RC22405 (W0204-086) April 24, 2002
Computer Science

IBM Research Report

Neighborhood Coherence and Edge Based Approaches for Film Scene Extraction

Ba Tu Truong[+], Svetha Venkatesh[+] , Chitra Dorai

Department of Computer Science[†]
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia

IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g. , payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

Neighborhood Coherence and Edge Based Approaches to Film Scene Extraction

Ba Tu Truong[†], Svetha Venkatesh[†], Chitra Dorai[‡]

Department of Computer Science[†]
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{truongbt, svetha}@cs.curtin.edu.au

IBM T. J. Watson Research Center[‡]
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

Abstract

In order to enable high-level semantics-based video annotation and interpretation, we tackle the problem of automatic decomposition of motion pictures into meaningful story units, namely scenes*. Since a scene is a complicated and subjective concept, we first propose guidelines from film production to determine when a scene change occurs in film. We examine different rules and conventions followed as part of Film Grammar to guide and shape our algorithmic solution for determining a scene boundary. Two different techniques are proposed as new solutions in this paper. Our experimental results on 10 full-length movies show that our technique based on shot sequence coherence performs well and reasonably better than the color edges-based approach.*

1. Introduction

Motion pictures, like other video genre are constructed from shots, which are contiguous sequences of images captured by a single camera. Until now, a large body of work has been devoted to solving the problem of automatic shot boundary detection using color, edges, object correlation and motion cues, singly or in combination. While shots are fundamental, they are analogous to letters or words in a written language (depending on their length and complexity) and convey little semantic information in isolation. Most meaningful information is embedded at a higher level of film structure, known as *scenes* or story units, which reflect the aggregation of multiple shots revolving around a single dramatic persona, incidence or location [2].

Since scenes are composed of many shots, automatic detection of scene boundaries requires high-level concepts spanning shot sequences and between shots analysis. There are currently two major trends in exploiting visual information for scene boundary extraction. The first trend groups shots into different clusters and interprets the temporal ordering of elements from each cluster [4]. Techniques from the second trend [3, 6] develop a memory viewer based model that is either causal/finite or non-causal/infinite to define a shot coherence measure based on shot recall ability and scenes are detected by searching for local minima on this coherence curve. The problem with techniques from the first trend is that they critically depend upon clustering parameters (threshold, number of clusters), as clustering is the fundamental step in this approach. Clustering also inhibits the ability to visualize the progression across shots that would be useful in locating scene boundaries. Techniques from the second trend require accurate local minima detection from a rather noisy signal (due to many parameters involved in computations – shot length, spacing, etc) which would be difficult. Further, this model draws guidance from subjective perception of viewers, as opposed to being based on knowledge of how scenes are constructed using cinematic devices by directors worldwide.

Our approach is based on careful examination of forces that go into crafting scenes in film. Leveraging *Film Grammar* as the means to inspire and shape our algorithmic solution to the problem of scene extraction, we investigate different rules and conventions that are employed in composing scenes in motion pictures. We propose two new approaches to extracting scene boundaries using visual content. The first detects scene transitions using a multiresolution edge detection scheme on the $\mathcal{H}$, $\mathcal{L}$, and $\mathcal{S}$ values of the shot signal, and is inspired by a recent work on movie tempo [1]. The second technique is based on computing a neighborhood visual coherence measure at each shot boundary. We also propose a new shot similarity measure that takes similarity of colors into account in the formulation of our color coherence measure. We report results from experiments on contemporary mainstream color motion pictures.

