RoleNet: Movie Analysis from the Perspective of Social Networks

Chung-Yi Weng, Wei-Ta Chu, *Member, IEEE*, and Ja-Ling Wu, *Fellow, IEEE*

***Abstract*—With the idea of social network analysis, we propose a novel way to analyze movie videos from the perspective of social relationships rather than audiovisual features. To appropriately describe role's relationships in movies, we devise a method to quantify relations and construct role's social networks, called RoleNet. Based on RoleNet, we are able to perform semantic analysis that goes beyond conventional feature-based approaches. In this work, social relations between roles are used to be the context information of video scenes, and leading roles and the corresponding communities can be automatically determined. The results of community identification provide new alternatives in media management and browsing. Moreover, by describing video scenes with role's context, social-relation-based story segmentation method is developed to pave a new way for this widely-studied topic. Experimental results show the effectiveness of leading role determination and community identification. We also demonstrate that the social-based story segmentation approach works much better than the conventional tempo-based method. Finally, we give extensive discussions and state that the proposed ideas provide insights into context-based video analysis.**

***Index Terms*—Community analysis, movie understanding, social network analysis, story segmentation.**

I. Introduction

THE flourishing movie industries produce more than 4500 movies every year. With the advance of digital technologies, movies are produced or disseminated digitally, and seeing movies has been one of the most popular entertainments. Explosive amounts of movie data not only impede efficient storage or dissemination but also burden users in information access. Therefore, techniques of automatic movie organization and indexing are urgently needed.

Over the past decade, researches on movie analysis attempt to solve the most notorious problem—the semantic gap. However, it seems that approaches based on audiovisual features face an unbreakable impediment. From the research trend of movie video analysis, we found that these studies come from "frame-level" analysis, which is based on shot change detection and keyframe selection [29], [30], to "event-level" analysis, which further considers the temporal context or objects in the scenes and achieves the detection of some important events such as dialog [27] and gunplay [28]. However, when we watch a movie, what we really see are the "stories" derived from the action or interaction between characters. Humans don't care about how shots change or whether a scene is a dialog in watching movies. Therefore, we argue that a movie analysis system should be advanced to "story-level" analysis. Easily accessing specific events may be beneficial to professional editors, but easily accessing stories is beneficial to all the audience.

From the literature of filmmaking, we found that space arrangement is an important factor to build stories. A director arranges people and objects in a 3-D space and transforms real entities into two-dimensional images via cameras, based on the guidelines of graphic arts or theatre arrangement. This process is called "mise en scene" [33]. Characters enter the same space and interact with each other to narrate a story segment. After this story segment ends, the camera switches to another space, where the same characters or other characters interact to start a new story segment. Usually, each scene consists of many shots. Each shot forms a unit of space arrangement, and each scene forms a unit of story segment. The relationships between characters are constructed based on their interaction in scenes [34].

Fig. 1 shows an example of mise en scene, which are snapshots of a scene in "My Blueberry Night." After the character A enters the space, she interacts with character B. The camera also captures the action of character C, who is put on the side and never talks to others throughout this scene. However, the audience can easily perceive that these three characters have some kind of relationship through this arrangement. Therefore, we can see that how characters arranged in the same space and the relationship between them are very important in narrating stories. In this work, we propose a story-level analysis system based on the social relationships between characters. Mutual relations between roles rather than audiovisual features are extracted and modeled to facilitate movie understanding.

The idea of this work originates from *social network analysis* (SNA) [11], which is one of the research fields in social science. In social science, interactions between entities are modeled as a complex network, and the techniques of SNA are designed to discover hidden structures/properties that cannot be directly perceived or measured by people. The ideas have been widely applied with success to topics about Internet structuring, human

Manuscript received December 13, 2007; revised October 06, 2008. Current version published January 16, 2009. This work was supported in part by the National Science Council of Taiwan under Grants NSC97-2221-E-194-050, NSC 97-2221-E-002-181-MY3, NSC 97-2622-E-002-010-CC2, and NSC 96-2622-E-002-002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

C.-Y. Weng and J.–L. Wu are with the Deartment of Computer Science and Information Engineering, Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 106, Taiwan (e-mail: chunye@cmlab.csie.ntu.edu.tw; wjl@cmlab.csie.ntu.edu.tw).

W.-T. Chu is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chia-Yi 621, Taiwan (e-mail: wtchu@cs.ccu.edu.tw).

Digital Object Identifier 10.1109/TMM.2008.2009684

1520-9210/$25.00 © 2009 IEEE

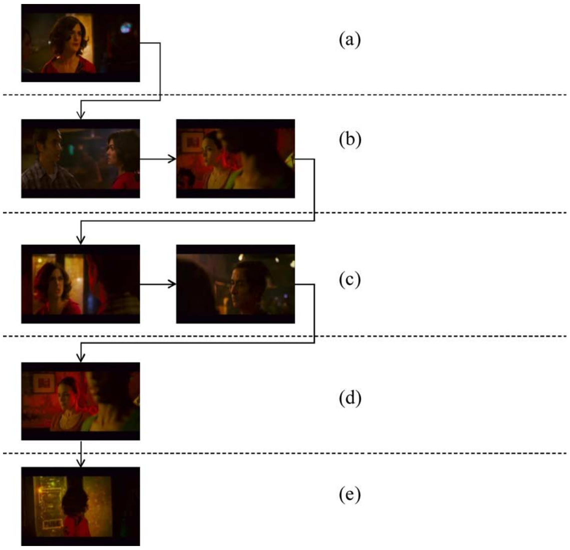


Fig. 1. Example of mise en scene. (a) Character A enters the space. (b) Character A talks to character B. Character C is put on the side to construct some relationship with them. (c) Character A argues with character B. (d) Character C is still put on the side. (e) Character A leaves the space.

interactions [12], epidemiology, ecosystems [13], and etc. Essentially, the gap between direct observations and hidden natures in social science is similar to the semantic gap in multimedia understanding.

In this work, we try to model relationships between computational observations as a complex network and introduce SNA techniques to discover hidden semantic information in movies. Humans understand the stories conveyed by a movie because they learn the mutual relations between characters (roles)[1]. How roles interact or conflict leads a story. Therefore, we treat a movie as a small society, which is constructed by roles and their interactions. We model the relationships between roles as a role's social network [20], [21], which is named as RoleNet. Based on RoleNet, we propose several SNA algorithms to discover hidden semantics. It's believed that the proposed method provides a novel viewpoint to analyze movies and is beneficial to bridge the semantic gap.

The contributions of this work are summarized as follows.

- The idea of SNA to do movie analysis: We elaborately introduce the concept of SNA to conduct semantic movie analysis. We realize the idea and practically bridge computational observations and the hidden semantic information.
- An approach to model roles' interrelationship as a network: We explicitly address how to evaluate roles' interrelationships and transform them into a network. The proposed construction methods are general enough to various types of movies.
- Novel algorithms to analyze social relationships in movies: Based on RoleNet, several algorithms are designed to achieve leading roles determination and community identification. They are social characteristics existing in movie videos, and are good clues to approach movie understanding.
- A social-relation-based story segmentation method: Instead of describing video scenes by audiovisual features, we represent scenes by role's context and devise a method to evaluate the progress of stories. We compare the proposed method with conventional approaches and demonstrate its effectiveness.

The rest of this paper is organized as follows. Section II provides a brief survey on movie video analysis. We also address the novelty of our work after describing other social-based methods. In Section III, we describe how to model roles' interrelationship and how to construct the RoleNet. Based on RoleNet, we perform community analysis for a pilot instance, i.e., the so-called bilateral movies, and further propose a generic model applicable to various types of movies in Section IV. In Section V, we describe video scenes based on role's context and propose a social-based approach for story segmentation. Section VI describes the evaluation results on community analysis and story segmentation. Some discussions and limitations of current works are stated in Section VII. Finally, the concluding remarks are given in Section VIII.

II. Related Work

Many studies have been proposed to analyze movies based on audiovisual features. They can be roughly categorized into the following categories: genre classification, story segmentation, and video abstraction. Rasheed *et al.* [1] exploited color, motion, and shot information to classify movies into comedies, action, dramas, or horror films. Adams *et al.* [2] evaluated video tempo on the basis of shot change frequency and motion information. For story segmentation, the idea of logical story units (LSU) [3] was proposed. An LSU contains a series of shots that convey a solid semantic meaning. Moreover, various video abstraction techniques [4]–[6] have been proposed to represent movie content in a compact manner, such as automatic summarization for action movies [7], [8].

Some studies were conducted for the so-called "affective content analysis." These works investigate human's perception drawn by audiovisual stimuli [9], [10]. On the basis of the knowledge from cinematography and psychology, human's emotion or affection is described by computational models. Stimuli derived from audiovisual features are still the focus of modeling.

Recently, few studies have been conducted to perform multimedia content analysis based on SNA. Vinciarelli *et al.* [22] is one of the first few researchers who investigate the usage of social relationship in segmenting radio programs. Radio programs with specific structure, i.e., each program has a "news" part and a "talk show" part, were processed. By identifying the occurrence of two anchormen and the end of first anchorman's talk, this work reported promising results in segmenting two parts of programs. They further extended their work to perform generic news story segmentation [23]. Similarly, they recognize each

[1] According to the definitions in Webster's dictionary, a character is "an imaginary person represented in a work of film", and a role is the "normal or customary activity of a person in a particular social setting." Actually, a character may play different roles according to different social situations. Using the word "character" is more precise. However, to concisely describe the proposed approach, i.e., RoleNet, we use character and role interchangeably in this work.

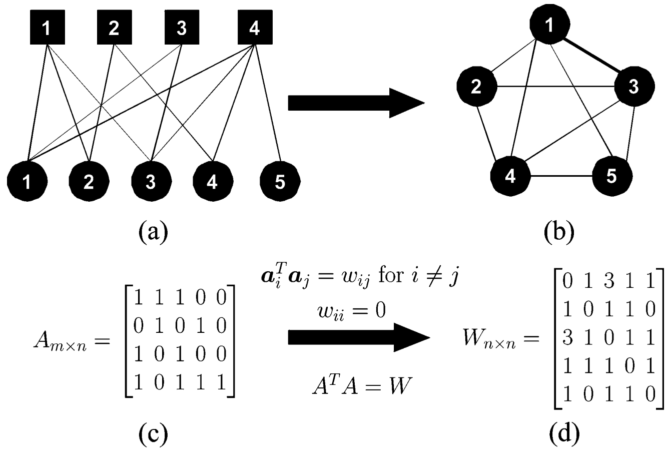

1
2
3
4
1
2
3
4
5
(a)
1
2
3
4
5
(b)
$A_{m\times n} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{bmatrix}$
$\boldsymbol{a}_i^T \boldsymbol{a}_j = w_{ij}$ for $i \neq j$
$w_{ii} = 0$
$A^T A = W$
$W_{n\times n} = \begin{bmatrix} 0 & 1 & 3 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 3 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$
(c)
(d)

Fig. 2. Graphical example to show the relationships between roles.

actor as in a "story" or as an "anchorman" from audio information, and then segment news programs into stories.

