadata, which can be either provided by content producers or, alternatively, by Video Content Analysis (VCA) algorithms. Today, the latter reaches even semantically meaningful levels (e.g. mood interpretation) enabled by the available processing power of current CE platforms. Consequently, those solutions have reached such a level of complexity, that modularization of the components into so-called Service Units (SU) is both required and desirable for reasons of reusability [1][2][3].

In this paper we describe and benchmark four CD algorithms jointly developed in e.g. [4]. In general there are two SBDs, the so-called CD, identifying abrupt cut transitions, and the so-called Gradual Transition Detector (GTD), for gradual transitions such as dissolves and fades. In literature CD instances are also often referred to as shot boundaries (SB) or shot cuts separating AV content items into individual video shots, which is further used for e.g. AV Scene Boundary Detection (ScBD) [5].

Despite a rich variety of CDs reported by instance in TRECVideo [6] the problem still remains open [7]. A literature survey of relevant AV-content segmentation methods - including CDs - can be found in [8].

This paper is structured as follows: Section 2 describes an CD based on a MacroBlock (MB) correlation factor called Mean Absolute Difference (MAD). Section 3 presents a field-difference-based CD. Section 4 introduces a spatio-temporal frame-segmentation-based CD. Section 5 describes a rough-indexing-based CD. Section 6 summarizes the detection results and error function graphs. Final conclusions are drawn in Section 7.

2. MACROBLOCK CORRELATION CD (MBC CD)

State-of-the-art compression systems such as AV encoders contain among others a video compression block with Motion Estimator (ME) as further explained in [9][10]. The ME identifies the best matching MB of the current frame in the successor (or predecessor) frame by means of minimizing the MAD value [10], which can be seen as motion compensated MB inter-frame correlation factor. Consecutively, the total sum (further called MAD_{total}) of all MBs of all slices over the entire frame is normalized with the maximal achievable value

$$MAD\_max\_frame=(nr\_MB/slice)*nr\_slices*MAD\_max\_MB \quad (1)$$

with *MAD_max_MB* representing the maximal reachable MAD value, *nr_MB/slice* the number of MBs per slice and *nr_slices* the number of slices per frame. Consecutively, *Norm_MAD* can be calculated by

$$Norm\_MAD = \frac{MAD_{total}}{MAD\_\max\_frame} \quad (2).$$

Hereon, *Norm_MAD* of the current frame ($Norm\_MAD_n$) is compared to a mean-value-based adaptive threshold A_Th_n,

$$A\_Th_n = T * \frac{1}{2(W-2)}\left(\sum_{i=n-W}^{n-x} Norm\_MAD_i + \sum_{i=n+x}^{n+W} Norm\_MAD_i\right) \quad (3)$$

with T being a factor by which $Norm\_MAD_n$ has to minimally exceed the averaged *Norm_MAD*, with W representing the window length in number of frames, n being the index of the current frame investigated and x being an inner window. Instances, at which $Norm\_MAD_n$ exceeds A_Th_n, are indexed as SBs.

0-7803-9332-5/05/$20.00 ©2005 IEEE

3. FIELD DIFFERENCE CD (FD CD)

The FD CD calculates for each field in an interlaced video signal, an Inter-Field Dissimilarity (*IFD[n]*) of the current field (*n)* and the predecessor field *n-1*. The luminance signal *I(x,y,n)*, with the spatial coordinates *(x,y)* and the field index *n*, cannot be directly compared with the predecessor-field luminance value *I(x,y,n-1)*, at the same spatial position *(x,y)*, due to the different interlace phases of the two fields. Instead, *I(x,y,n)* is compared with de-interlaced luminance value $I_{dei}(x,y,n-1)$, which is equal to

$$median\ \left(I(x,y,n), I(x,y+1,n-1), I(x,y-1,n-1)\right) \quad (4),$$

using vertical temporal median as de-interlacing method. The resulting *IFD[n]* is defined as

$$IFD[n] = \frac{1}{N}\left|\{(x,y) \in P[n] : |I(x,y,n) - I_{dei}(x,y,n-1)| > T_{dis}\}\right| \quad (5)$$

with *P[n]* representing a pixel set with size *N*, containing the spatial positions in field with index *n*, and where T_{dis} is a preset threshold. I.e., the number of dissimilar pixels is counted. Finally, instances at which the local *IFD[n]* exceeds the maximum IFD value of the past *W* fields, increased by a preset threshold value *T*, as defined in

$$IFD[n] > IFD[n-m] + T, \forall m \in \{1,2,3,\cdots,W-1,W\} \quad (6),$$

are marked as a cut. The result is a field accurate CD.