2. Film Grammar on Scene Composition

The term 'scene' is borrowed from the French classical theater that had a precise beginning and ending corresponding to the arrival and departure of characters. In the Film Encyclopedia, Katz states: "In the strictest sense, a scene is defined as a section of a motion picture which is unified in time and space. It is made up from a series of shots from varying angles and is usually filmed in one session. As a unit of language, the scene is intermediate between a shot and a sequence, being a larger unit than the first and smaller unit than the latter" [2]. Given this definition, we use the following rules as guidelines to set up the ground truth for our experiments with film and to define scenes in our work (see motivation and details in [7]):

1. When there are no two interwoven parallel actions, a change in location or time or both defines a scene change.
2. An establishing shot is considered part of that scene.
3. When parallel actions are present and interleaved, and there is a switch between one action to another, a scene boundary is marked if and only if the duration of the action is shown for at least 30 sec.
4. A montage sequence involving "dynamic cutting", where shots containing different spatial properties are rapidly joined to convey a single dramatic event, forms a single scene.

The two most commonly used scene structures are *Separation* and *Familar Image* as described by Sharff in his investigation of cinesthetic impact [5]. *Separation* is the arrangement of shots showing subjects individually and interwoven, creating the shot pattern **...ABABAB...**. In addition, in *separation*, any action involving motion should take place while the characters are involved in the screen, minimizing abrupt visual changes. The *familiar image* is a scene structure in which a picture reappears in a film with rhythmic frequency with approximately the same composition and framing [5]. It can be distributed throughout the film to create the effect of recall. Within a scene, it functions as a pivot image. Separation and familiar image are among the factors motivating the computation of shot neighborhood coherence presented in this paper.

It is also well understood that color plays an important role to connote a certain character, time, setting or mood. Therefore, transitions between scenes essentially involving changes in time and space are often associated with changes in characters and mood, resulting in great visual changes, and this is exploited in our computation.

3. Feature Extraction

Given a digital stream of a movie in the form of MPEG-1-encoded data, shot detection is first carried out. Shot attributes such as average $\mathcal{H}$, $\mathcal{L}$, $\mathcal{S}$ histograms are computed. The HLS space is quantized using 12 bins of hue, 5 bins of lightness, and 4 bins of saturation, resulting in a total of $113 = 1 + 1 + (5-2) + 12(4-1)(5-2)$ in our final quantized color palette. Using a normalization scheme [7] that weights the color content of a shot by the duration it remains in effect, we compute normalized shot color histograms.

4. Scene Transitions via Edge Detection

Changes in the overall color atmosphere in a film will result in changes, together or individually, in the $\mathcal{H}$, $\mathcal{L}$, $\mathcal{S}$ histograms of associated shots. These color transitions can be detected using an edge detector based on Deriche's recursive filtering algorithm using Gaussian kernels [1]. This is a multi-scale edge detector that is parameterized by σ, which determines the slope of target edges in the signal. Larger σ detects smaller slope of the target edges and vice versa. Two thresholds (τ, and -τ), one negative and one positive, can be applied to the output from this multi-resolution algorithm, known as the *edge signal*, to locate true edges: the higher the threshold, the larger the edges detected, and vice versa. The section above/below these thresholds means a negative/positive change in the original signal. Since there are three different edge signals in our feature set, one for each $\mathcal{H}$, $\mathcal{L}$, and $\mathcal{S}$ signal, they are added before applying thresholds to extract edges. Hence, for each σ and shot index i, the accumulated edge signal $\mathbb{E}_i^{\mathcal{HLS}}$ is computed from individual edge signals $\mathbb{E}_i^{\mathcal{H}}$, $\mathbb{E}_i^{\mathcal{L}}$ and $\mathbb{E}_i^{\mathcal{S}}$ as follows:

$$\mathbb{E}_i^{\mathcal{HLS}}(\sigma) = \alpha|\mathbb{E}_i^{\mathcal{H}}(\sigma)| + \beta|\mathbb{E}_i^{\mathcal{L}}(\sigma)| + \gamma|\mathbb{E}_i^{\mathcal{S}}(\sigma)|.$$

The weighting reflects different degrees of contribution of changes in $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{S}$ to a scene change. Currently, with the absence of specific statistical information about their proportional contributions, we set all weights to 1. Then, $\mathbb{E}_i^{\mathcal{HLS}}(\sigma)$ is thresholded to extract edges that denote true scene transitions. Figure 1 shows the computed edges on the $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{S}$ signals for a segment of the movie, Twelve Monkeys. Each column indicates the temporal extent of each detected edge.