Another interesting work concerning social relationships was proposed in [24]. By using features of speech behavior, interaction, and topics in group meeting videos, Rienks *et al.* analyzed the influence of each participant. They recognize which participant(s) are more influential than others, which is also one of the tasks in our work though the targeted media are movies. With audiovisual features such as speaking length and motion activity, Hung *et al.* [36] estimate the most dominant person in a group meeting. Also for meeting recordings, Garg *et al.* [35] perform role recognition based on lexical information and social network analysis.

The works described above either developed techniques of social network analysis or focused on social signals like speaker interaction. In these years, researchers try to bring social intelligence [37] into computer systems and achieve intelligent analysis. These works can be categorized into a growing research domain called social signal processing [38]. Complete introduction of this domain is beyond the scope of this paper, but a recent literature survey can be found in [39].

We proposed an SNA-based approach to analyze movies in [20]. Leading roles and corresponding communities can be automatically identified by checking the social relationships between characters. However, this approach can only be applied to a specific type of movie, say bilateral movies (see Section IV for detailed descriptions). In [21], we extended this approach to various kinds of movies. The number of leading roles and hierarchy of communities are automatically determined. In this paper, we further exploit social relationships to describe video scenes, and develop a story segmentation module that works from a different perspective from conventional approaches. We report comprehensive performance evaluation of leading role determination, community identification, and story segmentation for different types of movies and TV shows.

As compared to the works in [23], the novelty of our work is twofold: 1) more elaborate graph-based analysis and 2) story segmentation based on role's social context. Both works represent social relationship as a graph. However, Vinciarelli and Favre took how an actor appears in different news segments a feature vector, and identified each feature vector as representing an anchorman or an actor in a story. In our work, we further exploit the correlation between characters and perform leading role and community identification based on the graph. For the data with relatively simple structure, they can achieve story segmentation by identifying the property of each actor, i.e., a story role or an anchorman role. In our work, we determine the boundaries of stories based on the changes of mutual relationship shown in different video scenes. Simply identifying whether a character appears in a scene doesn't fulfill the need of movie story segmentation. We have to detect tuning points of stories, in which characters in different story segments have significantly different social context.

III. RoleNet

A. Definition of RoleNet

A model that is suitable to describe roles' relationship should possess the following characteristics.

- Representing relationships effectively: There are many roles in a movie, and relationships among them are often intricate. In addition, closeness between different pairs of roles varies. How to effectively represent these characteristics is the first key.
- Facilitating systematic analysis: We would like to design algorithms to automatically analyze these intricate relationships. Therefore, the devised model should be structurally well-defined and is able to facilitate systematic analysis.

With these requirements, a RoleNet is defined as follows.

Definition: A RoleNet is a weighted graph expressed by

$$G = \langle V, E, W \rangle$$

where $V = \{v_1, v_2, \ldots, v_n\}$ represents the set of roles in a movie, $E = \{e_{ij} | \text{if } v_i \text{ and } v_j \text{ have relationship}\}$, and the element w_{ij} in W represents the strength of the relationship between v_i and v_j .

To construct a RoleNet, we have to address how to quantify the "relationship" between roles, i.e., w_{ij}. In this work, the relationship between roles is developed when they interact with each other. More often two roles appear in the same scenes, more chances they can interact, and closer relationship is built between them. Therefore, we can quantify roles' relationship as the number of co-occurrence between roles.

B. Construction of RoleNet

At the first step of RoleNet construction, a movie is viewed as a bipartite graph (c.f. Fig. 2(a)). The square nodes denote scenes, and the circular nodes denote roles. The edge between the ith square node and the jth circular node represents that the jth role appears in the ith scene. For a movie that consists of m scenes and n different roles, we can express the status of occurrence by a matrix $A = [a_{ij}]_{m\times n}$, where the element

$$a_{ij} = \begin{cases} 1, & \text{if the } j\text{th role appears in the } i\text{th scene,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The matrix presentation of Fig. 2(a) is shown in Fig. 2(c). More specifically, the jth column vector, $\boldsymbol{a}_j = (a_{1_j}, a_{2_j}, \ldots, a_{m_j})$, of A denotes the scenes where the jth role appeared. Based on this occurrence matrix, we can identify the co-occurrence of the ith role and the jth role by

$$w_{ij} = \sum_{k=1}^{m} a_{ki}a_{kj} = \boldsymbol{a}_i^T\boldsymbol{a}_j, \text{for } i \neq j. \qquad (2)$$

The value of w_{ij} is actually the inner product of $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$. This measurement can be generalized to the whole matrix. The co-occurrence status among roles in a movie can be expressed by

$$W_{n\times n} = A^T A \qquad (3)$$

in which w_{ij} is especially set as 0 when $i = j$.

In the example of Fig. 2, the co-occurrence status among roles is expressed by Fig. 2(d), and the corresponding graphical representation, i.e., RoleNet, is shown in Fig. 2(b). The edge between two nodes denotes that these two roles once appeared in the same scene. Note that the edges are weighted according to the closeness between these two roles. The thicker (larger weight) an edge is, the closer the two roles are.

After the processes described above, we transform role's relationship into RoleNet. On the basis of this network, we elaborately perform analysis for different types of movies.

IV. Community Analysis

A. Bilateral Movie Analysis

To demonstrate the effectiveness of the proposed idea, we take a popular type of movie, named "bilateral movies," as a pilot instance. In a bilateral movie, there are two apparent leading roles. Other roles assist the progress of story and can be grouped into two communities, which are respectively led by these two leading roles. For example, there are often a justice group and an evil group in action movies. As for romance movies, it is common to find two groups that belong to the hero and the heroine, respectively.

Fig. 3 shows the RoleNet constructed from a typical bilateral movie—"You've Got Mail." We can roughly see closeness between roles via edge weights. However, there are actually finer structures hidden in this network. Table I shows the true casts and the corresponding positions in this movie. Roles 1 and 2 are the hero and the heroine, and other roles can be separated into two groups led by them, respectively. The RoleNet in Fig. 3 consists of intricate edges so that the finer structure cannot be directly observed.

To facilitate deeper investigation, we propose a process to perform community analysis, as depicted in Fig. 4. For a given RoleNet, we first determine the leading roles, and then identify the hidden communities. Leading roles are the persons who have the most significant impact and dominate the progress of stories in a movie. A community is a group of roles that relatively have similar relationships to a leading role. For example, the roles 3, 5, 6, 7 are the heroine's friends and colleagues. They live or work with the heroine, and the audience can clearly perceive that they are "at the same side."

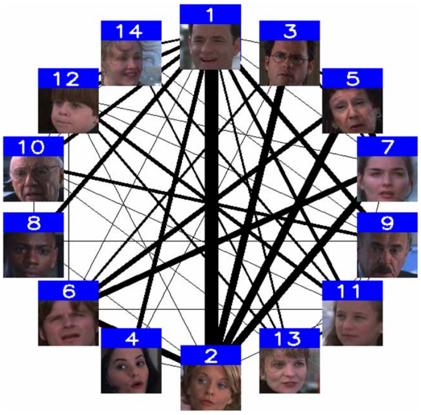


Fig. 3. RoleNet of the movie "You've Got Mail".

TABLE I
Roles in the Movie "You've Got Mail"

Node (Role)	Meaning of roles
1	The hero (Tom Hanks)
2	The heroine (Meg Ryan)
3, 5, 6, 7	The heroine's friends and colleagues
4, 8, 9, 10, 11, 12, 13, 14	The hero's friends, relatives, and colleagues.

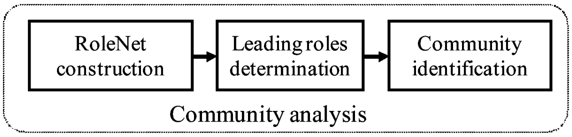

RoleNet construction
Leading roles determination
Community identification
Community analysis

Fig. 4. Process of community analysis based on RoleNet.

Leading Role Determination: In SNA, evaluating the impact of each individual is one of the earliest issues. It is known as the *centrality problem* [11]. One way to measure the centrality value of a node is to calculate the number of connected edges to it. However, this measurement doesn't faithfully reflect the information within a RoleNet, which is a weighted graph and conveys much information in edge weights. Therefore, based on RoleNet, we evaluate the centrality c_i of the node (role) i as

$$c_i = \sum_{j\neq 1} w_{ij} \qquad (4)$$

where w_{ij} is the edge weight defined in Section III-B.

In bilateral movies, we choose the nodes with the first two largest centrality values as the leading roles.

Community Identification: After determining the leading roles, we would like to investigate how other roles relate to them and identify which roles have similar characteristics, i.e., they form a community. From the perspective of SNA, communities are groups of nodes within which the connections are dense but between which the connections are sparse.

According to the characteristics of bilateral movies, there are two major communities led by two leading roles. Determining these two communities can be viewed as a binary labeling problem. We denote the first and the second leading roles as v_p and v_q. The problem of community identification can be expressed as follows.

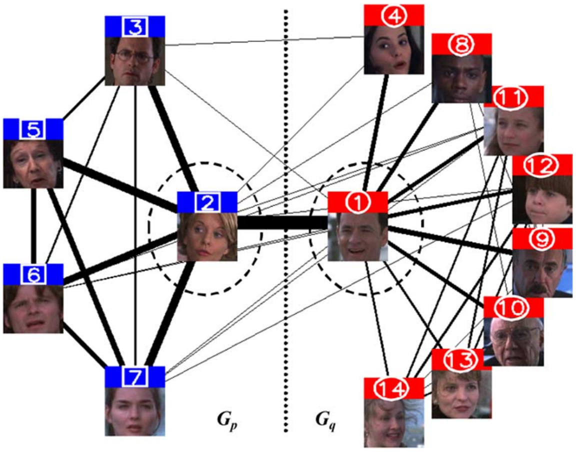


Fig. 5. Results of community identification for the movie "You've Got Mail".

