Video Temporal Segmentation Based on Color Histograms and Cross-Correlation

Anderson Carlos Sousa e Santos and Helio Pedrini(✉)

Institute of Computing, University of Campinas, Campinas, SP 13083-852, Brazil
helio@ic.unicamp.br

Abstract. Several fields of knowledge generate and consume massive volumes of videos, such as entertainment, telemedicine, surveillance and security. The rapid growth in the demand for multimedia content has driven the development of fast and scalable mechanisms for storing, retrieving and transmitting video sequences. The automatic temporal segmentation is a fundamental process in the analysis of video content. This work proposes and evaluates an adaptive video shot detection based on color histograms and normalized cross-correlation. Experiments conducted on several video sequences demonstrate that the combination of these two features achieve high accuracy rates.

Keywords: Multimedia content · Video transition · Shot detection · Temporal segmentation · Frame dissimilarities

1 Introduction

Advances in data acquisition technologies have enabled users to record and share videos through a number of portable devices, such as cell phones, tablets, and digital cameras. Due to this steady increase in multimedia contents, a challenging task is to develop efficient mechanisms for storing, indexing, retrieving and transmitting such large amounts of data.

Video summarization [3] consists in automatically generating a short version of a video sequence, allowing the user to quickly evaluate the relevance of its content by means of only a set of representative frames. As a temporal video segmentation process [4,6], some challenges associated with the video summarization include camera motion, varying lighting conditions, video genres, and subjectivity in the evaluation process.

The main contribution of this work is the proposition and evaluation of a video shot segmentation method based on the combination of inter-frame dissimilarity vectors of color histogram distances and block-based normalized cross-correlation between image pixel intensities. In addition, an adaptive local threshold strategy is defined to automatically detect the boundary frames. Experiments conducted on public video sequences demonstrate that the proposed method achieves high accuracy rates.

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 225–232, 2017.
DOI: 10.1007/978-3-319-52277-7_28

This paper is organized as follows. Section 2 briefly presents some relevant concepts and works related to the topic under investigation. Section 3 describes the proposed shot video detection methodology. Section 4 presents and discusses some of the results obtained with the proposed method. Finally, Sect. 5 concludes our work and includes some future work suggestions for improving the proposed method.

2 Background

Due to the advances in multimedia technology and large availability of digital content, there is an increasing demand for robust mechanisms for storing, indexing, browsing and retrieving video data. An open research problem is the automatic construction of a compact and meaningful representation of massive video sequences to help users understand the most important information of their content [9].

Temporal segmentation of a video into semantic units is a crucial stage in the analysis of video contents, whose process is known as shot boundary detection. A video shot consists of one or more frames generated contiguously to form a continuous action in time and space. A video summary can be constructed from a set of keyframes that represent the shots. In this context, two categories of transitions between shots are commonly defined: abrupt and gradual transitions. An abrupt transition corresponds to a cut between one frame of a shot and its adjacent frame in the next shot, whereas a gradual transition represents a smooth change over several frames.

Several video shot boundary detection approaches have been proposed in the literature [1,2,5,7]. Two main steps are commonly performed in the cut detection methods: (i) a similarity or dissimilarity measure is initially computed for each pair of consecutive frames and (ii) a cut is detected if the measure is higher than a specified threshold.

3 Methodology

The proposed video cut detection method is based on two different dissimilarities between consecutive frames: the Bhattacharyya distance between color histograms and the inverse normalized cross-correlation between the intensity image blocks. The resulting metrics are combined with a simple mean fusion and submitted to an adaptive thresholding technique that detects the relative high disparity and classifies the frames as part of a shot transition or not. These main steps are illustrated in Fig. 1.

3.1 Histogram-Based Dissimilarity

In order to calculate the inter-frame dissimilarity, a quantized color histogram (CH) is extracted from each frame and the distance between two consecutive frames is calculated with the Bhattacharyya distance, as defined in Eq. 1.

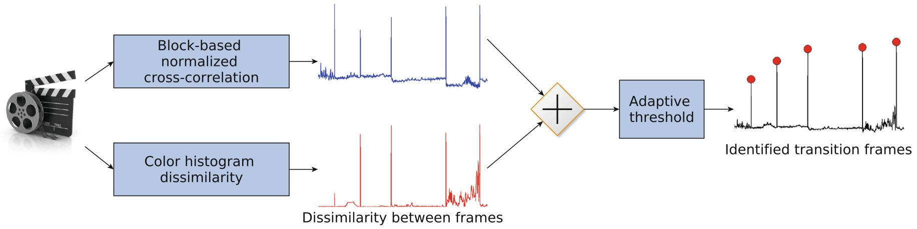

Block-based normalized cross-correlation
Color histogram dissimilarity
Dissimilarity between frames
Adaptive threshold
Identified transition frames

Fig. 1. A flowchart of the proposed video cut detection method.

$$d(H_i, H_{i-1}) = \sqrt{1 - \frac{1}{\sqrt{\overline{H}_i \cdot \overline{H}_{i-1} \cdot N^2}} \sum_{b}^{B} \sqrt{H_i(b) \cdot H_{i-1}(b)}} \tag{1}$$

where $\overline{H_k} = \frac{1}{N} \sum_j H_k(j)$ and $H_i(b)$ is the probability of frame i having a pixel that falls into the color bin b.