Visually, a shot should be more similar to other shots in the same scene than to shots in a different scene. Therefore, the exact location of a scene transition should maximize average inter-differences between shots of the previous scene and shots of the next scene. For a detected edge temporally spanning shots $\mathbf{S}_m$ to $\mathbf{S}_n$, the exact scene transition position $\mathbb{T}(m, n)$, (i.e., the index of the first shot of the next scene) in terms of shot indices can be determined as:

$$\mathbb{T}(m, n) = \{t \mid \min_{m<t<n} \frac{\sum_{i=m}^{t-1}\sum_{j=t}^{n} \mathbb{S}(\mathbf{S}_i, \mathbf{S}_j)}{(t-m)(n-t+1)}\},$$

where $\mathbb{S}$ is the similarity measure between two shots as defined in the next section.

5. Scene Transitions via Shot Neighborhood Coherence

Measuring the visual similarity between shots is the basis of most scene boundary detection techniques that use

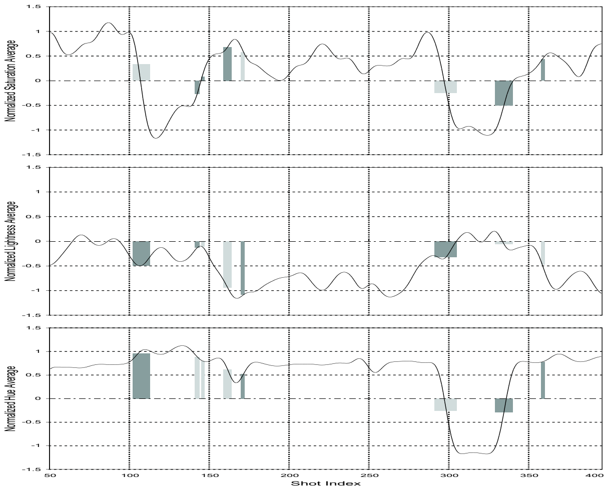

Figure 1. Scene tranistions from detected edges on average $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{S}$ from shots.

visual content. However, these techniques have failed to handle at least one of the following issues adequately, especially the first one:

1. The similarity between two colors should be modeled as a continuous, rather than a discrete value. Shots within a scene may exhibit different lighting or shading characteristics, due to different camera shooting angle and overall motion.

2. In fast motion scenes, or slow disclosure shots, only a part of a shot is similar to another shot of the same scene. Therefore, we should use as many frames as needed to evaluate the similarity between shots.

We address the first issue by moving away from traditional image/frame comparison using bin-wise histogram comparison, which does not take similarities between shades of colors into consideration. We propose a new similarity measure, which mimics the process of measuring the similarity between two images by gradually computing and excluding regions with the highest similarity. We first form the color similarity matrix **I** based on the Euclidean distance between colors in $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{S}$ space. Moreover, we set the distance to **INF** when the $\mathcal{H}$ components of two colors are more than two hue levels apart. This is to confirm the fact that it is neither possible to obtain both blue and red colors for the same object or background region by changing the camera angle nor to have both red, blue shading of the same color. Other values are normalized to within range [0-1]. Let $\mathbf{F}_i[u]$ denote the bin u of frame $\mathbf{F}_i$, and $\mathbf{I}_{ut}$, the similarity between colors u and t. We define component similarity between two frames $\mathbf{F}_i$, $\mathbf{F}_j$ and two bins u, t as:

$$\mathbb{P}(\mathbf{F}_i, \mathbf{F}_j, u, t) = \mathbf{I}_{ut} \min(\mathbf{F}_i[u], \mathbf{F}_j[t]).$$

