# Discovering Semantics from Visualizations of Film Takes

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 this paper, we study the application of a scene structure visualizing technique called Double-Ring Take-Transition-Diagram (DR-TTD). This technique presents takes and their transitions during a film scene via nodes and edges of a 'graph' consisting of two rings as its backbone. We describe how certain filmic elements such as montage, centre/cutaway, dialogue, temporal flow, zone change, dramatic progression, shot association, scene introduction, scene resolution, master shot and editing orchestration can be identified from a scene through the signature arrangements of nodes and edges in the DR-TTD.*

## 1 Introduction

One problem facing current multimedia content management systems is the large gap between the rich meaning that users want when they query and browse media and the low level nature of content descriptions that we can actually compute. A serious need therefore exists to develop algorithms and technologies that can automatically annotate content and establish semantic connections between form and function, allowing users to access and navigate the indexed media in many interesting ways. Upon recognizing this problem, Dorai and Venkatesh [1] have proposed the *Computational Media Aesthetics* (CMA) framework for high-level content analysis of media. It is defined as "the algorithmic study of a variety of image and aural elements with insights from film grammar. It is also the computational analysis of the principles that have emerged underlying their manipulation of creative art of clarifying, intensifying and interpreting an event for audience." As seen in Figure 1, CMA advocates drawing guidance from media production principles, namely Film Grammar, for systematic analysis. It aims at offering the user high level semantics as intended by the filmmaker, both structural and expressive, in browsing, searching and navigating film and video documents.

The aim of this work is to study the application of the CMA framework to discover semantics embedded in the editing patterns of a film scene. Instead of a fully automatic approach, we investigate how meanings can be uncovered from signature arrangements of film takes and their transitions in a Double-Ring Take-Transition-Diagram (DR-TTD) [2].

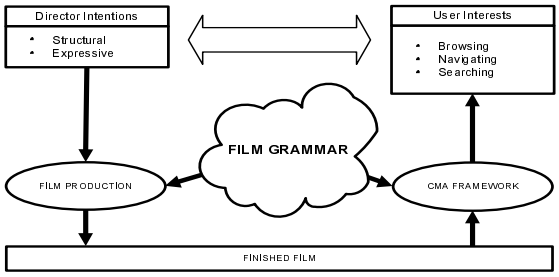



**Figure 1. The CMA framework.**

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[1] 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 in an edited film as generally understood in multimedia research literature. As seen in Figure 2, the filmmaker shoots many takes for a scene and during the film editing stage, different portions of selected takes are spliced together to produce the intended film scene. The term take used in this work literally means a set of 'edited' shots that belongs to the same production shot. We have 4 such takes in Figure 2, although 5 takes are produced during the film shooting.

Structurally, the take is the middle layer between the shot and the scene. This layer is rarely investigated but rich in semantics, as the arrangement of shots from takes shows the mediation level and narrative/dramatic intentions, in terms of editing, that the filmmaker applies on the film content. A signature arrangement in the DR-TTD exists for a given

[1] During film shooting, a (production) shot is a set of production takes and the notation "Shot X, Take Y" is used to distinguish between them.



Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

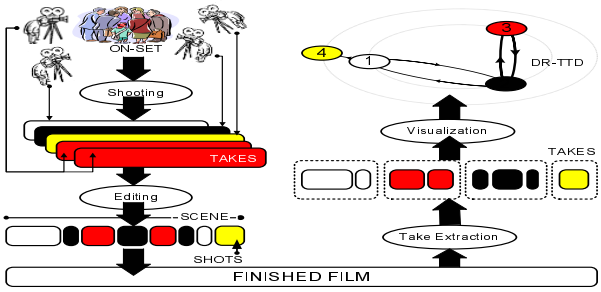



**Figure 2. Takes, shots, scenes and DR-TTD.**

semantics because (a) the filmmaker relies on film grammar and film syntax that guide how shots should be arranged to create certain dramatic impact or simply to avoid disorienting the viewer and (b) the construction of the DR-TTD is motivated from film grammar.

## 2 Previous work

An extracted take is essentially a cluster of 'identical' shots. Clustering of shots for the purpose of content browsing and presentation has been examined in [3] which clusters several visual features to create a hierarchical view of video content. Recently, we investigated the use of clustering to detect film scenes that are coherent in time/space or mood and have presented them in a Scene-Cluster Temporal Chart [4].

Shot clustering/grouping has been often used as an intermediate step in extracting scene boundaries [5, 6, 7]. These methods, therefore, do not demand that shots clustered together come from the same take, but from the same scene. They then use overlapping link reasoning to merge separated clusters into scenes. [6] proposes a technique called time-adaptive grouping to create a table-of-content for a video document. The authors attempt to incorporate shot length and shot activity into the shot similarity measure.

[8] studies the problem of mining video editing rules by performing row and column analysis on a matrix formed by shot indices and 3 shot attributes: distance, camera work and duration. [9] proposes a video editing support system that exploits film grammar related to shot size and camera work.

The DR-TTD visualization is based on the concept of Scene Transition Graph (STG) [5]. They both represent clusters and transitions among them via nodes and directed edges of a graph. However, a DR-TTD extensively exploits film grammar to express richer semantics and its purpose is not to detect the scene transitions (by searching for cut edges in the graph), but to show the internal structure of a scene and make certain semantics explicit.

