# FINDING THE OPTIMAL TEMPORAL PARTITIONING OF VIDEO SEQUENCES

*Ba Tu Truong and Svetha Venkatesh*

Department of Computing
Curtin University of Technology
GPO Box U1987 Perth, Western Australia 6845

## ABSTRACT

The existing techniques for shot partitioning either process each shot boundary independently or proceed sequentially. The sequential process assumes the last shot boundary is correctly detected and utilizes the shot length distribution to adapt the threshold for detecting the next boundary. These techniques are only locally optimal and suffer from the strong assumption about the correct detection of the last boundary. Addressing these fundamental issues, in this paper, we aim to find the global optimal shot partitioning by utilizing Bayesian principles to model the probability of a particular video partition being the shot partition. A computationally efficient algorithm based on Dynamic Programming is then formulated. The experimental results on a large movie set show that our algorithm performs consistently better than the best adaptive-thresholding technique commonly used for the task.

## 1. INTRODUCTION

Video segmentation, often performed by detecting transitions occurring between shots in a digital video stream, is a fundamental process in automatic video analysis since it results in disjoint contiguous video segments that can serve as basic units to be indexed, annotated, and browsed. A shot in a video is defined as an unbroken sequence of images of a real or animated world captured between a camera's "record" and "stop" operations [1][1]. Shots are dominantly joined together in the editing stage of video (post) production with sharp cuts between them to form a complete story sequence and to provide a certain narrative structure to events portrayed. Occasionally, shots are joined by gradual visual effects such as fades, dissolves and wipes.

Although numerous techniques have been proposed for this fundamental problem in video analysis, they are often threshold-based and exhibit several shortcomings. Early solutions are poorly formulated and strongly rely on the empirical observations on a limited data set. Recent advanced techniques have incorporated prior knowledge about the video structure, i.e., shot length distribution, to adapt the detection threshold according to the time elapsed since the last shot change. However, they are only locally optimal, as they proceed sequentially, and rely on a strong assumption of the correct detection of the last shot change.

In addressing these shortcomings, we propose a technique for finding the best possible shot partition in a global sense that takes information from the entire video sequence in order to decide if a boundary occurs at a particular frame. This technique is derived from Bayesian principles and validated using a large data set.

## 2. BACKGROUND

The cut detection problem can be formulated as follows. Let $\mathcal{V} = \{f_1, f_2, ..., f_n\}$ represent the video sequence, where $f_i$ denotes the $i$-th frame in the sequence. Let $\mathcal{S}_i$ represent a segment of continuous frames frame $\mathcal{V}$, i.e., $\mathcal{S}_i = \{f_{s_i}, f_{s_i+1}, ..., f_{e_i}\}$. A partition $\mathsf{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_N\}$ is valid for $\mathcal{V}$ if and only if $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for every $(i, j)$, and $\mathcal{S}_1 \cup \mathcal{S}_2 ... \cup \mathcal{S}_N = \mathcal{V}$, which is equivalent to $s_1 = 1, e_N = n, e_i + 1 = s_{i+1}$ for all $i \in [1, N)$. For a valid partition $\mathsf{S}$ of $\mathcal{V}$ to be a shot segmentation, the following condition needs to be satisfied: $f_{i+1}$ is not the next frame to $f_i$ in a production shot (i.e., captured from one single camera from one single operation) if and only if $(i = e_k)$ for some $k$.

Most of current techniques concentrate on defining a discriminating function that differentiates between the set of frame pairs, usually successive, that belong to the same shots and those spanning across two shots, i.e., $(f_{e_i}, f_{s_{i+1}})$. All exploit the following properties:

- Frames surrounding a shot boundary generally display a significant change in visual content.
- Frames within a shot are very similar.

A discontinuity-based feature vector $z_i$ is devised and computed at every frame transition $(f_i, f_{i+1})$, which should form two distinctive clusters of inter-shot and intra-shot frame pairs. A discrimination function $\mathcal{F}$ is then learned to detect these clusters. In practice, $z_i$ is one dimensional and results from applying a distance function (e.g., histogram intersection) to two feature vectors (e.g., color histogram, edge his-

[1] This definition applies for a production shot. A shot in an edited video sequence is in fact a portion of a production shot, defined by two editing points.

