OPTIMIZED VIDEO SCENE SEGMENTATION

Jingdong Wang† *Xinmei Tian‡ Linjun Yang† *Zheng-Jun Zha‡ Xian-Sheng Hua†

†Microsoft Research Asia ‡University of Science and Technology of China

ABSTRACT

In this paper, we propose an optimized video scene segmentation approach with considering both content coherence and temporally contextual dissimilarity. First, a chain structure is constructed by connecting temporally adjacent shots to represent a video. Then the chain is partitioned such that the content within a chain segment is coherent enough and the contextual similarity of temporally adjacent chain segments is small enough. This task is formulated as a ratio function of content coherence and contextual similarity. Finally, we present an effective and efficient hierarchical chain partitioning approach to find the optimal scene segmentation. Experimental results on a set of home videos and feature movies demonstrate the superiority of the proposed approach over several existing key approaches.

1. INTRODUCTION

Temporal video scene segmentation is always an important and fundamental problem in video processing and understanding, and plays a key role as video structuring for video indexing and searching. Scene segment can be defined as a syntactic organization of temporal adjacent shots to represent a continuous action, while a shot is a physical entity corresponding to camera on/off operations. Due to the scene complexity in the videos of different genres, video scene segmentation is still a challenging problem.

There is a large literature on video scene segmentation. Basically the methods can be divided into three categorizes: the merging based method, the splitting based method and the model based method. The merging based approaches perform a bottom-up procedure to gradually aggregate the visually similar and temporally adjacent shots to form a meaningful scene. The best-first model merging (BFMM) method proposed in [10] constructs a chain with the shots as nodes and the loops upon the boundaries between consecutive shots. In [2], scene segmentation is obtained by shot detection based on backward shot coherence and shots grouping based on motion analysis. The disadvantage of the merging based methods is that it takes no consideration of the overall content coherence within a scene.

The spitting based approaches adopt a top-down procedure to separate the video into scene segments. This category of methods usually build a weighted graph with the edges connecting the shots and the weights on the edges reflecting the temporal and visual similarities, and partition the graph for shot clustering. In [7], a complete-link method is used to partition this graph into several subgraphs such that each subgraph satisfies a color similarity constraint. In [3], a spectral clustering method (or normalized cuts) in [4] is adopted to recursively bipartition the graph for scene detection.

The model based approaches assume each scene segment satisfies essentially a latent generative model. In [5], the shots in each scene are assumed to be visually generated from a Gaussian distribution, then a Gaussian mixture model is adopted to cluster the shots based on the visual feature. The approach in [8] presented a statistical method, Markov chain Monte Carlo, to infer scene segmentation with combining the generative model and the temporal consistency. A similar combination method, the energy minimization based segmentation method in [1], used an iterative algorithm for scene segmentation. The model related approaches, such as [1, 5, 8], all assumed that the content of each scene is generated from a latent parametric probabilistic model. However, for videos of different genres, it is very difficult to model the content of a scene in a parametric model.

In this paper, we propose an optimized video scene segmentation approach with considering both content coherence and temporal contextual dissimilarity. We first fragment the video into a sequence of shots, then build a chain structure by connecting the temporally adjacent shots. The scene segmentation problem is formulated as a chain segmentation problem, whose goal is to maximize the contextual dissimilarity of adjacent scenes and simultaneously maximize the content coherence within each scene.

The main contributions lies in two aspects. The first is modeling the content coherence using the similarity within a scene instead of a parametric form such as a Gaussian distribution. The second is a hierarchical scene bisegmentation scheme for efficient and effective optimization, in which scene bisegmentation is obtained in a global optimal way and K-way segmentation is sub-optimally approximated.

By comparison, the method in [3] considering both content and temporal constraints adopts a similar recursive scheme, but the computation is very difficult and the result is not sat-
isfactory due to the impractical utility of the constraints. The method in [10] also uses a chain structure but it takes no consideration of content coherence within a scene.

