IDENTIFYING FILM TAKES FOR CINEMATIC ANALYSIS

Ba Tu Truong, Svetha Venkatesh

Department of Computer Science
Curtin University of Technology
GPO BOX U 1987 Perth 6001, Australia
{truongbt, svetha}@cs.curtin.edu.au

Chitra Dorai

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

ABSTRACT

In this paper, we focus on the ‘reverse editing’ problem in movie analysis, i.e., the extraction of film takes, original camera shots that a film editor extracts and arranges to produce a finished scene. The ability to disassemble final scenes and shots into takes is essential for nonlinear browsing, content annotation and the extraction of higher order cinematic constructs from film. In this work, we investigate agglomerative hierarchical clustering methods along with different similarity metrics and group distances for this task, and demonstrate our findings with 10 movies.

1. INTRODUCTION

Much of the work in CBVIR has focused on video segmentation including shot/scene extraction and effective algorithms have been reported in this area. In addition, increasingly popular DVD technology has allowed many features, including chapter/scene selection (manually labeled during DVD production) to be incorporated in an DVD release for consumer ease of access to content. The challenge in video analysis has now turned to developing technologies that take advantage of available shot, scene indices for content annotation and better semantic understanding of audio-visual materials to present useful modes to access and manipulate content. In this work, we study the problem of extraction of original film takes from produced video and examine the use of clustering techniques to detect film takes automatically.

A film take is defined as “one uninterrupted run of the camera to expose a series of frames,” according to the Dictionary of Film Terms. A film take is also known as a shot captured during the film shooting, and before the editing stage as opposed to shots in the finished film which are generally understood as the portion of the visual stream between two consecutive cut points, or in edited film, splice points. To avoid confusion, this paper always uses the term ‘shot’ in the context of the finished film.

The left side of Figure 1 shows the film production process from shooting raw takes to producing the final edited material. During the shooting, different takes of a scene are acquired from multiple camera setups, angles, and/or different filming sections. The editor creates the final shot sequence of the scene by selecting from, mixing, and alternating between different portions of these takes to achieve the desired emphasis. Scenes are then assembled into a finished film as seen by viewers. The right side of Figure 1 outlines the reverse editing process which uses keyframes/shot/scene indices previously extracted to detect takes that contributed to the final production. It shows, for example, shots of a scene being analyzed and grouped into 3 clusters which map to 3 takes captured during the film shooting.

Figure 1: Film production and reverse-editing process.

Custering of shots has been examined in some previous studies [1, 2, 3]. These studies tend not to explicitly specify what the extracted clusters represent, except describe them in terms of the results obtained (e.g., indoor, coffee shop scenes), and neither do they specify any consistent groundtruth nor measure the system performance on a large set of data. They often use shot clustering as a means for extracting scene boundaries. [1] proposes the notion of Scene Transition Graph which organizes clustered shots into a directed graph for compact representation and scene segmentation in a video. Our work alternatively uses scene indices available through other methods (some we have developed) as the temporal constraints in our clustering analysis.

2. UTILITY OF FILM TAKES

This section outlines how automated extraction of film plays an essential role in computing many higher order cinematic constructs and tasks in film analysis. For illustration purposes, we use a hypothetical scene. The annotated shot sequence of this scene and its shot/take transition graph are shown in Figure 2. The numbers denote the take indices, and shots of the same shading belong to the same take.
Content Summarization/Annotation: The identification of film takes will enable more compact representation of the video under analysis. Rather than being overwhelmed by all the shots of the scene, only one shot from each take needs to be presented to the user. The reduction factor for the above scene is 11/29. In real sequences, the reduction would be much higher. It also allows the user to browse the video content in a graph-like structure rather than linearly going through all shots. Many shot features can be annotated for the whole take and these include distance, angle, color, lighting, framing, and composition.

Editing Rhythm: Apart from movement and cutting rates, the repetition of certain shots in 2-beat, 3-beat patterns is an essential element in establishing the rhythm of the film. Shots in takes 4, 8, 9, 10, 11 form 2-beat patterns.