Shots, scenes and takes serve as the main input for our visualization process. The extraction of shot/scene indices is a well documented problem and many solutions are provided in the literature. Following is the summary of the four-steps in our take extraction technique described in [10]:

1. Check if the scene is action-driven or drama-driven by examining its tempo characteristics. Discard the scene if it is action driven, as detecting takes for these kinds of scenes is difficult and less useful.
2. Compute the proximity matrix that measures the similarity between all pairs of shots in the scene.
3. Create shot groups by using a conventional clustering method and the proximity matrix.
4. Employ rules and conventions in film editing to merge and split clusters to further improve the results.

Our experimental results on 10 movies indicate that it is useful to divide the frame into sub-blocks and to measure shot similarity as the maximum of keyframe similarities. The performance is also better for dramatic-oriented, well edited films than action based films [10].

## 3 Visualizing take transitions via DR-TTD

In this section, we describe four important elements of a DR-TTD [2]: Node, edge, sequence and unit. Understanding these elements is essential in understanding the usefulness of a DR-TTD. The following notations are used: $\{x|\mathbf{Cond}(x)\}$ for the set of elements $x$, where condition $\mathbf{Cond}(x)$ is satisfied; and $|\mathbf{X}|$ for the number of elements in list/set/vector $\mathbf{X}$.

Let $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_n\}$ denote the shot sequence of a scene in temporal order, and $\mathbf{T} = \{\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_m\}$ denote the the set of takes extracted from this scene. We have $\mathbf{T}_i \subset \mathbf{S}$, $\mathbf{T}_i \cap \mathbf{T}_j = \oslash$, and $\mathbf{T}_1 \cup \mathbf{T}_2 \cup ... \cup \mathbf{T}_m = \mathbf{S}$. Note that each take is numbered according to the order of the first shot in the take. Hence, for $\mathbf{T}_i = \{\mathbf{S}_{i_1}, \mathbf{S}_{i_2}, ..., \mathbf{S}_{i_m}\}$ and $\mathbf{T}_j = \{\mathbf{S}_{j_1}, \mathbf{S}_{j_2}, ..., \mathbf{S}_{j_m}\}$, we have $i < j \iff i_1 < j_1$.

### 3.1 Node

Each take is represented by a node in DR-TTD. The backbone of a DR-TTD consists of two circles where all nodes are placed.

#### 3.1.1 Definitions

There are two kinds of nodes:

- $\mathcal{I}$-nodes: These nodes represent takes with at least two member shots. The set of all $\mathcal{I}$-nodes is denoted by $\mathcal{I}$. $\mathcal{I} = \{\mathbf{T}_i \mid \mathbf{T}_i \in \mathbf{T}, |\mathbf{T}_i| > 1\}$. Nodes of this kind are placed on the inner circle of the DR-TTD.
- $\mathcal{O}$-nodes: These nodes represent takes with only one member shot. The set of all $\mathcal{O}$-nodes is denoted by $\mathcal{O}$. $\mathcal{O} = \{\mathbf{T}_i \mid \mathbf{T}_i \in \mathbf{T}, |\mathbf{T}_i| = 1\}$. Nodes of this kind are placed on the outer circle of the DR-TTD.



Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

Although, the use of these two rings is semantically motivated, it also simplifies many visualization aspects. Nodes on the $\mathcal{O}$-ring have functions that include adding drama, excitement, and highlights to the story. They tend to contain information less important for understanding the story. Nodes on the $\mathcal{I}$-ring are important because they are repeated and carry the weight of the plot.

Sharff [11] describes eight cinesthetic elements that provide aesthetic gratification in film. Except orchestration and parallel action, other cinesthetic elements are linked to the ring indices of the nodes. Separation, familiar image and master shot principle relates to nodes on the $\mathcal{I}$-ring as shots used to construct these elements are repeated throughout the scene. Slow disclosure, moving camera and multiangularity are assembled from fragmented shots and relate to nodes on the $\mathcal{O}$-ring.

#### 3.1.2 Representation

Nodes are represented by circles and two smaller half circles are used to indicate if the take contains the first/last shot. It is useful to further incorporate the visual characteristics of a take into the appearance of its node. The first method is to compute the take features using every shot belonging to the take. The feature normalization framework discussed in [10] can be used. Let $\mathbf{F}_j$ denote the $j$-th frame in the video sequence and $\mathbf{F}_{i_1}$, $\mathbf{F}_{i_2}$, ..., $\mathbf{F}_{i_u}$ be $u$ $\mathcal{R}$-frames of shot $\mathbf{S}_i$, with $\mathbf{F}_{k_1}$ and $\mathbf{F}_{k_u}$ being the first and last frame of this shot. The color feature $\chi$ of this shot can be normalized from these $\mathcal{R}$-frames as follows:

$$\mathbf{S}_i^{\chi} = \frac{\sum_{k=1}^{u-1} \mathbf{F}_{i_k}^{\chi} . (i_{k+1} - i_k)}{\mathbf{S}_i^{\mathcal{L}}},$$

where $\mathbf{F}_{k_i}^{\chi}$ is the measurement of frame $k_i$. In this scheme, each $\mathcal{R}$-frame accounts for the visual content of the shot from its position to the position of the next $\mathcal{R}$-frame. Similarly, the feature $\chi$ of a take $\mathbf{T}_i$ comprising of consecutive shots $\mathbf{S}_{i_1}$, $\mathbf{S}_{i_2}$, ..., $\mathbf{S}_{i_t}$ can be computed as:

$$\mathbf{T}_i^{\chi} = \frac{\sum_{k=1}^{t} \mathbf{S}_{i_k}^{\chi} . \mathbf{S}_{i_k}^{\mathcal{L}}}{\sum_{k=1}^{t} \mathbf{S}_{i_t}^{\mathcal{L}}},$$

where $\mathbf{S}_i^{\mathcal{L}}$ is the duration of shot $\mathbf{S}_i$.