TABLE II
FINER COMMUNITY STRUCTURE IN THE MOVIE "YOU'VE GOT MAIL"

Macro	Micro	Meaning of roles
1		The hero (Tom Hanks)
2		The heroine (Meg Ryan)
3, 5, 6, 7	3	The heroine's boy friend
	5, 6, 7	The heroine's colleagues
4, 8, 9, 10, 11, 12, 13, 14	4	The hero's girl friend
	8	The hero's assistant
	9, 10	The hero's father and grandfather. The hero and they are co-founders of a company.
	11, 12	The hero's niece and nephew. They just visit the hero at holiday.
	13, 14	The hero's stepmother and her servant

Given a RoleNet, find a labeling solution Δ^*:

$$\Delta^* = \arg\min_{\Delta} C(\Delta) \text{ subject to } \delta_p = 0 \text{ and } \delta_q = 1 \tag{5}$$

$$C(\Delta)=\sum_{i,j} |\delta_i - \delta_j| w_{ij} \tag{6}$$

$$\Delta = \{\delta_i, i = 1, \ldots, n\} \tag{7}$$

where n is the number of roles, Δ is a set of binary labels, $\delta_i = 0$ if v_i is assigned to the community led by v_p, and $\delta_i = 1$ if v_i is assigned to the community led by v_q. The value $C\Delta$ is the closeness between two communities, which is calculated by summing the weights between roles in two different communities. The first leading role v_p is labeled as 0 ($\delta_p = 0$), and the second leading role v_q is labeled as 1 ($\delta_q = 1$). Equation (5) expresses that the optimal solution is the labels causing the least closeness between two different communities. For a brief description, we use G_p to represent the roles in the community led by v_p, and use G_q to represent the roles in the community led by v_q.

This problem can be solved through finding the minimum cut between two leading roles. Therefore, we adopt the maximum-flow-minimum-cut algorithm [14] to find the optimal labeling.

Fig. 5 shows the result of community identification corresponding to Fig. 3. Node 1 and node 2 are correctly detected as the leading roles (with the dash-line circles). The roles identified as the members of G_p are marked by solid-line squares, and the roles identified as the members of G_q are marked by solid-line circles. These results exactly match the real cases listed in Table I.

B. Generalization

The bilateral example has shown that RoleNet well describes the relationship between roles and the proposed method effectively discovers the hidden structure. In this section, we would like to generalize the idea to various kinds of movies. The generalization process tackles with the following issues.

- Automatically determining the number of leading roles: The prescribed process is conducted when the number of leading roles is predetermined. Therefore, we try to develop a method that automatically determines the number of leading roles.
- Analyzing finer communities: There are actually finer structures within the identified communities. For example, although the roles no. 4 and no. 8 are both identified in the community led by the hero, as shown in Fig. 5, they have totally different positions in the movie. Table II shows the true information at finer granularity. In this work, we call the rough communities identified previously as macro-communities, and call the finer structure in Table II as micro-communities.

Leading Role Determination: An important observation can be utilized to automatically determine the number of leading roles. Leading roles make significantly larger impact than other roles. More specifically, there is a large gap between the impact of leading roles and that of supporting roles. Based on this observation, the problem of leading role determination can be mathematically expressed as follows:

$$\Gamma^* = \arg\min\nolimits_{\Gamma}(\min\Theta_1 - \max\Theta_0) \tag{8}$$

where $\Theta_1 = \{c_i|\ell_i = 1\}$, $\Theta_0 = \{c_i|\ell_i = 0\}$, and $\{\Gamma = \{\ell_i, i = 1, 2, \ldots, n\}$. Θ is a set of binary labels, in which $\ell_i = 1$ is the *ith* role is assigned as a leading role, and $\ell_i = 0$ otherwise. The set Θ_1 represents centrality values of the roles assigned to leading roles. The physical meaning of $(\min\Theta_1 - \max\Theta_0)$ is the difference of centrality values between the least important leading role and the most important supporting role. The result Γ^* we want is the labels that cause the largest centrality difference.

To solve this problem, still taking "You've Got Mail" as an example, we propose an automatic leading role determination method in the following.

Algorithm 1: Leading Role Determination

1. Calculate the centrality value of each role, as shown in (4). Fig. 6(a) shows a real example.
2. Sort the centrality values in descending order, as shown in Fig. 6(b).
3. Calculate the centrality difference between two adjacent roles. Fig. 6(c) shows the centrality difference distribution,

TABLE III
INFORMATION OF THE EVALUATION DATA

ID	Movie Title	Genre	# of leading roles	Length	# of story segments	% of scenes with faces
M1	The Devil Wears Prada (2006)	Comedy / Drama	1	109 min	31	100% (69/69)
M2	You've Got Mail (1998)	Comedy / Romance	2	119 min	28	100% (58/58)
M3	21 Grams (2003)	Crime / Thriller	3	124 min	106	97.5% (120/123)
M4	Catch Me If You Can (2002)	Crime / Drama	2	141 min	42	100% (101/101)
M5	The Lake House (2006)	Fantasy / Romance	2	105 min	35	98.7% (75/76)
M6	My Blueberry Night (2007)	Drama / Romance	1	90 min	28	79% (45/57)
M7	Broken Flowers (2005)	Comedy / Mystery	1	106 min	19	97.5% (39/40)
M8	Casino Royale (2006)	Action / Adventure	1	144 min	30	98.3% (59/60)
M9	Gladiator (2000)	Action / Adventure	1	155 min	28	96.7% (58/60)
M10	Stranger Than Fiction (2006)	Drama / Romance	1	113 min	23	100% (64/64)
S1	Sex and the City (Season 1, Episode 1, 1998)	Comedy / Romance	1	26 min	18	95.8% (23/24)
S2	Friends (Season 7, Episode 7, 2000)	Comedy / Romance	6	23 min	17	100% (17/17)
S3	Sex and the City (Season 3, Episode 1, 2000)	Comedy / Romance	4	29 min	23	96.5% (28/29)
Total				21 hr 24 min	428	97.2% (756/778)

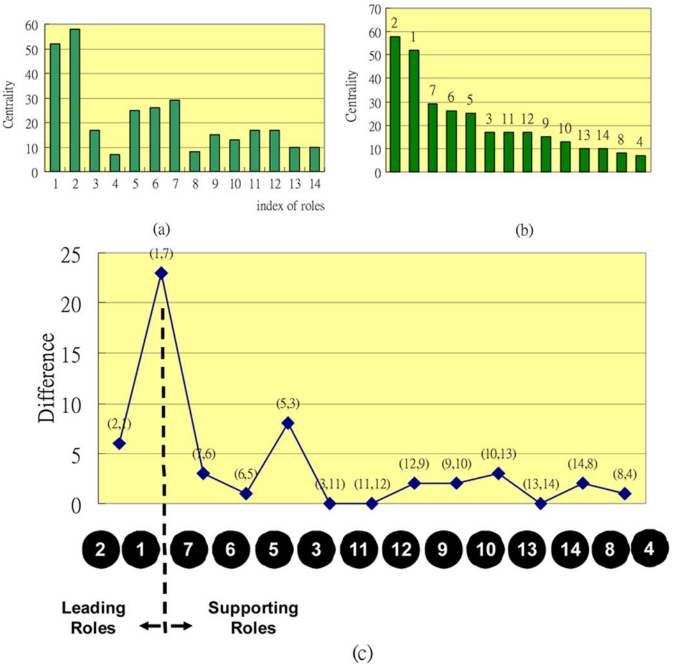


Fig. 6. From the movie "You've Got Mail": (a) Centrality value of each role; (b) sorted centrality values; and (c) difference of centrality values between adjacent roles.

in which each point represents the boundary characteristic between two roles.

4. Find the maximum point in the difference distribution, which represents the largest gap in centrality.

The selected boundary determines the number of leading roles. In the example of Fig. 6, this method automatically determines that the roles no. 2 and no. 1 should be leading roles. The complexity of this algorithm is bounded to the sorting of centrality values. If there are n roles in a movie, the complexity is $O(n \log n)$.

Community Identification:

Micro-Community Identification: After determining the leading roles, we devise a method to directly discover the hidden micro-communities. The idea is to appropriately group certain roles into a micro-community. Because a leading role may pass through several micro-communities, it's not reasonable to assign he/she into only one micro-community. Therefore, we first remove the leading roles and the edges linked to them from the RoleNet. Then, Algorithm 2 is applied to the modified RoleNet. We use the value t to index the community's evolution situation. The value t is initialized as 0 in the beginning and increases by one when the community situation changes.

Algorithm 2: Micro-Community Identification

```
1. Initialize every individual node as a micro-community.
   The set of micro-community is denoted as
   $\Pi_t = \{T_1^t, T_2^t, \ldots, T_n^t\}, t = 0$, if there are initially $n$
   individual nodes. The size of the $p$th community in $\Pi_t$ is
   denotes as $|T_p^t|$, which is the number of nodes included in
   this community.
2. From the modified RoleNet, find the edge that has the
   largest weight, say the edge $e_{ij}$ between the node $v_i$ and
   the node $v_j, v_i \in T_p^t$ and $v_j \in T_q^t$, then
   1) If $|T_p^t| \geq 1$ and $|T_q^t| = 1$, then $T_p^{t+1} =$
      $T_p^t \cup T_q^t, \Pi_{t+1} = \Pi_t - \{T_q^t\}$, and $t = t + 1$.
   2) If $|T_p^t| > 1$ and $|T_q^t| > 1$, then keep current
      community situation.
3. Remove the edge $e_{ij}$ from the modified RoleNet and go
   to *Step* 2 until all edges have been removed.
```

The progress of this algorithm can be illustrated as a dendrogram, which describes how we cluster communities at each step. For example, as shown in Fig. 7, the roles no. 6 and no. 7 are first categorized together $(t = 1)$, then the role no. 5 is merged into this community at the second level $(t = 2)$. (We say "level" but not "iteration" because the community situation may not change at each iteration.) The same process can be iteratively applied until all nodes have been examined.

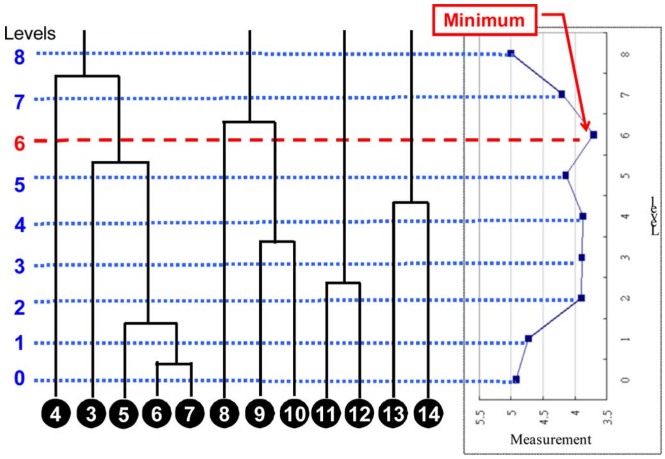


Fig. 7. Dendrogram of the clustering process.