4. COLOR SEGMENTATION CD (CS CD)

The MBC CD and FD CD compare frames on pixel level. Histogram-based approaches compare frames on frame level. The third CD resides on an intermediate level: it is based on color segmentation. Here, we use a watershed-like segmentation [11]. This is not intended to be object segmentation, as objects may have widely varying colors. As similar frames have similar segmentations, the dissimilarity of segmentations can be used in a CD. Figure 1 shows frame segmentations around an abrupt transition.

The similarity of consecutive segmentations is quantified by a *consistency measure*. It compares segment maps $S_{n-1}(x,y)$ and $S_n(x,y)$, where S_{n-1} and S_n are segment labels. $S_{n-1}(x,y)$ is motion compensated to handle object motion. We define an overlap matrix *A*, where A_{ij} is the number of pixels that were in segment *i* in the previous frame and are now in segment *j*:

$$A_{ij} = |\{(x,y) : S_{n-1}(x,y) = i \wedge S_n(x,y) = j\}| \quad (7).$$

For each segment in one frame, we define the segment in the other frame it *maps to* as that segment with which it has the most overlapping pixels: Segment *p* maps to segment *q* if $A_{pq} \geq A_{pr}$ for all *r*. We now define two consistency measures $C_{AND}(n)$ and $C_{OR}(n)$ as:

$$C_{AND} = \sum A_{pq} \forall \{p,q\} : A_{pq} \geq A_{pr} \forall r \wedge A_{pq} \geq A_{sq} \forall s \quad (8),$$

$$C_{OR} = \sum A_{pq} \forall \{p,q\} : A_{pq} \geq A_{pr} \forall r \vee A_{pq} \geq A_{sq} \forall s \quad (9).$$

Figure 1: Top : Frames around an abrupt transition.
Bottom : Corresponding segmentation.

C_{AND} measures the area of all segments which map bi-directionally onto each other, whereas C_{OR} measures the area of all unidirectional mappings. Notches in the consistency measures indicate a cut. C_{AND} proved to be the most accurate indicator of a cut. However, as any segmentation algorithm requires a threshold, image noise or texture may cause segments to be split up or merged in subsequent frames. This decreases C_{AND}, as for a split segment it only counts the area of the largest of the newly generated smaller segments. C_{OR} is insensitive to this effect as all smaller segments map uni-directionally to the big segment. Hence, we combine the two consistency measures:

$$\begin{aligned} \sigma(n) &= 1 + \left|C_{OR}(n) - mean\{C_{OR}(n-N),\ldots,C_{OR}(n-1)\}\right| \\ C(n) &= C_{AND}(n-1) + (C_{AND}(n) - C_{AND}(n-1)) \cdot \sigma(n) \end{aligned} \quad (10)$$

The notches in *C* representing transitions become more exposed in comparison to C_{AND}, but for notches caused by segment splitting, *C* remains more or less constant due to the insensitivity of C_{OR} to this particular cause. Especially for content with large textured areas, such as the grass in a football field or water surfaces such as the sea, the performance of the detector improves.

CS CD uses an adaptive threshold method [12] with a sliding window of size W+1 and checks for a cut transition in the middle of the window for each frame:

$$SC(i) = \begin{cases} 1 & \text{if} \quad C(i) < C(j)/T \ , \forall j \in \{i - \frac{W}{2}, \ldots, i + \frac{W}{2}\}, j \neq i \\ 0 & \text{else} \end{cases} \quad (11)$$

The parameter *T* determines the required depth of the notch. *W* should be smaller than the minimal duration between two consecutive cuts, as only one cut per window can be detected. An analysis of shot lengths of the corpus material [8] of Section 6 has shown that a window size of 10 frames does not lead to a large amount of missed transitions.

5. ROUGH INDEXING CD (RI CD)

The RI CD copes with cut- and gradual transitions. RI means using noisy and incomplete- "rough" -data for fast indexing of AV content. Such data can be extracted from MPEG streams when decoding, or collected during the MPEG encoding process. The method is based on two assumptions: presence of i) motion changes and ii) spatial content changes at cuts. The first one does not always hold in real content, but is realistic in MPEG encoded motion. RI CD consists of two cooperative processes running on an MPEG stream: change detection in P-frames and I-frames.

In P-frames, we suppose that the MB motion vectors $(dx_i, dy_i)^T$ follow a single affine motion model for the frame:

$$\begin{aligned} dx_i &= a_1 + a_2 x_i + a_3 y_i \\ dy_i &= a_4 + a_5 x_i + a_6 y_i \end{aligned} \qquad (12)$$

with (x_i,y_i) coordinates of the MB centers. Normalized absolute differences of estimated motion parameters for consecutive P-frames $\Delta^* a_n(t)$ and an absolute difference of the number of intra-coded MBs $\Delta Q(t)$ form a multiplicative mixture $D(t)$ used to detect a cut transition:

$$D(t) = (|\Delta Q(t)| + 1)^{\beta} \left(1 + \sum_{n=1}^{6} \Delta^* a_n(t)\right)^{1-\beta} \qquad (13)$$

with β between 0 and 1, and set to 0.8 by default. Supposing a Gaussian distribution $N(\mu,\sigma)$ of $D(t)$ inside each shot, we train it during the W first P-frames and compute a shot–adaptive detection threshold $\lambda = \mu(D,W) + T\sigma(D,W)$.