This method can be used to compute any feature (such as average color, histograms, etc). The feature can be used to "label" the node. An alternative is to "label" the node with the whole image representing that node. It is more appropriate to extract only one frame from the shot sequence to represent the take rather than combining many images into one. The use of iconic images is useful as they can give a user a rough idea about the take angle, distance and subjects. This method can proceed by first selecting a shot from all shots in the take and then selecting a representative frame ($\mathcal{R}$-frame) from the $\mathcal{R}$-frame list of the selected shot. The following three factors should be considered in selecting a shot:

- The distance from the shot to the centroid ($\mathcal{C}$) of all shots in the take. We approximate the distance by the average distance of a shot to other shots in the take. It is desirable to select the shot close to the centroid.

- The length of the shot ($\mathcal{L}$). It is important to select the shot that is most dominant for the take. Such shots are often indicated by their long duration.

- Motion level of the shot ($\mathcal{M}$). If the shot has a lot motion, it is maybe a transitional shot and thus not representative of the take. Hence, we prefer to select shots that have low motion.

Currently, we combine these factors linearly (after Gaussian-normalizing them) to select a shot to represent the take:

$$\mathcal{R}(\mathbf{T}_i) = (\mathbf{S}_{i_k} | 1 \leq k \leq t, \max(-\mathbf{S}_{i_k}^{\mathcal{C}} + \mathbf{S}_{i_k}^{\mathcal{L}} - \mathbf{S}_{i_k}^{\mathcal{M}}))$$

For a selected shot $\mathbf{S}_i$, we then select the most representative $\mathcal{R}$-frame from its $\mathcal{R}$-frame list by choosing the $\mathcal{R}$-frame that accounts for the longest duration of the shot:

$$\mathcal{R}(\mathbf{S}_i) = (\mathbf{F}_{i_k} | 1 \leq k < u, \max(i_{k+1} - i_k))$$

### 3.2 Edge

#### 3.2.1 Definition

In a STG, two nodes $\mathbf{T}_i$ and $\mathbf{T}_j$ are connected by a directed edge if there is an index $u$ such that $\mathbf{S}_u$ in $\mathbf{T}_i$ and $\mathbf{S}_{u+1}$ is in $\mathbf{T}_j$. We extend this by using the width of each edge to indicate how much interaction occurs between the two takes. The interaction level indicates whether the two shots are loosely or strongly tied as a semantic unit. The width $\mathcal{E}$ of an edge between two nodes $\mathbf{T}_i$ and $\mathbf{T}_j$ is calculated as:

$$\mathcal{E}(\mathbf{T}_i, \mathbf{T}_j) = |\{t | 1 \leq t \leq n-1, \mathbf{S}_t \in \mathbf{T}_i, \mathbf{S}_{t+1} \in \mathbf{T}_j\}|$$

An edge between them is claimed if and only if $\mathcal{E}(\mathbf{T}_i, \mathbf{T}_j) > 0$. Let $\mathbf{E}$ denote the set of all edges. $\mathbf{E}$ is comprised of 4 subsets: $\mathbf{E}_{\mathcal{I}\mathcal{I}}, \mathbf{E}_{\mathcal{I}\mathcal{O}}, \mathbf{E}_{\mathcal{O}\mathcal{I}}$ and $\mathbf{E}_{\mathcal{O}\mathcal{O}}$ where $\mathcal{I}$and $\mathcal{O}$are ring indices of the nodes.

#### 3.2.2 Representation

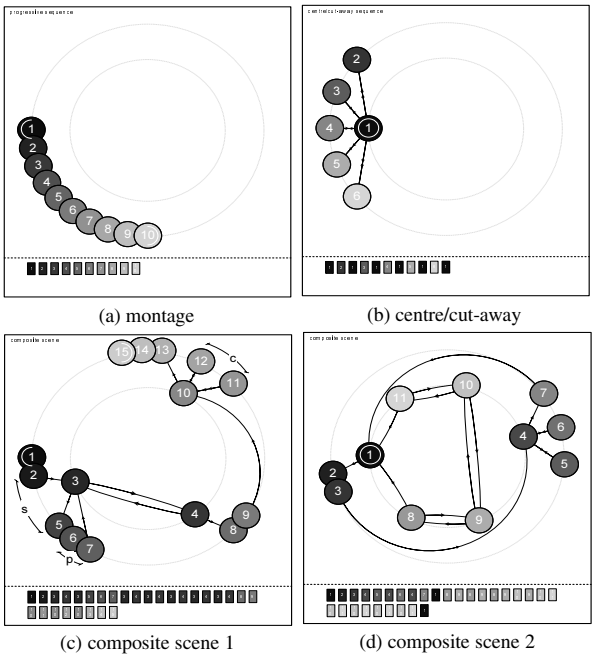
We represent different kinds of edges in a DR-TTD explicitly. An edge in $\mathbf{E}_{\mathcal{I}\mathcal{I}}$ is represented by an elliptical arc to show the interaction level of two $\mathcal{I}$-nodes. Circular arcs along the $\mathcal{O}$-ring are used for $\mathbf{E}_{\mathcal{O}\mathcal{O}}$-edges. $\mathbf{E}_{\mathcal{I}\mathcal{O}}$-edges are represented by a straight line. If an $\mathbf{E}_{\mathcal{O}\mathcal{I}}$ edge connects an $\mathcal{O}$-node with its linked $\mathcal{I}$-node (a link within an unit, see Step 2), a straight line is used, otherwise a circular arc is used (a link across units).



Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

### 3.3 Primitive sequences of $\mathcal{O}$-nodes

A non-dialogue sequence is often constructed from the following two primitive structural elements (See Figure 3). These elements can be deployed for the entire scene or combined together to construct the scene:



(a) montage (b) centre/cut-away

(c) composite scene 1 (d) composite scene 2

**Figure 3. Scene constructs and DR-TTDs.**

**Montage sequence:** In montage sequences, the filmmaker assembles non-repeated shots from multiple angles/time/places to create an unified dramatic concept. Each take therefore contains only one shot and is placed consecutively on the $\mathcal{O}$-ring (See Figure 3(a)).

**Centre-shot/Cut-away sequence:** A video sequence may be constructed by linking different actions to a central action. For example, in order to show a person looking around, the filmmaker shows the man's face repeatedly, following each shot by a shot of his current viewpoint. The DR-TTD of this sequence shows a $\mathcal{I}$-node connected to many $\mathcal{O}$-nodes (See Figure 3(b)).

One important property of a DR-TTD is that every $\mathcal{O}$-node is part of either a montage sequence or a centre sequence. For example, in Figures 3(c) and 3(d) (see [10] for its STG representation) which show two scenes that use more than one of the above two sub-constructs, the montage sequences are (1,2), (5,6,7), (8,9), (14,15) and (2,3) and the centre-shot sequences are (11,12,13) and (5,6,7). All $\mathcal{O}$-nodes in a given sequence are arranged in a anti-clockwise direction.

### 3.4 Unit

$\mathcal{O}$-node sequences can be linked backward directly or indirectly to $\mathcal{I}$-nodes. The exception is the first sequence that contains the starting shot, which can link forward to a $\mathcal{I}$-node. This linking forms an unit in a DR-TTD. We often see that the content of a $\mathcal{I}$-node is related to the content of its linked sequences. In Figure 3(c), the links are from sequence (1,2) (forward) to node 3, (5,6,7) to node 3, (8,9) to node 4, (11,12,13) to node 10 and (14,15) indirectly to node 10, whilst in Figure 3(d), the links are from sequence (2,3) to node 1 and (5,6,7) to node 4. In this way, each O-node sequence can be assigned to one and only one $\mathcal{I}$-node. For each $\mathcal{I}$-node, all its associated sequences are arranged in an anti-clockwise direction; the ordering also reflects their temporal order in the video sequence. An $\mathcal{I}$-node not linked to any $\mathcal{O}$-node forms a unit of itself. In Figures 3(c) and 3(d), we have 3 and 6 such units respectively. In Figure 3(d), $\mathcal{I}$-nodes 8,9,19,11 are units by themselves.

### 3.5 DR-TTD Construction

A DR-TTD is constructed via the following procedure that aims at achieving as much as possible the clarity and anti-clockwise temporal flow of all takes [2] in the diagram:

1. Number takes according to their temporal order.
2. Create sequences and units.
3. Find the orderings of $\mathcal{I}$-nodes that minimize number of crossings among $\mathcal{II}$-edges.
4. Among these orderings, find those minimizing number of crossings among $\mathcal{IO}$-edges and $\mathcal{OI}$-edges.
5. Find one ordering that most likely keeps the $\mathcal{I}$-nodes in an anti-clockwise direction.

All $\mathcal{O}$-nodes are placed on the $\mathcal{O}$-ring based on three spacing parameters (p, c, s) for spacing between two progressive nodes, two $\mathcal{O}$-nodes of the same centre sequence and two $\mathcal{O}$-node sequences linked to the same $\mathcal{I}$-node (see Figure 3(c)). $\mathcal{I}$-nodes are placed relative to its unit.

## 4 Recognizing film semantics from DR-TTDs

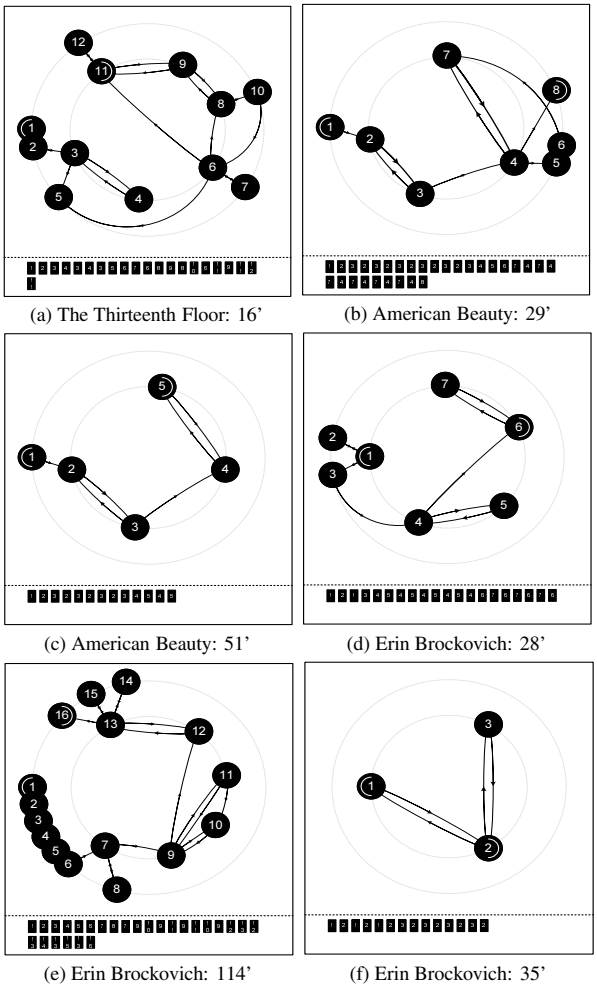
In this section, we describe various film semantics and show how they can be recognized from a DR-TTD. This discussion is based on the study of film grammar and detailed examination of the DR-TTDs for hundreds of scenes from 10 movies of all major genres including American Beauty, The Matrix, Truman Show, The Siege, 12 Monkeys, The Mummy, The 13th Floor, Sleepy Hollow, Chameleon and Erin Brockovich. Throughout the section, we will use different real film scenes as examples. Each example is annotated with its movie name and its time index (min). Unfortunately, we are unable to present representative images of the takes that are described in Section 3.1.2.



Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

### 4.1 Zone moving and dramatic progression

The location of a scene maybe divided into smaller zones. During the scene narrative, characters may move from zone to zone or the dramatic focus may change to a different group of characters in the scene. Filmmakers have different camera set-ups for different zones, resulting in different take sets. In addition, the editing moves forward temporally. Therefore, we have a cut edge[2] in DR-TTD when zone-to-zone transitions are made. We are primarily interested in cut edges that are also $\mathcal{II}$-edges or $\mathcal{OI}$-edges, since they connect different units of the scene. In addition, when a cut edge is an $\mathcal{OI}$-edge, we can generally conclude that the montage sequence linked to the starting node shows the transition between zones. Otherwise, the movement is often indicated in the last $\mathcal{I}$-node of the first zone.



(a) The Thirteenth Floor: 16' (b) American Beauty: 29'

(c) American Beauty: 51' (d) Erin Brockovich: 28'

(e) Erin Brockovich: 114' (f) Erin Brockovich: 35'

**Figure 4. Zone/drama change examples.**

[2]An edge of a graph is a cut-edge if its deletion increases the number of components in the graph

Figure 4(a) shows a scene in The Thirteenth Floor where detective Larry McBain meets Jason. They first talk outside, then move into the computer lab resulting in cut edge 5—6. Zone movement is shown in Take 5. The party scene of American Beauty is shown in Figure 4(b). Two zones, one showing the conversation between Carolyn and Buddy and the other showing the conversation between Ricky and Lester, are separated by cut edge 4—5.

Zone changes are not the only narrative event that triggers a cut edge. It may also arise from the dramatic progression of the scene. The most common device for increasing/decreasing the dramatic emphasis is varying shot sizes (also called shot distance). The filmmaker may start all shots as medium shots, then later change to close-ups to indicate the drama has heightened. For example, the dinner scene between Carolyn and Buddy shown in Figure 4(c) has a cut edge 3—4 separating medium shots and close-ups. The close-ups indicates increased intimacy toward the end of the scene. The zone change and dramatic progression may occur together in one single scene as seen in Figure 4(d) and 4(e). Figure 4(d) shows a scene in Erin Brockovich in which Erin first meets Donna Jensen at the door, and then moves to the lounge room, causing cut edge 3—4. As the conversation becomes more dramatic toward the end, close-ups shots are used and are separated from early medium shots via cut edge 4—6. Figure 4(e) shows a similar scene between these two characters occurring later in the film, the meeting outside is separated with the conversation inside via cut edge 7—9, while cut edge 9—12 indicates the increased drama in their conversation. Note that if the emotion changes in only one character, the DR-TTD would contain a triangle with an missing edge. Figure 4(f) shows one such scene in Erin Brockovich when Erin talks to Ed about being fired. Ed (Take 2) is calm for the entire scene while Erin's emotion is intensified from medium shots (Take 1) to close-ups (Take 3).

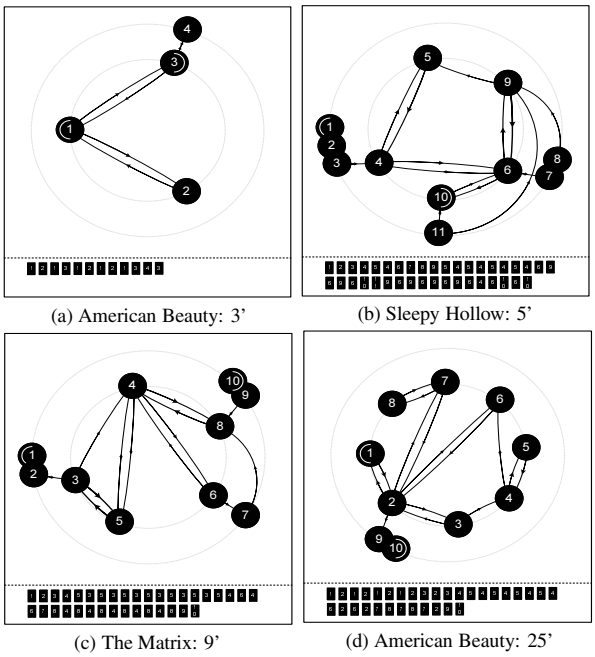
### 4.2 Shot association

As an edge indicates an interaction between two nodes, its absence may be meaningful too. If two nodes are linked via another node, we say they are 'indirectly' linked. The indirect link is a device for changing from one shot to another in which the same character is captured with a different camera setup. A direct link between two such nodes may result in discontinuity. Also, in dialogue sequences, dramatic emphasis may vary for only one character through a different camera setup, while others remain the same (e.g., the last example of the previous section). Therefore, if we have three $\mathcal{I}$-nodes heavily connected on two sides, with no edges on the remaining side, we can generally assume that the two unconnected $\mathcal{I}$-nodes are takes of the same character at different camera angles/distance.