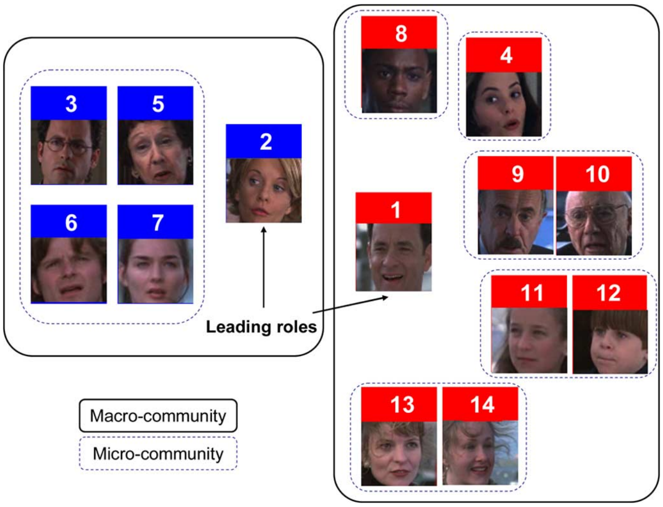


Fig. 8. Results of micro-community and macro-community identification.

Each level in the dendrogram represents a case of community situation. Now the problem is to determine which level in the dendrogram is the best. We design a measurement to evaluate the community case at different levels. For the level t, the measurement is defined as

$$AvgW_t = \frac{\sum w_{ij}}{||\Pi_t||}, \forall v_i \in T_p^t, v_j \in T_q^t, p \neq q \tag{9}$$

where Π_t denotes the community situation at the level t, and $||\Pi_t||$ denotes the number of communities in this case. The value $AvgW_t$ represents the average weight between different communities at level t. The right part of Fig. 7 shows the measures at different levels.

Conceptually, the value of $AvgW$ represents the closeness between communities. In community identification, we prefer that roles in different communities are least related. Therefore, we pick the community case that causes the minimal $AvgW$. In Fig. 7, the minimal $AvgW$ value occurs at the sixth level, in which the micro-communities includes roles $\{4\}, \{3,5,6,7\}, \{8\}, \{9,10\}, \{11,12\}$, and $\{13,14\}$, respectively. The roles in the same brace are classified into the same micro-communities. This result is very close to the true states listed in Table II.

Because we remove at least one edge at each iteration until all edges are removed, the complexity of the algorithm is bounded to the number of edges. The maximum possible edges is $n(n-1)/2$. Therefore, the complexity of micro-community identification is $O(n^2)$.

Macro-Community Identification: On the basis of micro-communities, we can aggregate them to construct macro-communities. Because a macro-community contains a leading role and his/her most related micro-communities, the problem of macro-community identification can be solved by assigning micro-communities to the most appropriate leading role.

Let L represent the set of leading roles. For the micro-community T_p, the assigning process is as follows:

$$v^* = \arg\max_{v_i \in L}(\max_{v_j \in T_p} w_{ij}) \tag{10}$$

where the value w_{ij} denotes the weight between the leading role v_i and the role v_j in T_p. We use the largest weight to represent the closeness between T_p and the leading role v_i. By checking the value with respect to every leading role, we finally assign T_p to the one that has the largest weight with T_p.

Fig. 8 simultaneously shows the results of micro-community and macro-community identification for the movie "You've Got Mail." There are two macro-communities (in solid-line squares). The micro-communities (in dash-line squares) in each macro-community are automatically determined by the proposed algorithm. Note that the result of macro-community identification is the same as that in Fig. 5. By comparing the results of micro-community with the true states listed in Table II, only the role no. 3 is erroneously identified.

In macro-community identification, we compute the weight of each edge between supporting roles and leading roles. If there are ℓ leading roles and $(n-\ell)$ supporting roles, we must compute $\ell(n-\ell)$ edges at most. Therefore, the complexity of this process is bounded to $O(n)$.

V. Story Segmentation

Automatic story segmentation is a widely studied topic in video analysis researches. After shot change detection, shots that have similar content-based characteristics and are temporally adjacent are clustered together to form a scene or a story [15], [17]. For news videos, audiovisual features are fused to model the boundaries of reported stories [18]. For movie videos, the so-called logical story units [3] are determined through checking shots' visual similarity and the pattern of shot changes. Based on tempo analysis [2] and computational media aesthetics [19], we developed a system to automatically perform story segmentation for action movies [8].

Numerous works that extract audiovisual features and fuse them by elaborate models have been proposed. However, for the task of story segmentation in movies, most works overlook the essence of stories. Although changes of stories usually accompany with significant variations in visual appearance, the main factor for humans to sense a story boundary is the essential change in semantics. Semantics represents what a director wants to describe and is often derived from the interaction between characters. Therefore, based on RoleNet and community analysis described in the previous sections, we propose a story segmentation method from a new perspective and demonstrate its superiority.

A. Scene Representation

Similar to conventional story segmentation works, we first define the representation of scenes. The major difference between the proposed representation and conventional ones is that we describe scenes by "the context of roles" rather than audiovisual features. Story segmentation is achieved by comparing the role's context in successive scenes.

Let $r(k)$ denote the kth character in a specific scene. The relationship between this role and others can be expressed by a "profile vector" $\boldsymbol{w}_{r(k)} = (w_{1r(k)}, w_{2r(k)}, \ldots, w_{nr(k)})$, which is the $r(k)$-th column vector of the matrix W in equation (3). The vector $\boldsymbol{w}_{r(k)}$ denotes the closeness between the role no. $r(k)$ and others. It is normalized into a unit vector for the following process. For the ith scene, we collect the profile vectors of the roles appearing in this scene to form a matrix CM_i that describes roles' context: $CM_i = [\boldsymbol{w}_{r(1)}\boldsymbol{w}_{r(2)} \cdots \boldsymbol{w}_{r(p)}]$. It is an n by p matrix if there are p roles in this scene and there are totally n roles in the movie. Similarly, the matrix CM_j for the jth scene is denoted by $CM_j = [\boldsymbol{w}_{r(1)}\boldsymbol{w}_{r(2)} \cdots \boldsymbol{w}_{r(q)}]$, if there are q roles in this scene. Note that the vector $\boldsymbol{w}_{r(k)}$ in the ith scene and the vector $\boldsymbol{w}_{r(k)}$ in the jth scene would be different, i.e., the identifications of the kth characters in these two scenes are different. Using the notations of $\boldsymbol{w}^i_{r(k)}$ and $\boldsymbol{w}^j_{r(k)}$ is more precise but dazzles readers. Therefore, we use the simplified notation in the following description.

Based on this information, the context-based similarity between "the sth role in the ith scene" and "the tth role in the jth scene" is defined as the inner product of two corresponding profile vectors: $\boldsymbol{w}_{r(s)} \cdot \boldsymbol{w}_{r(t)} = \boldsymbol{w}^T_{r(s)}\boldsymbol{w}_{r(t)}$. By calculating the context-based similarities between every two roles in two successive scenes, the similarity between the ith scene and the jth scene can be expressed in a matrix form: $CM_{ij} = CM_i^T CM_j$. Fig. 9 shows an example of calculating the context similarity between two scenes.

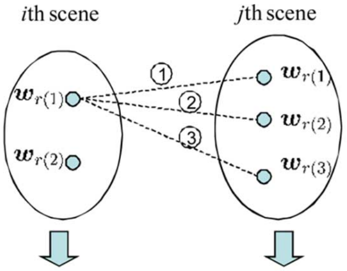
Finally, the context-based difference between the ith scene and the jth scene is defined as

$$d_{ij} = 1 - \frac{1}{pq}\sum_{s=1}^{p}\sum_{t=1}^{q} CM_{ij}(s,t) \tag{11}$$

which represents the average difference between pairs of roles in two different scenes. The value is between 0 and 1.

B. Story Segmentation

Going through the whole movie, we can plot a "difference curve" that represents the difference between adjacent scenes. Fig. 10 shows the curve of context-based difference for the movie "You've Got Mail." Variations of the heights of this curve implicitly present some clues for finding story boundaries. Based on this curve, the goal of story segmentation is to find appropriate scene boundaries that represent changes of stories. In this work, we propose a story segmentation method called "storyshed," which is motivated by the watershed algorithm for image segmentation but is modified to meet the need of this task.



$$CM_i = \begin{bmatrix}\boldsymbol{w}_{r(1)} & \boldsymbol{w}_{r(2)}\end{bmatrix} \quad CM_j = \begin{bmatrix}\boldsymbol{w}_{r(1)} & \boldsymbol{w}_{r(2)} & \boldsymbol{w}_{r(3)}\end{bmatrix}$$

Context-based similarity:

$$CM_i^T CM_j = \begin{array}{ccc} ① & ② & ③ \\ \end{array}$$
$$\begin{bmatrix} \boldsymbol{w}^T_{r(1)}\boldsymbol{w}_{r(1)} & \boldsymbol{w}^T_{r(1)}\boldsymbol{w}_{r(2)} & \boldsymbol{w}^T_{r(1)}\boldsymbol{w}_{r(3)} \\ \boldsymbol{w}^T_{r(2)}\boldsymbol{w}_{r(1)} & \boldsymbol{w}^T_{r(2)}\boldsymbol{w}_{r(2)} & \boldsymbol{w}^T_{r(2)}\boldsymbol{w}_{r(3)} \end{bmatrix}$$

Fig. 9. Example of calculating the context-based similarity between two scenes.

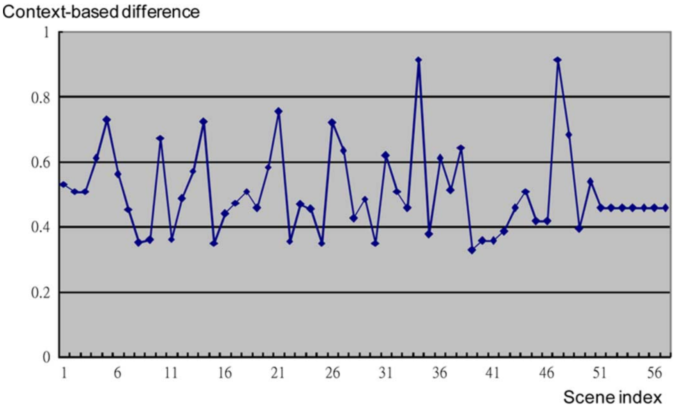


Fig. 10. Context-based difference curve for the movie "You've Got Mail".