Cinesthetic Elements — Separation: Stefan Sharff [4] states that cinema has its own unique method of providing aesthetic gratification and composing cinematic sentences, called ‘cinesthetic elements.’ Four of eight different elements described by Sharff can benefit from the extraction of film takes: separation, familiar image, orchestration, and multi-angularity. Separation is the fragmentation of a scene into single images, seen in alternation, A, B, A, B, A, B, etc. Separation is a particularly strong element in cinema. In the example, separation starts at the first shot of take 8 and ends just before the last shot. Separation would be detected based on the alternation between takes.

Cinesthetic Elements — Familiar Image: This element refers to the repetition of certain images which thus become familiar and are used as the means of keeping together continuity. Familiar images would be detected by looking for takes with at least 3 shots and not alternating with other takes. Takes 1 and 4 are familiar images in the sample scene.

Cinesthetic Elements — Multi-Angularity: This element conveys information in different sizes and in contrasting angles; it is the most common structure in cinema, responsible for creating three dimensionality on a flat screen. Multi-angular shots tends not to be repeated and often occur at the beginning of the scene and we can look for consecutive takes with only 1 shot. Takes 2 and 3 are likely to be part of a multi-angularity configuration.

Cinesthetic Elements — Orchestration: This is an open concept and includes symmetry of shot arrangements. Take 2 include the opening and ending shots and can be seen as an orchestration element.

Shot Flow: From takes and their alternation we can extract certain shot flow patterns. Takes 8 and 9, 10, and 11 alternate with each other suggesting a dialogue scene. Take 4 branches to shots of different takes suggesting that it is the centre of action around which other shots revolve. In addition, takes 2, 3, 4, 5, 6, 7 seems to be separated from takes 8, 9, 10, 11.

Shot Associations: By identifying film takes we have already detected the association between shots. There are also associations between different takes which can be inferred from shot flows. For example, since there is neither a flow from takes 10 and 8 nor from takes 11 and 9, we can generally deduce that the transition between them would break the flow of the story. Therefore, it is likely that that takes 10 and 8 (11 and 9) shoot the same character using different focal lengths.

Movement Within Scene: Certain aspects of character/camera movements within the scene can also be interpreted. Assume that there is some motion in the first shot of take 8 and last shot of take 10 and there is no motion in between; it is likely that these two shots involve characters entering/leaving the position of action.

Relative Difference/Constrast: If a detected take is in a cold tone while another is in a warm tone, we can conclude that there is different state of mind associated with characters in these takes. Likewise, if one contains motion and while another is static, we can conclude that one character is volatile or unsettled while the other is calm.

Measuring Shot/Take Importance: An essential component of the scene can be measured by how much the shot is repeated or how long the total duration of all shots of the same take is. Takes 1, 4, 8, 9, 10, 11 seem to contain essential story information while takes 2, 3, 5, 6, 7 are likely to be peripheral. Also, takes 10/11 are likely to be more important than takes 8/9.

Dramatic Shift: Certain shift in drama/action are reflected in the shift of shot patterns. There is a shift between takes 8/9 to takes 10/11. We can also interpret the kind of dramatic shift within the scene by measuring the shot distance (e.g., via face sizes). If take 8/9 is a medium-shot and take 10/11 is a close-up-shot, we can infer that the drama has increased toward the end of the scene.

3. ALGORITHMS USED IN TAKE DETECTION

Given a digital stream of a movie, shot detection is first carried out. Then, we extract scene indices and compute shot/scene features from those of representative keyframes.

3.1. Clustering and Validation Method

We examine traditional hierarchical clustering techniques including Complete Linkage (CL), Group-Average Linkage (GAL), Median (MED) and Ward’s Minimum Variance (WARD) for shot grouping. They all share the same basic operations as outlined in Algorithm 1[5]:

There are methods to extract a partition from a cluster hierarchy independently. At the current state of our work, we extract the partition that ‘best matches’ the ground truth.

Figure 2: A shot sequence and Shot Transition Graph.
Algorithm 1 Hierarchical clustering procedure
1. Form $n$ single-member clusters.
2. Find the closest pair of distinct clusters, merge them as a new cluster. Delete old clusters and decrement number of clusters by one.
3. Stop if number of clusters equals 1, else goto 2