3.2 Block-Based Cross-Correlation

The negative normalized cross-correlation (NCC) is a dissimilarity measure over the intensity image, as stated in Eq. 2.

$$d(f_i, f_{i-1}) = -\frac{1}{N} \sum_{x,y} \frac{(f_i(x,y) - \overline{f}_i)(f_{i-1}(x,y) - \overline{f}_{i-1})}{\sigma_{f_i} \sigma_{f_{i-1}}} \tag{2}$$

where $\overline{f_i}$ is the average of f_i and σ_{f_i} is the standard deviation.

In order to avoid sensitivity to local changes between frames and presence of noise, each video frame is divided into non-overlapping blocks and the negative cross-correlation is calculated for each pair of corresponding blocks. Algorithm 1 summarizes the main steps for the block-based cross-correlation.

The block with the minimum dissimilarity is chosen since a significant change in it implies that all other blocks also changed.

3.3 Fusion

A combination of the dissimilarity vectors of the histogram-based distance and the block cross-correlation is performed to minimize the individual errors and uncertainty. Prior to the fusion process, a z-score normalization followed by a min-max scaling is applied to both vectors. The resulting dissimilarity vector constitutes a weighted mean between each position. Equation 3 summarizes the process.

$$D = \omega \cdot D_{CH} + (1 - \omega) \cdot D_{B\text{-}NCC} \tag{3}$$

Algorithm 1. Block-based cross-correlation dissimilarity

```
input  : video $V$, number of blocks $K$
output: dissimilarity vector $D$
$D \leftarrow \emptyset$
for $f_i \in V$ do
    divide the video frame into $K$ blocks
    $NCC \leftarrow \emptyset$
    for $k \in K$ do
        $NCC_k \leftarrow d(f_{k_i}, f_{k_{i-1}})$    // Equation 2
    $D \leftarrow \min(NCC)$
return $D$
```

where D_{CH} and $D_{B\text{-}NCC}$ are the dissimilarity vectors for the color histogram and the block-based cross-correlation, respectively, D is the final vector of dissimilarities between frames, whereas ω is the weight applied to each dissimilarity measure.

3.4 Adaptive Thresholding

The thresholding over the dissimilarity vector is performed locally through a moving window. Since the goal is to find *peaks* in the frame dissimilarities, this stage is similar to an outlier detection process.

A local median M is calculated for each moving window of size m with center at i. The frames i and $i-1$ are considered as boundary transition frames if their dissimilarity is equal to or greater than the median plus an α value ($d_i \geq M+\alpha$). Furthermore, it needs to be the maximum point within the window to ensure that only the dominant peak is labeled as transition, avoiding redundancy. Figure 2 illustrates the behavior of the proposed thresholding method.

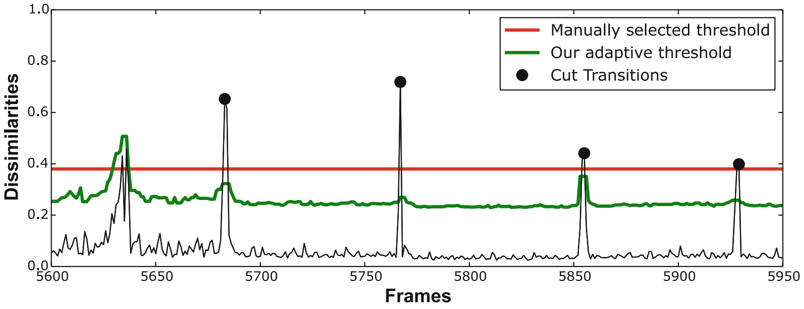


Fig. 2. Adaptive threshold over temporal dissimilarities.

4 Experimental Results

Experiments were conducted on two different annotated data sets. The first one, referred here to as VIDEOSEG'2004 [10], contains 10 video sequences within a diversity of genres, such as news, commercial, movies, cartoons, television shows, as well as other challenging scenarios with low quality digitization, low lighting conditions, fast motions and production effects. The second data set is a shot boundary test collection for the TRECVID'2002 [8]. It consists of 18 videos, where most of them are documentaries and amateur films with low quality, noise and production artifacts, varying in length, date of creation and production style.