The remainder of this paper is organized as follows. The formulation of video scene segmentation is given in Sec. 2. Then the optimization method is presented in Sec. 3. Next, the experimental results in Sec. 4 demonstrate the superiority of the proposed approach. Finally, Sec. 5 concludes this paper.

2. FORMULATION

The proposed video scene segmentation framework consists of two steps: shot detection and shot grouping. First, shots/key-frame detection and feature extraction are performed. Then, a chain is constructed by connecting the temporally adjacent shots, and a hierarchical chain partitioning approach is propesed to group the shots and consequently obtain the video scene segmentation.

We perform shot detection by following the method in [1]. A video is first parsed into \( N \) shots \( V = \{s_1, \cdots, s_N\} \), and a key frame is extracted to represent each shot using the color based method [9]. Next, we extract the color moment \( x_i \) in the Lab color space as the feature of each key frame. Finally, a video is represented by a sequence \( X = \{x_1, \cdots, x_N\} \). Given the sequence, the task of video scene segmentation is to group the shots into several shot sets \( \{V_1, \cdots, V_L\} \) with \( L \) the number of sets (scenes), or equivalently to assign scene label \( l_i \) to each shot \( s_i \).

2.1. Formulation

An ideal video scene segmentation usually satisfies two characters: the coherence within the same scene, and the distinctness between different scenes. The coherence is described as the similarities of the shots in the scene. Mathematically, this coherence is formulated to maximize the content similarity of scene \( V_i \), which is defined as

\[
S_{\text{content}}(V_i) = \sum_{s_i, s_j \in V_i} S(s_i, s_j),
\]

where \( S(s_i, s_j) \) is the similarity of two shots \( s_i \) and \( s_j \).

The distinctness of different scenes is described by the contextual dissimilarities between the shots of the different scenes. Mathematically, this distinctness is formulated to minimize the contextual similarity:

\[
S_{\text{context}}(V_i, V_j) = \sum_{s_m \in V_i, s_n \in V_j} S(s_m, s_n).
\]

The similarity of two shots is defined based on both the visual and temporal similarities:

\[
S(s_i, s_j) = \text{sim}(x_i, x_j) \times \delta_i, \quad \delta_i \leq 1,
\]

where \( \text{sim}(x_i, x_j) \) is the visual content similarity of two shots, and \( \delta_i \) is an indicator function to evaluate the temporal similarity with \( \delta_i \) the temporal neighborhood size. In this paper, we build one-dimensional chain,}

\[
\begin{align*}
S_1 & \quad S_2 \quad S_3 \quad S_4 \quad S_5 \quad S_6 \\
\end{align*}
\]

Fig. 1. Chain representation of a sequence of shots. The bottom chain is separated into two subchains [1, 3] and [4, 6] by removing the green edge (3, 4).

\[
\{s_1, s_2, \cdots, s_{t-1}, s_t, \cdots, s_{N-1}, N\} \text{ as shown in Fig. 1, to describe the contextual (temporal) relation. Then in the chain representation, the temporal contextual threshold } \delta_t \text{ is equal to 1.}
\]

The scene segmentation is a multi-objective task, i.e., to maximize the content similarity within each scene and simultaneously minimize the contextual similarities of different scenes. We adopt the ratio criterion, similar to the normalized cut criterion in [3, 4], to formulate the multi-objective problem into a single objective function

\[
f(V_l) = \sum_{l=1}^{L} S_{\text{context}}(V_l, V_l) / S_{\text{content}}(V_l),
\]

where \( V_l = V - V_l \) is equal to the difference between sets \( V \) and \( V_l \) with \( V = \bigcup_{l=1}^{L} V_l \).