3.2. Measuring Shot Similarity
Since the compositions of shots of the same logical take should largely match each other, pixel-by-pixel matching is appealing. However, we also need to tolerate variances due to motion and camera adjustments and avoid producing spurious clusters. We chose to match 4 blocks of the frames separately and combine the results. This essentially addresses the arrangement of colors on the left, right, top and bottom part of the frame. Each block is represented by a color histogram and their similarity is calculated using histogram intersection. The similarity between two frames $F_i, F_j$ can be measured as:

$$S(F_i, F_j) = S(H_{i}^{l}, H_{j}^{l}) + S(H_{i}^{r}, H_{j}^{r}) + S(H_{i}^{t}, H_{j}^{t}) + S(H_{i}^{b}, H_{j}^{b}).$$

We then formulate similarity between two shots as the maximum similarity between any pair of keyframes of these shots:

$$S(S_i, S_j) = \max_{F_{km_i} \in S_i, F_{km_j} \in S_j} S(F_{km_i}, F_{km_j}).$$

3.3. Clustering Refinements/PostProcessing
General clustering techniques do not take into account specific characteristics of the data domain. Based on the understanding of underlying film production process and film techniques, we devise algorithms for recursively merging and splitting clusters to further improve the results.

First, we need to deal with consecutive shots that are grouped into the same cluster. Other than for some rare ‘staccato’ effects, it is very unlikely that two consecutive shots are edited from the same take. The grouping of two consecutive shots into a cluster is either due to noisy shot indices or the failure of our similarity metric in discriminating these two shots. Errors of the first kind are rare due to the reliability of shot indexing process. Most errors are of the second type and these clusters need to be split.

The splitting algorithm is shown in Algorithm 2 and it proceeds by choosing two consecutive shots with the least similarity as seeds for two new clusters. The rest are assigned to the closest cluster while maintaining the minimum fusion level.

Secondly, movements within the scene may cause the clustering to be sparse. Since the camera follows character movements and actions so as to maintain continuity, the viewer is presented with cues to perceive that the shot sequences in old and new positions are of the same take. However, clustering techniques like CL and WARD may fail as they measure the distance using all shots in two clusters.

Algorithm 2 Splitting Clusters
1. Search all consecutive shots pairs of this cluster.
2. Select the least similar pair as seeds for two new clusters.
3. Stop if there are no remained shots, else select the next shot that have the smallest distance to either of the cluster and assign it to the closer cluster. Repeat 3.

The shot with movements is either the last shot of first cluster or the first shot of the second cluster. For the first case, the last shot of a cluster is most similar to shots of the other cluster. Algorithm 3 shows how these movements can be detected to merge the clusters.

Algorithm 3 Merging Clusters
1. Search all cluster pairs ($C_1, C_2$) satisfying that the last shot $C_1$ is 2 shots before the first shot of $C_2$, i.e., $C_2[1] = C_1[m] = 2$ and $m + n \geq 4$. Goto 4.
2. Compute:
   $$\alpha_1 = \min \{ S(C_1[m], C_2[i]) \} \quad \beta_1 = \max \{ S(C_1[i], C_2[j]) \}$$
   $$\alpha_2 = \min \{ S(C_1[i], C_2[1]) \} \quad \beta_2 = \max \{ S(C_1[i], C_2[j]) \}$$
3. If $(\alpha_1 < T$ and $\alpha_1 > \beta_1)$ or $(\alpha_2 < T$ and $\alpha_2 > \beta_2)$, merge $C_1$ and $C_2$.
4. Select the next cluster pair and goto 2, else stop.

A similar situation to movements within the scene is the use of fluid camera movements that span several shots for dramatic impact. For example, a zoom shot is cut to another shot and back to the old shot where the zooming is still on. Due to visible camera movements, those zoom shots are perceived as the same take; however, the differences between two images tends to be larger than the threshold set during the clustering as the zoom continues while the other shot is shown. Algorithm 4 outlines how clusters would be merged in this situation. Currently, fluid camera shots are manually identified to facilitate this step.

Algorithm 4 Merging Clusters (Fluid camera movements)
1. Search all cluster pairs ($C_1, C_2$) satisfying that the last shot $C_1$ is 2 shots before the first shot of $C_2$, i.e., $C_2[1] = C_1[m] = 2$ and $m + n \geq 4$. Goto 3.
2. If $C_1[m]$ and $C_2[1]$ are both classified ‘fluid’ and their difference is not too large merge $C_1$ and $C_2$.
3. Select the next cluster pair and goto 2, else stop.

Figure 3 shows how these methods have refined 6 raw clusters into 5 final clusters for a hypothetical scene.