Denote the set of scene boundaries as $B = \{b_1, b_2, \ldots, b_{N-1}\}$, in which the element b_i denotes the boundary between the ith and the $(i+1)$-th scenes, and N is the total number of scenes. The proposed storyshed algorithm is to determine whether a scene boundary is a story boundary, based on the context-based difference between adjacent scenes. The context-based difference corresponding to b_i is denoted by d_i. In the first step of the segmentation process, we first find the valleys and peaks from the difference curve by checking d_i. That is

$$\begin{cases} b_i \in Y, & \text{if } d_i < d_{i-\alpha 1} \text{ and } d_i < d_{i+\alpha 2} \\ b_i \in P, & \text{if } d_i > d_{i-\alpha 1} \text{ and } d_i > d_{i+\alpha 2} \\ b_i \in OT, & \text{otherwise} \end{cases}$$
$$\alpha_1 = \min\{j | j \in A_1\}$$
$$\alpha_2 = \min\{j | j \in A_2\}$$
$$A_1 = \{k | (d_i - d_{i-k}) \neq 0, 1 \leq k \leq i-1\}$$
$$A_2 = \{k | (d_i - d_{i+k}) \neq 0, 1 \leq k \leq (N-1-i)\} \tag{12}$$

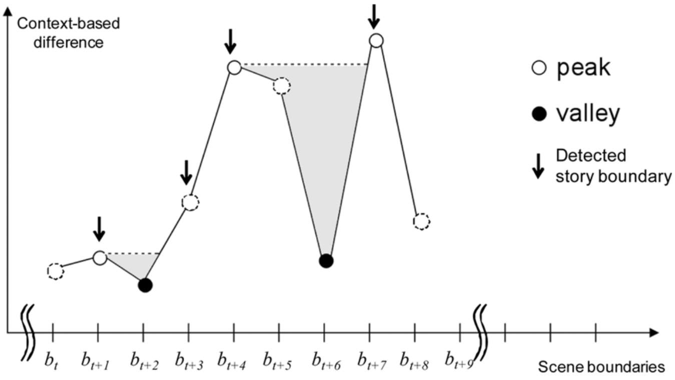

Context-based difference
○ peak
● valley
Detected story boundary
b_t b_{t+1} b_{t+2} b_{t+3} b_{t+4} b_{t+5} b_{t+6} b_{t+7} b_{t+8} b_{t+9}
Scene boundaries

Fig. 11. Example of the storyshed segmentation method without global threshold.

where $2 \leq i \leq N-2, Y$ denotes the set of scene boundaries in valleys, P denotes the set of boundaries in peaks, and the set OT includes all other boundaries. Thus, $Y \cup P \cup OT = B$.

Initialize an empty set SB that will store the story boundaries. For each valley y_j in Y, find the nearest peaks to it. Let's denote the left peak of y_j as p_1 and the right peak of y_j as p_2, $p_1 \in P$ and $p_2 \in P$. Fill water into this valley until the height of the horizontal just floods p_1 or p_2. Without loss of generality, assume that p_1 is flooded first and the height of p_1 is H. Pick the scene boundaries b_k between the peaks p_1 and p_2, for which the corresponding context-based difference d_k is no less than H. Therefore, the set of story boundaries would be $SB = SB \cup \{b_k\}$.

The storyshed algorithm is summarized as follows.

Algorithm 3: The Storyshed Algorithm

```
Input: The set of scene boundaries $B = \{b_1, b_2, \ldots, b_{N-1}\}$
and the corresponding context-based difference values
$D = \{d_1, d_2, \ldots, d_{N-1}\}$.
Output: The set of story boundaries.
  1. According to $D$, find the boundaries in the valley set $Y$
     and the peak set $P$.
  2. For each valley in $Y$, find the two nearest peaks from $P$
     that are respectively at the left and the right of it.
  3. For each valley $y_i$ in $Y$, fill water into each valley until
     the height of the water horizontal just floods one of the
     corresponding peaks.
  4. Pick the scene boundaries $b_k$ which have context-based
     difference values no less than the water horizontal and
     are located between $y_i$'s two peaks. The set of story
     boundaries $SB = SB \cup \{b_k\}$.
```

Fig. 11 shows an example of the processes described above. Two valleys (solid black circles) in this example are at b_{t+2} and b_{t+6}. The nearest peaks to b_{t+2} are at b_{t+1} and b_{t+4}, and that to b_{t+6} are at b_{t+4} and b_{t+7}. After the processes described above, the ones selected to be story boundaries are $b_{t+1}, b_{t+3}, b_{t+4}$, and b_{t+7}.

The results reveal the boundaries that the scenes around them have significantly different context information. However, the storyshed algorithm only considers local characteristics and may miss the ones that are indeed story boundaries. For example, the scene boundary b_{t+5} is actually a story boundary but is not detected by the prescribed approach. Basically, if the context-based difference value is large enough, the corresponding scene boundary is definitely a story boundary, no matter the difference value around this boundary (local context variation) is large or not.

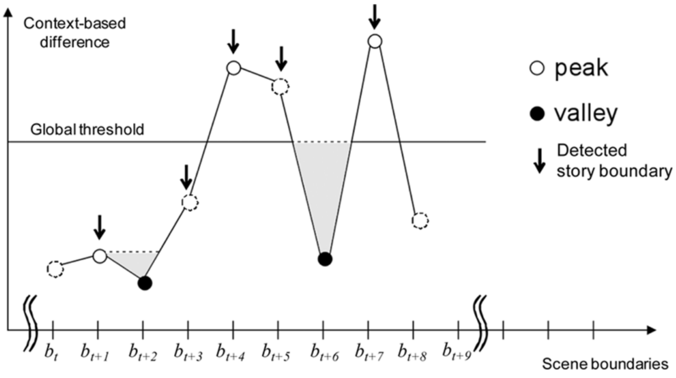

Context-based difference
Global threshold
○ peak
● valley
Detected story boundary
b_t b_{t+1} b_{t+2} b_{t+3} b_{t+4} b_{t+5} b_{t+6} b_{t+7} b_{t+8} b_{t+9}
Scene boundaries

Fig. 12. Example of the storyshed segmentation method with a global threshold.

To consider the global characteristics, we take the average of the context-based difference values of all peaks as a global threshold. If the difference value corresponding to a scene boundary is larger than this threshold, it's viewed as a story boundary. The global threshold is adaptively calculated according to the social relationships between roles in different movies. No manual tuning is needed.

Fig. 12 shows the storyshed segmentation results with the global threshold. Now the boundary b_{t+5} is detected as a story boundary as well. Processing with this global threshold is the same as applying a constraint in the third step of the segmentation algorithm so that the height of the water level is limited to be lower than the threshold.

The proposed story segmentation method considers role's context information between scenes and more appropriately represents the changes of stories. To our knowledge, this study is one of the first few works to perform movie story segmentation from the perspective of social relations.

The computation of story segmentation process consists of two parts: calculation of context-based difference and the storyshed algorithm. Because the complexity of the storyshed algorithm is linearly related to the number of scenes in a movie, the complexity of the whole process is bounded to the calculation of context-based difference. If there are p roles and q roles in two successive scenes, and $CM_i(n \times p)$ and $CM_j(n \times q)$ represent these two scenes' role context, the scene similarity is computed through $CM_i^T CM_j$. Therefore, the complexity is $O(pqn)$. The maximum possible number of p and q is n, thus the complexity of computing the context-based difference between two scenes is $O(n^3)$. There are $m-1$ differences needed to be computed. Therefore, the complexity of computing all context-based differences is $O(mn^3)$, and so is the whole story segmentation process.

TABLE IV
DIFFERENT SITUATIONS IN DEFINING THE COMMUNITY GROUND TRUTH

Subject A's opinion	Subject B's opinion	Subject C's opinion	Final ground truth
{1,2,3} {4,5,6}	{1,2,3} {4,5,6}	{1,2,3} {4,5,6}	{1,2,3} {4,5,6}
{1,2,3} {4,5,6}	{1,2,3} {4,5,6}	{1,2,3,4} {5,6}	{1,2,3} {4,5,6}
{1,2,3} {4,5,6}	{1,2} {3,4} {5,6}	{1,2,3,4} {5,6}	Anyone
{1,2,3} {4,5,6}	{1} {2,3} {4,5,6}	{1,2,3,4} {5,6}	Anyone

VI. EVALUATION

We use ten Hollywood movies and three TV shows to evaluate the proposed methods. The total length of the evaluation is over 21 h and 428 story segments are included. As shown in Table III, these movies belong to different genres and have different numbers of leading roles. We also show that over 97% of scenes actually contain face information, which provides a solid foundation for us to reveal role's social relationship. In order to faithfully show the effectiveness of the proposed RoleNet model, we first demonstrate the experimental results that are based on the manually-labeled data in Sections VI-A and VI-B. We then propose an implementation method in Section VI-C to automate the whole process and demonstrate the corresponding results.

A. Community Analysis

Ground Truth: For each movie, we asked three persons to manually label leading roles after watching movie. The ones that were labeled as the leading ones by all the three persons are treated as the ground truth. For macro-community, we asked three persons to label which roles belong to which macro-communities. Different annotators may have different preferences in grouping. However, the community situation is usually not intricate. Directors tend to attract the audience directly and make the scenario be understood easily. If the position of a certain role is controversial, we assign this role according to the result of majority voting.

To clarify different situations in defining community ground truth, let us assume that there are six roles (roles 1 ~6) to be categorized, and three subjects (A, B, C) are asked to give the community ground truth. Different identification situations are listed in Table IV. The numbers in braces denote the indices of roles. For example, $\{1,2,3\}$ in subject A's opinion represents that the roles no. 1, 2, and 3 are determined as in the same community. In the case that all subjects have different opinions, we claim that the community identification results are correct if they conform to any one of the opinions. Similar mechanism is applied to define the ground truth of micro-communities.

TABLE V
PERFORMANCE OF LEADING ROLE DETERMINATION AND MACRO-COMMUNITY IDENTIFICATION

Movie ID	Ground truth	Determined leading roles	# of roles categorized correctly / # of roles
M1	1	1	12 / 12
M2	1, 2	1, 2	14 / 14
M3	1, 2, 6	1, 2, 6	20 / 20
M4	1, 2	1, 2	15 / 15
M5	1, 2	1, 2	8 / 9
M6	1	1	9 / 9
M7	1	1	15 / 15
M8	1	1	15 / 15
M9	1	1	15 / 15
M10	1	1	10 / 10
S1	1	1	14 / 14
S2	1, 2, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7, 9	10 / 12
S3	1, 2, 3, 4	1, 2, 3	7 / 12

Performance of Leading Roles Determination: The performance of leading role determination is shown in the third column of Table V. The numbers in each cell denote the indices of roles. For example, the roles no. 1, 2, and 6 are determined as the leading roles in M3. We can see that almost perfect performance can be achieved.

The promising performance comes from two reasons.

1) Leading roles pass through most scenes in a movie and have close relationship with others. The trend of close relationships between roles is apparent.
2) The proposed method effectively captures the characteristics of leading roles. Based on the representation of RoleNet, leading roles can be clearly identified by measuring the impact of different roles.