0-7803-9332-5/05/$20.00 ©2005 IEEE

togram, motion vectors) extracted from $f_i, f_{i+1}$, and $\mathcal{F}$ is threshold-based. [2] describe several methods for adapting the threshold to the activity level in the scene.

**Statistical approach.** The shot boundary detection problem can be treated under the statistical framework as the problem of deciding between two hypotheses.

1. Hypothesis $\mathcal{H}_1$: Shot boundary present between two frames $k$ and $k+1$ $(\mathcal{S})$.

2. Hypothesis $\mathcal{H}_0$: No shot boundary present between two frames $k$ and $k+1$ $(\bar{\mathcal{S}})$.

Given the distribution $\mathsf{p}_{\mathcal{S}}(z)$ and $\mathsf{p}_{\bar{\mathcal{S}}}(z)$ of discontinuity values measured across two successive frames for the two hypotheses, the optimal threshold can be found using minimum-cost framework as shown in [3]. [4] use the prior knowledge about the shot length distribution to adjust the threshold based on how much time has elapsed since the previously claimed shot boundary. A similiar approach is adopted in [5], which incorporates a measure of the strength of the current frame being a local peak in $z$.

## 3. THE OPTIMAL SHOT PARTITIONING

### 3.1. Formulation

In this section, we outline a novel approach to the shot boundary detection problem that aims to detect a global optimal solution by searching for the most probable shot partitioning. This approach significantly departs from existing techniques, which either detect shot transitions sequentially or independently of one another. A valid partition $\mathsf{S}$ is treated as a random variable, consisting of the number of shots $N$ and the shot boundary locations $(s_i, e_i)$. The most probable shot partition is the one that produces the largest *posterior*:

$$\mathsf{S} = \arg\max_{\mathsf{S}} \mathsf{P}(\mathsf{S}|D) = \frac{\mathsf{P}(D|\mathsf{S})\mathsf{P}(\mathsf{S})}{\mathsf{P}(D)} \propto \mathsf{P}(D|\mathsf{S})\mathsf{P}(\mathsf{S}) \tag{1}$$

**Modelling $\mathsf{P}(\mathsf{S})$.** This quantity expresses the probability of an arbitrary partitioning being a shot partition, considering no evidence, except segment lengths $(l_i = e_i - s_i + 1)$, are observed or available. Under the assumption that all shots are independent of one another, we have:

$$\begin{aligned}\mathsf{P}(\mathsf{S}) &= \mathsf{P}(\mathcal{S}_1, \mathcal{S}_2, .., \mathcal{S}_N)\\ &= \mathsf{P}(s_1, s_2, s_N)\\ &\sim \mathsf{P}(e_1)\mathsf{P}(e_2|e_1)..\mathsf{P}(e_N|e_{N-1})\\ &= \prod_{i=1}^{N-1} \frac{\mathsf{p}^*(l_i)}{\mathsf{p}^*(0 < x \leq n - \sum_{j=1}^{i-1} l_j)}\end{aligned} \tag{2}$$

where $\mathsf{p}^*(.)$ denotes the pdf. of the shot length. This formulation means the shot length is normalized by its possible range and satisifies $\sum_{\mathsf{S}} \mathsf{P}(\mathsf{S}) = 1$ and can be adjusted to the case where there is a set of candidate boundary locations.