Suppose the labels of the scene segments are sorted according to the temporal order of the shots, i.e., each shot, \( s_i \in V_l \), appears temporally after each shot, \( s_j \in V_{l+1} \) \((i < j)\), and let \( V_l = \{s_{t_l + 1}, \cdots, s_{t_l}\} \). Then

\[
S_{\text{context}}(V_l, V_l) = \sum_{s_{t_l + 1} \in V_l} S_{\text{context}}(V_l, s_{t_l + 1}) + S_{\text{context}}(V_l, s_{t_l}) = S(s_{t_l - 1}, s_{t_l + 1}) + S(s_{t_l}, s_{t_l + 1}).
\]

3. OPTIMIZATION

In this section, we present a novel combinatorial optimization approach to minimize the objective function Eqn. (4), which is more effective and efficient than existing continuous relaxation based methods such as spectral relaxation in [3, 4] and semi-definite relaxation in [6]. In the following, we first present an optimal exact approach to the two-way video scene segmentation (scene bisegmentation), i.e., the minimization of Eqn. (4) with \( L = 2 \), then we propose a hierarchical chain partitioning approach to multiple-way scene segmentation.

3.1. Scene bisegmentation

Suppose \( V_1 = \{s_1, \cdots, s_t\} = [1, t] \) and \( V_2 = \{s_{t+1}, \cdots, s_{N}\} = [t + 1, N] \), then their temporal adjacent shots are \( s_t \) and \( s_{t+1} \), as depicted in Fig. 1. Then, their contextual similarity is calculated as the following:

\[
S_{\text{context}}(V_1, V_2) = S(s_t, s_{t+1}) = \text{sim}(x_t, x_{t+1}).
\]

The content similarities are calculated as \( S_{\text{content}}(V_1) = S_{\text{content}}([1, t]) \) and \( S_{\text{content}}(V_2) = S_{\text{content}}([t + 1, N]) \), with

\[
S_{\text{content}}([t_1, t_2]) = 2 \sum_{t = t_1}^{t_2 - 1} \text{sim}(x_t, x_{t+1}) \quad \text{and} \quad \text{sim}(x_t, x_{t+1}) = \exp(-|x_t - x_{t+1}|^2 / 2\sigma^2).
\]
Algorithm 1 EVALUATE-CONTENT-SIMILARITY(t_l, t_r, a[])
Input: t_l, t_r, a[]
// t_l, t_r are staring and ending indices of the input video segment
// a[] is the similarity array
Output: c[i] // c[] is the array of content similarities
1: for i = t_l + 1 to t_r do
2: c[i] = c[i - 1] + 2a[i]
3: end for
4: return c[]

Algorithm 2 VIDEO-BISEGMENTATION(t_l, t_r, a[])
Input: t_l, t_r, a[]
Output: t_{min}, m_{min} are optimal segmenting index and objective value
1: t_{min} = t_l + 1
2: t_{min} = -INFINITY
3: r_{min} = \min(r_{min}, r)
4: for i = t_l + 1 to t_r do
5: r = a[i] \times \frac{1}{\sum_{j=t_{l}+2}^{t_{r}} a[j]}
6: r_{min} = \min(r_{min}, r)
7: t_{min} = i
8: end for
9: return (t_{min}, r_{min})

Therefore, the objective function of two-way scene segmentation is rewritten as follows:
\[ f(V_1, V_2) = \frac{\text{sim}(x_t, x_{t+1})}{S_{\text{content}}([1, t])} + \frac{\text{sim}(x_{t+1}, x_t)}{S_{\text{content}}([t + 1, N])}. \] (5)

A naive method of minimizing Eqn. (5) is exhaustively to calculate all kinds of combinations \( V_1 \) and \( V_2 \), which will take \( O(n^2) \) time. However, we notice two properties about content similarity: similarity complementarity and evaluation overlapping. Mathematically, the complementarity property means
\[ S_{\text{content}}([1, t]) + S_{\text{content}}([t + 1, N]) = S_{\text{content}}([1, N]) - 2 \cdot \text{sim}(s_t, s_{t+1}). \] (6)

With this property, it is sufficient to calculate the content similarity \( S_{\text{content}}([1, t]) \). The overlapping region is represented by the following equation:
\[ S_{\text{content}}([1, t]) = S_{\text{content}}([1, t - 1]) + 2 \cdot \text{sim}(s_{t-1}, s_t). \] (7)

According to this property, we can calculate the content similarity \( S_{\text{content}}([1, t]) \) in a temporal order from \( t = 1 \) to \( t = N \). The pseudo code of evaluating the content similarity is presented in Alg. 1.