The performance in TV shows is worse than that in movies. TV shows often last for less than thirty minutes and have fewer than thirty scenes. The pace of shows is fast, because directors have to use short and fewer scenes to present stories. Moreover, the selected TV shows include many leading roles[2]. Their performance often ends before the relationships between roles are appropriately constructed. People can infer what happen and understand the subtle relationships between roles quickly, but the proposed method still appeals to the well-constructed relationships based on the frequent co-occurrence of roles.

Performance of Community Identification: The fourth column of Table V shows the performance of macro-community identification. It shows that the performance of the proposed community process is very promising for movies. Most roles are correctly assigned to the corresponding leading role. This result again confirms that the trend of mutual relationship is apparent, and the proposed method catches this characteristic.

The identification performance of TV shows is worse because we face the same situation as described in the previous section. To verify the length issue, we especially concatenate two episodes of "Sex and The City" (Season 2, Episodes 11 and 12) into a one-hour video and perform the same processes for leading role determination and community analysis. All four leading roles are correctly determined, and the results of community analysis are much better than that of analyzing one

[2] We generally know that there are four leading roles in "Sex and The City," and six leading roles in "Friends." However, the show S1 is an exception. It is the first episode of this series, and the most important leading role (Carrie Bradshaw) introduces the start of other leading roles in the following episodes.

TABLE VI
DETAILED PERFORMANCE OF MICRO-COMMUNITY IDENTIFICATION

Movie ID	Ground truth	Results of micro-community identification

M1	Leading roles:{1}
	Other Roles:
	{2,3,4,9,11},{5,6,7},{8},{10},{12}
		{2,3,4,9},{5,6,7},{8},{10},{11},{12}

M2	Leading roles:{1,2}
	Other Roles:
	{3},{4},{5,6,7},{8},{9,10},{11,12},{13,14}
		{3,5,6,7},{4},{8},{9,10},{11,12},{13,14}

M3	Leading roles:{1,2,6}
	Other Roles:
	{3,4,5},{7},{10,11,12,13,16},{8,9,18}, {14,17}, {15},{19},{20}
		{3,4,5},{7,10,11,12,13,16},{8,9,18},{14,17},{15},{19},{20}

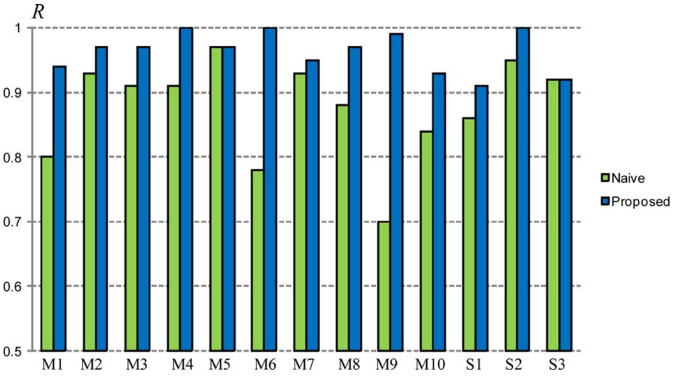


Fig. 13. Performance comparison of micro-community identification based on the proposed quantification method.

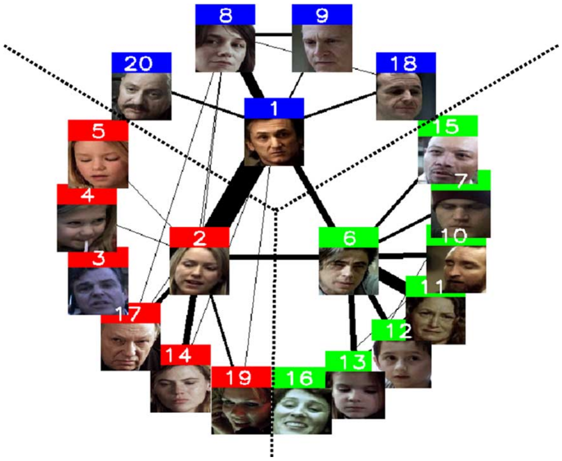


Fig. 14. Results of community identification for the movie "21 Grams".

episode only. This result verifies the length issue and reveals the limitation of the proposed methods as well.

Detailed results of micro-community identification for three movies, M1, M2, and M3, are listed in Table VI. We can see that only the role no. 11 in M1, the role no. 3 in M2, and the role no. 7 in M3 are erroneously categorized.

It is hard to directly perceive the performance from the detailed results as shown in Table VI. Therefore, we design a method to quantify the massive experimental results. We transform the ground truth and the identification results into the relationships between pairs of roles. If two roles v_i and v_j are in the same community, the indicative values ζ_{ij} and ζ_{ji} of the pair (v_i, v_j) are set as 1. Otherwise, they are set as 0. There are $(\begin{smallmatrix}k\\2\end{smallmatrix})$ possible pairs if there are k roles to be identified. Among the $(\begin{smallmatrix}k\\2\end{smallmatrix})$ possible pairs, we calculate how many of them are correctly labeled. The ratio of correctly labeled pairs to all possible ones is used to quantify community identification results. That is, the ratio R is calculated as

$$R = \frac{\sum_{i=1}^{k}\sum_{j\neq i}\delta_{ij}}{2\times\begin{pmatrix}k\\2\end{pmatrix}} \tag{13}$$

$$\begin{cases}\delta_{ij}=1, & \text{if}, \zeta_{ij}^{g}=1 \text{ and } \zeta_{ij}^{v}=1\\ \delta_{ij}=1, & \text{if}, \zeta_{ij}^{g}=0 \text{ and } \zeta_{ij}^{v}=0\\ \delta_{ij}=0, & \text{otherwise}\end{cases} \tag{14}$$

where ζ_{ij}^{g} and ζ_{ij}^{v} are pair relationships transformed from the ground truth and the identification results. The value δ_{ij} indicates whether the identified result between the roles i and j is the same as the ground truth. The larger the ratio is, more accurate the identification results are.

Based on this measurement, we can quantify the results of micro-community identification. In addition, we also take a naïve case to be the reference basis. In the naïve case, roles are crudely viewed to be independent, and each role forms a micro-community alone. Fig. 13 shows the performance comparison for all evaluation data based on the proposed measurement. From this figure, we can easily see the superiority of the proposed micro-community identification process. For the movie M7, most scenes contain only one character. The hero has brief reunions with four women he used to know, and each woman and her relatives surely form separate micro-communities. However, among the few scenes where the leading role co-occurs with others, the leading role recalls how he interacts with all other supporting roles. This contaminates the weak relationships between the leading role and all others and correlates different micro-communities. This special arrangement harms the proposed social-based process.

Performance in TV shows is not as good as in movies. Shorter video length and more leading roles weaken the relationships among roles. Many micro-communities have only one member. Therefore, the naïve method often works well.

Fig. 14 shows the results of community identification for the movie "21 Grams." Comparing Fig. 14 with Fig. 5, we can clearly see the difference of community structures between different types of movies. Nodes in Fig. 5 are separated into two sides, while nodes in Fig. 14 distribute much more like a triangle.

B. Story Segmentation

Ground Truth: Similar to community analysis, the ground truths of story boundaries were decided manually. We invited several subjects who were not familiar the goal and process of our work and knew nothing about the chapter information in advance. They were asked to examine every scene change boundary and decide whether it's a story boundary. For every decided story boundary, opinions from different subjects were compared. If different decisions were made among subjects, the final decision was made by majority voting.

Someone may argue that the chapter information provided in DVDs could be the ground truth of story boundaries. However, if we carefully examine this information, we can easily find that many chapter boundaries are not proper story boundaries. Chapter information just provides rough browsing, and a chapter often contains more than one story segments. According to our experiments, many user-defined story boundaries don't match the chapter boundaries. Therefore, we did not adopt the DVD information as the ground truth. The numbers of story segments are listed in Table III. The number of story segments in "21 Grams" is significantly larger than others because stories in this movie are derived from three leading roles and are switched frequently.

Performance of Story Segmentation: The results of story segmentation are presented in terms of purity [23]. Given the ground truth of stories $S = \{(s_1, \Delta t_1), \ldots, (s_{N_g}, \Delta t_{N_g})\}$ and the results of story segmentation $S^* = \{(s_1^*, \Delta t_1^*), \ldots, (s_{N_g}^*, \Delta t_{N_g}^*)\}$, the purity ρ is defined as

$$\rho = \left(\sum_{i=1}^{N_g} \frac{\tau(s_i)}{T} \sum_{j=1}^{N_v} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_i)} \right) \cdot \left(\sum_{j=1}^{N_v} \frac{\tau(s_i^*)}{T} \sum_{i=1}^{N_g} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_j)} \right) \qquad (15)$$

where $\tau(s_i, s_j^*)$ is the length of overlap between the stories s_i and $s_j^*, \tau(s_i)$ is the length of the story s_i, and T is the total length of all stories. In each parenthesis, the first term indicates the fraction of the current evaluated story, and the second term indicates how much a given story is split into smaller stories. The purity value ranges from 0 to 1. Larger purity value means that the result is closer to the ground truth.

We compare the performance of story segmentation based on four approaches—The tempo-based approach [8], the one with global thresholds, the storyshed algorithm, and the storyshed algorithm with thresholds. As described in [8], movie tempo is calculated based on motion activity, audio energy dynamics, and shot change frequency. According to the tempo curve corresponding to a movie, our previous work [8] can be applied to determine the boundaries of stories in movies. We compare the tempo-based approach with the newly proposed social-based approach to demonstrate the superiority of the latter.

Fig. 15 shows that the social-based approach works much better than the tempo-based one. Although the performance improvement varies in different movies, the storyshed algorithm has significantly better performance than thresholding. After combining storyshed with thresholding, the result is even slightly better than the storyshed algorithm, especially in M3, M7, and S2. Table VII shows the overall purity in different approaches. Generally, the proposed method has about 0.48 improvement in purity over the tempo-based approach.

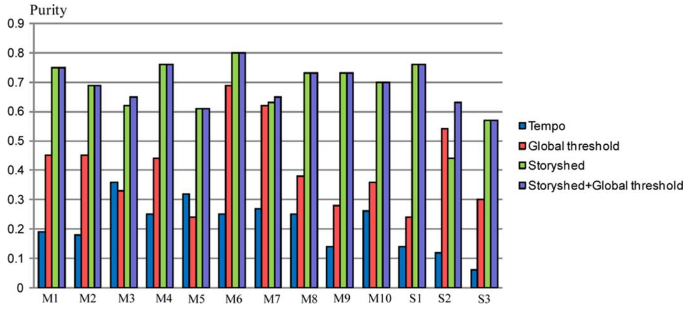


Fig. 15. Performance of story segmentation in terms of purity.