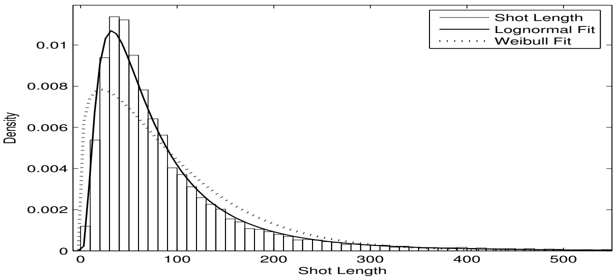

Figure 1: Modelling Shot Length Distribution

**Shot length distribution.** Previous work has modelled shot length using various standard distributions, notably Poisson [5], Weibull [4, 6], Erlang [4]. However, in our study of a large and diverse data set, we found that the shot length from an arbitrary movie is most appropriately modelled by a Lognormal or Loglogistic distribution. Figure 1 compares the Weibull fit and Lognormal fit produced by Matlab distribution fitting tool.

**Modelling $\mathsf{P}(D|\mathsf{S})$.** Let $\mathcal{N}(s, e)$ and $\mathcal{B}(f_t, f_{t+1})$ denote the event that no boundary is present in the segment $\{f_s, ..., f_e\}$ and that a boundary is present between frames $f_t, f_{t+1}$ respectively. After taking into account the shot length through $\mathsf{P}(\mathsf{S})$, we can consider these components as independent and $\mathsf{S}$ is given as: $\mathsf{S} = \mathcal{N}(s_1, e_1), \mathcal{B}(e_1, s_2), ..., \mathcal{B}(e_{N-1}, s_N), \mathcal{N}(s_N, e_N)$. In addition, the data for detecting a boundary is based on inter-shot discontinuity, but the data for detecting a segment with no boundary is based on intra-shot similarity, and thus they can be considered as independent components. Therefore, we have:

$$\begin{aligned}&\mathsf{P}(D|\mathsf{S}) = \mathsf{P}(D(s_1, e_1), D(e_1, s_2), ..., D(e_{N-1}, s_N), D(s_N, e_N)|\\ &\mathcal{N}(s_1, e_1), \mathcal{B}(e_1, s_2), \mathcal{N}(s_2, e_2), ..., \mathcal{B}(e_{N-1}, s_N), \mathcal{N}(s_N, e_N))\\ &= \mathsf{P}(D(s_1, e_1)|\mathcal{N}(s_1, e_1)).\mathsf{P}(D(e_1, s_2)|\mathcal{B}(e_1, s_2)) \ldots\\ &= \prod_{i=1}^{N} \mathsf{P}(D(s_i, e_i)|\mathcal{N}(s_i, e_i)) \prod_{i=1}^{N-1} \mathsf{P}(D(e_i, s_{i+1})|\mathcal{B}(e_i, s_{i+1}))\end{aligned} \tag{3}$$

**Computing discontinuity feature.** In order to take into account the spatial arrangement of color in measuring the difference $\mathcal{D}(.)$ between two frames, each frame is divided into four blocks and the sum of histogram difference of the four corresponding block pairs is computed. Unlike previous methods which measure using only two successive frames around $t$, we model $t$ as a function of frame discontinuity around $t$.

$$z_t^* = \min_{i \in (t-w, t-1), j \in (t, t+w-1)} \mathcal{D}(f_i, f_j), \tag{4}$$

where $w$ is the window size. This helps to eliminate some temporary noises in the discontinuity value within a shot. The window size should be chosen so that there is no shot with its length less than the window that is surrounded by two identical shots, otherwise these boundaries can be missed.

We consider $w = 15$ (at 25fps) as a safe choice for movie sequences; however, it can be much larger for other video genres.

In our implementation, $D(e_{i-1}, s_i)$ is represented by $z^*_{s_i}$, whilst $D(s_i, e_i)$ is represented by the maximum discontinuity value at every frame within the segment, and we call this intra-segment discontinuity. That is,

$$D(s_i, e_i) = \max_{s_i < t \leq e_i} z^*_t \qquad (5)$$

Figure 2 shows that the discontinuity values measured at shot boundaries in the training data set strongly follow a Gamma distribution, whilst intra-shot discontinuity can be modelled relatively well by a Lognormal distribution.