For example, missing link 2—3 in an American Beauty scene where the Burnham family go to work (Figure 5(a)) indicates two takes of the house pass-way at long (Take



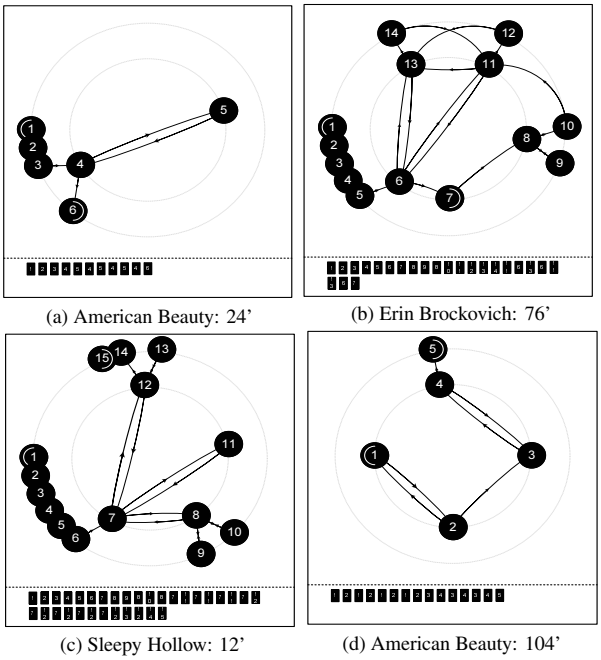
Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE



(a) American Beauty: 3’ (b) Sleepy Hollow: 5’

(c) The Matrix: 9’ (d) American Beauty: 25’

**Figure 5. Shot association examples.**



(a) American Beauty: 24’ (b) Erin Brockovich: 76’

(c) Sleepy Hollow: 12’ (d) American Beauty: 104’

**Figure 6. Introduction & resolution examples**

2) and medium-long shots (Take 3). The latter allows a closer view of characters’ faces and actions. Figure 5(b) shows the courtroom scene in Sleepy Hollow, missing edge 5—6 relates to takes 5 and 6 being the medium and close-up shots of Constable Crane. Missing edges 4—9, 4—10 and 9—10 relate to takes 4, 9 and 10 being the medium, profile medium and medium close-up shots of the judge. In The Matrix, when Neo meets Trinity at the night club (Figure 5(c)), takes 5, 6 and 8 are shots of Trinity that relate to the missing edges. Figure 5(d) shows the scene in American Beauty in which Ricky meets Jane at school. Missing edges 3—6, 3—7 and 6—7 link to takes 3, 6 and 7 being long, medium and close-up shots of Ricky, while missing edge 2—8 links to takes 2 and 8 being medium and medium close-up shots of Jane.

The confidence of the missing edge inference depends on the number and the size of indirect links. For example, it is more likely to be true for missing edge 4—9 than for 4—10 in Figure 5(b). Also note that the missing links in a scene involving many characters may merely indicate there is no narrative conflict involvement/interaction between them. For example, in Figure 5(d), two school girls in take 1 have left the scene when Ricky arrives, resulting in missing edges 1—3, 1—6 and 1—7.

### 4.3 Scene introduction and resolution

The existence of montage sequences at the start and/or the end of a scene indicates that the filmmaker has visually staged the introduction and/or resolution to the main narrative in the scene. The introduction to a scene is achieved by showing: (a) location and background activity, (b) relative positions between characters and (c) the current action of a character in close-ups. The resolution sequence at the end of a scene often provides: (a) links to actions in the next scenes and (b) gradual conclusion of the scene.

For example, Figure 6(a) shows the scene where Ricky talks to his dad in the car. The first 3 shots introduce the scene by showing the street, a hand over a notebook, and a shot of both Ricky and his dad. The resolution shot (Take 6) reveals what is written in Ricky’s book, indicating the fact that he is a drug dealer. The court hearing scene in Erin Brockovich (Figure 6(b)) shows a gradual approach to the main event. The first 5 shots show the parking lot, gate sign, hallway and wide shots of the whole court room from the back and front. Figure 6(c) shows a scene in Sleepy Hollow. There are various shots (Takes 1-6) of party activities preceding the meeting between Constable Crane and the Van Tassel family (Takes 7, 8, 11, 12), the final two shots introducing new characters into the story and connect to the next scene. The last conversation between Lester and Angela (Figure 6(d)) is concluded by Lester walking away (Take 5), linking to the next scene where he sits in a room.

The absence of introduction sequences indicates that the filmmaker would like to have the spectator primarily attending to the action (e.g., the dialogue between Lester and Angela in Figure 6(d)). Also, the scene may not resolve visually but through dialogue or characters’ expressions. For example, Erin smiles at the outcome of the hearing (Take 7, Figure 6(b)).