TABLE VII
OVERALL PURITY OF STORY SEGMENTATION IN DIFFERENT APPROACHES

	Tempo	Global threshold	Storyshed	Storyshed + Global threshold
Overall	0.21	0.41	0.68	0.69

C. Performance Based on Automatic Labeling

Automatic Labeling: The sections described above demonstrate the effectiveness of the proposed models and processes. In this section, we further propose a framework to conduct automatic scene boundary detection and labeling, and provide experimental results to show how the errors derived from automatic labeling affect the performance. Due to space limitation, we only describe the implementation in brief. For details of the implementation, please refer to [21].

In order to achieve automatic labeling, we first apply the method proposed in [15] to perform scene detection. We then exploit OpenCV face detector [25] to detect the locations and regions of faces in each scene. There may be many errors in face detection. Therefore, we have to process the face location data further to obtain more reliable bases. Two processes are developed for this task.

1) Noise filtering: Non-face objects would be misdetected as faces. Because a non-face object may look like a face in just a certain view, we check the locations and areas of detected faces in adjacent frames and filter out those without consistent areas or similar locations.
2) Grouping: After noise filtering, two face sequences are connected if faces in them are spatially close and the time distance between two face sequences is short.

Because the final result we really want is which roles appeared in which scene, we do not need to perfectly recognize every detected face in every frame. We can just sample several faces from a face sequence for face recognition, and vote to determine which face is presented. In this work, we adopt the face recognition module proposed in [26].

It is worth noting that although the faces are taken at varying lighting conditions, scaling, or poses, we can still successfully recognize them in most scenes. A character in a scene would show up in many shots, i.e., perhaps hundreds or thousands of frames. If the character really plays an important role, directors would not take him in side-view or in back-view forever. To determine whether a character appears in a scene, we just need to successfully match the character at more than one frame. If

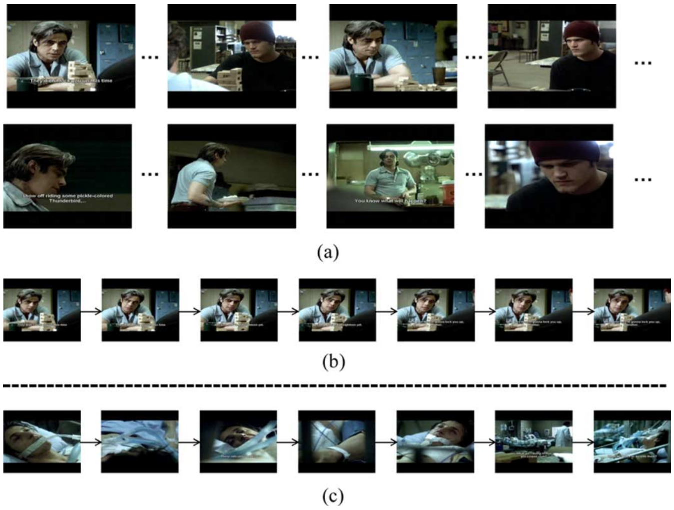


Fig. 16. (a) Some keyframes of the shots in a scene. (b) Some frames in the first shot in (a). (c) Some keyframes of the shots in a scene, in which we failed to correctly detect and match faces.

TABLE VIII
PERFORMANCE OF MACRO-COMMUNITY IDENTIFICATION BASED ON DIFFERENT LABEL DATA

Movie ID	# of roles labeled correctly / # of roles (manual)	# of roles labeled correctly / # of roles (automatic)
M1	12 / 12	12 / 12
M2	14 / 14	14 / 14
M3	20 / 20	18 / 20
M4	15 / 15	10 / 15
M5	8 / 9	7 / 9
M6	9 / 9	9 / 9
M7	15 / 15	15 / 15
M8	15 / 15	15 / 15
M9	15 / 15	15 / 15
M10	10 / 10	10 / 10
S1	14 / 14	7 / 14
S2	10 / 12	10 / 12
S3	7 / 12	3 / 12

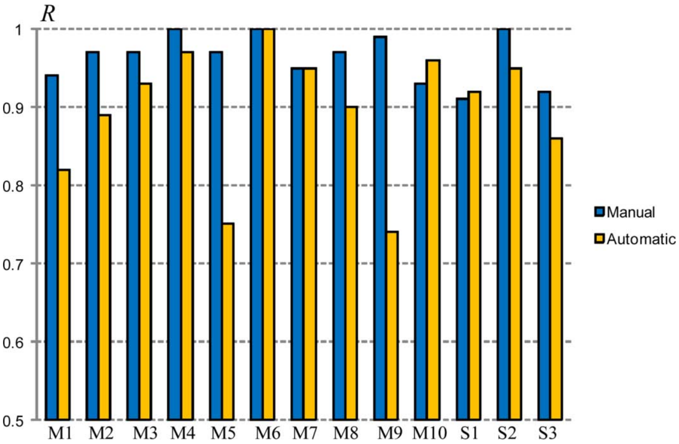


Fig. 17. Performance comparison of micro-community identification using different labeling data.

this character really appears in this scene, the probability of the face being detected/matched is very high.

Fig. 16(a) shows some keyframes of shots in a scene. It is obvious that faces would be taken at drastically different situations, and the face matching problem seems to be extremely difficult. However, if we carefully examine each shot, like the frames shown in Fig. 16(b), we can find that we have many chances to get the frontal faces and successfully recognize the character. Actually, there are 4350 frames in the scene, and we only have to successfully recognize these two characters at least once. In the real implementation, we do not even have to examine all frames but sample one per ten frames. Of course, we still failed in some cases, as shown in Fig. 16(c). In this scene, although it contains many shots and many frames, not anyone out of thousands of frames shows the frontal face. In this case, we may miss the chance to increase the weight values between some roles. For story segmentation, we view this scene as a single story.

Performance Based on Automatic Labeling: Although there would be errors in face recognition, we can obtain similar performance as manually-labeled data did in leading role determination. As we described in Section VI-A, the impacts of leading roles are significantly larger than others. Thus, we can still perform well even the labeling data are annoyed by recognition errors.

For macro-community identification, we list the results based on manually-labeled and automatically-labeled data in Table VIII. From this table, we see that the trend of macro-communities is also apparent, and the identification results are similar based either on manual or automatic labeling data. The performances of S1 and S3 are especially bad because poor video quality degrades the performance of face recognition. Worse face recognition results cause the errors of leading role determination, and accordingly affect the subordination between leading roles and macro-communities.

For micro-community identification, performance difference in terms of the ratio described in equation (13) is depicted in Fig. 17. It's not unexpectable that errors of automatic labeling degrade the performance. The degree of degradation depends on the visual situations in different movies. However, the degradation is acceptable, and the results of micro-community identification are still useful in developing applications. The performance is expected to be improved when we incorporate state-of-the-art scene boundary detection and face recognition modules in the future.

VII. Discussion

We emphasize the value of analyzing movies by the proposed approaches. Although the leading roles are often tagged when movies were produced, how and where they present in scenes are missing. The scenes which different leading roles appear in construct different storylines, which are not considered in conventional approaches. In addition, knowing the communities of a movie is beneficial for browsing and organization. Recall that conventional approaches view a movie as a hierarchical structure (frames, shots, and scenes), and people conduct hierarchical browsing based on the keyframes they are seeing. This browsing scheme may be good in movies with apparent visual arrangement, such as action movies, but may not provide much information for dramatic movies. In many cases, accessing a movie based on characters, stories, or communities, are more intuitive

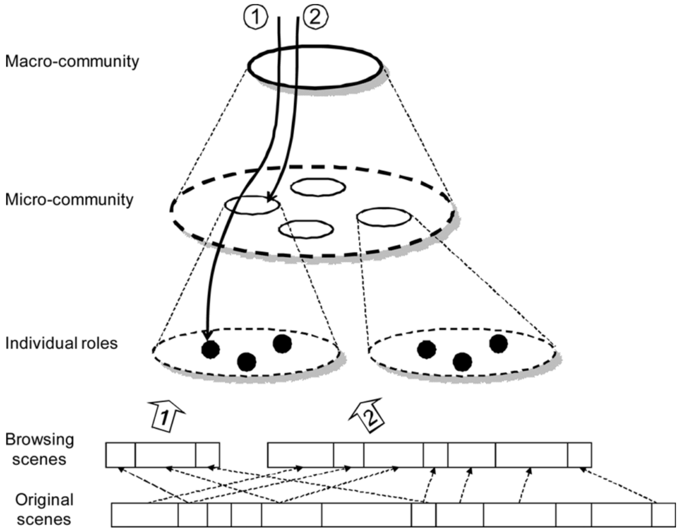


Fig. 18. Community-based hierarchical browsing.

and comprehensible to users. For example, a user can easily access "the action of justice side before the battle starts." We describe the extension and limitation of the proposed approach in the following.

A. Extension

Community-Based Hierarchical Browsing System: As we know the community structures and leading roles, a community-based browsing system can be built. This kind of browsing scheme is totally different from hierarchical shot-based browsing as presented in [16]. Users can browse the stories made by the hero and his relatives, for example, by exploring the corresponding macro-communities. More specifically, users can explore deeper to see the stories related to the hero's family or the story specifically related to an individual.

Fig. 18 illustrates the community-based browsing system. At the first access, a user selects a specific role's story, e.g., the hero's father, and the scenes this role ever appears are returned. At the second access, a user selects the micro-community representing the hero's family (accessing a micro-community), and the scenes where the family members appear are returned. Note that this kind of browsing follows the "hierarchy of social relationships" rather than the "hierarchy of content-based similarity."

We can further find the hierarchical structure of a storyline based on micro-communities. For example, all the stories containing members of the macro-community led by the hero form a storyline. With the aids of micro-communities, we can further divide the storyline into the stories involving the hero's family, the stories involving the hero's work, and the stories involving the hero's friends.

Scene Description: We can develop more elaborate systems based on the results of community identification. According to the characters involved in scenes, we not only can segment different stories but also may identify the property of stories. For example, in an action movie containing two macro-communities, the stories where members in both macro-communities are involved may cause "conflict" situations. On the contrary, the stories where only members in the same macro-communities appear may be viewed as a "developing" part of a scenario. Many studies have been done to represent stories or scenes by keyframes. Techniques of semantic concept detection [31], [32] can be applied to annotate objects, events, or sites in stories. However, semantic presentation that represents cinematic properties was rarely addressed in literature.

Compatibility: It's important to point out that the proposed method and existing content-based analysis methods are not irreconcilable. They can be integrated to achieve finer movie understanding. For example, existing scene importance measures can be integrated with the RoleNet-based measures to facilitate advanced highlight extraction. Combining these two approaches will be conducted in the future.