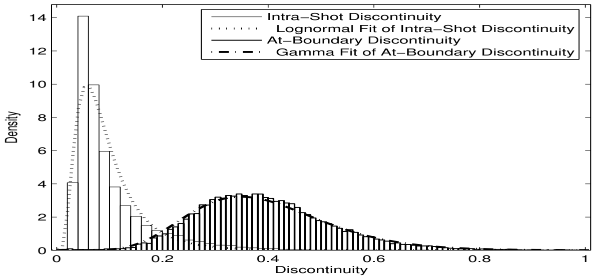

Figure 2: Intra-Shot & At-Boundary Discontinuity

### 3.2. The Search Algorithm

From Equations (1), (2) and (3), we have:

$$\begin{aligned}\log(\mathsf{P}(D|\mathsf{S})\mathsf{P}(\mathsf{S})) \propto \sum_{i=1}^{N-1} \mathsf{P}(D(e_i, s_{i+1})|\mathcal{B}(e_i, s_{i+1})) \\ + \sum_{i=1}^{N} (\log \mathsf{P}(D(s_i, e_i)|\mathcal{N}(s_i, e_i)) + \log \mathsf{P}(e_i|e_{i-1}))\end{aligned}, \qquad (6)$$

and we need to maximize the RHS to find the optimal solution. There are a total of $2^n$ valid partitions of $\mathcal{V}$, therefore the exhaustive-search for the optimal solution has an exponential complexity. However, if we consider $n$ frames as $n$ vertices of a directed left-to-right graph, where each node and edge has a weight assigned as follows:

$$\begin{aligned} w_n(f_i) &= \log \mathsf{P}(D(i, i+1)|\mathcal{B}(i, i+1)) \\ w_e(f_i, f_j) &= \log \mathsf{P}(D(i, j)|\mathcal{N}(i, j)) + \log \mathsf{P}(e_i|e_{i-1})\end{aligned},$$

then the problem of searching for the optimal paritioning is equivalent to the problem of searching for the longest (including the weights of middle vertices) left-to-right path from vertex $f_1$ to vertex $f_n$ in this graph, which can be solved by dynamic programming techniques (DPT). This is possible due to the property that if $\{f_{k_1}, ..., f_{k_t}\}$ ($k_1 = 1, k_t = n$) is the longest path from vertex $f_1$ to $f_n$ then $\{f_{k_1}, ..., f_{k_{t-1}}\}$ is the longest path from vertex $f_1$ to $f_{k_{t-1}}$.

Let $\mathbb{P}(f_i, f_j), \mathbb{L}(f_i, f_j)$ denote the longest path from vertex $f_i$ to $f_j$ and its length respectively. The following DPT procedure is guarantee to find the longest path from $f_1$ to $f_n$ with a complexity of $\varphi(n)$, where $O(\varphi(n)) = O(n^2)$. The procedure recursively computes the longest path from $f_1$ to $f_k$ utilizing information about the longest paths from $f_1$ to $f_2, ..., f_{k-1}$.

**Algorithm 3.1:** FINDTHELONGESTPATH($\mathcal{V}$)

```
local $\Delta, i, j$
$\mathbb{P}(f_1, f_1) \leftarrow \{f_1\}$
$\mathbb{L}(f_1, f_1) \leftarrow 0$

for $i \leftarrow 2$ to $n$
  do
    $\mathbb{P}^{(1)}(f_1, f_i) = \{f_1, f_i\}$
    $\mathbb{L}^{(1)}(f_1, f_i) = w_e(f_1, f_i)$

    for $j \leftarrow 2$ to $i - 1$
      do
        $\mathbb{P}^{(j)}(f_1, f_i) = \mathbb{P}(f_1, f_j) \cup \{f_i\}$
        $\mathbb{L}^{(j)}(f_1, f_i) = \mathbb{L}(f_1, f_j) + w_n(f_j) + w_e(f_j, f_i)$

    $\Delta = \arg\max_{j \in (1, i-1)} \mathbb{L}^{(j)}(f_1, f_i)$
    $\mathbb{L}(f_1, f_i) = \mathbb{L}^{(\Delta)}(f_1, f_i)$
    $\mathbb{P}(f_1, f_i) = \mathbb{P}^{(\Delta)}(f_1, f_i)$

return $(\mathbb{P}(f_1, f_n), \mathbb{L}(f_1, f_n))$
```

### 3.3. Search Space Reduction

The above DPT procedure has reduced the exponential complexity to $O(n^2)$. However, considering that a two-hour movie has 180,000 frames at 25fps, the number of partitions to be searched is still very large. Fortunately, there are some simple techniques that can further reduce the search space to a computationally manageable size.