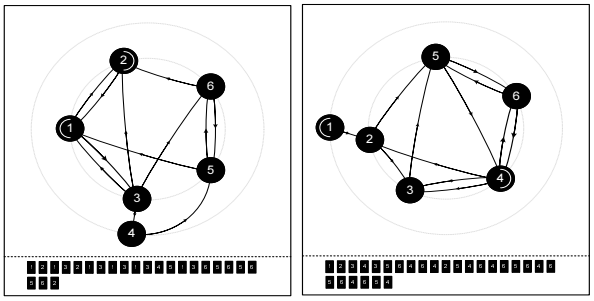
Proceedings of the 10th International Multimedia Modelling Conference (MMM’04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

### 4.4 Master shot

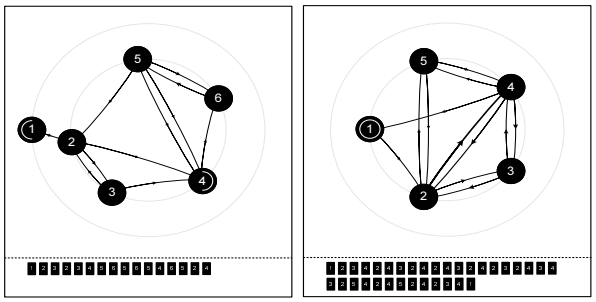
In classical Hollywood style, a master shot is a filmic recording of an entire scene, from start to finish, and taken from an angle that keeps all the players in view. It is ordinarily supplemented with other shots such as close-ups of individuals. It establishes an objective and stable perspective on a given situation [11]. The master shot gives a broad view and occurs at least twice in the course of a scene. The master shot is often interleaved with shot-reverse-shot sequences and the transition between a master shot to close shots of any individual character is natural and not disorientating. Therefore, in a DR-TTD a master shot often links with many takes of shot-reverse-shot configurations.

For example, in Figure 7(a), Take 2 of the scene in Sleepy Hollow where Constable Crane starts performing autopsy is a master shot which shows the table and all characters. It links with shot-reverse-shot sequences 1—3 and 5—6. Figure 7(b) shows the scene outside the basketball court in American Beauty with Take 2 being the master shot showing all characters talking in a group. It is also linked with other shots in shot-reverse-shot sequences. Similarly, the master shot of the scene where the Fitts family are having breakfast (Figure 7(c)) is Take 4.

Many master shots may occur within a single scene as shown by Takes 1 and 5 in the dinner scene of the Burnham family in American Beauty (Figure 7(d)). They show the dinner table in long and medium long shots. The conclusion of a take being a master shot is reinforced if it occurs at the start/end of the scene (see Figure 7(a), 7(c) and 7(d)).



(a) Sleepy Hollow: 27’ (b) American Beauty: 16’



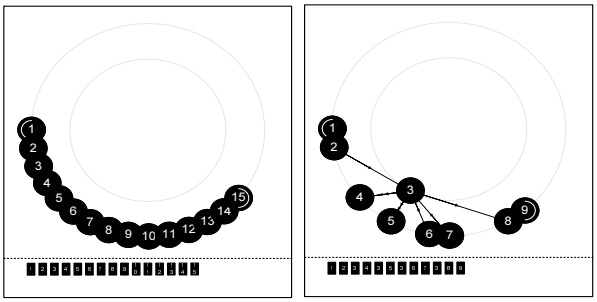
(c) American Beauty: 23’ (d) American Beauty: 62’

**Figure 7. Master shot xamples.**

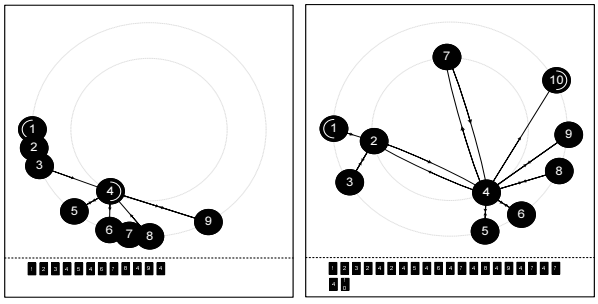
### 4.5 Non-dialogue narration & familiar images

Our analysis so far has mainly focused on dialogue-based scenes. These scenes are characterized by dominant shot-reverse-shot sequences (two $\mathcal{I}$-nodes linked by two thick edges in DR-TTDs). However, drama-driven scenes do not necessarily contain dialogues. One way to construct such scenes is to use a series of progressing shots (montage sequence). These shots, each showing a fragmented part of the scene, are unified in dramatic incidence. For example, in order to create the scene where people are looking for Truman in The Truman Show (Figure 8(a)) the filmmaker uses crowd shots of different street corners at different angles and distances.

Any picture that reappears in a film with approximately the same framing and composition is called familiar image. It is one of the devices for linking fragmented shots of a scene. The familiar image plays the role of a pivotal image around which a scene or part of a scene is constructed [11]. In the DR-TTD, familiar images show up as centre shots. For example, the scene where Constable Crane travels to the town in Sleepy Hollow is constructed through this device (Figure 8(b)). The shot of Constable Crane sitting in the horse carriage (Take 3) links all images of the road, forest, etc. Figure 8(c) shows a scene in The Mummy where High Priest Imhotep tries to use Evelyn to resurrect his lover and the shot of Evelyn tied on a table (Take 4) is used as the familiar image. At a briefing in The Siege, shown in Figure 8(d), the shot of Agent Hubbard (Take 4) is the centre of action that links various shots of people listening.