Extensibility: The proposed model can not only be applied to movies. Although we only provide a few sample results in the evaluation section, we can see that the proposed method can be effectively applied to TV shows. We believe that other kinds of story-oriented videos that consist of roles' interaction can be effectively modeled and analyzed in a similar way.

B. Limitation

Appropriate entities for modeling social relations may vary for different applications or in different domains of digital content. In this work, modeling of social relationship is based on the co-occurrence of characters, which is automatically inferred from the results of face detection and face registration. Although related techniques are widely studied for many years, reliable face detection/registration techniques that are invariant to lighting conditions, poses and large motion are not well developed currently. Therefore, the performance of finer analysis such as micro-community identification is influenced by the detection/registration performance.

Face recognition is not the only way to construct RoleNet. Other techniques, such as speaker identification, can also be adopted. We can identify whether a character appears in a scene based on speech information, as the works done in [22] and [23]. However, we know that both face clustering and speaker identification are not perfect currently. Therefore, we just apply a handy tool (face detection and clustering) to develop this work. Combining audio with face information to construct RoleNet is reasonable and would be developed in the future.

The proposed processes appeal to the co-occurrence frequency of roles. The performance is limited if the relationships between roles are subtle, or if the length of videos is not sufficient to clearly represent relationships. This also implies that the proposed analysis methods are not suitable to "alternative movies" or some "artistic movies," in which the directors would fully represent his/her idiosyncrasy and ignore conventional space arrangement as adopted in most Hollywood movies.

VIII. Conclusion

We have introduced the idea of social networks to movie video analysis. Instead of utilizing audiovisual features, we treat a movie as a small society and analyze it through role's relationships. Movies are elaborately transformed to a role's social network, called RoleNet. Based on RoleNet, we develop a generic method to automatically determine the number of leading roles and identify macro- and micro-communities. Moreover, the context information is used to describe video scenes. Difference between scenes is, therefore, evaluated and used to perform story segmentation.

In the experiments, we extensively evaluate the performance of the proposed methods for different genres of movies and TV shows. The proposed modeling and processes work effectively in identifying leading roles and communities based on manually-labeled data. We also propose an implementation framework to accomplish automatic scene labeling. For automatically-labeled data, the performance degradation is expectable and is acceptable. For story segmentation, we compare the proposed algorithm with a conventional tempo-based approach. The proposed social-based segmentation method achieve about 0.69 in purity, which is much superior to that of the tempo-based approach.

With the aid of RoleNet, we approach movie understanding from a perspective different from traditional audiovisual features. We also discuss the extension and limitation of the proposed idea. In the future, we will step further to investigate and realize the potential research topics described in the discussion section. For example, by combining feature-based methods, more interesting and intelligent applications will be built. Furthermore, speaker identification techniques will be incorporated into the process of RoleNet construction, in the future.

References

[1] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, Jan. 2005.

[2] B. Adams, C. Dorai, and S. Venkatesh, "Toward automatic extraction and expression of expressive elements from motion pictures: Tempo," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 472–481, 2002.

[3] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 492–499, 2002.

[4] Y. Li, S.-H. Lee, C.-H. Yeh, and C.-C. J. Kuo, "Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 79–89, 2006.

[5] B.T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun., and Applic.*, vol. 3, no. 1, 2007.

[6] B. Jung, T. Kwak, J. Song, and Y. Lee, "Narrative abstraction model for story-oriented video," in *Proce. ACM Multimedia Conf.*, 2004, pp. 828–835.

[7] A.F. Smeaton, B. Lehane, N.E. O'Connor, C. Brady, and G. Craig, "Automatically selecting shots for action movie trailers," in *Proc. ACM Int. Workshop on Multimedia Information Retrieval*, 2006, pp. 231–238.

[8] H.-W. Chen, J.-H. Kuo, W.-T. Chu, and J.-L. Wu, "Action movies segmentation and summarization based on tempo analysis," in *Proc. ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, 2004, pp. 251–258.

[9] H.L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[10] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[11] J. Scott, *Social Network Analysis: A Handbook*. : Newbury Park, 1991.

[12] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Phys. Rev.*, vol. 68, p. 065103(R), 2003.

[13] A.E. Krause, K.A. Frank, D.M. Mason, R.E. Ulanowicz, and W.W. Taylor, "Compartments revealed in food-web structure," *Nature*, vol. 426, pp. 282–285, 2003.

[14] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[15] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Proc. IEEE Computer Society Confe. Computer Vision and Pattern Recognition*, 2003, vol. 2, pp. 343–348.

[16] X. Zhu, A.K. Elmagarmid, X. Xue, L. Wu, and A.C. Catlin, "InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 648–666, Aug. 2005.

[17] M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Comput. Vis. Image Understand.*, vol. 71, no. 1, pp. 94–109, 1998.

[18] W. H. Hsu and S.-F. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2003, vol. 2, pp. 413–416.

[19] C. Dorai and S. Venkatesh, "Computational media aesthetics: Finding meaning beautiful," *IEEE Multimedia*, vol. 8, no. 4, pp. 10–12, 2001.

[20] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "Movie analysis based on roles' social network," in *Proc. IEEE Int. Conf. Multimedia & Expo.*, 2006, pp. 1403–1406.

[21] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: Treat a movie as a small society," in *Proc. ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, 2007, pp. 51–60.

[22] A. Vinciarelli, F. Fernandez, and S. Favre, "Semantic segmentation of radio programs using social network analysis and duration distribution modeling," in *Proc. IEEE Int. Conf. Multimedia & Expo.*, 2006, pp. 779–782.

[23] A. Vinciarelli and S. Favre, "Broadcast news story segmentation using social network analysis and hidden Markov models," in *Proc. ACM Multimedia*, 2007, pp. 261–264.

[24] R. Rienks, D. Zhang, and W. Post, "Detection and application of influence rankings in small group meetings," in *Proc. Int. Conf. Multimodal Interfaces*, 2006, pp. 257–264.

[25] *Open Source Computer Vision Library*, [Online]. Available: http://www.intel.com/technology/computing/opencv/

[26] A. V. Nefian, M. H. , and III. Hayes, "An embedded HMM-based approach for face detection and recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1999, vol. 6, pp. 3553–3556.

[27] Y. Li, S. Narayanan, and C.-C. J. Kuo, "Content-based movie analysis and indexing based on audiovisual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 8, pp. 1073–1085, Aug. 2004.

[28] W.-T. Chu, W.-H. Cheng, J. Y.-J. Hsu, and J.-L. Wu, "Towards semantic indexing and retrieval using hierarchical audio models," *ACM Multimedia Syst. J.*, vol. 10, no. 6, pp. 570–583, 2005.

[29] H.J. Zhang, C.Y. Low, S.W. Smoliar, and J.H. Wu, "Video parsing, retrieval and browsing: An integrated and content-based solution," in *Proc. ACM Multimedia Conf.*, 1995, pp. 15–24.

[30] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Proc. SPIE*, 1999, vol. 3656, pp. 290–301.

[31] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1678–1689, Oct. 2006.

[32] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia Mag.*, vol. 13, no. 3, pp. 86–91, 2006.

[33] L.D. Giannetti, *Understanding Movies*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[34] D. Bordwell, *Narration In The Fiction Film*. Madison, WI: Univ. Wisconsin Press, 1985.

[35] N.P. Garg, S. Favre, H. Salamin, D. Hakkani Tur, and A. Vinciarelli, "Role recognition for meeting participants: An approach based on lexical information and social network analysis," in *Proc. ACM Multimedia Conf.*, 2008, pp. 693–696.

[36] H. Hung, D. Jayagopi, C. Yeo, G. Friendland, S. Ba, J. Ramchandran, N. Mirghafori, and D. Gatica-Perez, "Using audio and video features to classify the most dominant person in a group meeting," in *Proc. ACM Multimedia Conf.*, 2007, pp. 835–838.

[37] K. Albrecht, *Social Intelligence: The New Science of Success*. New York: Wiley, 2005.

[38] A. Pentland, "Social signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 108–111, Jul. 2007.

[39] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," in *Proc. ACM Multimedia Conf.*, 2008, pp. 1061–1070.

Chung-Yi Weng received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2005 and 2007, respectively.

His current research interests include social network analysis, multimedia applications, and multimedia content analysis.

Mr. Weng received the Best Full Technical Paper Award from ACM Multimedia in 2006.

Wei-Ta Chu (M'04) received the B.S. and M.S. degrees from National Chi Nan University, Taiwan, R.O.C., in 2000 and 2002, respectively, and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, in 2006.

Since 2007, he has been the Assistant Professor in the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. He was a visiting scholar at the Digital Video and Multimedia Laboratory, Columbia University, New York, during July-August 2008.

His research interests include digital content analysis, multimedia indexing, digital signal process, and pattern recognition.

Dr. Chu won the Best Full Technical Paper Award in ACM Multimedia 2006.

Ja-Ling Wu (SM '98–F'08) received the Ph.D. degrees in electrical engineering from Tatung Institute of Technology, Taipei, Taiwan, R.O.C., in 1986.

From 1986 to 1987, he was an Associate Professor of the Electrical Engineering Department, Tatung Institute of Technology. In 1987, he transferred to the Department of Computer Science and Information Engineering (CSIE), National Taiwan University (NTU), Taipei, where he is presently a Professor. From 1996 to 1998, he was assigned to be the first Head of the CSIE Department, National Chi Nan University, Puli, Taiwan. During his sabbatical leave (from 1998 to 1999), he was invited to be the Chief Technology Officer of the Cyberlink Corp. In this one-year term, he was involved with the developments of some well known audio-video software, such as the PowerDVD. Since August 2004, he has been appointed to head the Graduate Institute of Networking and Multimedia, NTU. He has published more than 200 technique and conference papers. His research interests include digital signal processing, image and video compression, digital content analysis, multimedia systems, digital watermarking, and digital right management systems.

Dr. Wu was the recipient of the Outstanding Young Medal of the R.O.C. in 1987 and the Outstanding Research Award of the National Science Council, R.O.C., in 1998, 2000, and 2004, respectively. In 2001, his paper "Hidden Digital Watermark in Images" (co-authored with C.-T. Hsu), published in IEEE TRANSACTIONS ON IMAGE PROCESSING, was selected to be one of the winners of the Honoring Excellence in Taiwanese Research Award, offered by ISI Thomson Scientific. Moreover, his paper "Tiling Slideshow" (co-authored with his students) won the Best Full Technical Paper Award in ACM Multimedia 2006. He was selected to be one of the lifetime Distinguished Professors of NTU in November 2006. He was elected as an IEEE Fellow in 2008 for his contributions to image and video analysis, coding, digital watermarking, and rights management.