**Breaking down the search range.** If we can identify correctly, via an independent method, that a shot boundary is present between two frames $(f_k, f_{k+1})$, the partitioning of the video sequence $\mathcal{V}$ can be done, with a potentially significant reduction in complexity, by applying the above DPT procedure to two sub-sequences $\mathcal{V}_1 = \{f_1, ..., f_k\}$ and $\mathcal{V}_2 = \{f_{k+1}, ..., f_n\}$ and merging the results.

The basic idea is to find a region $Z_{\mathcal{S}}$ in the frame discontinuity feature space $Z$ such that if we decide the hypothesis $\mathcal{H}_1$ for this region, the probability for false detection $\mathsf{P}_F$ is almost zero.

$$\mathsf{P}_F = \int_{Z_{\mathcal{S}}} \mathsf{p}(z|\bar{\mathcal{S}})dz \simeq 0. \qquad (7)$$

We observe that false shot boundaries produced by the conventional thresholding approach (see Section 1) are normally the result of object and/or camera movements, which by their nature continue for a number of frames. Thus, a shot boundary can be reliably identified between two frames $(f_i, f_{i+1})$ if it coincides with an isolated peak in the discontinuity feature curve, characterized by the following two conditions:

$$\begin{cases} z_i > \mathcal{T}_1 \\ z_j < \mathcal{T}_2, \forall j \in [i - \mathsf{W}, i + \mathsf{W}], j \neq i \end{cases}, \qquad (8)$$

where $\mathcal{T}_1, \mathcal{T}_2$ are two discontinuity thresholds and W is a window size that can be determined quickly from the training data.

Assuming this technique divides the video sequence $\mathcal{V}$ into $m$ sub-sequences with duration $n_1, .., n_m$, the complexity of the dynamic procedure is now $\sum_{i=1}^{m} \varphi(n_i)$ instead of $\varphi(n) = \varphi(\sum_{i=1}^{m} n_i)$.

**Reducing the number of search points.** With the DPT procedure described above, we still need to look through every frame location to find the optimal solution. Fortunately, most frames can be identified as non-boundary frames by a simple, computationally efficient method also based on the discontinuity measure. The idea is to find a region $Z_{\mathcal{S}}$ in the frame-discontinuity feature space $Z$ such that if we decide the hypothesis $\mathcal{H}_1$ for this region, the probability for miss detection $\mathsf{P}_M$ is almost zero.

$$\mathsf{P}_M = \int_{Z - Z_{\mathcal{S}}} \mathsf{p}(z|\mathcal{S}) dz \simeq 0. \tag{9}$$

By plotting values of two different discontinuity measures, $z^1$ and $z^2$, at boundary locations, we can easily identify a posible region, characterized by the following constraints:

$$\begin{cases} z_i^{(1)} > \mathcal{T}_3 \\ z_i^{(2)} > \mathcal{T}_4, \\ az_i^{(1)} + bz_i^{(2)} + c > 0 \end{cases} \tag{10}$$

The complexity of the DPT procedure when searching only on potential boundary locations are $\varphi(h)$ instead of $\varphi(n)$, where $h$ denotes the number of potential boundary locations extracted as above.