The evaluation protocol follows the TRECVID guidelines, such that the results are assessed in terms of precision, recall and their harmonic mean (F_{score}). Equations 4 and 5 express the precision and recall measures, respectively, for a video V with a detection set S.

$$\text{Precision} = \frac{\displaystyle\sum_{f_i \in V} S(i) \in \mathit{Cut} \wedge i \in \mathit{True\ Cut}}{\sum_{f_i \in V} S(i) \in \mathit{Cut}} \tag{4}$$

$$\text{Recall} = \frac{\displaystyle\sum_{f_i \in V} S(i) \in \mathit{Cut} \wedge i \in \mathit{True\ Cut}}{\sum_{f_i \in V} i \in \mathit{True\ Cut}} \tag{5}$$

The F_{score} measure is defined as

$$F_{score} = 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

The adaptive thresholding parameters were empirically determined and applied to all videos in both data sets. In our experiments, values of $\alpha = 0.2$ and window size $m = 7$ achieved the best performance. For the histogram, a quantization with 32 bins for each RGB channel (totalizing 32,768 colors) was defined. The number of blocks K applied to the cross-correlation was set to 16, generating a 4×4 grid on the frames. For the fusion, a constant weight $\omega = 0.5$ demonstrated to be the best overall value in both data sets.

Table 1 shows the results for the VIDEOSEG'2004 data set. The described approaches and a baseline available for the data set [10] based on feature tracking were compared with the proposed fusion method.

It is noticeable that our fusion strategy for color histogram and block-based cross-correlation outperforms the respective individual methods and the provided baseline. The block-based normalized cross-correlation performs poorly in comparison to the global approach, once the videos in the VIDEOSEG'2004 data set have different frame dimensions and the block partitioning can discard important information. Nevertheless, such factors did not affect the performance of our fusion method.

Table 2 shows the results for the TRECVID'2002 data set. It is possible to observe that the proposed fusion outperforms all other approaches. Moreover,

Table 1. Video cut detection results (VIDEOSEG'2004).

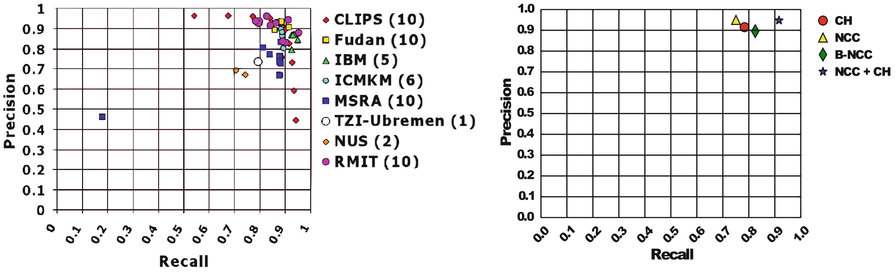
Method	Precision (%)	Recall (%)	F_{score} (%)
Color Histogram (CH)	84.66	85.71	84.94
Normalized Cross-Correlation (NCC)	88.94	90.96	88.72
Block-based NCC (B-NCC)	97.44	69.80	75.35
Feature tracking [10]	87.40	96.10	90.80
Proposed fusion (B-NCC + CH)	98.21	91.60	**94.17**

Table 2. Video cut detection results (TRECVID'2002).

Method	Precision (%)	Recall (%)	F_{score} (%)
Color Histogram (CH)	78.34	91.50	83.72
Normalized Cross-Correlation (NCC)	75.05	94.99	80.45
Block-based NCC (B-NCC)	82.46	89.60	85.08
Proposed fusion (B-NCC + CH)	91.60	94.76	**92.93**

the block-based cross-correlation achieves superior performance than the global cross-correlation.

Figure 3 presents the official results, provided by TRECVID'2002, for each participant in the competition. Number in parentheses represent the number of submissions for each team. Our results are presented in a similar manner to allow an adequate comparison.