For I-frames, the change can be detected by matching of spatial content. To do this, we warp consecutive I-Frames by motion compensation with estimated models [13]. Here only rough low-resolution versions of images, the "DC frames" consisting of DC coefficients of DCT in MPEG I-frames, are used. After warping we use a mean squared error as a similarity measure weighted by the inverse of the energy of the local image gradient in order to reduce the contribution of errors on image contours $WMSE(k)$. An extended experimental study of $WMSE$ in TRECVideo experiments [13] allowed us to establish a threshold based on a histogram. If $WMSE(k) > \alpha\mu(WMSE, W)$(α=2,3,4,5 for the experiments in this paper relatively to T=1.5,1.8,2.2,2.5), then a SB is detected at I-frame $I(t_{k+1})$. In this paper SBD has been restricted to video cuts only, therefore detected changes had to be classified as "gradual" or "cut". At cut instances, $D(t)$ exhibits narrow peaks with consistency check of the sign of $\Delta Q(t)$ - from positive to negative.

6. COMPARISON OF CDs

All four CDs were tested on a corpus of 8 hours [8] of AV content captured from TV broadcast. The content was chosen carefully to adequately represent real-world broadcasting material containing a variety of genres, such as Series, Magazines, Commercials and Sports. The CD's performance was evaluated for each genre separately in terms of precision (related to false positives) and recall (related to false negatives) [6] to assess the CD's effectiveness according to the different video features of each genre. The performance of the CDs relies on two parameters: a threshold T, defining the (relative) difference of the sample with neighboring samples required to be indexed as a cut, and a window size W, defining the number of neighboring samples. Each detector was tested with several settings, where one of the parameters was fixed at an optimal default value and the other one was varied. Figure 2 displays the CD's performance on three of the tested genres. By selecting the adequate detector or settings, one can tune the recall and precision according to one's preferences.

We now compare the CDs on several criteria, which may indicate their suitability for specific situations.

Performance. All CDs reach comparable levels. The RI CD scores lower on recall (Figure 2a), whereas the MBC CD scores lower on precision (e.g. Figure 2c). The CS CD can reach high precision, but has a lower recall limit (e.g. Figure 2b). The FD CD is overall the most reliable. Only RI CD was tested on the TREC Video Corpus in the TRECVID2004 campaign. As it uses motion vectors and spatial information extracted from compressed streams, its performance depends on the accuracy of the encoder's motion compensation. On TRECVID2004, RI CD showed lower performance with recall of 86.8 and precision of 77.8, as the TRECVID Corpus was MPEG1 encoded.

Complexity. The FD CD has the lowest complexity, as for each pixel only a median and an absolute difference value have to be computed. The MBC CD computes an absolute difference as well, but it requires motion estimation. The RI CD has intermediate complexity, as it requires motion vectors and it makes a robust estimate of an affine motion model. The measure ΔQ requires only computations on MB resolution. The CSB CD exceeds all others in complexity, as it requires image segmentation.

Latency. Both the FD and RI CD have zero-frame latency, as the cut detection result is immediately available for the current frame, making them suitable for on-line detection. The other two CDs use a symmetric window surrounding the current frame, resulting in latencies of several frames.

Robustness. All CDs except the FD CD require motion estimation (ME). Depending on the amount of motion and the quality of the ME, ME errors may propagate into the CD result. As the FD CD does not require ME, it is the most robust one. Furthermore, it requires no parameters other than T and W . Through the use of segmentation, the CS CD is relatively robust to small errors in ME. The other two CDs are more critically relying on accurate ME.

Overall. Considering the properties of the CDs, the FD CD is the best general solution in terms of trade-off between complexity and performance.