(a) The Truman Show: 75’ (b) Sleepy Hollow: 8’



(c) The Mummy: 97’ (d) The Siege: 25’

**Figure 8. Non-dialogue narration examples.**

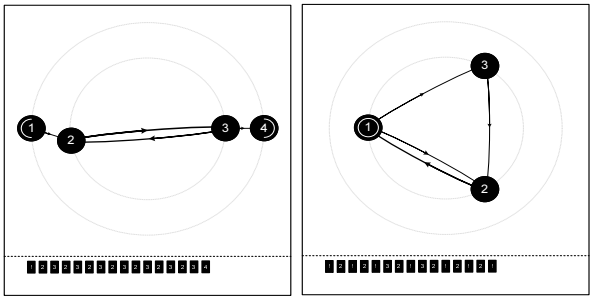


Proceedings of the 10th International Multimedia Modelling Conference (MMM’04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE

### 4.6 Orchestration in film editing

[11] defines orchestration as the arrangement of various elements of structure throughout a scene or entire film, which includes the symmetry in the editing pattern for a film scene. This can be quickly recognized through the symmetry of the DR-TTD itself.

For example, Figure 9(a) shows a simple scene (dinner between Agent Hubbard and Elise) in The Siege symmetric via shot-reverse-shot (Takes 2,3) pattern. Another scene in The Siege showing the meeting between Elise, Agent Hubbard and General Devereaux (Figure 9(b)) also has a symmetric editing pattern around Take 3, the medium shot of Agent Hubbard. Figure 9(c) shows the phone conversation between Erin and George in Erin Brockovich. Each take is used with rhythmic frequency and creates the symmetry in its DR-TTD. The scene in American Beauty when Lester and Ricky are smoking pots (Figure 9(d)) has a symmetry around Take 2, a two-shot take of Ricky and Lester. It also starts and ends with the same take.



(a) The Siege: 33' (b) The Siege: 44'

(c) Erin Brockovich: 66' (d) American Beauty: 32'

**Figure 9. Editing orchestration examples.**

## 5 Conclusions

We have studied the application of DR-TTD representation in uncovering semantics embedded in the editing patterns of a film scene. We have described DR-TTD signatures for the following elements:

- **zone change and dramatic progression** : cut edges.
- **shot association** : missing edges of a triangle.
- **visual introduction/resolution** : montage sequences at the start/end of the scene.
- **master shot** : $\mathcal{I}$-nodes link with shot-reverse-shot sequences via thin edges.
- **non-dialogue narration and familiar images** : montage sequence, and centre shots/$\mathcal{I}$-nodes linking to many $\mathcal{O}$-nodes.
- **editing orchestration** : symmetry of DR-TTDs.

Above applications of DR-TTD should not only be understood from content analysis and annotation perspective. DR-TTD can also aid in the authoring of film and video content. For example, an amateur filmmakers can use DR-TTDs to verify if the editing of a scene has matched their dramatic/narrative intentions and adjust it accordingly.

## References


[1] Chitra Dorai and Svetha Venkatesh, "Computational Media Aesthetics: Finding meaning beautiful," *IEEE Multimedia*, vol. 8, no. 4, pp. 10–12, October-December 2001.

[2] Ba Tu Truong, S. Venkatesh, and C. Dorai, "DR-TTD: A technique for visualizing film editing patterns," Tech. Rep., School of Computing, Curtin University of Technology, Perth, Western Ausralia, May 2003.

[3] D. Zhong, H.J. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," in *Storage and Retrieval for Still Image and Video Databases IV*, 1996, pp. 239–246.

[4] B. T. Truong, S. Venkatesh, and C. Dorai, "Application of computational media aesthetics methodology to extracting color semantics in film," in *ACMMM'02*, France Les Pins, Oct 2002.

[5] M. Yeung, B-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 7, no. 1, pp. 94–109, July 1998.

[6] Yong Rui, Thomas S. Huang, and Mehrotra S., "Constructing table-of-content for videos," *ACM Multimedia System Journal: Special Issue in Multimedia Systems on Video Libraries*, vol. 7, no. 5, pp. 359–368, 1999.

[7] Emmanuel Veneau, Re'mi Ronfard, and Patrick Bouthemy, "From video shot clustering to sequence segmentation," in *ICPR'00*, Barcelona, sep 2000, vol. 4, pp. 254–257.

[8] Yuya Matsuo, Miki Amano, and Kuniaki Uehara, "Mining video editing rules in video streams," in *ACMMM'02*, France Les Pins, Oct 2002.

[9] Masahito Kumano, Yasuo Ariki, Miki Amano, Kuniaki Uehara, Kenji Shunto, and Kiyoshi Tsukada, "Video editing support system based on video grammar and content analysis," in *ICPR'02*, 2002, pp. 1031–1036.

[10] Ba Tu Truong, S. Venkatesh, and C. Dorai, "Idenfitying film takes for cinematic analysis," in *ICME'03*, Baltimore, 2003, vol. 2, pp. 405–408.

[11] Stefan Sharff, *The elements of Cinema: Towards a cinesthetic impact*, Columbia Uni. Press, New York, 1982.




Proceedings of the 10th International Multimedia Modelling Conference (MMM'04)
0-7695-2084-7/04 $ 20.00 © 2004 IEEE