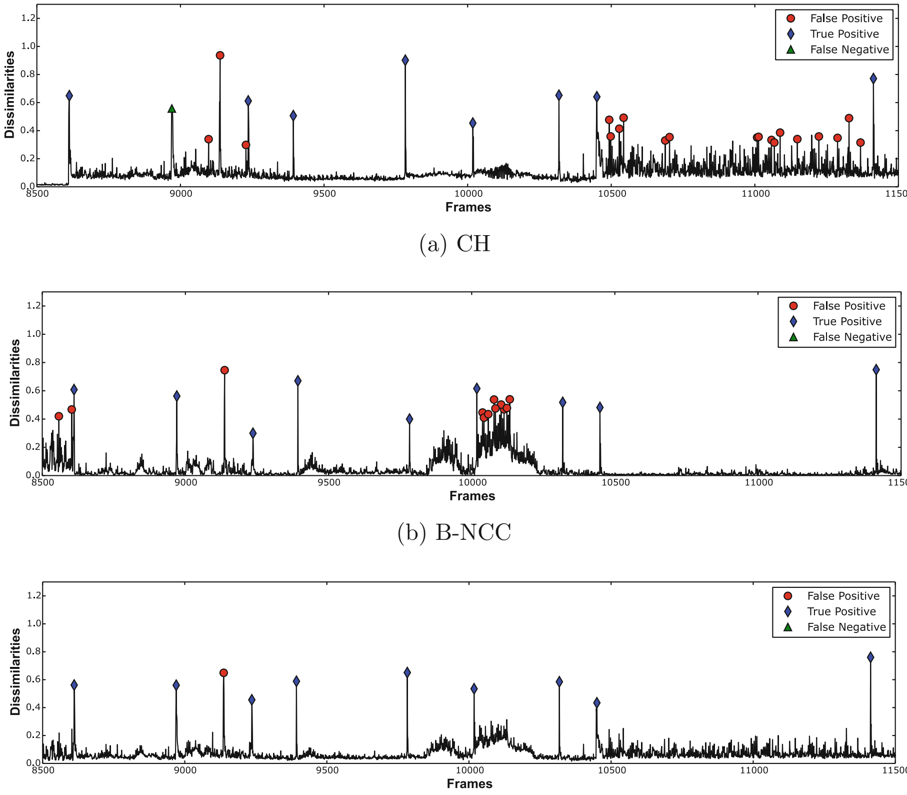
(a) TRECVID'2002 results

(b) Our results

Fig. 3. Precision/Recall performance for (a) participants in the TRECVID'2002 and (b) methods described in this work.

Our proposed method outperforms the majority of the submissions both in precision and recall. Furthermore, even the methods without combination are competitive to those submitted to TRECVID'2002. This corroborates the advantage of our adaptive thresholding method.

Figure 4 illustrates the inter-frame dissimilarities for a video section from TRECVID'2002. It is possible to observe that the number of false positives detected by the color histogram and block-based normalized cross-correlation are reduced through the fusion strategy.



(a) CH

(b) B-NCC

(c) B-NCC + CH

Fig. 4. Frame dissimilarities for a video section from TRECVID'2002.

5 Conclusions and Future Work

This work proposed an adaptive video cut detection method based on the combination of color histograms and cross-correlation. Furthermore, a local thresholding strategy is used to search for relative significant peaks.

Although the inter-frame dissimilarity measures are simple, both the fusion and adaptive thresholding approaches produced significant improvements in the experimental results obtained on two different data sets containing challenging videos. The proposed method also proved to be very competitive compared to the best submissions to TRECVID'2002 shot boundary competition.

As directions for future work, we intend to extend the method to address video gradual transitions and the automatic determination of weights in the fusion process.

Acknowledgments. The authors are thankful to FAPESP (grant #2011/22749-8) and CNPq (grant #307113/2012-4) for their financial support.

References

1. Apostolidis, E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6583–6587. IEEE (2014)
2. Birinci, M., Kiranyaz, S.: A perceptual scheme for fully automatic video shot boundary detection. Sig. Process.: Image Commun. **29**(3), 410–423 (2014)
3. Cirne, M.V.M., Pedrini, H.: Summarization of Videos by Image Quality Assessment. In: Shao, L., Shan, C., Luo, J., Etoh, M. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Advances in Pattern Recognition, pp. 901–908. Springer, Heidelberg (2014). doi:10.1007/978-1-84996-507-1_10
4. Jiang, H., Zhang, G., Wang, H., Bao, H.: Spatio-temporal video segmentation of static scenes and its applications. IEEE Trans. Multimedia **17**(1), 3–15 (2015)
5. Jiang, X., Sun, T., Liu, J., Chao, J., Zhang, W.: An adaptive video shot segmentation scheme based on dual-detection model. Neurocomputing **116**, 102–111 (2013)
6. Petersohn, C.: Temporal Video Segmentation. Jörg Vogt Verlag, Niederwaldstr (2010)
7. Tippaya, S., Sitjongsataporn, S., Tan, T., Chamnongthai, K., Khan, M.: Video shot boundary detection based on candidate segment selection and transition pattern analysis. In: IEEE International Conference on Digital Signal Processing, pp. 1025–1029, July 2015
8. TRECVID: TRECVID Data Availability (2016). http://trecvid.nist.gov/trecvid.data.html
9. Veltkamp, R., Burkhardt, H., Kriegel, H.P.: State-of-the-Art in Content-based Image and Video Retrieval, vol. 22. Springer Science & Business Media, Heidelberg (2013)
10. Whitehead, A., Bose, P., Laganiere, R.: Feature based cut detection with automatic threshold selection. In: Enser, P., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A.F., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 410–418. Springer, Heidelberg (2004). doi:10.1007/978-3-540-27814-6_49