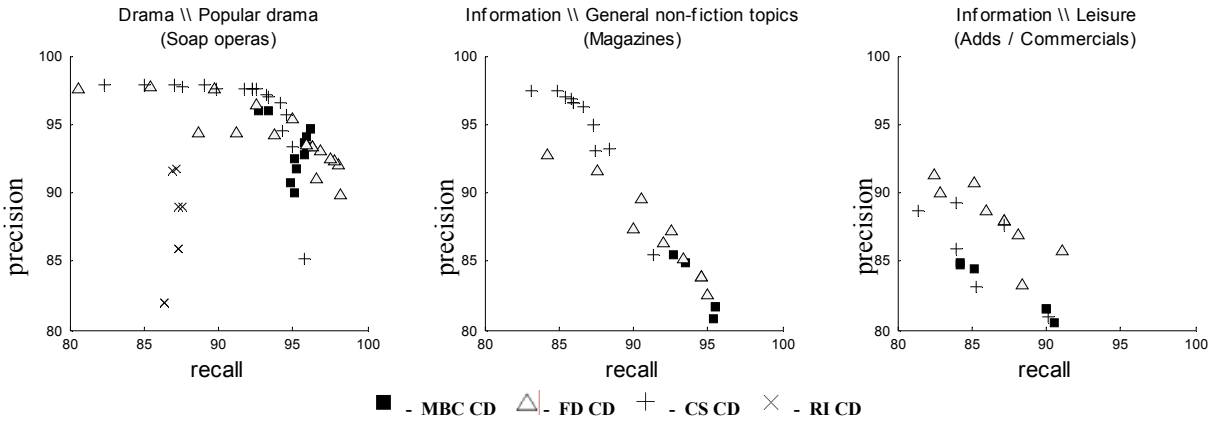

Figure 2: Precision / recall performance of all four CDs according to the genre of the analyzed video content

Depending on requirements (e.g. low latency, low complexity) and available side information (motion vectors, segmentation, etc.), one of the other CDs may be optimal for specific applications.

7. CONCLUSIONS

We have benchmarked four CDs in terms of precision and recall. All CDs are obtained as byproducts of other video processing operations such as MPEG encoding. The results on an AV corpus show that all CDs reach comparable levels for detecting video cuts. Differences across genres are more pronounced than differences within genres for different detectors. FD CD shows a good performance and has low complexity. Hence, the FD CD is in general preferred. The user can control the precision/recall trade-off through a combination of threshold T and window size W (tuneable precision). A topic for future research is a SBD for gradual transitions. Whereas the MBC CD and FD CD detectors compare subsequent frames, the CS CD (through segment tracking) and RI CD can handle larger inter-frame spacing, and as such may be more suitable for GTD.

Acknowledgement. We thank Ardjan Dommisse for the work on the FD CD.

8. REFERENCES

[1] J. Nesvadba, P. Fonseca, et al., "Face Related Features in Consumer Electronic (CE) device environments", Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, pp 641-648, The Hague, Netherlands, 2004.

[2] F. de Lange, J. Nesvadba, "A Networked Hardware/Software Framework for the Rapid Prototyping of Multimedia Analysis Systems", Proc. Int. Conf. on Web Information Systems and Technologies, Miami, USA, 2005.

[3] J. Nesvadba, et al., "Real-Time and Distributed AV Content Analysis System for Consumer Electronics Networks", Proc. IEEE Int. Conf. for Multimedia and Expo, Amsterdam, The Netherlands, 2005.

[4] Cassandra:www.research.philips.com/technologies/storage/cassandra/
MultimediaN:www.multimedian.nl/
Candela:www.hitech-projects.com/euprojects/candela/

[5] J. Nesvadba, et al., "Low-level cross-media statistical approach for semantic partitioning of audio-visual content in a home multimedia environment", Proc. IEEE Int. Workshop on Systems, Signals and Image Processing, pp. 235-238, Poznan, Poland, 2004.

[6] http://www-nlpir.nist.gov/projects/tv2004/tv2004.html

[7] A. Hanjalic, "SBD: Unraveled and Resolved?", IEEE Trans on CSVT, v.12, N2, pp. 90-104, 2002.

[8] J. Nesvadba et al., "CANDELA: Literature survey, D1.1B, State-of-the-art Report, Annex", www.hitech-projects.com/euprojects/candela/deliverables/candela-wp1-d11b-signoff.pdf

[9] N. Dimitrova, S. Jeannin, J. Nesvadba, et al., "Real-time commercial detection using MPEG features", Proc. 9th Int. Conf. on information processing and management of uncertainty in knowledge-based systems, pp. 481-486, Annecy, France, 2002.

[10] G. de Haan, et.al., "True-Motion Estimation with 3-D Recursive Search Block Matching", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 3, No. 5, pp. 368-379, 1993.

[11] F. Ernst, et. al., "Dense structure-from-motion: an approach based on segment matching", Proc. ECCV, LNCS 2531, pp. II-217-II-231, Springer, Copenhagen, 2002.

[12] B.-L. Yeo, B. Liu, "Rapid scene analysis on compressed video", IEEE Trans, Circuits Syst. Video Technology, vol. 5, pp. 533-544, 1995.

[13] L. Primaux, J. Benois-Pineau, et. al., "Shot Boundary Detection In The Framework of Rough Indexing Paradigm", TRECVID'04 Workshop, Gaithersburg, 2004. www-nlpir.nist.gov/projects/tvpubs/tvpapers04/ubordeaux.pdf