After the content similarities are evaluated, we traverse all shot stamps \( t \) to find \( t_{min} \) corresponding to the minimum objective value. Then video bisegmentation is obtained. The pseudo code is shown in Alg. 2.

3.2. \( K \)-way scene segmentation

In this subsection, we present a best-first hierarchical bisegmentation scheme, called hierarchical chain partitioning (HCP), to approximate \( K \)-way scene segmentation when \( K > 2 \). Since there are \( O(nK-1) \) combinations of segmentation cases, exact minimization of Eqn. (4) requires \( O(nK-1) \) time, which is too expensive. Therefore, a practical algorithm is essential for video scene segmentation.

The proposed best-first hierarchical bisegmentation scheme is a recursive algorithm. First, the video is segmented into two coarse scene segments \( V_1 = [t_1, t_2] \) and \( V_2 = [t_2 + 1, N] \). Then, on each coarse scene segment \( V_i = [t_{i-1}, t_i] \), the bisegmentation scheme is performed to provide a candidate segmentation \( [t_{i-1}, t_{i}], [t_{i}, t_{i+1}] \). Next, considering all the candidate bisegmentation, the candidate bisegmentation with the smallest objective value is chosen as the next segmentation. The above two steps, providing candidate segmentation and selecting best candidate, are iterated until the convergence condition is reached. The pseudo code is given in Alg. 3.

3.3. Complexity analysis

Line 2 in Alg. 1 for evaluating content similarity is run in \( O(t_r - t_l) = O(N - 1) = O(N) \). Similarly, we can easily analyze that its time complexity is \( O(N) \) for Alg. 2. In Alg. 3, VIDEO-BISEGMENTATION() is called in \( O(K) \) times and VIDEO-BISEGMENTATION() takes \( O(N) \) time. Therefore, the total time complexity is \( O(KN) \).

4. EXPERIMENT

In this section, we demonstrate the proposed video scene segmentation approach on two sets: home videos and feature movies. To measure the segmentation performance, we use the balanced \( F \)-score, defined as
\[ F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \]
We adopt the method in [3] to give a statistical result. We use the offset of detected boundary as the "tolerance", and the detected boundary is viewed as true if its offset is within the tolerance. For comparison, we also present several existing key approaches: the energy minimization segmentation (EMS) based method [1], the best-first merging model (BFMM) based method [10], the recursive graph splitting (RGS) based method [3], and the MCMC based method [8].
In the experiment, the number $K$ is set by hand as the ground truth scene number, the length scale in similarity evaluation is set as the average distance $\sigma = E(\| x_i - x_j \|)$. In addition, the number of clusters is set by hand as the ground truth scene number. In MCMC, the mean parameters are well tuned. For BFMM, the number of clusters is set as the ground truth scene number. In EMS that requires a training scheme to improve performance, HCP without training gets competitive performance at low tolerance and better performance at high tolerance. For RGS that uses a similar recursive scheme, the proposed approach performs well on videos of various domains and is superior over several existing approaches.

5. CONCLUSION

In this paper, we propose a novel video scene segmentation approach. It aims to maximize the content coherence within a scene segment and maximize the contextual dissimilarity between temporally adjacent scene segments. A combinatorial optimization algorithm in linear time complexity is presented for a global optimal scene bisegmentation. Moreover, we present an effective and efficient hierarchical chain partitioning method to $K$-way scene segmentation. The experimental results on home videos and feature movies demonstrate that the proposed approach performs well on videos of various domains and is superior over several existing approaches.

6. REFERENCES