The overall similarity of two frames can be calculated as the recursive sum of individual component similarities:

$$\mathbb{S}(\mathbf{F}_i, \mathbf{F}_j) = \begin{cases} \mathbb{S}(\mathbf{F}_i^*, \mathbf{F}_j^*) + \mathbb{P}(\mathbf{F}_i, \mathbf{F}_j, u_0, t_0) & \exists (u_0, t_0) \\ 0 & \text{otherwise.} \end{cases}$$

with

$$(u_0, t_0) = \{(u, t) \mid (\max_{1 \le u, t \le N} \mathbf{I}_{ut}), \mathbb{P}(\mathbf{F}_i, \mathbf{F}_j, u, t) > 0\}$$

$$\mathbf{F}_i^*[u] = \begin{cases} \mathbf{F}_i[u] - \min(\mathbf{F}_i[u_0], \mathbf{F}_j[t_0]) & \text{if } u = u_0 \\ \mathbf{F}_i[u] & \text{otherwise.} \end{cases}$$

$$\mathbf{F}_j^*[t] = \begin{cases} \mathbf{F}_j[t] - \min(\mathbf{F}_i[u_0], \mathbf{F}_j[t_0]) & \text{if } t = t_0 \\ \mathbf{F}_j[t] & \text{otherwise.} \end{cases}$$

This means that after taking the component similarity of the two most similar colors, the component similarities of the remaining parts of the two histograms are recursively extracted until there does not exist any pair of colors from each histogram that have component similarity greater than zero, i.e., two colors are not similar at all or the bin size of at least one of the colors is zero. The value of $\mathbb{S}$ is then normalized by the total number of pixels in a frame. It should be noted that this measure will become essentially the bin-wise intersection metric when **I** is the identity matrix.

In order to handle the second issue, we formulate the similarity between two shots as the maximum similarity between any pair of representative frames, $\mathcal{R}$-frames of these shots. In our $\mathcal{R}$-frame representation of a shot [7], the number of frames needed for each shot is proportional to the magnitude of visual changes within the shot. If two shots are static, only their first frames are used. On the other hand, during a chase scene, many frames are used for each shot to adequately cover the continuously changing visual space.

$$\mathbb{S}(\mathbf{S}_i, \mathbf{S}_j) = \max_{\mathbf{F}_{k_{mi}} \in \mathbf{S}_i^{\mathcal{R}}, \mathbf{F}_{k_{nj}} \in \mathbf{S}_j^{\mathcal{R}}} \mathbb{S}(\mathbf{F}_{k_{mi}}, \mathbf{F}_{k_{nj}}).$$

We then define the *shot neighborhood coherence* measure at the shot $\mathbf{S}_m$ as the maximum similarity $\mathbb{S}$ between neighborhood shots preceding and succeeding this shot and within a window τ.

$$\mathbb{C}(\mathbf{S}_m) = \max_{i < m \le j, j < i + \tau} \mathbb{S}(\mathbf{S}_i, \mathbf{S}_j).$$

The low visual variance within a scene as well as the shot repetition (separate, dialogue, familiar images) means that the shots within a scene are highly coherent. On the other hand, we are unlikely to find two similar shots between two scenes, where the coherence between shots across the scene boundaries is typically low. In our implementation, we use both the number of frames and number of shots in defining the temporal window as it would cater to both long and short shots. The shot coherence can now be thresholded to extract scene indices.

	Ground Truth			Edge Based		Coherence Based	
Movie	Dur	Scenes	Genre	*Rec*	*Prec*	*Rec*	*Prec*
Star Wars I	134	97	Scifi/Adv/Act	79.4	67.0	87.7	73.4
The 13th Floor	100	65	Mys/Scifi/Thrl	70.8	80.7	84.7	83.4
The Matrix	136	60	Act/Thrl/Scifi	90.0	71.1	86.7	91.3
Sleepy Hollow	103	61	Fts/Hrr/Mys	83.7	68.0	80.4	83.1
Tall Tale	98	51	Adv/Family	78.5	74.1	84.4	71.7
Chameleon	120	46	Scifi/Thrl	82.7	82.7	93.5	81.2
12 Monkeys	130	68	Drm/Thrl/Scifi	79.5	75.0	92.7	78.8
The Mummy	124	52	Adv/Act/Hrr	67.4	67.4	63.5	75.0
American Beauty	121	74	Drm/Cmd	75.7	84.9	90.6	80.8
The Siege	115	72	Act/Thrl/Drm	79.2	76.0	93.1	78.0
Overall	1181	646		78.7	73.9	86.3	79.2

Table 1. Experimental data and performance statistics.