## 4. EXPERIMENTAL RESULTS

Our algorithm is tested on 12 full-length movies, consisting of 18911 shots. It is compared against the best one in a family of adaptive thresholding methods reported in [2] using two common metrics: Recall (R) and Precision (P). This method normalizes the histogram difference between two frames by the mean and variance of surrounding values. We use the same metric in Equation 4 and the search space reduction technique described in Section 3.3. Table 1 shows that our algorithm consistently outperforms the best adaptive thresholding method across all movies. The improvement in precision and recall is roughly 1% each, which translates to around 15-20 false positives and 15-20 false negatives in each movie. Considering that we have not yet adapted discontinuity values to the visual activities surrounding the present frame, the results are very promising.

Results in Table 1 clearly show that the shot partition can be done much more reliably on drama-based movies such as *American Beauty* and *Erin Brockovich* than on action-based movies such as *The Matrix*, *The Mummy* and *Star Wars I*. A close examination of errors revealed that although our algorithm has managed to eliminate several false positives and false negatives through the identification of the longest path, it still has problems dealing with sequences of intense motion, e.g., the racing sequence in *Star Wars I* and the fighting sequences in *The Matrix*. In addition, close-up shots of moving characters and/or objects tend to produce false positives, as they produce relatively large discontinuity values.

| Movie | shots | Adaptive | | Proposed | |
|---|---|---|---|---|---|
| | | R(%) | P(%) | R(%) | P(%) |
| *Star Wars I* | 2106 | 93.8 | 95.6 | 94.2 | 95.9 |
| *The 13th Floor* | 1340 | 96.4 | 96.3 | 97.2 | 96.6 |
| *The Matrix* | 2400 | 94.5 | 96.1 | 96.0 | 96.6 |
| *Tall Tale* | 1219 | 95.6 | 96.2 | 97.5 | 96.7 |
| *Chameleon* | 965 | 96.0 | 97.4 | 97.2 | 97.6 |
| *12 Monkeys* | 1309 | 98.0 | 97.9 | 98.6 | 98.0 |
| *The Mummy* | 1847 | 93.1 | 94.3 | 94.7 | 95.2 |
| *American Beauty* | 1068 | 99.7 | 98.2 | 99.8 | 97.9 |
| *The Siege* | 1971 | 97.0 | 97.4 | 97.8 | 98.0 |
| *Truman Show* | 1371 | 98.5 | 94.1 | 99.1 | 95.1 |
| *Titanic* | 1975 | 98.6 | 97.0 | 99.3 | 97.1 |
| *Erin Brockovich* | 1340 | 99.7 | 98.8 | 99.9 | 99.0 |

Table 1: Performance Results.

## 5. SUMMARY AND CONCLUSIONS

In this paper, the video segmentation problem is solved by searching for the optimal shot partitioning in a global sense. We showed how the problem can be modelled using Bayesian principles and the optimal solution can be found by DPT. We demonstrated the validity of the technique using 12 full-length movies. Further improvements can be achieved by building a multivariate model of shot length and visual dynamics and adapting the frame-discontinuity measure according to the surrounding visual dynamics.

## 6. REFERENCES

[1] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proceedings of the Second ACM International Conference on Multimedia (MULTIMEDIA '94)*, New York, Oct. 1994, pp. 357–364.

[2] Y. Yusoff, W. Chrismas, and J. Kittler, "Video shot cut detection using adaptive threshold," in *Bristish Machine Vision Conference (BMVC'00)*, Bristol, The Eleventh British Machine Vision Conference 11-14 Sept. 2000.

[3] Yucel Altunbasak, "A statistical approach to threshold selection in temporal video segmentation algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, Istanbul, Turkey, Jun 2000, vol. 6, pp. 2421–2424.

[4] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 3–14, jan 2000.

[5] Alan Hanjalic and HongJiang Zhang, "Optimal shot boundary detection based on robust statistical models," in *Proceedings of IEEE International Conference on Multimedia and Systems*, Florence, Italy, June 1999, vol. 2, pp. 710–714.

[6] C. Taskiran and E. J. Delp, "A study on the distribution of shot lengths for video analysis," in *SPIE Conference on Storage and Retrieval for Media Databases*, San Jose, CA, jan 2002, vol. 4315.