4. EXPERIMENTAL RESULTS
We set up a data set consisting of 10 full-length movies of all major genres including action, horror, science fiction, thriller, fantasy, drama, and comedy. We have previously developed algorithms for extracting scene indices automatically. However, in this work we use scene indices that
are manually labeled. They reflect the ideal case of non-noisy input data. We also remove those scenes in which the extraction of takes is difficult and less useful. They are typically high-tempo action scenes with higher motion and shorter shot lengths. We also exclude ‘montage’ scenes without repeated shots from analysis as the ‘best match’ method always returns the perfect results for these scenes. Adjusted Rand Index (R^*), cluster recall (CR) and cluster precision (CP) [6] are used to measure the performance. It should be noted that the expected value for R^* is 0.

While groundtruthing, we use the following guidelines to decide if two shots belong to the same take: (a) Both shots must belong to the same scene; (b) the last frame of the first shot must have similar camera parameters (framing, angle, composition) as the first frame of the second shot; (c) special case with fluid camera movements: The filmmaker did indeed signal to the viewer that two shots are from the same take through continuous zooming. Table 1 shows different relations between a shot A and a shot B during the groundtruthing process. Each case is represented by its first and last frames. Case 1 shows two identical static shots. Case 2 describes two motion shots where continuity is maintained. Two shots of Case 3 are from a ‘fluid’ zooming. The distances of two shots in Case 4 are different. Two shots of the last case have different initial framings but then pan to the same framing.

Table 1: Do two shots belong to the same take?

<table>
<thead>
<tr>
<th>Shot A</th>
<th>Shot B</th>
<th>Same Take?</th>
</tr>
</thead>
<tbody>
<tr>
<td>first frm.</td>
<td>last frm.</td>
<td>first frm.</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>NO</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>NO</td>
</tr>
</tbody>
</table>

Figure 3: Cluster merging and cluster splitting.

Figure 4 shows the performance (R^*) on all movies for 4 clustering techniques (CL, GAL, MED, WARD) and different configurations regarding (a) dividing/not dividing the frame into 4 sub-blocks and (b) calculating the shot similarity as the average/maximum similarity of keyframes. The best results are obtained for GAL and WARD. It is also evident that the division of the image into sub-blocks and the ‘maximum’ approach offers better performance.

The performance with the best configuration and WARD method for individual movies is shown in Table 2. The first column shows the number of scenes being clustered in each movie after filtering. Lower results are from dark and/or dynamic movies such as The Mummy, Sleepy Hollow, 12 Monkeys, and The Siege, while better results are obtained for bright, dialogue-oriented films such as American Beauty, Chameleon, Truman Show, and Erin-Brockovich.

Table 2: Take extraction results.

<table>
<thead>
<tr>
<th>Movie</th>
<th>Scenes</th>
<th>CP</th>
<th>CR</th>
<th>R^*</th>
</tr>
</thead>
<tbody>
<tr>
<td>The 13th Floor</td>
<td>43</td>
<td>0.97</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>The Matrix</td>
<td>50</td>
<td>0.97</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>Sleepy Hollow</td>
<td>42</td>
<td>0.94</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>Erin Brockovich</td>
<td>62</td>
<td>0.99</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>12 Monkeys</td>
<td>29</td>
<td>0.95</td>
<td>0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>American Beauty</td>
<td>65</td>
<td>0.98</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>The Siege</td>
<td>38</td>
<td>0.93</td>
<td>0.92</td>
<td>0.85</td>
</tr>
<tr>
<td>Truman Show</td>
<td>43</td>
<td>0.96</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>Chameleon</td>
<td>28</td>
<td>0.99</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>The Mummy</td>
<td>23</td>
<td>0.92</td>
<td>0.88</td>
<td>0.77</td>
</tr>
<tr>
<td>Average</td>
<td>41.2</td>
<td>0.96</td>
<td>0.93</td>
<td>0.87</td>
</tr>
</tbody>
</table>

5. CONCLUSION

We have described our investigation of techniques for extracting film takes, a cinematic element with many useful applications. We combine traditional hierarchical clustering algorithms with three different postprocessing methods that handle different aspects of film editing. Our experimental results on 10 movies show the usefulness of dividing the frame into sub-blocks and measuring shot similarity as the maximum of keyframe similarities.

6. REFERENCES