6. Experimental Results

In order to ensure that the overall performance measures of the algorithm are not biased toward a specific movie type, we set up a data set consisting of 10 full-length movies of all major genres including action (Act), horror (Hrr), science fiction (Scifi), adventure (Adv), thriller (Thrl), fantasy (Fts), family (Fml), drama (Drm), comedy (Cmd) and mystery (Mys). Information about each movie is represented in Table 1. The genre classification is taken from The Internet Movie Database Web site (IMDB) (www.imdb.com).

Table 1 shows the results of edge based and coherence based approaches after being refined by the same post-processing techniques detailed in [7]. Overall, the shot neighborhood coherence approach yields better precision (*Prec*) and recall (*Rec*) (86.3, 79.2) compared to the edge based approach (78.7, 73.9). This also persists across individual movies except for the recall values of The Matrix, Sleepy Hollow and The Mummy and the precision values of Chameleon, Tall Tale and American Beauty. The reason could be that the colors of The Matrix, Sleepy Hollow and The Mummy are tinted towards specific tones (grayish green, grayish light blue, and earthy respectively). This lowers shot coherence values across scenes in these movies. The normalization of the average $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{S}$ values magnifies slight changes in color atmosphere, hence allowing more indices to be retrieved by the edge based approach. On the other hand, Chameleon and American Beauty contain a wide range of colors making the coherence based method more sensitive to false alarms than the edge based method, as the normalization of average $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{S}$ now would suppress small changes in color atmosphere. This also suggests that the coherence based method could be further improved if such normalization processes were incorporated. The highest results for the edge based approach are obtained with Chameleon (82.7, 82.7), The Matrix (90.0, 71.1) and American Beauty (75.7, 84.9). Similarly, the shot coherence based approach performs best on Chameleon (93.5, 81.2), The Matrix (86.7, 91.3), American Beauty (90.6, 80.8) and The Siege (93.1, 78.0).

7. Conclusion

We have addressed the problem of automatically determining scene boundaries in motion pictures. We investigate different rules and conventions used in Film Grammar that underpin and shape an algorithmic solution to this problem. We proposed two different techniques. The first is based on detecting scene transitions using a multi-resolution edge detection mechanism on color signals formed by normalized average Hue, Lightness and Saturation. The second approach estimates the coherence level at each shot by computing weighted color similarity of neighborhood shots. Experimental results on 10 full-length movie data show that our technique based on the shot coherence performs well and reasonably better than the color edges-based approach.

References

[1] B. Adams, S. Venkatesh, and C. Dorai. Studying shot length and motion as contributing factors to movie tempo. In *ACM Multimedia 2000*, pages 353–355, Los Angeles, October 2000.

[2] E. Katz. *The Film Encyclopedia*. Harper Collins Publishers, 2nd edition, 1994.

[3] J. R. Kender and B.-L. Yeo. Video scene segmentation via continuous video coherence. In *1998 IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

[4] Y. Rui, T. S. Huang, and S. Mehrotra. Constructing table-of-content for videos. *ACM Multimedia System Journal: Special Issue in Multimedia Systems on Video Libraries*, 7(5):359–368, 1999.

[5] S. Sharff. *The elements of Cinema: Toward a cinesthetic impact*. Columbia University Press, New York, 1982.

[6] H. Sundaram and S.-F. Chang. Determining computable scenes in films and their structures using audio visual memory models. In *ACM Multimedia 2000*, CA, October 2000.

[7] B. T. Truong, C. Dorai, and S. Venkatesh. Automatic scene extraction in motion pictures. Technical Report 1/2002, Curtin University of Technology, 2002.